UNIVERSITÀ DI PISA

DIPARTIMENTO DI LETTERATURA, FILOLOGIA E LINGUISTICA

Corso di Laurea Magistrale in Informatica Umanistica

TESI DI LAUREA

Costruire un sistema di *Visual Question Answering* per l'italiano con la traduzione automatica: esperimenti ed analisi

Relatore:
Prof. Alessandro Lenci

Seconda Relatrice:
Dott.ssa Lucia C. Passaro

Candidata:
Elisa Barisani

ANNO ACCADEMICO 2020/2021

Indice

Introduzione **3**

1 Stato dell'arte **5**

1.1 Lavori correlati 5

1.1.1 *Natural Language Processing* e *Question Answering* testuale 5

1.1.2 *Computer Vision* e *Object Detection* 8

1.1.3 Multimodalità, *Image Captioning* e *Visual Question Answering* . . 10

1.2 Risorse e strumenti per il *Visual Question Answering* 14

1.2.1 Dataset di *Visual Question Answering* 15

1.2.2 *Neural Machine Translation* per colmare la mancanza di risorse in italiano 29

1.3 Modelli e baseline 32

1.3.1 *Joint embedding approaches* 33

1.3.2 *Attention mechanism* 35

1.3.3 Modelli composizionali 36

1.3.4 *Knowledge base-enhanced approaches* 37

1.3.5 LXMERT 39

2 Materiali e metodi **43**

2.1 Esperimenti con un modello addestrato sull'italiano: NMT + LXMERT-it 45

2.1.1 Creazione di GQA-it 46

2.1.2 Traduzione dei *scene graph* di GQA 50

2.1.3 Applicazione del modello LXMERT-it 53

2.2 Esperimenti con un modello pre-addestrato in inglese: LXMERT-pretrained + NMT 54

2.2.1 Nuovo dataset 55

2.2.2 Campione di GQA 57

2.2.3 Applicazione del modello LXMERT 59
2.2.4 Valutazione 61

3 Valutazione **64**
3.1 Valutazione degli esperimenti NMT + LXMERT-it 64
3.1.1 Errori nelle risposte 66
3.1.2 Analisi dei *pattern* delle domande 67
3.2 Valutazione degli esperimenti con LXMERT- pretrained + NMT 71
3.2.1 Errori del modello 73
3.2.2 Errori nella traduzione automatica delle domande 74
3.2.3 Errori nella traduzione automatica delle risposte 76
3.3 Confronto tra l'utilizzo di NMT + LXMERT-it e LXMERT-pretrained + NMT 80

Conclusioni **83**

Bibliografia **84**

Elenco delle figure

1.1 Esempio di *scene graph* creato a partire da *region graph* (Krishna et al. 2017, pag. 4). 21
1.2 Funzionamento del *framework* LXMERT (Tan et al. 2019, pag. 3). 40

2.1 Diagramma di flusso degli esperimenti. 44
2.2 Distribuzione della tipologia di risposte del *test set gold* di GQA-it. 50
2.3 Distribuzione del/i soggetto/i delle domande del *test set gold* di GQA-it. . 50
2.4 Distribuzione della tipologia di risposte del nuovo dataset. 56
2.5 Distribuzione del/i soggetto/i delle domande del nuovo dataset. 56
2.6 Immagine 2 estratta dal nuovo dataset. 57
2.7 Distribuzione della tipologia di risposte del campione di GQA. 58
2.8 Distribuzione del/i soggetto/i delle domande del campione di GQA. 58
2.9 Immagine n457744 estratta dal campione di GQA. 58
2.10 *Workflow* degli esperimenti LXMERT-pretrained + NMT. 60

3.1 Distribuzione degli errori nelle risposte del *test set gold* di GQA-it. 67
3.2 Distribuzione degli errori nella NMT delle domande del *test set gold* di GQA-it. 69
3.3 Distribuzione degli errori del modello per i due dataset. 74
3.4 Distribuzione degli errori nella traduzione automatica delle domande per i due dataset. 76
3.5 Distribuzione degli errori nella traduzione automatica delle risposte per i due dataset. 78

Elenco delle tabelle

1.1 Caratteristiche dei dataset presentati. La tabella è ripresa e modificata da Wu, Teney et al. (2016), pag. 11. 17

2.1 Statistiche di GQA-it (Croce, Passaro et al. 2021, pag. 4.) 47
2.2 Statistiche della resa italiana dei *scene graph* di GQA. 51
2.3 Domande in inglese e in italiano per l'immagine 2 del nuovo dataset. 57
2.4 Domande in inglese e in italiano per l'immagine n457744 estratta dal campione di GQA. 59
2.5 Caratteristiche dei dataset utilizzati per gli esperimenti. 63

3.1 Risultati di LXMERT e LXMERT-it su 3.000 domande di GQA e GQA-it (Croce, Passaro et al. 2021). 65
3.2 Accuratezza di LXMERT-pretrained sul campione del *balanced test* set di GQA. 72
3.3 Accuratezza di LXMERT-pretrained sul nuovo dataset. 72
3.4 Occorrenze degli errori del modello LXMERT pre-addestrato sui due dataset analizzati. 73
3.5 Occorrenze degli errori nella traduzione automatica delle domande dall'italiano all'inglese sui due dataset analizzati. 75
3.6 Occorrenze degli errori nella traduzione automatica delle risposte dall'inglese all'italiano sui due dataset analizzati. 78
3.7 Differenze nell'accuratezza dei due dataset considerando gli errori non gravi. 79
3.8 Confronto della perdita di accuratezza utilizzando NMT + LXMERT-it e LXMERT-pretrained + NMT. 81

Introduzione

I passi avanti compiuti negli ultimi anni dall'intelligenza artificiale, e in particolare lo sviluppo di algoritmi basati sulle reti neurali, hanno aumentato l'interesse nei confronti di *task* multimodali che completano e arricchiscono l'interazione uomo-macchina.
In particolare, sono aumentate le ricerche che riguardano la creazione di sistemi multimodali che combinano input visivi e linguistici con lo scopo di garantire nel prossimo futuro un modo naturale di interrogare il contenuto visivo.
Ogni giorno si entra in contatto con un gran numero di immagini, provenienti da varie fonti, come il web, articoli, grafici ecc. Sia per gli utenti umani sia per i sistemi automatici, si rivela fondamentale inferire informazioni che consentano di comprendere le immagini, individuando al loro interno gli oggetti, i loro attributi e le relazioni che li legano.
L'estrazione di queste informazioni semantiche dalle immagini ha numerose applicazioni che spaziano dal supporto a persone non vedenti e ipovedenti alla possibilità di interrogare, utilizzando il linguaggio naturale, grandi banche dati di immagini senza utilizzare meta-dati o tag.
Un importante *task* multimodale per l'unione di input visivi e linguistici è il *Visual Question Answering*, proposto da Antol et al. nel 2015, che ha lo scopo di rispondere correttamente a domande poste in linguaggio naturale su un'immagine (Antol et al. 2015). Questo *task* integra la comprensione generale delle immagini e del linguaggio naturale e la necessità di rispondere a domande basate sul senso comune.
Il *task* è ampiamente affrontato per l'inglese, lingua per la quale sono presenti numerosi dataset (Antol et al. 2015, Krishna et al. 2017, Hudson et al. 2019) e modelli (Tan et al. 2019). Tuttavia, la ricerca non si è ancora concentrata su altre lingue, come l'italiano.

L'obiettivo di questo elaborato è quello di esplorare come le risorse inglesi esistenti possano essere utilizzare per il VQA in italiano. In particolare, l'ipotesi è che i passi avanti compiuti dalla traduzione neurale automatica siano tali da consentire di trasporre dataset e sistemi inglesi in italiano, permettendo di evitare il costo della creazione di risorse *ex novo*.

Vengono analizzate due possibili modalità per l'utilizzo della traduzione automatica per la costruzione di sistemi di VQA in italiano: la prima è quella di fornire una traduzione automatica parzialmente validata di dataset esistenti e utilizzarla per riaddestrare modelli esistenti sull'italiano (Croce, Passaro et al. 2021); la seconda è quella di utilizzare modelli pre-addestrati in inglese attraverso la traduzione automatica della domanda in input e della risposta in output.
Entrambi gli approcci consentono di trasporre dall'inglese all'italiano risorse già esistenti attraverso la traduzione automatica, ma portano a una perdita di qualità nel modello linguistico. L'obiettivo degli esperimenti è indagare quanto e come il rumore introdotto dalla traduzione automatica porti a una perdita di accuratezza dei sistemi.

Il primo capitolo dell'elaborato è dedicato alla rassegna dello stato dell'arte del *Visual Question Answering*, partendo da un'analisi del contesto in cui questo *task* si è sviluppato e proseguendo con un'analisi dei principali dataset e modelli con cui viene affrontato.
Nel secondo capitolo, vengono presentati gli strumenti e le metodologie utilizzati per gli esperimenti. Vengono descritte le risorse e vengono spiegati i passaggi necessari per la creazione di sistemi di *Visual Question Answering* che riaddestrano un modello in italiano e che utilizzano un modello pre-addestrato in inglese su un input italiano. Viene anche trattata la creazione di dataset per l'italiano attraverso la traduzione automatica di risorse in inglese e la loro validazione parziale o totale.
L'ultimo capitolo della trattazione è dedicato all'analisi qualitativa e quantitativa dei risultati ottenuti attraverso gli esperimenti effettuati.

1. Stato dell'arte

Il *Visual Question Answering* (VQA) è un *task* di Intelligenza Artificiale (IA) introdotto da Antol et al. nel 2015 e definito come segue: "Given an image and a natural language question about the image, the task is to provide an accurate natural language answer" (Antol et al. 2015).

Il VQA collega due campi appartenenti al dominio dell'Intelligenza Artificiale, ovvero il *Natural Language Processing* e la *Computer Vision*, al fine di stimolare la ricerca e spingere i confini di entrambi i campi.

Da un lato, il *Natural Language Processing* si occupa di permettere interazioni tra computer e umani utilizzando il linguaggio naturale, dall'altro, la *Computer Vision* si occupa dei metodi per acquisire, elaborare e comprendere le immagini e ha lo scopo di insegnare alle macchine come "vedere". Anche se questi due campi condividono metodi simili radicati nell'apprendimento automatico, storicamente si sono sviluppati separatamente e hanno visto progressi significativi verso i loro rispettivi obiettivi negli ultimi decenni. La crescita combinata di dati visivi e testuali ha spinto ad un'unione degli sforzi (Wu, Teney et al. 2016).

1.1 Lavori correlati

1.1.1 *Natural Language Processing* e *Question Answering* testuale

Il *Natural Language Processing* (NLP) è il campo dell'IA che si occupa di studiare come i computer possono essere utilizzati per comprendere e manipolare il linguaggio naturale, testuale o parlato, per diverse applicazioni (Chowdhury 2003). Lo scopo dei ricercatori che si occupano di NLP è quello di raccogliere conoscenza su come gli umani comprendono e utilizzano il linguaggio al fine di sviluppare gli strumenti e le tecniche per far sì che i computer comprendano e manipolino le lingue per performare i *task* designati (Chowdhury

2003). Come le altre aree dell'IA, il NLP si basa su un gran numero di discipline, come la linguistica, la scienza, l'intelligenza artificiale, le scienze informatiche e dell'informazione ecc. (Chowdhury 2003).

Il NLP si sviluppa negli anni Sessanta e Settanta con la crescita dell'interesse nei confronti dell'Intelligenza Artificiale (IA); inizialmente si basa su metodi logico-deduttivi in linea con la metodologia razionalista e simbolica tipica dell'IA e condivisa dalla linguistica generativa che in quegli anni è dominante (Lenci et al. 2005). Contemporaneamente in ambito britannico si sviluppa la linguistica dei *corpora* che utilizza prevalentemente strumenti di analisi quantitativa e statistica ed esplora in maniera empirica e deduttiva le regolarità che emergono dalle raccolte di grandi quantità di testo (Lenci et al. 2005). Questo approccio prende piede, tuttavia, solo negli anni Ottanta e Novanta, quando viene superata la visione generativa e chomskyana della linguistica, i cui fondamenti teorici si basano sull'utilizzo di metodi scientifici standard, ovvero sul metodo induttivo piuttosto che sulla deduzione a partire dai *corpora*. Con lo sviluppo del *Machine Learning* (ML) e dei *corpora* linguistici, si comprendono le potenzialità dell'utilizzo del ML applicato a *corpora* di esempi reali e il NLP passa all'utilizzo di strumenti di analisi statistica e quantificativa (McEnery et al. 2003, M. Johnson 2009).

Con l'affermarsi delle reti neurali per numerosi *task* di IA, dagli anni Dieci del 2000 vengono sviluppate reti neurali che raggiungono ottime prestazioni in molte applicazioni NLP, come la traduzione automatica (Cho et al. 2014), il riconoscimento delle *Named Entities*, la *Sentiment Analysis* e il *part-of-speech tagging*. Il vantaggio principale di questi modelli neurali è la loro capacità di apprendere rappresentazioni utili senza richiedere uno sforzo di ingegnerizzazione delle *feature* come invece accade nei modelli di *Machine Learning* tradizionali: infatti, tra le tecniche più popolari rientrano l'uso di *word embeddings* per catturare le *feature* semantiche delle parole (Devlin et al. 2018) e l'aumento dell'apprendimento *end-to-end* di *higher-level task* (come la risposta automatica alle domande) invece di concentrarsi su una *pipeline* di compiti intermedi separati (ad esempio, *part-of-speech tagging* e *dependency parsing*) (Gridach 2020).

Le applicazioni del NLP sono numerose e includono diversi campi di studio, come la traduzione automatica, il *processing* del linguaggio naturale, lo *speech recognition*, le in-

terfacce uomo-macchina in linguaggio naturale ecc. (Chowdhury 2003). In particolare, il NLP è applicato sia al linguaggio parlato sia a quello scritto e riguarda tutti i livelli di analisi linguistica: analisi morfologica, con *task* quali il *part-of-speech-tagging* e la lemmatizzazione, l'analisi sintattica, con *task* quali il *parsing*, l'analisi semantica, con *task* quali l'*Entity Linking* e la *Sentiment Analysis*, l'analisi relazionale, con *task* quali l'estrazione delle relazioni in un testo, e l'analisi del discorso, con *task* quali l'analisi del discorso e l'estrazione e il riconoscimento dei *topic*. Come anticipato in precedenza, con lo sviluppo delle reti neurali il NLP inizia a concentrarsi su *higher-level task*, come la traduzione automatica e il *Question Answering* testuale (QA).

Il QA testuale è il *task* di NLP che riguarda la risposta automatica in linguaggio naturale a domande poste in linguaggio naturale; è un *task* di grande interesse per la comunità scientifica e negli ultimi due anni è stato molto studiato portando alla nascita di 80 nuovi dataset (Rogers et al. 2021).

Gli elementi fondamentali dei dataset di QA sono tre: la tipologia delle domande che contengono, la tipologia delle risposte e la fonte delle risposte.
Le domande dei dataset di QA testuale possono essere classificate in base alla loro forma in (Rogers et al. 2021): (*i*) linguaggio naturale, ovvero come verrebbero poste da un parlante umano, (*ii*) *query*, ovvero pezzi di informazione strutturati in modo logico che possono essere interpretati come domande, (*iii*) *story completion*, ovvero la scelta della conclusione di un passaggio testuale, e (*iv*) *cloze format*, si basano su affermazioni piuttosto che su domande o *query*.
La tipologia delle risposte si divide invece in (Rogers et al. 2021): (*i*) estrattive, richiedono di completare parte di una dimostrazione date una dimostrazione e una domanda, (*ii*) a scelta multipla, richiedono un algoritmo per scegliere tra una lista predefinita di risposte candidate e (*iii*) *free-form* o a risposta aperta, richiedono una risposta libera.
Per quanto riguarda la fonte delle risposte per le molte applicazioni del QA testuale esse possono essere suddivise in (Hirschman et al. 2001): (*i*) dati strutturati, ovvero database, (*ii*) dati semi-strutturati, per esempio, campi di commento nei database e (*iii*) testo

libero. Recentemente, gli sviluppi compiuti nel QA testuale portano ad un interesse per la sua estensione ad altre modalità, e quindi a fonti non testuali (Rogers et al. 2021). In particolare, viene esplorato il QA su immagini (Antol et al. 2015), video (Z. Zhao et al. 2017) e audio (Fayek et al. 2020); ciò porta alla creazione di numerosi dataset specifici e alla combinazione di più domini diversi (Rogers et al. 2021).

1.1.2 *Computer Vision* e *Object Detection*

La *Computer Vision* (CV) è il *task* di IA che si occupa dei metodi per acquisire, elaborare e comprendere le immagini e ha lo scopo di insegnare alle macchine come "vedere" (Szeliski 2010). Questo *task* ha numerose applicazioni, tra le quali, per esempio, l'*Optical Character Recognition* (OCR), per il riconoscimento automatico dei caratteri scritti a mano, il *Face Detection*, per consentire ai dispositivi fotografici di mettere a fuoco al meglio e per una ricerca per immagini mirata, l'*Automotive Safety*, per identificare ostacoli inaspettati come pedoni sulla strada, e molti altri (Szeliski 2010).

I principali problemi affrontati nella letteratura di CV sono *Image Classification* e *Object Recognition* (Forsyth et al. 2011).

L'*Image Classification* è il *task* di CV che richiede la classificazione dell'oggetto dominante in un'immagine senza conoscenze sulla sua posizione spaziale o il suo ruolo nell'immagine generale (D. Lu et al. 2007).

L'*Object Detection* è il *task* di CV che si occupa di individuare istanze di oggetti appartenenti a una determinata classe (ad esempio, uomini, edifici o auto), caratteristiche (ad esempio, colori o materiali) e attività degli oggetti (ad esempio, l'azione compiuta da un oggetto) all'interno di immagini digitali e video (Dasiopoulou et al. 2005). Rispetto all'*Image Classification*, l'*Object Detection* è un *task* più complesso, poiché implica la localizzazione di specifici oggetti appartenenti ad una determinata classe semantica, oltre al loro rilevamento. La posizione di un oggetto è tipicamente rappresentata attraverso una *bounding box* (Dasiopoulou et al. 2005).

Le prime e più popolari applicazioni si concentrano prevalentemente sul rilevamento dei volti (M. B. Lewis et al. 2003 e Hjelmås et al. 2001) e dei pedoni utilizzando vari dataset

ad hoc (Yang et al. 2016, Dollár et al. 2012).

Recentemente, per risolvere i *task* di *Image Classification* e *Object Detection* vengono utilizzate reti neurali.

Per il *task* di *Image Classification* generalmente vengono utilizzati *Convolutional Neural Network* (CNN), un algoritmo molto popolare per la classificazione delle immagini che tipicamente comprende *layer* di convoluzione, *layer* di funzioni di attivazione, *layer* di *pooling* per ridurre la dimensionalità senza perdere molte *feature* (Sultana et al. 2018). Sono stati sviluppati numerosi modelli pre-addestrati, come ResNet, VGG-16 e AlexNet.

Il *task* di *Object Detection* viene affrontato con algoritmi di *Region-based Convolutional Neural Network* (R-CNN) che apprendono da mappe di *feature* a generare *bounding box* che racchiudono gli oggetti delle immagini (Girshick et al. 2014). Le R-CNN estraggono *Region Proposal* per ipotizzare le posizioni degli oggetti (Uijlings et al. 2013) a partire dall'immagine in input e etichettano le loro classi e le corrispettive *bounding box*.

Successivamente, si sviluppano reti neurali Fast R-CNN che si basano su una *forward propagation* indipendente della CNN per ogni *region proposal*. Poiché queste regioni di solito hanno sovrapposizioni, le estrazioni di *feature* indipendenti portano a molti calcoli ripetuti. Uno dei principali miglioramenti di Fast R-CNN rispetto a R-CNN è che la propagazione *CNN forward* viene eseguita solo sull'intera immagine (Girshick 2015).

Un'altra soluzione più recente è la Faster R-CNN: il modello Fast R-CNN solitamente deve generare molte *region proposal* nella ricerca selettiva. Per ridurre le *region proposal* senza perdere precisione, il modello Faster R-CNN propone di sostituire la ricerca selettiva utilizzando un *Region Proposal Network* (RPN) che utilizza le *convolutional feature* dell'immagine intera attraverso un *detection network*, consentendo così una *region proposal* (S. Ren et al. 2015). Una RPN è una rete che predice simultaneamente i *bound* dell'oggetto e i punteggi relativi agli oggetti in ogni posizione. Le RPN vengono addestrate *end-to-end* per generare *proposal* che vengono utilizzate da Fast R-CNN (Girshick 2015) per il rilevamento. Attraverso un'ottimizzazione alternata, RPN e Fast R-CNN possono essere addestrati a condividere *convolution feature* (S. Ren et al. 2015).

1.1.3 Multimodalità, *Image Captioning* e *Visual Question Answering*

Ogni giorno si trova un gran numero di immagini presenti su varie fonti, come il web, articoli, diagrammi e pubblicità, che contengono immagini che gli utenti devono interpretare da soli. La maggior parte delle immagini non dispone di una descrizione, ma l'uomo è in grado di comprenderle senza le loro didascalie dettagliate. Tuttavia, la macchina ha bisogno di interpretare qualche forma di informazione semantica delle immagini, per esempio in forma di didascalia, per comprendere che oggetti e relazioni sono presenti nelle immagini e fornire agli umani didascalie o risposte automatiche su di esse (Hossain et al. 2019).

Grazie ai recenti progressi compiuti da CV e NLP, negli ultimi anni sono stati proposti numerosi *task* di IA per sfidare i sistemi automatici a unire CV e NLP.

Il primo *task* multimodale che combina NLP e CV è l'*Image Captioning*. Questo *task* nasce dalla necessità nel campo della CV e dell'*information retrieval* di arricchire l'informazione semantica associata alle immagini; sono infatti frequenti le applicazioni di recupero dati e indicizzazione delle immagini che richiedono di individuare all'interno di un grande database tutte le immagini che contengono un determinato oggetto o di descrivere il contenuto di un'immagine nuova (Pan et al. 2004).

L'*Image Captioning* è il *task* multimodale che ha lo scopo di assegnare automaticamente una didascalia in linguaggio naturale ad un'immagine (Hirschman et al. 2001 e Rogers et al. 2021). Assegnare una didascalia all'immagine richiede di riconoscere gli oggetti importanti, i loro attributi e le loro relazioni al suo interno e di generare frasi sintatticamente e semanticamente corrette (Hossain et al. 2019). Ad oggi, l'*Image Captioning* si basa su sistemi di *deep learning* che sono in grado di gestire le complessità del *task* e di integrare sistemi di NLP e CV (Hossain et al. 2019).

Un *task* multimodale che si sviluppa successivamente all'*Image Captioning* e che allo stesso modo combina NLP e CV è il *Visual Question Answering* (VQA), che nasce

dall'esigenza del NLP di estendere il QA testuale ad informazioni visive di supporto e dall'esigenza della CV di arricchire l'informazione semantica delle immagini.
Come indicato all'inizio del capitolo, il *task* di VQA viene proposto per fornire una risposta accurata in linguaggio naturale date un'immagine e una domanda sull'immagine formulata in linguaggio naturale (Antol et al. 2015).

La proposta di Antol et al. è che il VQA venga utilizzato in forma libera (*free-form*) e aperta (*open*), con domande in forma libera e aperte nel linguaggio naturale al fine di simulare al meglio scenari del mondo reale (Antol et al. 2015).

Ci sono molte potenziali applicazioni reali per il VQA. L'applicazione più immediata è l'aiuto alle persone non vedenti e ipovedenti: il VQA permetterebbe loro di ottenere attivamente informazioni sulle immagini sia sul web che nel mondo reale. Per esempio, mentre un utente non vedente scorre il suo feed dei social media, un sistema di didascalie può descrivere l'immagine e poi l'utente può usare il VQA per interrogare l'immagine e ottenere maggiori informazioni sulla scena (Kafle et al. 2017). Un esempio concreto è VizWiz, un'applicazione per dispositivi mobili che ha come target ipovedenti e offre in tempo quasi reale delle risposte a delle *visual question* (Bigham et al. 2010).
Più in generale, il VQA può essere usato al fine migliorare l'interazione uomo-macchina verso un modo naturale di interrogare il contenuto visivo. Un sistema VQA può anche essere usato per il recupero delle immagini, senza usare meta-dati o tag dell'immagine. Per esempio, per trovare tutte le immagini scattate in un ambiente piovoso, si può semplicemente interrogare con la domanda "Sta piovendo?" il dataset delle immagini (Kafle et al. 2017).

Il VQA si rivela inoltre interessante poiché coinvolge *task* diversi appartenenti al campo dell'IA e fornisce un equilibrio tra il progresso nello stato dell'arte e l'essere abbastanza accessibile per iniziare a ottenere miglioramenti (Agrawal, J. Lu et al. 2016).

Il *task* è ideato per rispondere a domande aperte che richiedono un insieme vasto di capacità IA per rispondere, e in particolare (Agrawal, J. Lu et al. 2016):

- Riconoscimento di oggetti a grana fine (ad esempio, "Che tipo di formaggio c'è sulla

pizza?");

- Rilevamento di caratteristiche degli oggetti (ad esempio, "Di che colore sono i suoi occhi?", "Di cosa sono fatti i suoi baffi?", "Questa persona aspetta compagnia?");
- Rilevamento di attività (ad esempio, "Cosa sta facendo quell'uomo?", "Quell'uomo sta mangiando?");
- Ragionamento sulla base della conoscenza (ad esempio, "Questa è una pizza vegetariana?");
- Ragionamento di senso comune (ad esempio, "Questa persona ha una vista perfetta?", "Questa persona sta aspettando compagnia?").

Il VQA eredita, quindi, la forma delle domande aperte dal QA testuale, e mutua anche tre varianti nelle risposte alle domande:

- Domande a risposta binaria ("sì" o "no") (Agrawal, J. Lu et al. 2016);
- Domande a risposta multipla, che richiedono un algoritmo per scegliere da una lista predefinita di risposte candidate (Agrawal, J. Lu et al. 2016);
- Domande a risposta aperta, che richiedono una risposta libera (Agrawal, J. Lu et al. 2016 e Hudson et al. 2019).

Rispetto al QA testuale l'aggiunta delle immagini comporta un importante aumento di complessità, poiché le immagini sono entità multidimensionali e tipicamente più rumorose del testo puro. Inoltre, le immagini mancano della struttura e delle regole grammaticali del linguaggio, e non esiste un equivalente diretto agli strumenti NLP, come i *parser sintattici* e la corrispondenza delle espressioni regolari. Infine il linguaggio naturale presenta un livello più alto di astrazione rispetto alle immagini che catturano maggiormente il mondo reale. Per esempio, confrontando la frase "un cappello rosso" con la molteplicità delle rappresentazioni che si possono immaginare, ci si può rendere conto che molti stili non potrebbero essere descritti in una breve frase (Wu, Teney et al. 2016).

Per quanto riguarda le differenze tra i *task* di CV e il VQA, la principale è che la domanda a cui rispondere non è determinata fino al momento dell'esecuzione. Nei problemi tradizionali di CV, come l'*Object Detection*, la singola domanda a cui un algoritmo deve rispondere è predeterminata e cambia solo l'immagine di input. Nel VQA, invece, la forma che la domanda assume non è nota, così come l'insieme delle operazioni richieste per rispondere. In questo senso, il VQA riflette più da vicino la sfida della comprensione generale delle immagini (Kafle et al. 2017). Inoltre il VQA, si rivela più complesso poiché è collegato al compito di rispondere a domande testuali, in cui la risposta deve essere trovata in una specifica narrazione testuale (implica la comprensione della lettura) o in grandi basi di conoscenza (implica il recupero delle informazioni).

Anche se l'*Image Captioning* è il primo *task* multimodale che porta all'unione dei campi di NLP e CV (Veit et al. 2016 e Thomee et al. 2015) e ha portato allo sviluppo di strumenti efficaci per l'apprendimento congiunto da immagini e testo per formare rappresentazioni di livello superiore (Wu, Teney et al. 2016), il VQA presenta una serie di sfide ulteriori.

La risposta a domande visive è un problema significativamente più complesso, poiché richiede spesso informazioni che non sono presenti nell'immagine e variano dal senso comune alla conoscenza enciclopedica su un elemento specifico. Rispetto al compito di *image captioning*, in cui un algoritmo deve generare una descrizione testuale in forma libera per una data immagine, il VQA può coinvolgere una gamma più ampia di conoscenze e capacità di ragionamento (Wu, Teney et al. 2016). Infatti, se un algoritmo di *image captioning* ha la libertà di scegliere le descrizioni più facili e pertinenti dell'immagine, per rispondere a una domanda è necessario trovare la risposta corretta a quella domanda specifica (Wu, Wang et al. 2016). Inoltre, gli algoritmi per il VQA sono tenuti a rispondere a tutti i tipi di domande che le persone potrebbero fare sull'immagine, alcune delle quali potrebbero essere pertinenti al contenuto dell'immagine, come "Quali libri ci sono sotto la televisione?" o "Qual è il colore della barca?", mentre altre potrebbero richiedere conoscenze o ragionamenti al di là del contenuto dell'immagine, come "Perché il bambino sta piangendo?" o "Quale sedia è la più costosa?" (Wu, Wang et al. 2016).

A differenza dell'*image captioning*, il VQA offre, inoltre, il vantaggio di una metrica

di valutazione più semplice. Mentre per l'*image captioning* la valutazione automatica è ancora un problema di ricerca difficile e aperto poiché le didascalie *gold* delle immagini sono lunghe e più difficili da confrontare con quelle previste, le risposte tipiche del VQA tendono a consistere di poche parole. Le domande sulle immagini, infatti, tendono spesso a riguardare informazioni specifiche e sono generalmente sufficienti semplici risposte composte da una a tre parole. In questi scenari, un algoritmo può essere valutato facilmente in base al numero di domande a cui risponde correttamente (Agrawal, J. Lu et al. 2016).

I due punti caratterizzanti del VQA sono dunque la multimodalità ottenuta dalla combinazione di NLP e CV e la possibilità di fornire metriche di valutazione automatica precise poiché molte risposte aperte contengono solo poche parole o un insieme chiuso di risposte che possono essere fornite in un formato a scelta multipla.

Un *task* di IA completo ideale dovrebbe (Agrawal, Batra e Parikh 2016): (i) richiedere una conoscenza multimodale al di là di un singolo sottodominio e (ii) avere una metrica di valutazione quantitativa ben definita per tracciare i progressi.
Il VQA costituisce, in questo senso, un *task* di IA completo: da un lato richiede una conoscenza multimodale al di là di un singolo sottodominio, dall'altro fornisce un metodo per valutare i progressi verso sistemi IA capaci di ragionamenti avanzati che combinano il linguaggio e la comprensione delle immagini (Agrawal, Batra e Parikh 2016).

1.2 Risorse e strumenti per il *Visual Question Answering*

In questa sezione vengono presentati e descritti i principali dataset esistenti per il VQA e le loro criticità. Viene compiuta una descrizione più dettagliata del dataset GQA (Hudson et al. 2019) che viene utilizzato per gli esperimenti presentati nel capitolo 2. Vengono anche descritti lo stato dell'arte della traduzione automatica e in che modo si può utilizzare per colmare la mancanza di risorse in lingue diverse dall'inglese.

1.2.1 Dataset di *Visual Question Answering*

Per la ricerca sul VQA vengono proposti numerosi dataset specifici che contengono, come minimo, delle triple composte da un'immagine, una domanda e la risposta corretta corrispondente. Talvolta vengono fornite annotazioni aggiuntive, come didascalie dell'immagine, regioni dell'immagine che supportano le risposte, o risposte candidate alla scelta multipla (Wu, Teney et al. 2016).

I dataset e le domande al loro interno variano ampiamente sulla base della loro complessità e della quantità di ragionamento e di informazioni non visive (ad esempio, conoscenze di senso comune) necessarie per fornire la risposta corretta. Come proposto da Wu et al. (2016), i dataset vengono classificati in base al tipo di immagini che contengono, come naturali (Agrawal, J. Lu et al. 2016, Krishna et al. 2017, Hudson et al. 2019 e Veit et al. 2016), clipart (Agrawal, J. Lu et al. 2016) o sintetiche. Le caratteristiche chiave sono (Wu, Teney et al. 2016):

- Fonte dell'immagine;
- Numero di immagini;
- Numero di domande;
- Numero di domande e numero di immagini;
- Numero di categorie per domanda;
- Collezioni di domande;
- Lunghezza media di una domanda;
- Lunghezza media di una risposta;
- Metriche di valutazione.

Un dataset viene tipicamente usato sia per l'addestramento sia per la valutazione di un sistema VQA.

La natura aperta del compito suggerisce, tuttavia, che altre fonti di informazione su larga scala potrebbero essere utili e probabilmente necessarie per addestrare sistemi VQA. Alcuni dataset, infatti, affrontano specificamente questo aspetto attraverso annotazioni di elementi di supporto in basi di conoscenza strutturate non visive (Krishna et al. 2017 e Hudson et al. 2019).

Di seguito vengono presentati alcuni tra i principali dataset di GQA che presentano caratteristiche diverse, riportate nella tabella 1.1.

Dataset	COCO QA	VQA-real	VQA-abstract	VQA 2.0	Visual Genome	Visual7w	GQA
Fonte immagini	COCO	COCO	Clipart	COCO	COCO + Flikr	COCO	VG + Crowd sourcing
Tot. immagini	117.648	204.721	50.000	204.721	108.000	47.300	113.018
Tot. domande	117.648	614.163	150.000	1M	1.445.322	327.939	22M
Num. domande / Num. immagini	1.0	3.0	3.0	4.88	13.4	6.9	194.67
Categorie domande	4	20+	20+	20+	7	7	10
Collezione domande	Automatica	Umana	Umana	Automatica	Umana	Umana	Automatica
Lunghezza media domande	8.6	6.2	6.2	6.2	5.7	6.9	7.9
Lunghezza media risposte	1.0	1.1	1.0	-	1.1	1.1	-
Metriche valutazione	Acc.	Acc contro 10 persone	Acc.	Acc.	Acc.	Acc.	Consistenza + Distribuzione + Validità e plausibilità + Distribuzione

Tabella 1.1: Caratteristiche dei dataset presentati. La tabella è ripresa e modificata da Wu, Teney et al. (2016), pag. 11.

MS COCO

MS COCO (Microsoft COCO) è un dataset su larga scala per l'*Object Detection*, *Object Segmentation* e il *Captioning* che include 328.000 immagini reali (suddivise in *test* e *train*) contenenti 91 categorie di oggetti distinti, 82 delle quali hanno più di 5.000 istanze all'interno del dataset (Veit et al. 2016).

Dal momento che questo dataset nasce per il *Captioning* e prevede annotazioni di didascalie e non di domande, viene esteso da COCO-QA creato da Ren, Kiros e Zemel (2015) che annota automaticamente 123.287 immagini (72.783 per il *train* e 38.948 per il *test*) con una coppia domanda/risposta trasformando le descrizioni delle immagini del dataset COCO originale in forma di domanda/risposta (Veit et al. 2016). Le domande sono categorizzate in quattro tipi in base al genere di risposta attesa: oggetto, numero, colore e posizione (Veit et al. 2016). Un effetto collaterale della conversione automatica delle didascalie è un alto tasso di ripetizione delle domande (Veit et al. 2016). Tra le 38.948 domande del *test set*, 9.072 appaiono anche come domande di *training* (M. Ren et al. 2015). Il dataset MS COCO è la principale fonte di immagini reali per i dataset di VQA.

VQA

VQA è un dataset suddiviso in due parti: *VQA-real*, che contiene immagini naturali, e *VQA-abstract*, che contiene clipart (Agrawal, J. Lu et al. 2016).

VQA-real comprende 123.287 immagini di *training* e *validation* e 81.434 immagini di *test* provenienti da MS COCO (Veit et al. 2016).
Le domande sono *open ended* e risultano in un insieme di risposte possibili diverse. Vengono annotate manualmente e gli annotatori vengono incoraggiati a fornire domande diverse. Molte risposte sono binarie (cioè "sì" o "no"), ma altre domande richiedono una risposta sotto forma di breve frase (Agrawal, J. Lu et al. 2016).
Per il *testing*, oltre alle domande *open ended*, vengono valutate risposte a scelta multipla,

fornendo 18 risposte candidate per ogni domanda (Agrawal, J. Lu et al. 2016).
Complessivamente, *VQA-real* contiene 614.163 domande e 7.984.119 risposte fornite dagli annotatori. Il vantaggio di fornire più di una risposta è quello di consentire una valutazione più bilanciata e comprendere forme diverse per una stessa risposta (Agrawal, J. Lu et al. 2016).

VQA-abstract contiene clipart annotate con coppie di domande/risposte come *set* separato e complementare alle immagini reali (Agrawal, J. Lu et al. 2016). Le clipart vengono create manualmente e riproducono scene realistiche, poiché l'obiettivo è quello di permettere una ricerca focalizzata sul *reasoning* di alto livello, eliminando la necessità di analizzare le immagini reali (Agrawal, J. Lu et al. 2016). Sono possibili due tipi di scene, interne ed esterne, ognuna delle quali permette un diverso set di elementi, inclusi animali, oggetti e persone con pose regolabili (Agrawal, J. Lu et al. 2016).
Vengono generate un totale di 50.000 scene, e vengono raccolte 3 domande per scena per un totale di 150.000 domande, in modo simile a quello delle immagini reali del dataset VQA (Agrawal, J. Lu et al. 2016). Ad ogni domanda rispondono 10 soggetti che forniscono anche un punteggio di fiducia. Le domande sono etichettate con un tipo di risposta: "sì/no", "numero" e "altro" (Agrawal, J. Lu et al. 2016).

Visual Genome

Visual Genome è un dataset contenente 108.077 immagini reali ottenute dall'intersezione di Flikr (Thomee et al. 2015) e MS COCO (Veit et al. 2016).
Questo dataset contiene 1.7 milioni di coppie domanda-risposta ed è il primo dataset a includere annotazioni strutturate dei contenuti delle immagini sotto forma di *scene graph* che descrivono gli elementi visivi delle scene, i loro attributi e le loro relazioni (Krishna et al. 2017).

Uno *scene graph* fornisce una rappresentazione formalizzata dell'immagine: ogni nodo denota un oggetto, ovvero un'entità visiva all'interno dell'immagine, come per esempio una persona, una mela, dell'erba o delle nuvole. Ogni oggetto è collegato a una *bounding*

box che specifica la sua posizione e dimensione, ed è annotato con circa 1-3 attributi che identificano proprietà dell'oggetto, per esempio il suo colore, forma, materiale o l'attività che sta svolgendo (Krishna et al. 2017). Gli oggetti sono collegati da archi di relazione, che possono rappresentare azioni (ad esempio verbi) o relazioni spaziali (ad esempio preposizioni) e comparative (Krishna et al. 2017).
Ogni immagine contiene la descrizione delle regioni che descrivono una porzione localizzata dell'immagine (Krishna et al. 2017).
Ogni regione è convertita in una rappresentazione del *region graph* di oggetti, attributi e relazioni a coppie e ogni *region graph* viene combinato per formare un *scene graph* con tutti gli oggetti legati all'immagine, come illustrato nella figura 1.1 (Krishna et al. 2017).

I *scene graph* sono annotati in linguaggio naturale. Le domande sono state composte manualmente e (*i*) iniziano con una delle "six Ws" ("who", "what", "where", "when", "why e "how"), (*ii*) sono non ambigue, (*iii*) sono precise e (*iv*) correlano la domanda all'immagine in modo che per fornire una risposta sia necessario guardare l'immagine (Krishna et al. 2017).
Vengono raccolti due tipi di coppie di domande e risposte (QA): QA *free-form* e QA basate su regioni (Krishna et al. 2017). Per le QA *free-form*, l'annotatore fornisce 8 coppie domanda/risposta usando 3 diverse "W" per avere una diversità, mentre per le QA basate su regioni, l'annotatore deve fornire domande/risposte relative a una specifica regione dell'immagine (Krishna et al. 2017).
Un grande vantaggio del dataset Visual Genome per VQA è il potenziale che offre l'utilizzo delle annotazioni strutturate delle scene che forniscono informazioni per aiutare la progettazione e l'addestramento dei sistemi VQA (Krishna et al. 2017).

Zhu et al. (2016) creano, a partire da un sottoinsieme di Visual Genome, il dataset Visual7w inserendo annotazioni aggiuntive. In particolare, viene stabilito un legame semantico tra le descrizioni testuali e le regioni delle immagini corrispondenti attraverso un *object-level grounding* per risolvere problemi di ambiguità dovuti alla coreferenza e comprendere al meglio la distribuzione degli oggetti (Zhu et al. 2016). La valutazione avviene su domande a risposta multipla in cui è necessario sceglie tra quattro risposte candidate

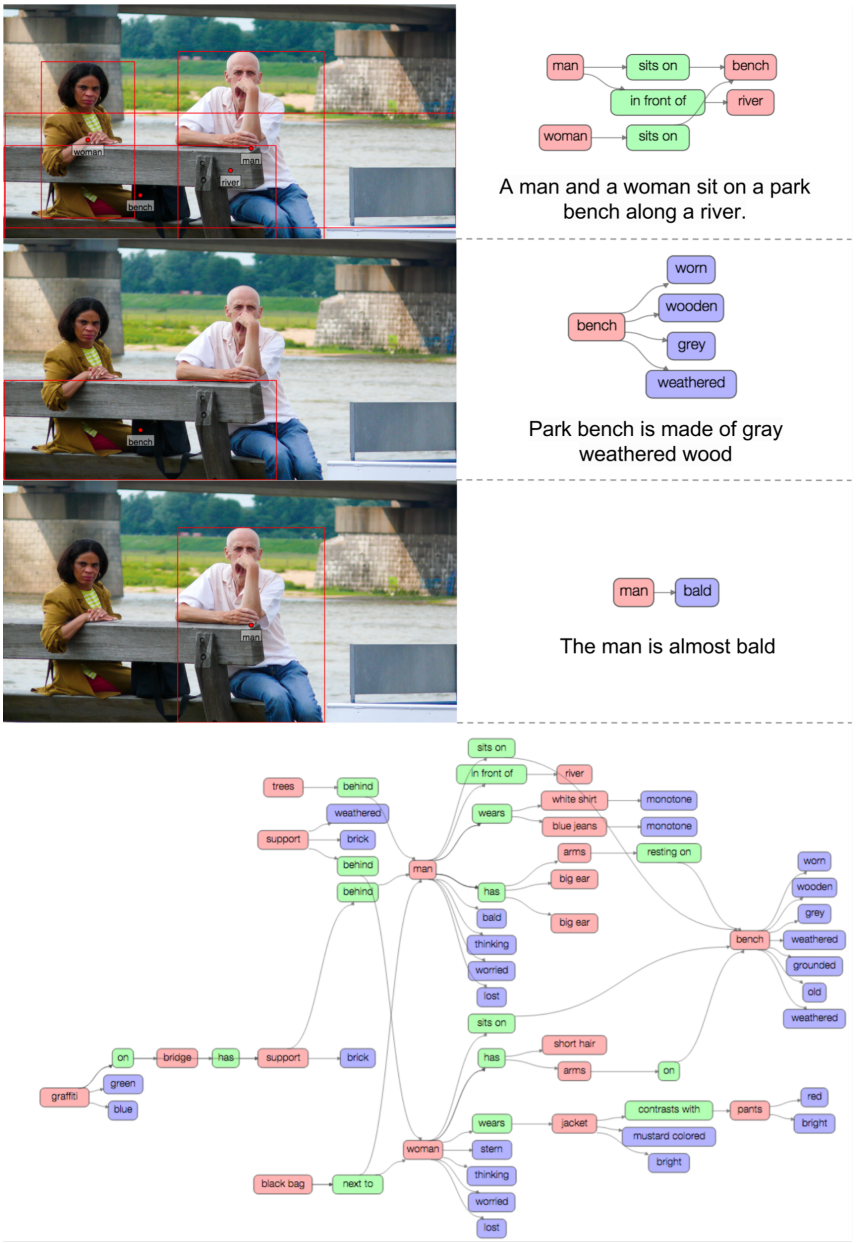

Figura 1.1: Esempio di *scene graph* creato a partire da *region graph* (Krishna et al. 2017, pag. 4).

(Zhu et al. 2016).

Bias nei dataset

Numerosi studi dimostrano che i dataset di VQA tendono a presentare *bias* nella suddivisione dei dati in *training* e *test* o nella distribuzione delle domande e delle risposte; questo fa sì che i modelli dipendano da queste tendenze statistiche piuttosto che da una reale comprensione della scena e che si dimostrino poco capaci di generalizzare su nuove istanze (Hudson et al. 2019, Agrawal, Batra e Parikh 2016, P. Zhang et al. 2016, Goyal et al. 2017 e B. Zhao et al. 2020).

Il *task* di VQA, con la sua multimodalità, dà vita a una serie di problemi legati all'intersezione tra visione e linguaggio di notevole importanza sia dal punto di vista della ricerca, sia per le numerose applicazioni. I *bias* nel linguaggio tendono ad essere un segnale più semplice per l'apprendimento automatico rispetto alle modalità visive e questo ha come risultato che i modelli tendano ad ignorare le informazioni visive, portando ad una "percezione gonfiata" della loro efficienza (Agrawal, J. Lu et al. 2016).
Diversi studi (P. Zhang et al. 2016, Zhou et al. 2015, Agrawal, Batra e Parikh 2016 e Agrawal, Batra, Parikh e Kembhavi 2018) evidenziano infatti come il linguaggio fornisca segnali superficiali che possono portare a buone prestazioni, senza che i modelli sottostanti comprendano veramente il contenuto visivo.

Agrawal et al. (2016) confrontano due modelli, uno con meccanismi di attenzione e uno senza, e il modello vincitore della VQA Challenge 2016 (MCB) (Fukui et al. 2016). I modelli presi in analisi hanno un'accuratezza tra il 50% e il 60% sul *validation* set di VQA. I comportamenti dei tre modelli vengono misurati (*i*) sulla loro capacità di generalizzare su nuove istanze e in particolare su coppie immagine-domanda del *test* set diverse da quelle del *training*, (*ii*) sulla comprensione completa della domanda, per investigare se la domanda viene analizzata nel complesso o vengono prese in considerazione solo le prime parole, e (*iii*) sulla comprensione completa dell'immagine, per analizzare se le predizioni dei modelli cambiano per una data domanda su immagini diverse. L'analisi rivela che,

nonostante i progressi, i modelli di VQA sono "myopic", ovvero tendono a fallire su istanze sufficientemente nuove, "jump to conclusions", ovvero convergono su una risposta predetta dopo aver analizzato solo la prima parte della domanda, e sono "stubborn", ovvero non cambiano la risposta tra le immagini (Agrawal, Batra e Parikh 2016).

Questo dimostra che i modelli tendono a basare le predizioni su correlazioni superficiali, senza tenere in considerazione le relazioni tra gli elementi visualizzati nell'immagine. Ciò porta a errori nella risposta a domande su elementi che sono meno frequenti nei dati di addestramento, anche se essi sono ben annotati (Goyal et al. 2017, J. Johnson et al. 2017 e Agrawal, Batra, Parikh e Kembhavi 2018).

Zhang et al. (2016) analizzano il dataset di VQA notando come nel *training* set siano presenti importanti *bias*: la risposta più comune alle domande sullo sport è "tennis", che è corretta per il 41% delle domande che iniziano con "What sport is"; allo stesso modo, "white" è corretto per il 23% delle domande che iniziano per "What color are the" e "2" è corretto per il 39% delle domande che iniziano con "How many".
Quasi metà delle domande nel dataset di VQA può ricevere una risposta corretta da una rete neurale che ignora completamente l'immagine e usa solo la domanda basandosi su regolarità sistematiche nei tipi di domande che vengono poste e sulle risposte che tendono ad avere. Questo si verifica anche per domande binarie, dove la risposta è "yes" o "no". Il 68% delle domande binarie possono, infatti, ricevere una risposta corretta rispondendo semplicemente "yes".
Inoltre, una rete neurale che si basi solo sul linguaggio potrebbe rispondere correttamente al 78% delle domande binarie senza tenere in considerazione l'immagine ma basandosi solo sul testo; questo può dare la falsa impressione che un sistema stia facendo progressi verso l'obiettivo di comprendere le immagini correttamente (P. Zhang et al. 2016).

Per risolvere i *bias* presenti in VQA vengono proposte nuove segmentazioni di *test* e *training* set dei dataset esistenti e *augmentation* dei dati visivi (P. Zhang et al. 2016, Goyal et al. 2017 e Agrawal, Batra, Parikh e Kembhavi 2018).

Zhang et al. (2016) prendono in considerazione il problema delle domande binarie e implementano il dataset di VQA con delle scene astratte complementari, in modo che quasi tutte le domande nel dataset bilanciato abbiano una risposta "sì" per una scena e "no" per una il più simile possibile. Propongono, inoltre, un approccio che estrae un riassunto conciso della domanda in forma di tupla, identifica la regione nella scena su cui dovrebbe concentrarsi, e verifica l'esistenza del concetto visivo descritto nella tupla per rispondere, prestando attenzione alle parti importanti di una domanda (P. Zhang et al. 2016).
Sebbene questo approccio sia un importante passo in avanti per la risoluzione dei *bias* nei dataset di VQA, le immagini utilizzate sono clipart e vengono prese in considerazione solo risposte binarie.

Goyal et al. (2017) estendono l'approccio precedente alle immagini reali del dataset di VQA per bilanciare domande e risposte proponendo una nuova versione di VQA, VQA 2.0. Per bilanciare il dataset VQA esistente vengono raccolte immagini complementari in modo che quasi ogni domanda nel dataset bilanciato sia associata non a una singola immagine, ma a una coppia di immagini simili che risultano in due risposte diverse alla domanda (una domanda ha due risposte diverse per due immagini diverse). Il risultato è un set di dati VQA più bilanciato che consiste in più di 443K coppie domanda-immagine per il *train*, 214K per il *validation* e 453K per il *test*. L'ipotesi è che un dataset bilanciato costringa i modelli a concentrarsi sull'informazione visuale.
Gli autori valutano inoltre i modelli VQA allo stato dell'arte sul dataset bilanciato, mostrando che i modelli addestrati sull'esistente dataset VQA "sbilanciato" si comportano male sul nuovo dataset bilanciato. Questo risultato conferma l'ipotesi che questi modelli sfruttano i bias linguistici nel dataset VQA esistente per ottenere una maggiore precisione (Goyal et al. 2017).

Agrawal et al. (2018) presentano una nuova suddivisione dei dataset di VQA 1.0 e 2.0 chiamate *Visual Question Answering under Changing Priors* (rispettivamente VQA-CP v1 e VQA-CP v2), create riorganizzando le suddivisioni di *train* e *validation* dei rispettivi dataset VQA in modo tale che la distribuzione delle risposte per tipo di domanda ("How

many", "What color is", ecc.) sia diversa nella suddivisione *test* rispetto a quella del *train*.
La scoperta chiave è che le prestazioni di tutti i modelli esistenti testati calano significativamente quando vengono addestrati e valutati sulle nuove suddivisioni rispetto alle suddivisioni originali. Questo risultato fornisce un'ulteriore conferma sul comportamento dei modelli VQA (Agrawal, Batra, Parikh e Kembhavi 2018).
Gli autori propongono anche un nuovo modello *Grounded Visual Question Answering* (GVQA) che contiene *bias* induttivi e restrizioni nell'architettura specificamente progettate per evitare che "imbrogli" affidandosi principalmente ai priori nei dati di addestramento. GVQA è motivato dall'intuizione che le domande in VQA forniscono due informazioni chiave: (*i*) cosa dovrebbe essere riconosciuto o su quali concetti visivi nell'immagine si deve ragionare per rispondere alla domanda (ad esempio, "Di che colore è il piatto?" richiede di guardare il piatto nell'immagine), (*ii*) cosa dovrebbe essere detto o qual è lo spazio delle risposte plausibili (ad esempio, "Di che colore...?" si deve rispondere con nomi di colori). L'ipotesi è che i modelli che non differenziano esplicitamente tra questi due ruoli tendano a confondere questi due segnali e imparino sulle coppie domanda-risposta che un colore plausibile di un piatto è il bianco, e al momento del test, si basano su questa correlazione più che sul piatto specifico nell'immagine di cui si tratta la domanda. GVQA separa esplicitamente il riconoscimento del concetto visivo dalla previsione dello spazio di risposta (Agrawal, Batra, Parikh e Kembhavi 2018).

Un ulteriore problema dei principali dataset di VQA è che utilizzano un linguaggio non composizionale e si basano prevalentemente sull'identificazione delle proprietà degli oggetti piuttosto che sulle relazioni spaziali; questo, insieme ai *bias* presenti (Ferraro et al. 2015), fa sì che non ci sia necessità di considerare la composizionalità del linguaggio e motiva la creazione di dataset che necessitino di *visual reasoning* composizionale (Suhr et al. 2018).
Suhr et al. (2018) propongono NLVR2, ovvero un *corpus* di fotografie pensato per il *Natural Language Visual Reasoning* che si concentra sulla diversità semantica, la composizionalità e sulle sfide del ragionamento visivo. I dati contengono 107.292 esempi di frasi

inglesi abbinate a fotografie del web. Il compito è prevedere se una frase in linguaggio naturale è vera su una coppia di fotografie (Suhr et al. 2018).

Un'altra problematica dei dataset esistenti è che non contengono annotazioni della struttura, del tipo e del contenuto della domanda, che aiuterebbero a risolvere i *bias* e gli errori dei modelli (Hudson et al. 2019).
Johnson et al. (2017) propongono di risolvere questo problema presentando un dataset che testa una serie di abilità di ragionamento visivo attraverso l'annotazione delle informazioni di ogni immagine in modo che esse siano complete facendo sì che fonti esterne di informazione, come la conoscenza del senso comune, non possano aumentare la possibilità di rispondere correttamente alle domande (J. Johnson et al. 2017). In altre parole, si cerca di evitare che il modello risponda a una particolare domanda come "Da cosa è coperto il terreno?" basandosi su dati appresi precedentemente e non sull'osservazione della scena relativa alla domanda (J. Johnson et al. 2017). La presenza di annotazioni riguardanti la struttura, il tipo e il contenuto delle domande può aiutare e correggere le cause alla radice degli errori commessi dai modelli (J. Johnson et al. 2017). Viene controllato il *bias* domanda-condizione e vengono evitate domande apparentemente complesse ma con semplici scorciatoie per la risposta corretta (J. Johnson et al. 2017). Infine, vengono utilizzate rappresentazioni strutturate di *ground-truth* sia per le immagini che per le domande: le immagini sono annotate con le posizioni e gli attributi degli oggetti, e le domande sono rappresentate come programmi funzionali che possono essere eseguiti per rispondere alla domanda. Queste rappresentazioni facilitano analisi approfondite non possibili con i tradizionali set di dati VQA (J. Johnson et al. 2017).

GQA

GQA (Hudson et al. 2019) è un dataset multimodale per il VQA che consiste di 113.018 immagini e circa 22M domande di vario tipo e con diversi gradi di composizione.

Lo scopo principale di GQA è quello di risolvere alcuni dei problemi presenti negli altri dataset, illustrati in 1.2.1, come la presenza di bias linguistici che fanno sì che i modelli siano dipendenti da correlazioni statistiche nei dati testuali piuttosto che da reali capacità

di comprensione della scena, la mancanza di composizionalità linguistica e la mancanza di una semantica chiara e della struttura delle domande.

In GQA, le immagini e le coppie domanda-risposta sono accompagnate da rappresentazioni semantiche corrispondenti: ogni immagine è annotata con un *dense scene graph*, che rappresenta gli oggetti, gli attributi e le relazioni che contiene, ogni domanda è associata a un programma funzionale che elenca la serie di passaggi di ragionamento che devono essere fatti per arrivare alla risposta e ogni risposta è accompagnata da giustificazioni sia testuali che visive, che indicano la regione corrispettiva all'interno dell'immagine (Hudson et al. 2019).

Questi elementi consentono di bilanciare al meglio la distribuzione di domande e risposte (Hudson et al. 2019).

I *scene graph* vengono ricavati dalle annotazioni dei *Visual Genome scene graph*, descritti nella sezione 1.2.1 e rappresentati nella figura 1.1, che vengono normalizzati al fine di convertire le annotazioni formulate in linguaggio naturale *free-form* in un'ontologia semantica non ambigua (Hudson et al. 2019). L'ontologia finale di GQA contiene 1.740 oggetti, 620 attributi e 330 relazioni distinti, raggruppati in una gerarchia che consiste di 60 diverse categorie e sottocategorie (Hudson et al. 2019).

Per la generazione delle domande viene utilizzato un *question engine*, non solo per evitare *bias* e fare sì che il dataset sia bilanciato, ma anche per fare in modo che le domande richiedano capacità di ragionamento, comprensione spaziale e inferenza *multi-step*[1].

Il *question engine* sfrutta (Hudson et al. 2019): (*i*) il contenuto, ovvero le informazioni su oggetti, attributi e relazioni fornite attraverso i *Visual Genome scene graph*, e (*ii*) la struttura, ovvero una grammatica che accoppia centinaia di *pattern* strutturali e risorse semantiche lessicali dettagliate (informazioni su oggetti, attributi e relazioni create a partire dal dataset VQA). La combinazione dei due elementi risulta in più di 22 milioni di domande diverse dotate di rappresentazioni strutturate sotto forma di programmi funzionali che ne specificano il contenuto e la semantica, fondate visivamente nei *scene graph*

[1] *GQA: Visual Reasoning in the Real World.* Consultato in data 05 settembre 2021. https://cs.stanford.edu/people/dorarad/gqa/about.html.

delle immagini (Hudson et al. 2019). Il dataset dispone di circa 1.7 milioni di domande sia binarie sia aperte che vengono bilanciate applicando tecniche di *smoothing* che rendono omogenee le risposte per i gruppi di domande (Hudson et al. 2019). Inoltre, le domande si concentrano sia sull'aspetto semantico e linguistico sia sull'inferenza multimodale (Hudson et al. 2019).

Il *question engine* opera su 524 pattern di domande che sono articolati in 117 gruppi; ogni gruppo è associato a 3 componenti: (*i*) un programma funzionale che rappresenta la sua semantica, (*ii*) un set di frasi testuali che la esprimono in linguaggio naturale, e (*iii*) una coppia di risposte corte e lunghe (Hudson et al. 2019).

I *pattern* sono ricavati a partire da 250 ricavati manualmente con l'aggiunta di 274 che derivano da VQA 1.0 (Hudson et al. 2019). Ogni *pattern* è associato a una rappresentazione strutturata sotto forma di un programma funzionale. Per esempio, la domanda "What color is the apple on the white table?" è semanticamente equivalente al programma: *select*: table → *filter*: white → *relate(subject,on)*: apple → *query*: color (Hudson et al. 2019, pag. 6). Queste rappresentazioni semanticamente non ambigue offrono numerosi vantaggi rispetto alle domande *free-form* in linguaggio naturale (Hudson et al. 2019). In primo luogo facilitano la diagnosi del successo o del fallimento dei metodi di VQA consentendo di dissezionare la loro *performance* su diversi assi (tipo, topologia, lunghezza della domanda ecc.). In secondo luogo, semplificano il bilanciamento della distribuzione del dataset al fine di evitare *bias* linguistici. Infine, consentono di identificare le relazioni di implicazione e di equivalenza tra domande diverse, soprattutto per quanto riguarda le domande logiche (Hudson et al. 2019).

Il dataset viene bilanciato attraverso la rappresentazione funzionale semantica: a ogni domanda viene assegnata un'etichetta globale che riguarda il tipo di risposta, e un'etichetta locale che considera il/i soggetto/i delle domande (Hudson et al. 2019).

Il tipo strutturale di una domanda deriva dall'operazione finale nel programma funzionale delle domande e può essere (Hudson et al. 2019): (*i*) di verifica per le domande binarie con risposta sì o no, (*ii*) di *query* per tutte le domande aperte, (*iii*) di scelta per le domande che presentano due alternative, (*iv*) logico che riguarda l'inferenza logica e (*v*) di confronto

per le domande che confrontano due o più oggetti.
Il tipo semantico si riferisce invece al principale soggetto della domanda (Hudson et al. 2019): (*i*) oggetto per le domande di esistenza, (*ii*) attributo, considera le proprietà o la posizione di un oggetto, (*iii*) categoria, considera l'identificazione delle proprietà o della posizione di un oggetto, (*iv*) relazione, per le domande riguardo il soggetto o l'oggetto della relazione descritta, e (*v*) globali, riguardano le proprietà generali della scena come il meteo o il luogo.
Attraverso queste etichette le domande vengono suddivise in gruppi e viene fatto uno *smoothing* della distribuzione delle risposte (Hudson et al. 2019).

Il dataset di GQA riesce quindi ad offrire da un lato rappresentazioni semantiche chiaramente definite e nitide, e dall'altro la ricchezza semantica e visiva delle immagini del mondo reale. Riesce inoltre a sviluppare un metodo efficace per generare un ampio numero di domande semanticamente varie, che sposa le rappresentazioni dei *scene graph* con i metodi linguistici computazionali (Hudson et al. 2019).

1.2.2 *Neural Machine Translation* per colmare la mancanza di risorse in italiano

La creazione di dataset annotati per il VQA è molto costosa e fino ad oggi la ricerca si è concentrata soprattutto su domande del linguaggio naturale formulate in inglese (Wu, Teney et al. 2016); di conseguenza, addestrare un modello a produrre risposte in un'altra lingua si rivela complesso.

Per quanto riguarda il VQA multilingue vengono fatti alcuni studi (D. Gupta et al. 2020), ma nessuno si concentra sull'italiano.

Recentemente, i metodi di traduzione neurale automatica, e in particolare quelli basati su architetture *Transformer* (Firat et al. 2017), hanno raggiunto risultati molto promettenti nel risolvere *task* diversi (Bojar, Chatterjee, Federmann, Fishel et al. 2019).

Dal momento che non esistono risorse e dati in italiano per il VQA, al fine di poter rispondere a domande in italiano sulle immagini una possibile strada da percorrere è quella di sfruttare i progressi ottenuti dai sistemi di traduzione automatica per utilizzare dataset e modelli che raggiungono lo stato dell'arte per l'inglese adattandoli all'italiano.

La *Machine Translation* (MT) è la traduzione automatica da una lingua ad un'altra utilizzando un *computer software* (Cheng 2019) e include una grande varietà di attività *computer-based* che riguardano la traduzione (Somers 1996). In un'era digitale si rivela fondamentale la possibilità di utilizzare sistemi di traduzione automatica che vengono utilizzati da milioni di persone ogni giorno, online o con applicazioni mobili, per comunicare superando le barriere linguistiche (Datta et al. 2020). Le applicazioni della MT sono numerose e spaziano dal supporto ai traduttori all'aiuto nella didattica multilingue (Way 2018).

Lo sviluppo recente dell'IA, dovuto all'applicazione di reti neurali, porta all'utilizzo di sistemi di *training end-to-end* basati sulle reti neurali come paradigma anche per la MT (Cheng 2019). Prendono piede, infatti, sistemi di *Neural Machine Translation* (NMT) che mirano a costruire una singola rete neurale che può essere sintonizzata congiuntamente per massimizzare le prestazioni di traduzione che è basata sul *framework encoder-decoder* (Bahdanau et al. 2014). Inizialmente vengono utilizzate principalmente *Recurrent Neural network* (RNN) con un *encoder* sotto forma di *bidirectional recurrent neural network* per codificare una frase in input da tradotte e un *decoder* sotto forma di RNN per predire le parole nella seconda lingua (Datta et al. 2020), e vengono utilizzati meccanismi di attenzione (Tu et al. 2016) per consentire al *decoder* di concentrarsi su parti diverse dell'input (Bahdanau et al. 2014 e Bojar, Chatterjee, Federmann, Graham et al. 2016).
Ad oggi, lo stato dell'arte viene raggiunto da sistemi che utilizzano un'architettura *Transformer* e che si rivelano molto efficienti su numerosi *task* (Firat et al. 2017 e Bojar, Chatterjee, Federmann, Fishel et al. 2019).

Anche se Hassan et al. (2018) sostengono che la NMT possa ottenere traduzioni di

qualità paragonabile a quelle operate da professionisti e addirittura superiori a quelle di non professionisti (Hassan et al. 2018), uno studio successivo di Toral et al. (2018) dimostra che la NMT non ottiene ancora una qualità pari alla traduzione umana (Toral et al. 2018).

La valutazione automatica dei risultati dei sistemi di MT si basa generalmente sul confronto tra la traduzione automatica e quella umana (Way 2018, Papineni et al. 2002 e Cer et al. 2010). La metrica di valutazione automatica più utilizzata è BLEU (*Bilingual Evaluation Understudy*) (Papineni et al. 2002, Coughlin 2003, Cer et al. 2010, Tavosanis 2019, Croce et al. 2020), che valuta le frasi tradotte automaticamente confrontandole con una o più traduzioni manuali e assegna un punteggio da 0 a 1 sulla base della loro similarità (1 è la totale coincidenza con la traduzione manuale presa come riferimento e 0 la totale differenza).

Le metriche di valutazione automatica vengono ampiamente utilizzate, ma la valutazione manuale si rivela spesso la soluzione migliore poiché consente di eseguire alcuni compiti che sono ancora al di fuori della portata della NMT. Una ragione a priori per preferire la valutazione umana a quella automatica è che gli esseri umani possono riconoscere tra due traduzioni ugualmente buone, mentre le metriche automatizzate si limitano a semplici confronti tra stringhe con una o più traduzioni di riferimento (Coughlin 2003, Tavosanis 2019): nella valutazione della correttezza della traduzione automatica è talvolta presente uno scarto significativo tra BLEU e la valutazione umana (Callison-Burch, Osborne et al. 2006, Callison-Burch, Fordyce et al. 2008 e Tavosanis 2019 per l'italiano).

Coughlin (2003) evidenzia che spesso le abilità linguistiche dei valutatori umani non vengono utilizzate nella valutazione che richiede un rapido confronto tra due stringhe fuori contesto. I valutatori umani, infatti, compiono un'analisi superficiale come quella di BLEU: nel tentativo di valutare la qualità della traduzione a livello di *corpus*, i risultati indicano che BLEU è un'alternativa altamente affidabile (Coughlin 2003). Ciononostante, Callison-Burch, Fordyce et al. (2008) e Tavosanis (2019) sottolineano l'importanza e l'efficacia della valutazione manuale (Callison-Burch, Fordyce et al. 2008, Tavosanis 2019) e Tavosanis (2019) dimostra che per la traduzione automatica dall'inglese all'italiano è

ancora significativo lo scarto tra BLEU e la valutazione manuale che si rivela più accurata (Tavosanis 2019).

Un approccio esplorato per colmare la mancanza di risorse per *task* multimodali in italiano senza dover creare dataset *ad hoc* è quella di tradurre le annotazioni di una risorsa esistente e addestrare un modello su essa.

Antonio, Croce e Basili (2019) creano MSCOCO-it e MSR-VTT-it, due risorse pensate per il *captioning* di immagini e video in italiano a partire dalla traduzione automatica degli omonimi dataset inglesi (Masotti et al. 2018 e Scaiella et al. 2019). Lo studio dimostra come questo approccio possa portare ad una risorsa che non necessita del costo dell'annotazione manuale e che può essere utilizzata per addestrare modelli ottenendo una buona accuratezza (Scaiella et al. 2019) .

Per quanto riguarda il *task* di VQA, Croce et al. (2021) propongono GQA-it, un dataset per il VQA in italiano creato utilizzando la NMT del dataset inglese (Croce et al. 2021). L'obiettivo del progetto è quello di generare un dataset su larga scala in cui *train* e *validation* set sono ottenuti tramite la traduzione automatica e il *test* set è convalidato manualmente (Croce, Passaro et al. 2021).

1.3 Modelli e baseline

In questa sezione vengono presentati i principali modelli utilizzati per il VQA, suddivisi in base alla loro architettura. In particolare, viene descritto il modello LXMERT (Tan et al. 2019) che è utilizzato per gli esperimenti nel capitolo 2.

Negli ultimi anni vengono proposti numerosi algoritmi di VQA; tutti gli approcci esistenti consistono in (Kafle et al. 2017): (*i*) estrazione delle *feature* dell'immagine (*image featurization*), (*ii*) estrazione delle *feature* della domanda (*question featurization*), e (*iii*) un algoritmo che combina queste *feature* per produrre una risposta.

Per le *feature* dell'immagine, la maggior parte degli algoritmi usa *Convolutional Neural Network* (CNN) pre-addestrati su modelli, come ImageNet (Deng et al. 2009).

Per quanto riguarda il *question featurization* sono esplorate diverse possibilità, tra cui *bag-of-words* (BOW) e codificatori LSTM (Long Short Term Memory).
Per generare una risposta, l'approccio più comune è quello di trattare il VQA come un problema di classificazione. In questo contesto, le *feature* dell'immagine e della domanda sono l'input del sistema di classificazione e ogni risposta unica è trattata come una categoria distinta (Kafle et al. 2017).

In questo capitolo vengono riportati i principali approcci al VQA seguendo la suddivisione riportata in Wu et al. (2016):

- *Joint embedding approaches*: usano reti neurali convoluzionali e ricorrenti;
- *Attention mechanism*: si concentrano su una specifica parte dell'immagine;
- Modelli composizionali: si basano su architetture modulari e comportano la connessione di moduli distinti progettati per ogni problema preso in istanza;
- *Knowledge base-enhanced approaches*: utilizzano dei dati esterni interrogando basi di conoscenza strutturate.

La maggior parte dei modelli, tuttavia, utilizza approcci combinati e appartiene quindi a più categorie.

1.3.1 *Joint embedding approaches*

Dal momento che il *task* di VQA viene proposto quando gli approcci di *deep learning* hanno già raggiunto una grande popolarità grazie alla loro *performance state-of-the-art* in numerosi *task* visivi e di NLP, la maggior parte dei lavori sul VQA si basa su essi (A. K. Gupta 2017).
L'approccio *joint embedding* viene inizialmente esplorato per l'*Image Captioning* ed è motivato dal successo dei metodi di *deep learning* sia nel NLP sia nella *Computer Vision*. Per quanto riguarda l'estrazione delle *feature* delle immagini, la maggior parte degli algoritmi utilizza *Convolutional Neural Network* (CNN) pre-addestrati all'*object detection*.

Le *feature* del testo sono ottenute con *embedding* di parole pre-addestrate su grandi *corpora* di testo. I *word embedding* rappresentano le parole in uno spazio in cui le distanze riflettono le somiglianze semantiche. Gli *embedding* delle singole parole di una domanda sono poi tipicamente dati in input a un *Recurrent Neural Network* (RNN) per catturare modelli sintattici e gestire sequenze di lunghezza variabile (Agrawal, J. Lu et al. 2016 e Kafle et al. 2017).
Anche se questo approccio viene spesso utilizzato dagli algoritmi di VQA, può generare solo risposte osservate nella fase di *training*.
In questi approcci, le *feature* delle immagini e delle domande sono l'input del sistema di classificazione e ogni risposta unica è trattata come una categoria distinta. La combinazione delle *feature* può essere ottenuta (Wu, Teney et al. 2016 e Kafle et al. 2017):

- Combinando le *feature* di immagini e domande utilizzando meccanismi semplici, come la concatenazione ecc., e fornendo la combinazione in input a un classificatore lineare o a un *neural network* (Gao et al. 2015);

- Combinando le *feature* di immagini e domande con metodi di *pooling bilineare* o schemi simili all'interno del *framework* di una rete neurale (Fukui et al. 2016);

- Utilizzando modelli bayesiani per sfruttare le relazioni sottostanti tra le distribuzioni delle *feature* domanda-immagine-risposta (Malinowski et al. 2014, Kafle et al. 2016).

Gao et al. (2015) propongono il modello *mQA* che contiene quattro componeneti: (*i*) *Long Short-Term Memory* (LSTM), per estrarre la rappresentazione delle domande in forma di *word embedding*, (*ii*) un CNN per estrarre la rappresentazione visuale, (*iii*) un LSTM per immagazzinare il contesto linguistico in una domanda e (*iv*) una componente di fusione per combinare l'informazione delle prime tre componenti e generare la risposta (Gao et al. 2015).
Fukui et al. (2016) propongono un metodo per sfruttare i *joint embedding visual and text features* che performano la loro "Multimodal Compact Bilinear pooling" (MCB) proiettando randomicamente l'immagine e le *feature* del testo in un spazio con più dimensionalità e poi convolvendo i vettori in un *Fast Fourier Transform (FFT) space* (Fukui et al. 2016).

Malinowski et al. (2014) propongono un metodo che utilizza il *parsing* semantico e la segmentazione delle immagini con un approccio bayesiano (Malinowski et al. 2014).

1.3.2 *Attention mechanism*

Una limitazione della maggior parte dei modelli che utilizzano *joint embedding approaches* è quella di utilizzare *feature* globali (a livello di immagine) per rappresentare l'input visivo. Questo può fornire informazioni irrilevanti o rumorose nella fase di predizione. Lo scopo dei meccanismi di attenzione è quello di affrontare questo problema utilizzando *feature* locali dell'immagine e permettendo al modello di assegnare un'importanza diversa alle caratteristiche provenienti da regioni diverse (Wu, Teney et al. 2016 e Vaswani et al. 2017). L'idea alla base di questi modelli è che alcune regioni visive in un'immagine e alcune parole in una domanda sono più informative di altre per rispondere a una data domanda. Per esempio, per un sistema che risponde a "Di che colore è l'ombrello?" la regione dell'immagine che contiene l'ombrello è più informativa di altre regioni dell'immagine. Allo stesso modo, 'colore' e 'ombrello' sono gli input testuali che devono essere affrontati più direttamente degli altri (Wu, Teney et al. 2016).
Le caratteristiche globali dell'immagine indagate dagli approcci *joint embedding* possono non essere abbastanza granulari per affrontare le domande specifiche della regione: per usare meccanismi di *attention* a livello spaziale, un algoritmo deve rappresentare le caratteristiche visive in tutte le regioni spaziali, invece che solo a livello globale (Kafle et al. 2017). In seguito, alle caratteristiche locali delle regioni rilevanti può essere data maggiore importanza in base alla domanda posta (Kafle et al. 2017).

Una prima applicazione dei meccanismi ai compiti visivi è proposta nel contesto della didascalia delle immagini da Xu e Saenko (2016). La componente di *attention* del modello identifica le regioni salienti in un'immagine, e un'ulteriore elaborazione focalizza la generazione della didascalia su quelle regioni. Questo concetto si traduce facilmente nel *task* di VQA per concentrarsi sulle regioni dell'immagine rilevanti per la domanda. In un certo senso, il processo di attenzione costringe a un esplicito passo aggiuntivo nel processo di ragionamento che identifica "dove guardare" prima di eseguire ulteriori calcoli (Xu et al.

2016).
Lu et al. (2016) propongono un nuovo modello di co-attenzione per VQA che ragiona congiuntamente sull'attenzione a livello di testo e di immagine. Inoltre, il modello ragiona sulla domanda (e di conseguenza sull'immagine attraverso il meccanismo di co-attenzione) in modo gerarchico attraverso una nuova rete neurale convoluzionale (CNN) (J. Lu et al. 2016).

La maggior parte dei meccanismi convenzionali di attenzione visiva utilizzati nella didascalie di immagini e nel VQA, come quelli illustrati in precedenza, sono di tipo *top-down*: utilizzano come contesto una rappresentazione di una didascalia parzialmente completata, o una domanda relativa all'immagine; questi meccanismi sono tipicamente addestrati a partecipare selettivamente all'output di uno o più strati di un CNN. Tuttavia, questo approccio non attribuisce importanza a come vengono individuate le regioni dell'immagine soggette all'attenzione (Anderson et al. 2018).
Anderson et al. (2018) propongono un approccio che oltre a meccanismi di attenzione visiva *top-down*, tipicamente usati per consentire una comprensione profonda delle immagini con un'analisi a grana fine, combina anche meccanismi *bottom-up*; ciò consente di calcolare l'attenzione a livello di oggetti e altre regioni importanti dell'immagine (come presentato anche in J. Lu et al. 2016). In questo approccio il meccanismo *bottom-up* (basato sul lavoro di S. Ren et al. 2015) propone delle regioni delle immagini, ciascuna delle quali è associata a un *feature vector*, mentre il meccanismo *top-down* determina il peso delle *feature*. Questo approccio rivela un'alta accuratezza e vince la sfida di VQA del 2017.

1.3.3 Modelli composizionali

I metodi discussi finora presentano limitazioni legate alla natura monolitica delle CNN e RNN utilizzate per estrarre rappresentazioni di immagini e frasi.
Nel VQA, spesso le domande richiedono più passaggi di ragionamento per rispondere correttamente (Kafle et al. 2017). Una direzione di ricerca sempre più popolare nella progettazione di reti neurali artificiali è quella di considerare architetture modulari che comportano la connessione di moduli distinti progettati per specifiche capacità desiderate.

Un potenziale vantaggio è un migliore uso della supervisione (Wu, Teney et al. 2016). Da un lato, viene infatti facilitato il *transfer learning*, poiché uno stesso modulo può essere usato e addestrato all'interno di diverse architetture e compiti complessivi; dall'altro lato, permette di usare la "supervisione profonda", cioè l'ottimizzazione di un obiettivo che dipende dagli output dei moduli interni (per esempio su quali fatti di supporto dovrebbe concentrarsi un meccanismo di attenzione) (Wu, Teney et al. 2016).
I modelli principali che utilizzando questo approccio sono (Wu, Teney et al. 2016 e Kafle et al. 2017):

- *Neural Module Network* (NMN), un *framework* che usa *parser* esterni della domanda per individuare il *sub-task* nella domanda (Andreas et al. 2016);
- *Recurrent Answering Units* (RAU), addestrato *end to end*, apprende implicitamente i *sub-task* senza necessitare di addestramento e si basa su un *recurrent deep neural network* dove ogni modulo corrisponde a un'unità completa di risposta con il suo meccanismo di attenzione (Noh et al. 2016);
- *Dynamic Memory Networks* (DNM), un *framework* di reti neurali composto da quattro moduli. Processa le sequenze di input e le domande, utilizza un modulo di memoria episodica e genera le risposte rilevanti (Kumar et al. 2016).

1.3.4 *Knowledge base-enhanced approaches*

Anche se il VQA implica la comprensione del contenuto delle immagini, spesso richiede informazioni non visive, come il senso comune, la conoscenza specifica di un argomento o addirittura una conoscenza enciclopedica. Una conoscenza degli usi e del contesto tipico degli oggetti presenti in un'immagine può essere utile per il VQA. Per esempio, per rispondere alla domanda "Quanti mammiferi appaiono in questa immagine?", bisogna capire la parola "mammifero" e sapere quali animali appartengono a questa categoria. Un possibile approccio per risolvere questo problema è quello di disaccoppiare il ragionamento (ad esempio come rete neurale) dall'effettiva memorizzazione dei dati o della conoscenza; per esempio, un sistema di VQA che ha accesso a una base di conoscenza potrebbe usarla per rispondere a domande su particolari animali, come i loro habitat, colori, dimensioni e

abitudini alimentari (Kafle et al. 2017 e Wu, Teney et al. 2016). Questo consentirebbe di risolvere alcuni problemi degli approcci *joint embedding* che possono catturare solo la conoscenza che è presente nel *test* set e sono quindi dipendenti da *bias* e poco generalizzabili.

Questa idea viene esplorata in Wu et al. (2016) combinando rappresentazioni interne del contenuto delle immagini con informazioni provenienti da una base di dati esterna e dimostrando che in questo modo le prestazioni migliorano. In questo caso le informazioni vengono estratte da DBpedia, un dataset che estrae conoscenza strutturata da Wikipedia (Auer et al. 2007), ma è possibile che l'uso di una fonte più adattata al VQA possa produrre un miglioramento maggiore (Wu, Wang et al. 2016).

Uno dei principali problemi dei sistemi di apprendimento automatico, e specialmente dei modelli che utilizzano reti neurali, è che necessitano di grandi quantità di dati e le prestazioni scalano bene con la quantità di dati, come dimostra l'applicazione dei modelli a dataset di maggiori dimensioni (Deng et al. 2009).
Tuttavia, la quantità di dati puliti annotati dall'uomo nei dataset esistenti è limitata ed è lontana dal saturare la capacità del modello. Per far fronte a questa carenza di dati, viene utilizzato l'approccio di "pre-addestramento e *fine-tuning*". Secondo questo paradigma, il modello viene prima pre-addestrato su dati su larga scala meno annotati o non annotati con "*pretext task*" (He, Fan et al. 2020) e viene poi messo a punto (*fine-tuned*) su compiti a valle, di solito attraverso una mole di dati molto inferiore.

Dal 2017 si diffondono anche modelli con architettura *Transformer*, che raggiungono ottimi risultati sia in *task* di NLP (Vaswani et al. 2017) sia in *task* di CV (He, Fan et al. 2020).

Questi modelli vengono introdotti da Vaswani et al. (2017) e si basano su un'architettura encoder-decoder fondata sull'*attention*. L'encoder consiste in *layer* di codifica che elaborano iterativamente l'input mappandolo in una rappresentazione continua che contiene tutte le informazioni apprese, mentre il decoder consiste in *layer* di decodifica che prendono la rappresentazione continua generata dall'encoder e passo dopo passo generano

un singolo output utilizzando anche l'output precedente (Vaswani et al. 2017).

La differenza principale con le reti RNN come LSTM (Agrawal, J. Lu et al. 2016, Kafle et al. 2017, e Gao et al. 2015) riguarda il fatto che i *Transformer* sono reti *feed-forward*, che non utilizzano meccanismi ricorsivi per elaborare l'input. In una RNN, tipicamente l'input viene processato in maniera ricorsiva, per esempio a ciascun neurone della sequenza viene fornito l'input corrente (per esempio il token della frase) e l'output dei neuroni precedenti. Nel caso dei *Transformer*, invece, l'intero input è fornito alla rete. Questa utilizza due meccanismi per gestirlo in contemporanea. Da un lato, un *encoding* posizionale, che marca la posizione di ciascun elemento dell'input all'interno della sequenza, e dall'altro il meccanismo di *attention*, che fornisce il contesto per qualsiasi posizione nella sequenza, ovvero gli altri elementi considerati come "importanti" per elaborare lo specifico token. Per esempio, se i dati di input sono una frase in linguaggio naturale, il *Transformer* non ha bisogno di elaborare l'inizio della frase prima della fine, ma identifica il contesto che conferisce significato ad ogni parola nella frase (Vaswani et al. 2017). Questa caratteristica permette una maggiore parallelizzazione rispetto alle RNN e quindi riduce i tempi di formazione (Vaswani et al. 2017).

La parallelizzazione aggiuntiva dell'addestramento consente l'addestramento su set di dati più grande (Vaswani et al. 2017).

1.3.5 LXMERT

Dal momento che preaddestramento e *fine tuning* e i modelli con architettura *Transformer* ottengono risultati promettenti sia per quanto riguarda i *task* legati alla visione (He, X. Zhang et al. 2016 e G. Huang et al. 2017) sia per quelli legati al linguaggio (Devlin et al. 2018, Peters et al. 2018 e Radford et al. 2018), Tan et al. (2019) propongono il *framework* LXMERT (Tan et al. 2019) per applicare tali soluzioni anche al *task* multimodale di VQA.

LXMERT (*Learning Cross-Modality Encoder Representations from Transformers*) (Tan et al. 2019) è un *framework* per l'apprendimento di connessioni multimodali visione-linguaggio.

Consiste in un modello *transformer* (Vaswani et al. 2017) costituito da tre *tranformer*

encoder: un *encoder* per oggetto-relazione, un *encoder* per il linguaggio e un *encoder* per la multimodalità (modalità incrociata).

Per consentire l'allineamento tra visione e linguaggio, il modello viene preaddestrato su cinque *task* rappresentativi diversi: (*i*) *masked cross-modality language modeling* (Devlin et al. 2018), (*ii*) previsione di *masked object* tramite regressione delle *feature* RoI (Cores et al. 2020), (*iii*) previsione di *masked object* tramite classificazione delle etichette rilevate, (*iv*) corrispondenza multimodale e (*v*) risposta a domande sulle immagini. Il pre-addestramento multimodale consente al modello di dedurre le *masked feature* sia dagli elementi visibili nella stessa modalità, sia dalle componenti allineate nella multimodalità. In questo modo, aiuta a costruire relazioni sia intra-modali che multimodali.

I dati per il pre-addestramento sono stati ricavati da cinque dataset visuali e linguistici: MS COCO (Veit et al. 2016), Visual Genome (Krishna et al. 2017), VQA 2.0 (Agrawal, J. Lu et al. 2016), GQA (Hudson et al. 2019) e VG-QA (Zhu et al. 2016). Di ogni dataset vengono utilizzati solo il *train* e il *dev* set, evitando i dati dei *test* set. Sui dataset viene condotta una pre-elaborazione per creare coppie di frasi e immagini allineate, utilizzando le domande come frase delle coppie immagine-frase e le risposte come etichette per il *task* di pre-addestramento di risposta a domande sulle immagini. Il dataset finale consiste in 9.18 milioni di coppie immagine-frase e 180k immagini distinte. In termini di token, i dati di pre-addestramento contengono circa 100 milioni di parole e 6.5 milioni di oggetti (Tan et al. 2019).

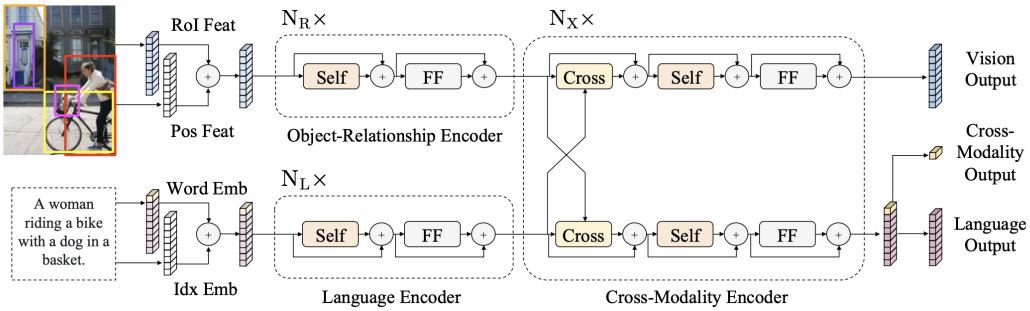

Figura 1.2: Funzionamento del *framework* LXMERT (Tan et al. 2019, pag. 3).

Come illustrato nella figura 1.2, il modello prende due input: un'immagine e la frase

associata. Ogni immagine è rappresentata come sequenza di oggetti e ogni frase è rappresentata come sequenza di parole.
Gli input *embedding* trasformano gli input, ovvero le immagini o le frasi, in due sequenze di *feature*: (*i*) un *embedding* delle frasi a livello della parola (*word-level sentence embedding*), con la suddivisione della frase in parole (usando il Word Tokenizer in Wu, Wang et al. 2016), la proiezione della parola e del suo indice in vettori e la loro aggiunta a *embedding* di parole *index aware*, e (*ii*) un *embedding* delle immagini a livello dell'oggetto (*object-level image embedding*), che tiene conto delle informazioni sulla posizione oltre che della regione di interesse (seguendo l'approccio in Anderson et al. 2018). Gli *embedding* ricavati vengono poi processati dai *layer* di *encoding* e in particolare, prima dagli *encoder* a modalità singola (visuale e linguistica), poi dall'*encoder* multimodale.

I tre *encoder* vengono costruiti utilizzando due tipologie di *attention layer*, i *self-attention layer* e i *cross-attention layer*, che hanno lo scopo di estrarre delle informazioni a partire da un set di vettori contesto rispetto a un vettore *query*. Questi *layer* calcolano il punteggio di corrispondenza tra il vettore *query* e ogni vettore contesto; l'output di un *attention layer* è la somma pesata dei vettori contesto in riferimento al punteggio normalizzato per la *softmax*.

Dopo i *layer* di *embedding* vengono applicati due *Transformer Encoder* agli input visuali e linguistici: i *Single Modality Encoder*, che consistono in un *Language Encoder* e in un *Object-Relationship Encoder*.
Ogni layer nei *Single Modality Encoder* ha un *self-attention sub-layer* e un *feed-forward sub-layer* composto da 2 *fully connected sub-layer*.
Successivamente viene applicato il *Multimodality Encoder*, in cui ogni *layer* per la multimodalità contiene: (*i*) due *self-attention sub-layer*, (*ii*) un *bidirectional cross-attention sub-layer*, usato per scambiare informazioni e allineare le entità tra le due modalità per apprendere rappresentazioni multimodali, e (*iii*) due *feed-forward sub-layer* che producono gli output.

Il modello multimodale di LXMERT fornisce 3 output: linguistico, visuale e multimodale.

Gli output linguistico e visuale sono due vettori generati dal *cross-modality encoder*, mentre all'output della multimodalità viene aggiunto un token specifico [CLS] prima delle parole della frase, e il corrispondente vettore di *feature* di questo token nelle sequenze di *feature* linguistiche viene usato come output di multimodalità.

Il modello viene valutato su 3 dataset: VQA v2.0 (Goyal et al. 2017), GQA (Hudson et al. 2019) e NLVR (Suhr et al. 2018).
Sui dataset di VQA 2.0 e GQA il modello presenta un miglioramento dell'accuratezza generale rispetto ai risultati ottenuti precedentemente da altri modelli. Il miglioramento nell'accuratezza ottenuto sul dataset NLVR dimostra anche buone capacità di generalizzazione per il *task* di *Visual Reasoning*. Tutti i *task* di pre-addestramento si rivelano fondamentali per raggiungere la migliore accuratezza del modello sui dataset.

2. Materiali e metodi

Come anticipato nella sezione 1.2.2, per risolvere la mancanza di risorse e dati per il VQA in lingue diverse dall'inglese e costruire un sistema per rispondere a domande in italiano è possibile sfruttare i progressi ottenuti dalla traduzione automatica.

L'ipotesi della tesi e degli esperimenti è, infatti, che i passi avanti compiuti dalla traduzione neurale automatica offrano la possibilità di sfruttare sistemi e dati esistenti, senza la necessità di creare dataset *ad hoc* e nuovi modelli per lingue diverse.

Ci sono più approcci praticabili per creare sistemi per il VQA in italiano utilizzando la traduzione automatica, in questa tesi vengono prese in considerazione (i) la traduzione delle annotazioni di una risorsa esistente e addestrare un modello su essa (Croce et al. 2021) e (ii) l'utilizzo di un modello pre-addestrato in inglese traducendo solo la domanda in input e la risposta in output.

Entrambe le soluzioni sono meno costose che annotare manualmente le immagini e ideare modelli *ex novo*, ma portano a una perdita di informazione dal punto di vista linguistico: lo scopo degli esperimenti è quello di indagare quanto e come il rumore introdotto dalla traduzione automatica porti ad una perdita di accuratezza.

Le risorse utilizzate sono GQA come dataset di riferimento, per i vantaggi illustrati nella sezione 1.2.1, e LXMERT come modello computazionale, dal momento che raggiunge lo stato dell'arte e presenta una buona capacità di generalizzazione, come illustrato nella sezione 1.3.5.

Per la traduzione automatica vengono utilizzati due strumenti di NMT basati su *transformer*: Google Translate (Google 2017) e Opus NMT (Tiedemann et al. 2020).

Google Translate (GNMT) è il traduttore automatico di Google disponibile per più di 100 lingue. Dal 2016 Google sfrutta la NMT per utilizzare un contesto più ampio che consente di creare la traduzione più pertinente, poi riorganizzata per essere simile a un discorso umano con una grammatica appropriata. Questo rende più facile capire ogni frase

e i paragrafi e gli articoli tradotti sono molto più fluidi e facili da leggere[1].

Opus NMT è un traduttore NMT basato su un'architettura *Transformer* che supporta traduzioni bilingue e multilingue ed è addestrato su *corpora* allineati di lingue (Tiedemann et al. 2020).

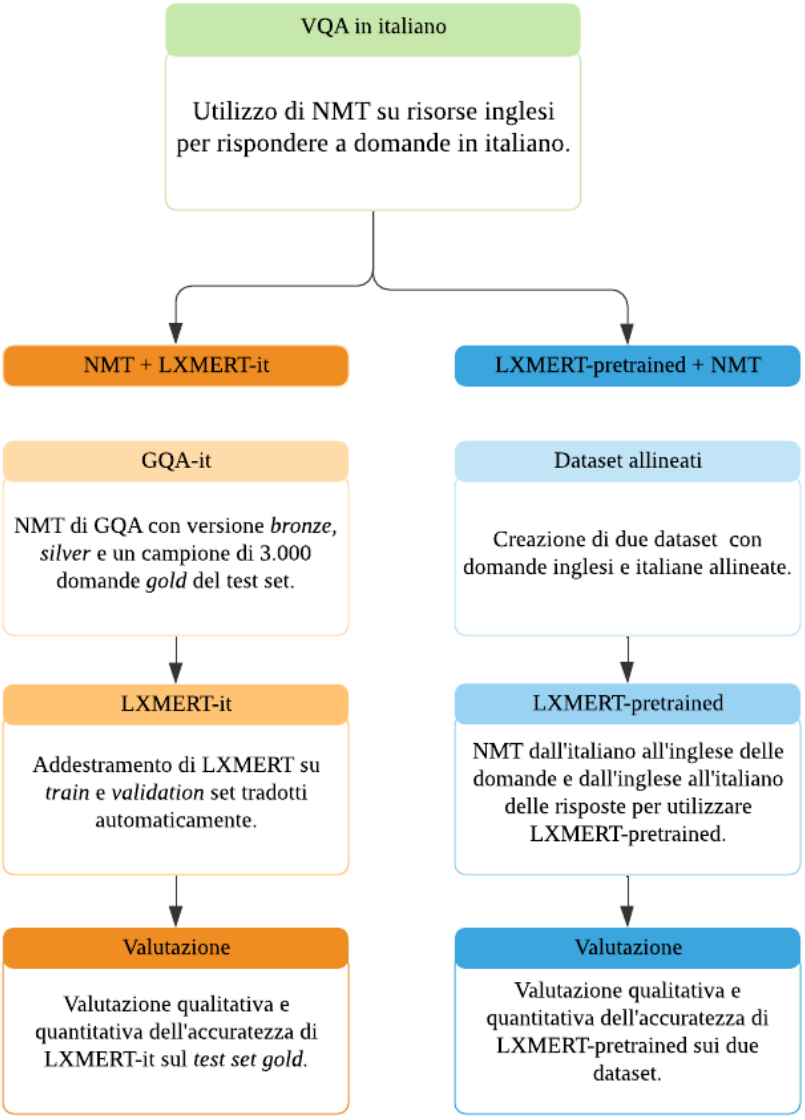

Figura 2.1: Diagramma di flusso degli esperimenti.

Per la valutazione dei due sistemi vengono condotte due serie di esperimenti, il cui

[1]Google Translate: *https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/*, consultato in data 1 settembre 2021.

diagramma di flusso è illustrato nella figura 2.1: la prima, NMT + LXMERT-it, è volta a valutare l'efficienza di LXMERT-it, ovvero del *framework* LXMERT (Tan et al. 2019) addestrato su GQA-it, una risorsa tradotta automaticamente in italiano a partire dall'omonimo dataset inglese (Croce, Passaro et al. 2021); la seconda, LXMERT-pretrained + NMT, sfrutta al massimo le potenzialità del *framework* LXMERT pre-addestrato in inglese, LXMERT-pretrained, utilizzando la traduzione automatica della domanda in input e della risposta in output. Per gli esperimenti sul secondo sistema proposto viene anche presentata una nuova risorsa per il VQA italiano che allinea annotazioni di domande in italiano e in inglese composte manualmente.
L'efficienza dei due sistemi viene valutata qualitativamente e quantitativamente in base alla perdita di accuratezza rispetto ai modelli sull'inglese. Le caratteristiche dei dataset utilizzati per gli esperimenti sono riassunte nella tabella 2.5.

2.1 Esperimenti con un modello addestrato sull'italiano: NMT + LXMERT-it

Croce et al. (2021) creano GQA-it, un dataset per il VQA in italiano a partire dalla NMT del dataset inglese GQA (Croce, Passaro et al. 2021).
L'obiettivo è quello di generare un dataset su larga scala in cui *train* e di *validation* sono ottenuti tramite la traduzione automatica e il *test* è convalidato manualmente (Croce, Passaro et al. 2021).
Il dataset è anche utilizzato per addestrare il *framework* LXMERT e valutare l'effetto del rumore prodotto dalla traduzione automatica sull'accuratezza. Questo studio fornisce quindi due contributi: (*i*) la creazione di un *test* set di riferimento in italiano e (*ii*) la misurazione della sensibilità al rumore introdotto dalla traduzione automatica del sistema (Croce, Passaro et al. 2021).

2.1.1 Creazione di GQA-it

La prima versione di GQA-it segue le specifiche proposte da Tan e Bansal (2019) per la suddivisione del dataset di GQA. Il dataset italiano consiste in 72.140 immagini e 943.000 coppie di domanda/risposta che compongono il dataset di *training*, e in 10.234 immagini e 132.062 coppie di domanda/risposta che compongono il dataset di *validation*. Le coppie domanda/risposta del *test* set non sono disponibili pubblicamente, poiché vengono utilizzate solo nelle competizioni. Per superare questo problema, viene utilizzato il *test-dev*, un sottoinsieme del materiale del *test set*, rappresentativo di diversi fenomeni linguistici e concettuali e consistente in 398 immagini e 15.578 coppie domanda/risposta. Questa suddivisione consente di riprodurre i risultati del modello LXMERT implementato da Tan e Bansal (2019).

Le tre partizioni vengono tradotte automaticamente utilizzando OPUS-NMT addestrato su un sottoinsieme di documenti italiano/inglese (Tiedemann et al. 2020). Un sottoinsieme di 3.000 coppie domanda/risposta del *test-dev* set originale viene validato manualmente da due annotatori di madrelingua italiana a cui viene richiesto di controllare e correggere le traduzioni, ottenendo una versione *gold* del dataset.

Il dataset risultante, GQA-it, è composto da più di 1,08 milioni di coppie domanda/risposta su più di 80mila immagini con una validazione parziale del *test* set e comprende una parte *bronze*, consistente nella semplice traduzione automatica, una parte *silver*, consistente nella normalizzazione delle risposte, e una parte *gold*, consistente nella validazione manuale di 3.000 coppie domanda/risposta del *dev-test* set set (Croce, Passaro et al. 2021), come illustrato nella tabella 2.1.

Dataset	Immagini	Coppie dom./risp.
train	72.140	943.000
valid	10.234	132.062
test-dev (silver)	398	12.578
test-dev (gold)	398	3.000

Tabella 2.1: Statistiche di GQA-it (Croce, Passaro et al. 2021, pag. 4.)

La qualità delle domande tradotte è valutata validando manualmente 500 elementi di traduzione e utilizzando la metrica BLEU (Papineni et al. 2002) che risulta in punteggio alto, ovvero 0,82 (Croce, Passaro et al. 2021). Dal momento che le domande sono frasi con una lunghezza media limitata, le traduzioni sono generalmente di buona qualità, per esempio "Is the surfer that looks wet wearing a wetsuit?" tradotto automaticamente in "Il surfista che sembra bagnato indossa una muta?" (Croce, Passaro et al. 2021). Inoltre, l'alto punteggio BLEU dipende anche dall'impostazione utilizzata per la traduzione: agli annotatori non viene richiesto di scrivere le traduzioni senza conoscere gli output del sistema, ma di validare, ovvero di correggere, gli output errati; questo implica che sarà presente un maggior numero di sequenze comuni tra le domande di input e quelle validate, con un conseguente punteggio BLEU alto (Scaiella et al. 2019). In ogni caso, questo punteggio BLEU suggerisce che il materiale italiano è caratterizzato da un basso livello di rumore dovuto al processo di traduzione automatica e consente di presumere che la traduzione delle domande sia di buona qualità.

Nonostante questo, la traduzione automatica compie alcuni errori ricorrenti e alcuni *pattern* di domande poco si prestano ad una traduzione in italiano; per esempio, "What kind of animal is the fence behind of?" è tradotto automaticamente in "Di che tipo di animale è il recinto dietro?" o "What is that table in front of?" è tradotto automaticamente in "Che cosa c'è il tavolo davanti?". In alcuni casi di ambiguità lessicale, inoltre, il senso delle parole viene assegnato in maniera errata, come nel caso di "What is the item of furniture that the cord is on called?" che è tradotto automaticamente in "Come si chiama il mobile

su cui si trova la corda?" (in questo caso, osservando l'immagine è possibile notare che il termine polisemico "cord" si riferisce a un cavo). Un'analisi qualitativa e quantitativa di questi errori di traduzione delle domande è fornita per il *test set gold* nel paragrafo 3.1.2.

La traduzione delle risposte è più complessa rispetto a quella delle domande e presenta alcuni problemi, molti dei quali dipendono dall'ambiguità lessicale: sono frequenti termini polisemici in inglese che non presentano in italiano un corrispettivo che ne catturi tutti i sensi senza disporre del contesto-immagine. In questi casi il senso della parola tende ad essere assegnato in maniera errata, ad esempio, la risposta "bat" può essere tradotta come l'animale "pipistrello" o l'oggetto "mazza" (Croce, Passaro et al. 2021).

Come suggerito in Croce, Zelenanska e Basili (2019), per ridurre l'ambiguità lessicale una risposta viene tradotta utilizzando come contesto la domanda corrispondente (Croce, Zelenanska et al. 2019). Per esempio, la risposta "mouse" viene tradotta correttamente se accoppiata alla domanda "What's next to the keyboard?", mentre traduzioni generiche, come "topo", vengono preferite quando non viene reso disponibile alcun contesto (Croce, Passaro et al. 2021).

Un altro problema nella traduzione automatica delle risposte consiste nella variabilità lessicale: il materiale inglese iniziale è caratterizzato da 1.842 tipi di risposte possibili (Croce, Passaro et al. 2021). Dopo la traduzione automatica, il numero diventa 3.306. Questo è in parte dovuto ai casi in cui il contesto non migliora la traduzione, ad esempio, la domanda "What's on top of the photo?" che non è molto utile per disambiguare la risposta "mouse". In altri casi, esistono più modi di tradurre lo stesso elemento lessicale, ad esempio, "aircraft" è tradotto sia come "aeromobile" che come "aeroplano" (Croce, Passaro et al. 2021).

Infine, mentre il dataset GQA mantiene una distinzione tra le risposte che coinvolgono espressioni singolari e plurali, generalmente i sostantivi e gli aggettivi inglesi non presentano una distinzione di genere tra il maschile e femminile, mentre una traduzione sensibile al contesto inflette la traduzione in maschile e femminile. Per esempio, "little" è stato tradotto in "piccola", "piccolo", "piccole" e "piccoli" a seconda degli elementi coinvolti nella foto (Croce, Passaro et al. 2021).

Per ridurre la variabilità lessicale, viene applicata una normalizzazione manuale alle risposte associate a più di due domande (Croce, Passaro et al. 2021). Ogni risposta originale inglese viene accoppiata con quelle tradotte, al fine di normalizzare manualmente le traduzioni. Mentre questo tipo di valutazione manuale è generalmente inefficace quando si tratta di traduzione automatica, GQA inglese mostra una polisemia ridotta, poiché le domande, le risposte e le annotazioni grafiche sono state normalizzate automaticamente per ridurre l'ambiguità linguistica (Hudson et al. 2019). La normalizzazione mantiene per i sostantivi e gli aggettivi la distinzione tra forme singolari e plurali; le azioni, che nel dataset presentano sempre la forma in *-ing*, per esempio, "skating", "jumping" o "sleeping", vengono tradotte nelle forme gerundive "sta facendo skateboard", "sta saltando" e "sta dormendo". Per quanto riguarda gli aggettivi, il rumore introdotto durante la loro traduzione rende problematico gestire il genere di tali parole, per cui le forme vengono normalizzate al genere maschile. Dopo questa normalizzazione manuale, il numero di risposte possibili nel dataset è 1.701 (Croce, Passaro et al. 2021).

La normalizzazione consente di ottenere una versione *silver* contenente le risposte del dataset (Croce, Passaro et al. 2021). La distribuzione delle 50 risposte più frequenti è generalmente conservata (Croce, Passaro et al. 2021).

Al fine di garantire la valutazione corretta dei sistemi addestrati su coppie domanda/risposta possibilmente rumorose, viene convalidato manualmente un sottoinsieme di 3.000 coppie domande/risposte del *balanced dev-test* set, selezionate in modo da preservare l'equilibrio dei dati. In particolare, viene ripristinata l'inflessione di genere, persa durante la fase di normalizzazione. Questa validazione consente di ottenere una versione *gold* del dataset che comprende un sottoinsieme di coppie domanda/risposta del *test* set.

Il campione estratto dal *dev-test* set, rispetto alla distribuzione del dataset GQA (Hudson et al. 2019), si presenta sbilanciato in favore di domande che richiedono una risposta binaria ("sì" o "no"), come è possibile osservare nella figura 2.2.

Sulla base della distinzione delle domande sulla base delle risposte, rispetto al dataset originale di GQA (Hudson et al. 2019), la distribuzione viene mantenuta con l'eccezione

della categoria delle domande di verifica che sono presenti in maniera notevolmente maggiore a scapito delle categorie di scelta e logica; questo rispecchia in generale lo sbilanciamento dei dati del *dev-test* set.

Per quanto la classificazione dipendente da il/i soggetto/i delle domande la distribuzione è raffigurata nell'immagine 2.3 e non presenta cambiamenti significativi rispetto al dataset non bilanciato di GQA (Hudson et al. 2019).

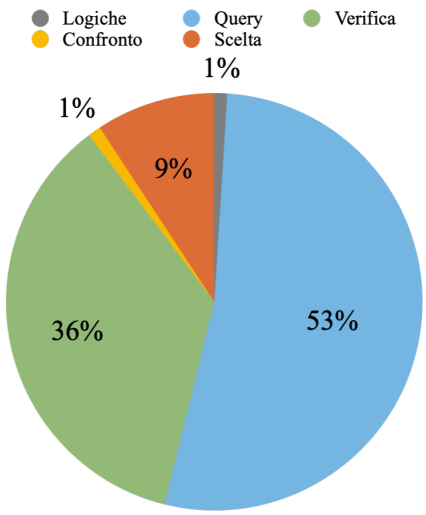

Figura 2.2: Distribuzione della tipologia di risposte del *test set gold* di GQA-it.

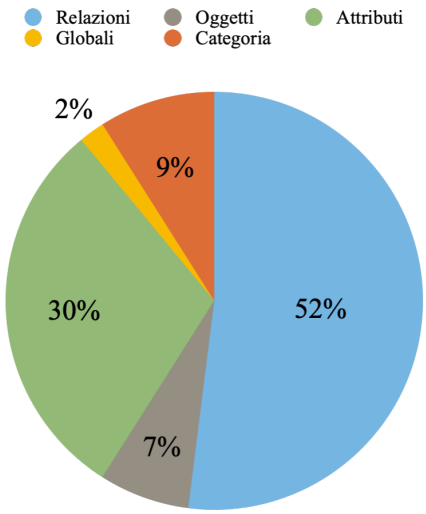

Figura 2.3: Distribuzione del/i soggetto/i delle domande del *test set gold* di GQA-it.

2.1.2 Traduzione dei *scene graph* di GQA

Per la traduzione completa della risorsa GQA, viene anche proposta una versione italiana dei *scene graph* di *train* e *validation* set di GQA; i *scene graph* del *test* set non sono stati ancora resi disponibili.

La traduzione di questa risorsa consente di disporre delle immagini annotate con etichette in italiano e di poter utilizzare la composizionalità per la creazione delle domande, senza dover dipendere da *pattern* strutturali pensati per l'inglese. La traduzione dei *scene graph* riguarda tre livelli: (*i*) la traduzione degli oggetti, ovvero degli elementi all'interno delle *bounding box* che sono i nodi dello *scene graph*, (*ii*) la traduzione delle relazioni,

ovvero degli archi che collegano i nodi-oggetti, e (*iii*) la traduzione degli attributi, che aggiungono informazioni sugli oggetti.

I tre elementi che caratterizzano i *scene graph* appartengono a tre classi morfologiche diverse: gli oggetti sono sostantivi, le relazioni sono verbi in *-ing* o preposizioni, e gli attributi sono aggettivi o verbi in *-ing* utilizzati in funzione attributiva.

I *scene graph* del *train* set di GQA riguardano 74.942 immagini e contengono 1.231.134 oggetti, 682.679 attributi e 1.005.600 relazioni, ma solo 1.702 oggetti distinti, 615 attributi distinti e 310 relazioni distinte. I *scene graph* del *validation* set di GQA riguardano invece 10.568 immagini e contengono 174.331 oggetti, 97.057 attributi e 142.210 relazioni, ma solo 1.536 oggetti distinti, 599 attributi distinti e 295 relazioni distinte.

La traduzione dei *scene graph* fornisce la versione *silver* di tutti gli elementi, e la versione *gold* di un campione dei *scene graph* di 100 immagini.

La versione *silver* del dataset viene creata attraverso la validazione manuale della traduzione automatica ottenuta con Google Translate di tutti gli elementi distinti fuori contesto, mentre la versione *gold* viene realizzata validando manualmente la traduzione *silver* utilizzando le immagini come contesto.

Dataset	**Immagini**	**Oggetti**	**Relazioni**	**Attributi**
train (silver)	74.842	1.229.294	1.004.893	681.739
validation (silver)	10.568	174.331	142.210	97.057
train (gold)	100	1.840	707	940

Tabella 2.2: Statistiche della resa italiana dei *scene graph* di GQA.

Il primo passo per la realizzazione della versione *silver* è quello di prendere in considerazione oggetti, attributi, e relazioni distinti e tradurli automaticamente fuori contesto.

La traduzione automatica viene poi normalizzata per l'italiano, utilizzando la forma maschile per i sostantivi (gli oggetti) e aggettivi (gli attributi) e il gerundivo per i verbi in *-ing* (le relazioni), come indicato per la creazione di GQA-it. Se in inglese non esiste una distinzione di genere per sostantivi e aggettivi, in italiano si pone il problema all'interno

del contesto. Nella traduzione automatica è possibile notare un'ambiguità lessicale consistente delle traduzioni. Dall'analisi dei risultati ottenuti tramite traduzione automatica si ottengono alcuni elementi lessicali ambigui per le tre classi di elementi del dataset:

- 363/1702 oggetti ambigui;
- 240/617 attributi ambigui;
- 234/310 relazioni ambigue.

L'ambiguità è data dalla polisemia dei termini inglesi che possono avere più di una traduzione italiana.

In tutti questi casi, con la validazione manuale per la creazione della versione *silver* si è cercato di individuare un iperonimo appropriato che contenesse in italiano tutti i significati possibili in inglese.

In questo modo è possibile ridurre l'ambiguità, che non è possibile però eliminare totalmente per sostantivi e aggettivi. Sono presenti, infatti, casi di mancata disambiguazione all'interno del dataset, ad esempio "brush" è utilizzato per i lemmi italiani "cespuglio", "spazzolino da denti", "spazzola" e "pennello".

Gli oggetti che non è stato possibile ricondurre a un iperonimo comune sono 17, e sono "bike", "biker", "brush", "calf", "card", "cards", "collar", "dip", "drum", "glass", "marker", "nut", "paddle", "pitcher", "plain", "trunk" e "trunks". Esiste un unico attributo per il quale non è stato possibile individuare un iperonimo, ovvero "pepper", utilizzato per riferirsi ai lemmi italiani "pepato" e "brizzolato".

Per la normalizzazione di ognuno di questi elementi è stato estratto un campione di 30 *scene graph* ed è stato osservato il senso più frequente. Chiaramente questo approccio non è ottimale, ma l'unico contesto disponibile per i *scene graph* sono le immagini e questo rende estremamente costosa la validazione manuale[2].

[2]Le traduzioni *silver* dei termini non disambiguati sono le seguenti: "bike"→, "biker"→, "brush"→"boscaglia", "calf"→"vitello", "card"→"foglietto", "cards"→"foglietti", "collar" "colletto"→, "dip"→"intingolo", "drum"→"tamburo", "glass"→"bicchiere", "marker"→"segnale", "nut"→ "chiodo", "paddle"→"pagaia", "pitcher"→ "brocca", "plain"→"pianura", "trunk"→"tronco", "trunks"→"tronchi", e "pepper"→"pepato".

Per la realizzazione della versioni *gold* del dataset sono state validati manualmente i *scene graph silver* di cento immagini del dataset. Questa annotazione consente, attraverso la traduzione contestuale di tutti gli elementi dei *scene graph*, di risolvere le problematiche dovute alla traduzione normalizzata.

La validazione *gold* prevede la flessione degli attributi e delle relazioni in base al genere e numero dell'oggetto a cui sono legate, e la risoluzione dell'ambiguità di aggettivi e sostantivi non disambiguati.

In particolare, vengono modificati il 4,73% degli oggetti totali e il 3,40% degli attributi totali del campione *gold*. Per quanto riguarda il genere di modifiche legate alla necessità di concordanza di sostantivo (oggetti) con attributi e relazioni, queste riguardano il 17,01% dell'intero campione.

In totale, il 65% delle immagini necessita di almeno una modifica: questo dimostra che la traduzione automatica e la validazione manuale non sono ideali per ottenere una versione italiana dei *scene graph* di GQA, che si rivela, in questo caso, caratterizzata da un rumore consistente.
Differentemente dalla traduzione automatica di coppie di domande e risposte per la realizzazione di GQA-it, la traduzione automatica di singole parole senza alcun contesto necessita di una validazione manuale.

2.1.3 Applicazione del modello LXMERT-it

Croce et al. (2019) utilizzano il dataset creato per valutare quanto e come il rumore prodotto dalla traduzione automatica influisca sull'accuratezza di un modello. Addestrano, quindi, il *framework* LXMERT (Tan et al. 2019) sul dataset GQA-it tradotto automaticamente e lo valutano sul *test* set convalidato manualmente di 3.000 coppie domanda/risposta. Il *framework* LXMERT di Tan e Bansal (2019) ottiene i risultati migliori senza utilizzare modelli BERT pre-addestrati esistenti, utilizzando un *task* di pre-addestrameto: i pesi dell'*encoder* linguistico sono inizializzati in modo casuale e pre-addestrati (insieme ai pesi dell'*encoder* per la multimodalità) utilizzando un dataset dedicato composto da didascalie di immagini e relative domande di circa 9 milioni di frasi, come illustrato in 1.3.5. Tuttavia, questa fase di pre-addestramento non è possibile per l'italiano, dal momento che i

dataset utilizzati da Tan e Bansal (2019) per creare il dataset di *pre-training* linguistico che accoppia immagini a didascalie non sono disponibili in una versione italiana.
Nonostante ciò, i risultati sperimentali mostrano che anche adottando un modello BERT pre-addestrato si possono ottenere buone prestazioni. Per addestrare efficacemente LXMERT su GQA-it, viene utilizzato il modello specializzato inglese con un modello BERT standard pre-addestrato, in particolare, BERT multilingue (Pires et al. 2019), che è disponibile anche per l'italiano.
Per gli altri due *encoder* non è necessario modificare il *framework* dal momento che non sono dipendenti dalla lingua: l'*encoder* originale di oggetti/relazioni viene conservato, e l'*encoder* multimodale viene inizializzato in modo casuale (Croce, Passaro et al. 2021).

2.2 Esperimenti con un modello pre-addestrato in inglese: LXMERT-pretrained + NMT

Un sistema per il VQA in italiano basato sul riaddestramento di un modello su una risorsa tradotta automaticamente consente di disporre di un modello addestrato sulla lingua di interesse, come dimostrato dagli esperimenti in 2.1. Tuttavia questo approccio presenta alcuni limiti. In primo luogo, anche se il rumore presente in GQA-it è ridotto, la traduzione automatica implica la presenza di qualche errore nel dataset non validato a livello di traduzione automatica delle domande. La normalizzazione delle risposte, inoltre, fa sì che un modello osservi solo risposte al maschile nella fase di addestramento e che non possa quindi fornire risposte flesse al femminile. Infine, il costo computazionale del *pre-training* aggiuntivo sui dati multi-modali in italiano ha reso impossibile l'esperimento, dunque il *framework* non viene utilizzato nella sua versione con performance migliori (Tan et al. 2019).

Un altro approccio possibile per il VQA in italiano consiste nell'utilizzare un modello pre-addestrato in inglese che raggiunge lo stato dell'arte attraverso la traduzione automatica della domanda in input e della risposta in output. Questo consente di sfruttare al

massimo le potenzialità del *framework* LXMERT pre-addestrato su dataset inglesi.
Per valutare l'efficacia del modello nel rispondere a domande sull'italiano vengono prese in considerazione le domande in inglese a cui il modello risponde correttamente.

Successivamente, il modello viene applicato alle domande in italiano, utilizzando la traduzione automatica delle domande dall'italiano all'inglese (mtit→en) per poter nuovamente interrogare il modello e ottenere la risposta prodotta in inglese che viene tradotta automaticamente in italiano (mten→it).

Al fine di poter valutare l'efficienza del *framework* per rispondere a domande sulle immagini in italiano viene fatto un confronto tra i risultati del modello su due dataset: (*i*) un campione di 100 immagini estratto dal *balanced test* set di GQA per valutare le differenze nelle prestazioni del modello in inglese e in italiano, e (*ii*) un dataset creato *ad hoc* per l'esperimento contenente 100 immagini annotate con 10 domande tradotte in italiano e in inglese e composte manualmente.

2.2.1 Nuovo dataset

Per valutare l'efficacia del modello LXMERT pre-addestrato in inglese si è deciso di utilizzare, oltre a un campione estratto da *test set* di GQA, una nuova risorsa che allinea domande in inglese e in italiano. Lo scopo è valutare il potere di generalizzazione del modello su dati diversi da quelli di GQA, ma anche fornire una nuova risorsa, annotata manualmente da annotatori madrelingua italiani.

Il nuovo dataset è creato a partire da 100 immagini selezionate da due annotatori sulla base della loro similarità a quelle del dataset GQA dal sito *Pexels*[3] che mette a disposizione fotografie ad alta definizione prive di diritto d'autore.

Ogni immagine viene annotata con 10 domande tradotte in italiano e in inglese e formulate manualmente da due annotatori madrelingua di italiano, adattando i *pattern* strutturali in inglese di GQA.

La creazione di questo dataset garantisce, inoltre, un'annotazione delle domande più accurata per l'italiano. Sono stati presi in considerazione, infatti, i *pattern* strutturali di

[3] *Pexels*: *https://www.pexels.com/it-it/*, consultato in data 23 settembre 2021.

GQA che sono maggiormente adattabili all'italiano e che si prestano maggiormente ad una traduzione multilingue inglese-italiano. Questo dataset, quindi, permette un'analisi linguistica e degli errori più accurata, poiché consente di partire da domande più standard per l'italiano rispetto a quelle ottenute attraverso la validazione della traduzione automatica delle domande che seguono i *pattern* composizionali di GQA.

Si è tentato di mantenere il più possibile la distribuzione delle domande per tipologia di risposta e per soggetto/i delle domande del dataset di GQA e di includere *pattern* appartenenti a tutte le tipologie.
La distribuzione delle domande sulla base della tipologia di risposta del nuovo dataset e sulla base del/i soggetto/i delle domande è rappresentata nelle figure 2.4 e 2.5.

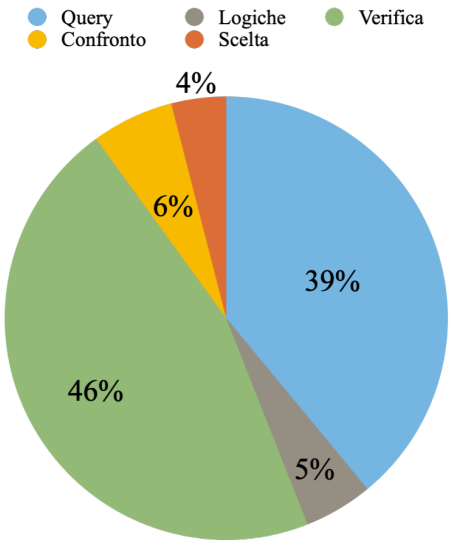

Figura 2.4: Distribuzione della tipologia di risposte del nuovo dataset.

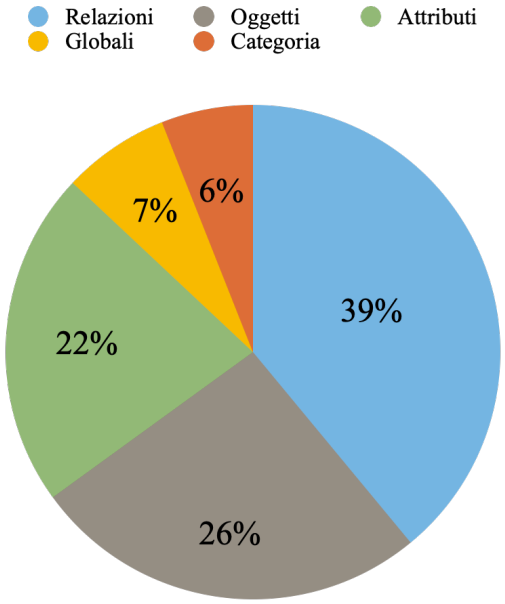

Figura 2.5: Distribuzione del/i soggetto/i delle domande del nuovo dataset.

Le risposte vengono annotate, ma non vengono utilizzate per la valutazione che avviene manualmente, dal momento che asserire una sola risposta come corretta si rivela complesso e richiede spesso una decisione arbitraria. Inoltre, nella fase di valutazione, si è osservato che anche per le domande di GQA più di una risposta spesso è ammissibile e corretta. Per esempio, nel caso di un'immagine che rappresenta una persona intenta a eseguire un

salto con lo skateboard, alla domanda “What is the person doing?” sarebbero corrette entrambe le risposte “jumping” e “skating”.

Figura 2.6: Immagine 2 estratta dal nuovo dataset.

questionEng	**questionIta**
Who is wearing a necklace?	Chi sta indossando una collana?
What is the person to the left of the woman waring?	Che cosa sta indossando la persona alla sinistra della donna?
Is there a plate in the picture that is not red?	C’è un piatto nell’ immagine che non è rosso?
Is the woman’s hair brown and curly?	I capelli della donna sono marroni e ricci?
On which side of the image is the man?	Da che lato dell’immagine è l’uomo?
Who is eating sushi?	Chi sta mangiando il sushi?
What is the woman to the right of the man holding?	Cosa sta tenendo la donna alla destra dell’uomo?
Is there any food inside the plate?	C’è del cibo nel piatto?
What is the orange sushi inside of?	Dentro cosa è il sushi arancione?
Is there any glass on the table?	C’è qualche bicchiere sul tavolo?

Tabella 2.3: Domande in inglese e in italiano per l’immagine 2 del nuovo dataset.

2.2.2 Campione di GQA

Il campione di GQA utilizzato comprende 100 immagini estratte dal *balanced test* set (al fine di evitare di utilizzare dati su cui il modello è addestrato) e 10 domande associate ad ogni immagine; le immagini e le domande associate vengono estratte casualmente. A ogni domanda in inglese viene associata la traduzione manuale in italiano.

La distribuzione delle domande in base alla tipologia di risposta di questo campione si rivela abbastanza sbilanciata, similmente al campione estratto per l’esperimento pre-

cedente: in questo caso, le domande di *query* e di verifica sono preponderanti, a scapito soprattutto delle domande logiche, come è possibile osservare nella figura 2.7.

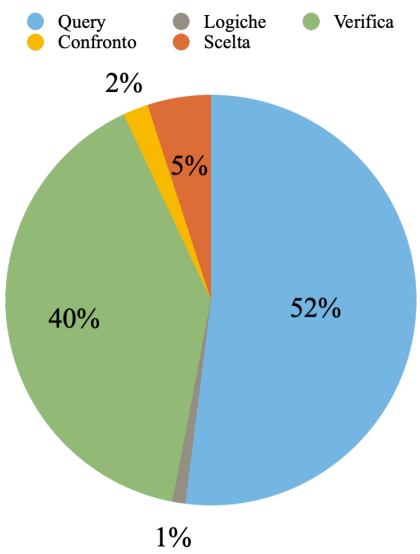

Figura 2.7: Distribuzione della tipologia di risposte del campione di GQA.

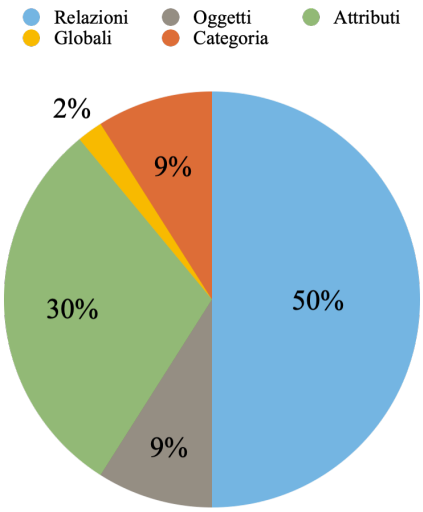

Figura 2.8: Distribuzione del/i soggetto/i delle domande del campione di GQA.

Figura 2.9: Immagine n457744 estratta dal campione di GQA.

questionEng	questionIta
What is in front of the people?	Cosa c'è di fronte alle persone?
Which kind of furniture is to the right of the pizza boxes?	Che tipo di arredamento c'è a destra delle scatole per pizza?
Are the pizza boxes to the left of a cup?	Le scatole della pizza sono a sinistra di una tazza?
What is the pizza inside of?	A cosa è dentro la pizza?
On which side is the glass plate?	Da che parte è il piatto di vetro?
Is the shirt short sleeved or long sleeved?	La maglia è a maniche corte o lunghe?
Who is standing?	Chi è in piedi?
What's the plate made of?	Di cosa è fatto il piatto?
Is the food to the left of the plate square or round?	Il cibo a sinistra del piatto è quadrato o rotondo?
Are these pizza boxes made of cardboard?	Queste scatole per pizza sono di cartone?

Tabella 2.4: Domande in inglese e in italiano per l'immagine n457744 estratta dal campione di GQA.

Anche per quanto riguarda la distribuzione delle domande sulla base del/i soggetto/i delle domande è possibile notare uno sbilanciamento: sono prevalenti domande sulle relazioni e sono rare domande globali, come è possibile osservare nella figura 2.8.

2.2.3 Applicazione del modello LXMERT

Come per l'esperimento precedente, viene utilizzato il *framework* LXMERT. In particolare, la versione utilizzata è quella rilasciata dagli autori che ottiene l'accuratezza maggiore nella *GQA Challenge 2019*[4]. Il modello in questo caso è pre-addestrato in inglese su tutti i cinque *task* descritti nella sezione 1.3.5. Per testare le prestazioni del *framework* LXMERT pre-addestato vengono valutati e confrontati i risultati ottenuti a partire dalle domande in inglese e quelli ottenuti a delle corrispettive domande in italiano, per poter comprendere quali errori dipendono dal modello e quali dal rumore provocato dalla traduzione automatica.

In primo luogo viene compiuta una valutazione sull'inglese, dando in input al modello le domande in inglese. In secondo luogo viene fatta una valutazione dell'efficienza del modello sulle domande in italiano che vengono tradotte automaticamente in inglese in modo da poter essere utilizzate dal modello come input.

[4] *GQA Challenge 2019*: *https://eval.ai/web/challenges/challenge-page/225/leaderboard/733*, consultato in data 23 settembre 2021.

Per la traduzione automatica viene utilizzata l'API di Google Translate[5] attraverso la libreria di Python *deep-translator*[6].

Viene calcolata la metrica BLEU (Papineni et al. 2002) per avere un primo dato sulla qualità della traduzione dalle domande dall'italiano all'inglese. Sulla traduzione di Google Translate dall'italiano all'inglese per le domande, il punteggio BLEU è di 0,71 sul dataset italiano e di 0,73 sul campione di domande estratte dal test set di GQA. Entrambi i punteggi sono abbastanza alti e in linea con la lunghezza media delle domande.

Oltre al punteggio BLEU, viene fatta anche una valutazione manuale dell'accuratezza della traduzione automatica; Tavosanis (2019) dimostra infatti che il punteggio di BLEU ottenuto nella traduzione dall'inglese all'italiano utilizzando Google Translate si discosta da quello ottenuto attraverso una valutazione manuale (Tavosanis 2019).

In ogni caso, questo punteggio BLEU suggerisce che il materiale tradotto in inglese è caratterizzato da un basso livello di rumore dovuto al processo di traduzione automatica per entrambi i dataset.

Una volta fornite le domande in input, il modello produce come output le risposte in inglese. Queste vengono utilizzate come risposta per l'esperimento sull'inglese e tradotte automaticamente dall'inglese all'italiano per le domande in italiano.

Il *workflow* del funzionamento del modello è illustrato nella figura 2.10.

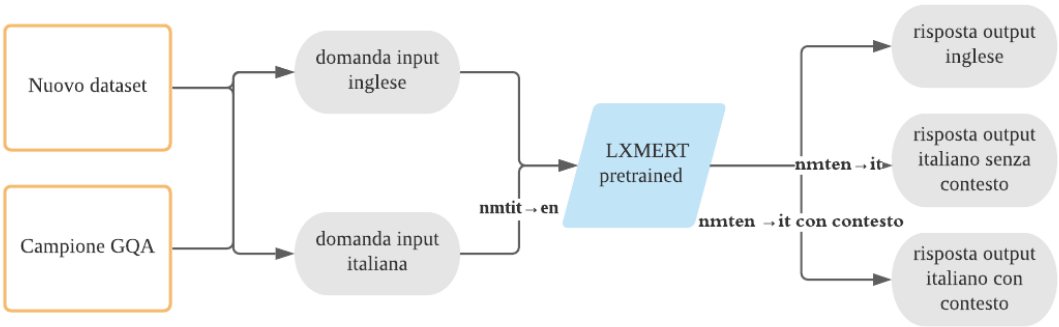

Figura 2.10: *Workflow* degli esperimenti LXMERT-pretrained + NMT.

[5] *Google Translate API*: *https://cloud.google.com/translate/docs/languages*, consultato in data 30 settembre 2021.

[6] *deep-translator documentation*: *https://deep-translator.readthedocs.io/en/latest/*, consultato in data 30 settembre 2021.

La traduzione automatica delle risposte dall'inglese all'italiano si rivela complessa. Sono infatti presenti numerosi casi in cui la risposta è un termine polisemico in inglese che non presenta un corrispettivo italiano che ne catturi tutti i sensi. In questo caso, dal momento che non sono presenti il contesto o l'immagine di riferimento, il traduttore fornisce spesso una traduzione errata. La traduzione delle risposte in italiano viene quindi fornita sia senza contesto sia utilizzando la domanda come contesto, come proposto in Croce (2019).

L'applicazione del modello produce quindi tre output di coppie domanda/risposta per ogni domanda dei due dataset: (*i*) la risposta in inglese originale alla domanda in inglese per valutare gli errori del modello, (*ii*) la traduzione automatica in italiano della risposta del modello alla domanda tradotta automaticamente dall'italiano all'inglese, e (*iii*) la risposta tradotta automaticamente in italiano utilizzando la domanda come contesto.

2.2.4 Valutazione

L'obiettivo della valutazione è quello di compiere un'analisi qualitativa e approfondita dei risultati per mettere in luce i punti di forza e di debolezza dei modelli pre-addestrati usati per il VQA e per la traduzione automatica.
A partire dagli output ottenuti vengono proposte una classificazione degli errori con la relativa quantificazione e una valutazione complessiva dei risultati ottenuti dal modello.

La classificazione degli errori avviene manualmente a opera di due annotatori e porta all'individuazione di tre classi: errori del modello, errori nella traduzione della domanda ed errori nella traduzione della risposta. All'interno di ogni classe vengono valutate le differenze qualitative e quantitative degli errori nei dataset utilizzati.

La valutazione confronta l'accuratezza delle risposte prodotte dall'input in inglese e dall'input in italiano, prima relativamente al *test* set di GQA, poi attraverso un confronto tra un campione di quest'ultimo e il dataset italiano prodotto per l'esperimento. Viene effettuata anche una valutazione su quanto il contesto-domanda risolva la traduzione delle

risposte e migliori l'accuratezza. Il calcolo dell'accuratezza avviene manualmente a opera dei due annotatori che osservano le immagini e le coppie domanda/risposta prodotte dal *framework* utilizzato. La scelta della valutazione manuale è motivata da diversi fattori: al momento non sono ancora state rilasciate le risposte del *balanced test* set di GQA e sono disponibili solo le domande, e come illustrato precedentemente, per il nuovo dataset dell'esperimento non è stato possibile definire una singola risposta alle domande.

Dataset	**Campione test-dev GQA**	**Nuovo dataset**	**Campione balanced-test set GQA**
Fonte immagini	Test-dev GQA	Crowd-sourcing	Balanced test set GQA
Tot. immagini	398	100	100
Tot. domande	3.000	1.000	1.000
N. domande/N. immagini	7,54	10	10
Categorie domande	10	10	10
Collezione domande	Automatica	Umana	Automatica
Traduzione in italiano	Automatica revisionata	Manuale	Automatica
Metriche di valutazione	Accuratezza	2 annotatori	2 annotatori

Tabella 2.5: Caratteristiche dei dataset utilizzati per gli esperimenti.

3. Valutazione

3.1 Valutazione degli esperimenti NMT + LXMERT-it

Per valutare l'accuratezza di LXMERT-it sul *test set gold*, vengono confrontate l'accuratezza ottenuta da modelli pre-addestrati in inglese, quella ottenuta una *baseline*, quella ottenuta da un *workaround* e quella ottenuta dal modello riaddestrato sul dataset tradotto, come indicato nel capitolo precedente.
Un sistema *baseline* che assegna la risposta più frequente (in questo caso "yes"/"sì") ottiene un'accuratezza del 18,5% (Croce, Passaro et al. 2021).
Il modello pre-addestrato migliore di Tan e Bansal (2019) ottiene un'accuratezza del 60,0% sul *test-dev* e del 59,0% sul sottoinsieme *gold*. L'accuratezza scende al 56,2% utilizzando BERT originale pre-addestrato e al 55,3% utilizzando BERT multilingue. Questo calo di prestazioni conferma i risultati di Tan e Bansal (2019) e rappresenta una sorta di *upper bound* per gli esperimenti in italiano, in quanto tutti i *setup* precedenti non sono influenzati dal rumore introdotto nel materiale di allenamento di GQA-it (Croce, Passaro et al. 2021).

Per quanto riguarda i risultati del modello LXMERT-it, la prima misura di accuratezza utilizzata come *baseline* per l'italiano è quella del *workaround* che traduce automaticamente input e output per poter utilizzare il modello pre-addestrato in inglese. Questo esperimento ottiene un'accuratezza del 47,1% con BERT pre-addestrato e del 44,8% con BERT multilingue (Croce, Passaro et al. 2021).

Questo calo è parzialmente dovuto alla traduzione automatica dall'italiano all'inglese, poiché le prestazioni del modello *en-pretrain* scendono dal 59,0% al 54,5% quando viene applicato a domande inglesi derivate dalla traduzione automatica, mentre il *bert-multilingue* scende dal 55,3% al 51,3%: questo indica che il modello non è robusto al rumore ottenuto

	Modello	Acc.
	baseline (most frequent answer)	18,5%
en	en-pretrain	59,0%
	bert-multilingual	55,3%
it	mtit→en + **en-pretr.** + mten→it	47,1%
	mtit→en + **en-multi.** + mten→it	44,8%
	bert-multilingual (gold ans.)	51,0%
	bert-multilingual (silver ans.)	52,6%

Tabella 3.1: Risultati di LXMERT e LXMERT-it su 3.000 domande di GQA e GQA-it (Croce, Passaro et al. 2021).

introdotto dalla NMT (Croce, Passaro et al. 2021).
Il restante calo di prestazioni è tuttavia dovuto alla traduzione dall'inglese all'italiano, principalmente a causa della polisemia e degli altri fenomeni discussi nella sezione precedente (Croce, Passaro et al. 2021).

Al contrario, il modello addestrato su GQA-it, cioè LXMERT-it, raggiunge il 51,0% di accuratezza, ovvero un risultato migliore di quelli ottenuti con il *workaround* e più che in linea con i risultati ottenuti con BERT multilingue in inglese. Valutando LXMERT-it con le risposte generate con la normalizzazione proposta (cioè le risposte *silver*), l'accuratezza sale al 52,6%. Un'analisi manuale delle differenze rivela che esse sono dovute principalmente a inflessioni di genere (ad esempio, "alto" e "alta", in inglese "tall"). Sfortunatamente, questi casi saranno inevitabilmente classificati male da LXMERT-it poiché durante l'addestramento ha osservato solo forme maschili (che sono state introdotte durante la fase iniziale di normalizzazione).

3.1.1 Errori nelle risposte

Un'analisi quantitativa degli errori sulle risposte *silver* su un campione casuale del 10% dataset di *test* riporta che nel complesso il 44% delle domande produce un errore. Dal punto di vista qualitativo vengono identificate 6 classi di errore principali (Croce, Passaro et al. 2021):

- Errori del modello nel riconoscimento oggetto/i: il 31% degli errori è dovuto all'identificazione di un oggetto sbagliato (ad esempio, "tavola" al posto di "sedia");
- Errori del modello nel riconoscimento attributo/i: il 14% degli errori è dovuto all'identificazione degli attributi dei vari oggetti (ad esempio,"blu" al posto di "nero" "chiuso" al posto di "aperto");
- Errori che riguardano caratteristiche spaziali: il 2% degli errori è dovuto a problemi relativi alle caratteristiche spaziali degli oggetti (ad esempio, "destra" al posto di "sinistra");
- Errori di sinonimia o iperonimia: il 17% degli errori è dovuto a sinonimi e iperonimi della risposta attuale (ad esempio "persona" al posto di "donna');
- Errori morfologici di genere e numero: il 3% degli errori è dovuto a caratteristiche morfologiche errate (ad esempio, "bella" al posto di "bello" o "persona" al posto di "persone");
- Errori nella resa verbale: il 3% degli errori è dovuto a un modo errato di esprimere le azioni (ad esempio, "sta dormendo" al posto di "sta sdraiato");
- Errori residuali: il 31% degli errori sono lasciati in una classe residuale (ad esempio "sì" al posto di "no").

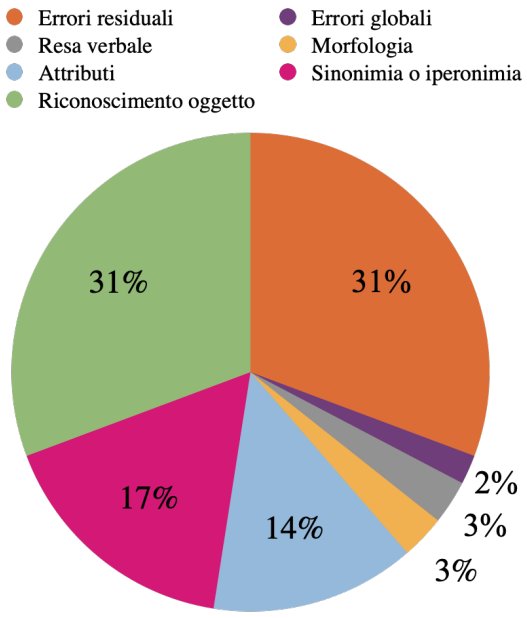

Figura 3.1: Distribuzione degli errori nelle risposte del *test set gold* di GQA-it.

3.1.2 Analisi dei *pattern* delle domande

Durante la fase di revisione manuale delle traduzioni automatica per la costruzione del *dev-test set gold* viene compiuta un'analisi qualitativa e quantitativa sui *pattern* che vengono sbagliati in maniera ricorrente dalla traduzione automatica con Opus NMT e su quelli la cui traduzione, seppur non errata, non può rispecchiare una forma standard di domanda in italiano a causa delle differenze linguistiche tra italiano e inglese.
Sulle 3.000 coppie domanda/risposta del dataset sono corrette 2.712 domande (il 90,4%) e 2.867 risposte (il 95,56%).
Il punteggio BLEU delle domande è elevato, ovvero dello 0,93, grazie alla metodologia utilizzata per la validazione che consiste nella correzione degli errori e non nella traduzione *ex novo* e evidenzia che il dataset è caratterizzato da un basso rumore dovuto alla traduzione automatica. Tuttavia, nella fase di revisione manuale, vengono rilevati alcuni *pattern* di domande che sono sbagliati in maniera ricorrente dalla NMT, diversi errori dovuti all'ambiguità lessicale dell'inglese e alcuni errori che derivano dalla perdita di elementi nella traduzione in italiano. In particolare gli errori nella traduzione automatica delle domande

dall'inglese all'italiano possono essere classificati in:

- Errori di ambiguità lessicale: la domanda non fornisce sufficiente contesto per risolvere la polisemia del termine inglese, per esempio "Is the floor tan and dirty?" tradotto in "Il pavimento è abbronzato e sporco?" invece che in "Il pavimento è marroncino e sporco?", oppure casi di iponimia e iperonimia, per esempio "What do you think is the watercraft that is lying next to the beach?" tradotto in "Quale pensi che sia la moto d'acqua che giace vicino alla spiaggia?" invece che in "Cosa pensi che sia l'imbarcazione che giace vicino alla spiaggia?". L'ambiguità lessicale riguarda tutte le classi morfologiche, ovvero sostantivi, aggettivi e verbi;

- Errori nei *pattern*: alcuni pattern presentano traduzioni automatiche errate ricorrenti, si presentano in una forma non *standard* o non si prestano a una traduzione in italiano, per esempio il *pattern* "*What + oggetto + preposizione articolata*", ad esempio "What animal is the couch behind of?", viene tradotto in modo errato in maniera ricorrente: nel caso dell'esempio precedente la traduzione automatica è "Cosa è il tavolo di fronte?" invece che "Di quale animale è fatto il divano?";

- Omissione di uno o più termini nella traduzione: in particolare, la traduzione automatica tende a omettere "seem" e "seems" e "look" e "looks" utilizzati come "sembra"/"sembrano" e gli aggettivi dimostrativi. Per esempio, "Do the shoes look black or pink?" è tradotto in "Le scarpe sono nere o rosa?" invece che in "Le scarpe sembrano nere o rosa?" o "Is this a skateboarder or a snowboarder?" è tradotto in "È uno skateboarder o uno snowboarder?" invece che in "Questo è uno skateboarder o uno snowboarder?". Un'altra omissione frequente è quella che riguarda alcuni aggettivi che vengono incorporati al termine a cui si riferiscono, per esempio "Is the young child to the right or to the left of the person that is wearing a shirt?" è tradotto in "Il bambino è a destra o a sinistra della persona che indossa una camicia?";

- Errori residuali: traduzioni errate che non è possibile ricondurre a un fenomeno linguistico preciso. All'interno di questa classe vengono fatti rientrare anche i casi in cui la traduzione automatica non è propriamente errata, ma non è totalmente

conforme a come verrebbe prodotta da un parlante umano, per esempio "What sits on top of the desk?", tradotto automaticamente in "Cosa si trova in cima alla scrivania?" e validato in "Cosa si trova sopra la scrivania?".

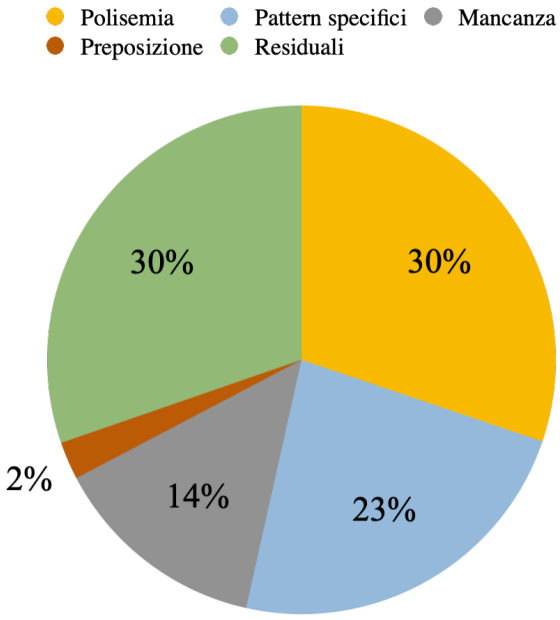

Figura 3.2: Distribuzione degli errori nella NMT delle domande del *test set gold* di GQA-it.

Gli errori più gravi tra quelli riscontrati sono quelli che rendono incomprensibili o errate le domande, e quindi gli errori di polisemia e quelli che riguardando i *pattern* delle domande. Se la mancanza di un aggettivo dimostrativo o la traduzione non precisa di una preposizione non impediscono la corretta interpretazione di una domanda, la traduzione errata di un termine o della struttura della domanda rende incomprensibili o sbagliate le domande.

Per quanto riguarda gli errori di polisemia, è interessante notare che il 9% di essi è dovuto alla traduzione errata del termine "sitting" che viene utilizzato in inglese per indicare non solo il senso dell'espressione italiana "essere seduto" ma anche il senso di "essere posato". Il traduttore automatico, tuttavia, non coglie questa sfumatura di significato e traduce tutto con la prima accezione; per esempio "What is the gray towel sitting on?" è tradotto in "Su cosa è seduto l'asciugamano grigio?" invece che con "Su cosa è posato l'asciugama-

no grigio?". Questa tipologia di polisemia non è tanto grave da portare a un'impossibilità nella comprensione della domanda, ma deve comunque essere validata per l'italiano.
Il fenomeno di errori più interessante riguarda sicuramente i *pattern* tradotti sistematicamente in modo errato. Le domande di GQA sono infatti create composizionalmente, a partire da *pattern* pensati per l'inglese. Inoltre, la natura composizionale delle domande risulta spesso in frasi lunghe e sintatticamente complesse; questo fa sì che anche la traduzione italiana risulti innaturale.
In particolare, il *pattern* "*What + oggetto + preposizione articolata*", ad esempio "What is that table in front of?", viene tradotto in modo errato in maniera ricorrente: nel caso dell'esempio precedente la traduzione automatica è "Cosa è il tavolo di fronte?" invece che "A cosa è davanti il tavolo?". Come si può osservare dalla figura 3.2, questo errore è secondo per frequenza solo agli errori di polisemia e a quelli residuali, e risulta in una traduzione in italiano totalmente incomprensibile per un parlante nativo, per il quale si rivela spesso impossibile individuare il soggetto o il senso della traduzione automatica senza avere a disposizione la domanda in inglese originale.
Altri *pattern*, invece, si rivelano non adatti a una traduzione italiano per le differenze linguistiche tra italiano e inglese. In particolare, un *pattern* di domande molto frequente in inglese è "*What/Which type + is + object?*", ad esempio "What type of animal is behind the person on the right?" che in italiano è tradotto letteralmente in "Che tipo di animale è dietro alla persona sulla destra?". In italiano la domanda tenderebbe ad essere posta con la forma di "Quale animale è dietro alla persona sulla destra?", ma è necessario mantenere la traduzione letterale per avere una corretta corrispondenza tra italiano e inglese.
Un'ulteriore differenza deriva dal fatto che in inglese è necessario specificare il soggetto della domanda anche in forma pronominale, mentre in italiano questo non è frequente. Per esempio la domanda inglese: "What is she kicking?" viene tradotta automaticamente come "Cosa sta calciando?". In questo caso, la traduzione è corretta, ma si presenterebbe ambigua per un modello per la risposta automatica su una frase: la mancanza di specificazione del soggetto, infatti, può portare a un'interpretazione errata della domanda e alla mancanza di distinzione del genere del soggetto.
Il rumore nella traduzione delle domande è comunque contenuto, come dimostrano il pun-

teggio di BLEU e l'alta percentuale di domande per le quali non è stata necessaria una validazione.

3.2 Valutazione degli esperimenti con LXMERT- pre-trained + NMT

Per la valutazione degli esperimenti LXMERT-pretrained + NMT, che utilizzano LXMERT pre-addestrato in inglese attraverso traduzione automatica di input e output, viene compiuta una valutazione dell'accuratezza del campione estratto dal *balanced test* set di GQA e quella del nuovo dataset creato per l'esperimento confrontando i risultati con quelli ottenuti utilizzando il modello direttamente in inglese.
Vengono inoltre valutati dal punto di vista qualitativo e quantitativo gli errori presenti nella traduzione automatica delle domande, dall'italiano all'inglese, e delle risposte, dall'inglese all'italiano. Le prestazioni del modello vengono valutate manualmente da due annotatori.

L'accuratezza delle risposte alle domande in inglese del campione del *balanced test* set GQA è di 80,1% (801 risposte su 1.000 vengono valutate corrette). Sullo stesso campione, l'accuratezza scende al 71,4% per la risposta alle domande in italiano (in questo caso, vengono valutate corrette 714 risposte su 1.000). Questo indica un calo dell'8,7% dell'accuratezza (in particolare, c'è una differenza di 87 risposte corrette).

Per quanto riguarda il nuovo dataset creato per gli esperimenti, su un totale di 1.000 domande, il modello ottiene un'accuratezza del 75,7% sull'inglese, con la risposta corretta a 757 domande su 1.000. L'accuratezza del modello scende al 64,8%, con la risposta corretta a 648 domande su 1.000, per la risposta alle domande in italiano. Questo indica un calo del 10,9% nell'accuratezza (in particolare, c'è una differenza di 109 risposte corrette).

La differenza di accuratezza sull'inglese tra i due dataset è del 4,4% ed è spiegabile con la differenza dei dati. Nel caso del campione di immagini e domande estratte dal *balanced*

test set di GQA i dati sono più simili a quelli su cui il modello è stato addestrato. Inoltre, la distribuzione delle domande per tipologia della risposta (figure 2.7 e 2.4) e per soggetto/i delle risposte (figure 2.8 e 2.5) è diversa per i due dataset.

Il calo dell'accuratezza sulle domande a partire dall'italiano riguarda entrambi i dataset e dipende dal rumore introdotto dalla traduzione automatica. La differenza nel calo dell'accuratezza tra i due dataset è minima (il 2,2%) e interessa maggiormente il nuovo dataset. Il rumore può verificarsi nella traduzione automatica della domande dall'italiano all'inglese, impedendo al modello di fornire la risposta adeguata, o nella traduzione automatica della risposta dall'inglese all'italiano, compromettendo il funzionamento corretto del sistema inglese.

	Occorrenza	Percentuale di accuratezza
Risposte corrette all'input in inglese originale	801	80,1%
Risposte corrette all'input in inglese tradotto	714	71,4%

Tabella 3.2: Accuratezza di LXMERT-pretrained sul campione del *balanced test* set di GQA.

	Occorrenza	Percentuale di accuratezza
Risposte corrette all'input in inglese originale	757	75,7%
Risposte corrette all'input in inglese tradotto	648	64,8%

Tabella 3.3: Accuratezza di LXMERT-pretrained sul nuovo dataset.

Per la valutazione dell'accuratezza viene fatta un'analisi sugli errori del modello (ovvero quelli presenti nelle risposte sull'inglese), e vengono poi valutati gli errori che si aggiungono a questi nella risposta a domande in italiano, e che derivano quindi dal rumore generato dalla traduzione automatica con Google Translate.

3.2.1 Errori del modello

Con errori del modello si intendono le domande in inglese originale a cui il modello LXMERT risponde in maniera errata.
Il calcolo degli errori del modello consente di osservare in quali casi la risposta errata dipende dal modello e in quali da errori di traduzione automatica a livello della risposta o della domanda.
Gli errori del modello vengono divisi in classi in base alla classificazione semantica delle domande di GQA in:

- Errori globali: errori nella risposta a domande globali che riguardano le condizioni atmosferiche o il luogo della scena;
- Errori sulle relazioni: errori nella risposta a domande sulla relazione tra oggetti nella scena;
- Errori sugli attributi: errori nella risposta a domande sugli attributi (per esempio, azioni, colori, materiali ecc.);
- Errori sugli oggetti: errori nella risposta a domande sul riconoscimento di uno o più oggetti.

Tipologia di errore	Campione di GQA	Nuovo dataset
Globale	3	19
Riconoscimento relazione	25	23
Riconoscimento oggetto	87	112
Riconoscimento attributi	84	89

Tabella 3.4: Occorrenze degli errori del modello LXMERT pre-addestrato sui due dataset analizzati.

Come si può osservare dalla tabella 3.4, il modello è più accurato sui dati ottenuti da GQA, che sono più simili ai dati osservati nella fase di addestramento e validazione.

Infatti, gli errori nelle risposte alle domande in inglese sono il 2,4% nel per il dataset in italiano, e l'1,9% per il campione di GQA.
Osservando gli errori è inoltre possibile notare che sono coerenti con la distribuzione dei soggetti delle di entrambi i dataset presi in esame (figure 2.8 e 2.5).
Per il campione di GQA sono più frequenti gli errori che riguardano le relazioni, seguiti da quelli che riguardano gli attributi, e poi quelli che riguardano domande globali, mentre per quanto riguarda il nuovo dataset dopo gli errori nel riconoscimento delle relazione sono più frequenti gli errori sul riconoscimento degli oggetti, seguiti da quelli sugli attributi e quelli globali, coerentemente con le diverse distribuzioni delle domande.

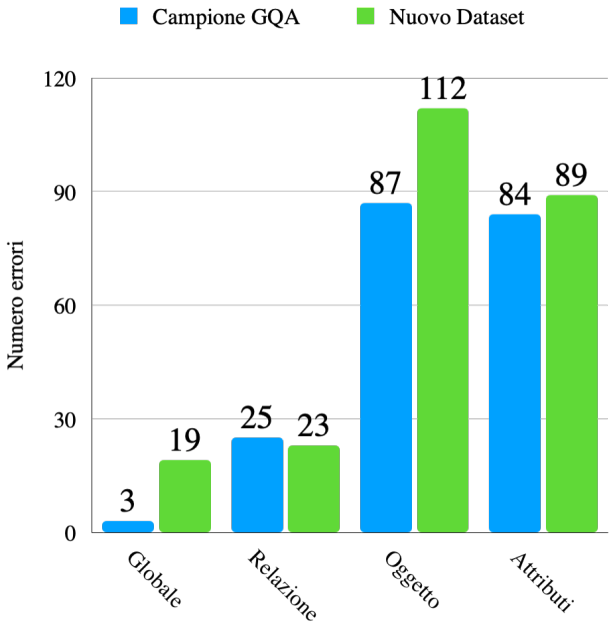

Figura 3.3: Distribuzione degli errori del modello per i due dataset.

3.2.2 Errori nella traduzione automatica delle domande

Vengono presi in considerazione gli errori nella traduzione automatica delle domande dall'italiano all'inglese. Nella fase di valutazione è stato possibile individuare due classi di errore: errori di polisemia e errori di traduzione automatica errata.

Gli errori di polisemia indicano i casi di ambiguità lessicale in cui un termine è polisemico in italiano e non esiste un corrispettivo in inglese che ne catturi tutti i sensi. Come

Tipologia di errore	Campione di GQA	Nuovo dataset
Polisemia	2	8
Traduzione errata	2	13

Tabella 3.5: Occorrenze degli errori nella traduzione automatica delle domande dall'italiano all'inglese sui due dataset analizzati.

visto precedentemente per gli errori nella traduzione dall'inglese all'italiano, il traduttore automatico non dispone del contesto e assegna spesso un significato errato. Per esempio, la domanda in inglese del campione di GQA "What is tied to the pants?", validata dagli annotatori in "Cosa è legato ai pantaloni?", viene tradotta automaticamente in "What is related to the pants?". In questo caso l'ambiguità lessicale riguarda "tie", che viene interpretato con il senso di "relazionato" e non di "allacciato". In questi casi il modello è in grado di rispondere correttamente all'input inglese originale ma non a quello tradotto dall'italiano.

Gli errori di traduzione automatica errata comprendono quei casi in cui avviene una traduzione grossolana della domanda che non è spiegabile con motivazioni linguistiche. Per esempio la domanda in inglese appartenente al campione di domande di GQA "Which kind is the food?", tradotta manualmente con il corrispettivo italiano "Di che tipo è il cibo?" viene tradotta da Google Translate come "What is the food like?", in questo caso, come indicato nel paragrafo 3.1.2, il *pattern* della domanda non si presta a una traduzione italiana letterale, e il traduttore non riesce a fornire una traduzione che ne catturi il senso corretto. Il modello, infatti, risponde correttamente all'input tradotto dall'italiano che si presenta come una domanda completamente diversa.

Differentemente dalle risposte, in cui per cercare di risolvere i casi di polisemia può essere utilizzata la domanda come contesto, in questo caso non è possibile aggiungere ulteriore contesto linguistico che consenta di individuare il senso corretto di un termine. Come si può osservare nella tabella 3.5, sul *balanced test set* di GQA gli errori nella traduzione della domanda sono solo 4, costituendo lo 0,4% del totale, mentre nel nuovo

dataset sono 21, costituendo il 2,1% del totale delle domande.
Questi errori non incidono quindi particolarmente sul calo dell'accuratezza causato dalla traduzione automatica; la maggior parte degli errori non dipendenti dal modello, dunque, riguarda la traduzione automatica delle risposte dall'inglese all'italiano.

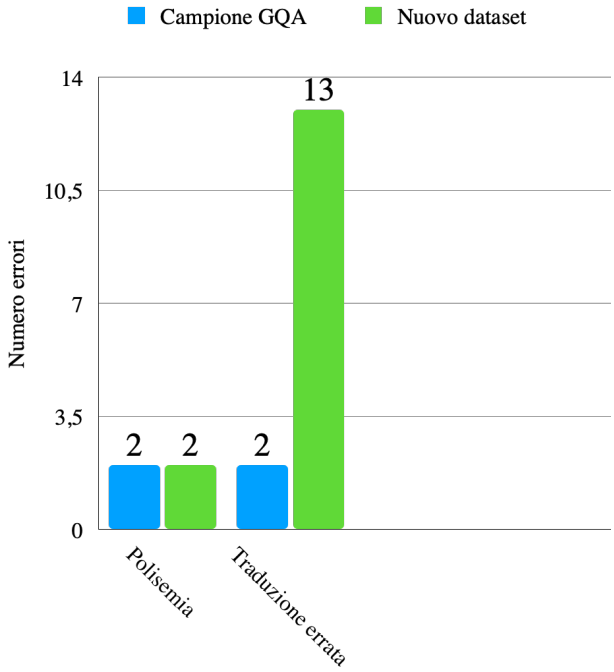

Figura 3.4: Distribuzione degli errori nella traduzione automatica delle domande per i due dataset.

3.2.3 Errori nella traduzione automatica delle risposte

Gli errori nella traduzione in italiano delle risposte originate in inglese come output del modello sono ricorrenti a livello morfologico e semantico per tutte le classi morfologiche.
Al fine di poter compiere una valutazione qualitativa e quantitativa della loro occorrenza, gli errori di traduzione delle risposte vengono analizzati in cinque classi: (*i*) errori di polisemia, (*ii*) errori di genere, (*iii*) errori di numero, (*iv*) errori di numero e genere e (*v*) errori nella resa verbale.
Gli errori di polisemia sono dovuti all'ambiguità lessicale di un termine inglese del quale non esiste un corrispettivo in italiano che ne catturi tutti i sensi; in questo caso la traduzione automatica assegna spesso un senso non corretto. L'esempio più frequente di questa

tipologia di errori riguarda la risposta "right", che è molto frequente come risposta alle domande globali o di relazione. Il termine inglese è polisemico e può essere tradotto in due lemmi italiani: "destro" e "giusto". Dal momento che viene questa risposta è utilizzata nelle domande di localizzazione, il senso corretto in italiano sempre "destro". Il traduttore automatico, tuttavia, traduce sempre la parola in "giusto". Errori dovuti alla polisemia si verificano non solo a livello di sostantivi, ma anche di aggettivi e verbi. Per esempio, la domanda del nuovo dataset "What color is the liquid inside the glass?" ha come risposta "dark" che viene tradotto automaticamente in "buio" invece che in "scuro". Un esempio di polisemia verbale riguarda la riposta "running", che viene tradotta in "in esecuzione" invece che in "correre", per esempio "What is the horse doing? Running" viene tradotto automaticamente in "Cosa sta facendo il cavallo? In esecuzione" invece che in "Cosa sta facendo il cavallo? Sta correndo".

Oltre agli errori semantici, nella traduzione automatica della risposta sono molto frequenti anche errori morfologici.

Gli errori morfologici più frequenti sono gli errori nella traduzione del genere; dal momento che in inglese non esiste il genere per sostantivi e aggettivi, il traduttore automatico spesso non identifica la corretta concordanza del genere con la domanda tradotta manualmente in italiano. Per esempio, la risposta alla domanda del nuovo dataset "Who is wearing the dress" è "baby", tradotto in italiano come "bambino".

Gli errori nel numero indicano il caso in cui non è presente una concordanza tra la domanda e in numero della risposta. Per esempio, la risposta alla domanda del campione di GQA "How large are the glasses the man is to the right of?", in italiano "Quanto sono larghi gli occhiali alla destra dei quali è l'uomo?", è "small", che viene tradotto in italiano con "piccolo" al posto di "piccoli".

Gli errori nel genere e numero includono entrambi gli errori precedenti.

Gli errori nella resa verbale indicano un errore nella traduzione del modo o della persona verbale: in inglese le risposte verbali del modello presentano sempre la forma in *-ing*. In italiano non esiste una forma verbale universale corrispondente che sia priva di genere e numero: il traduttore automatico traduce utilizzando modi diversi, come infinito (per

Tipologia di errore	Campione GQA	Nuovo dataset
Polisemia	64	37
Genere	8	29
Numero	1	0
Genere e numero	1	0
Resa verbale	5	22

Tabella 3.6: Occorrenze degli errori nella traduzione automatica delle risposte dall'inglese all'italiano sui due dataset analizzati.

esempio, "eating" viene tradotto in "mangiare"), participio passato (per esempio, "sitting" viene tradotto in "seduto") e presente (per esempio,"playing" viene tradotto in "gioca").

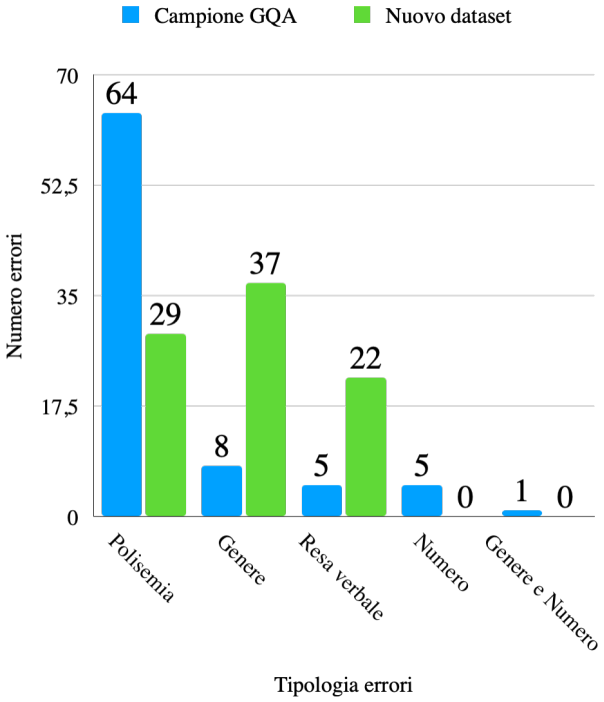

Figura 3.5: Distribuzione degli errori nella traduzione automatica delle risposte per i due dataset.

Come è possibile osservare nella tabella 3.6, il rapporto tra gli errori si mantiene lo stesso infatti: nel campione estratto da GQA il 7,9%, delle risposte è errato, con 79 risposte

	Campione di GQA	Nuovo dataset
Accuratezza sull'input inglese	80,1%	75,7%
Accuratezza sull'input italiano	71,4%	64,8%
Accuratezza con errori non gravi	73,3%	69,9%

Tabella 3.7: Differenze nell'accuratezza dei due dataset considerando gli errori non gravi.

errate su 1.000, mentre nel nuovo dataset l'8,8% delle risposte è errato, con 88 risposte errate su 1.000.

Anche in questo caso, tuttavia, ci sono differenze nella distribuzione degli errori: in particolare, gli errori di polisemia sono molto più frequenti nel campione di GQA rispetto al nuovo dataset, mentre nel nuovo dataset sono più frequenti gli errori di genere e di resa verbale ma non sono presenti errori di numero o di genere e numero.
La natura dei risultati, in questo caso, non è spiegabile con la distribuzione del soggetto/i delle domande nei campioni presi a riferimento: nel nuovo dataset, infatti, sono presenti meno domande su attributi, che richiedono una flessione per genere, e dunque ci si aspetterebbe che siano anche presenti meno errori appartenenti a questa categoria.
Per quanto riguarda la gravità degli errori della traduzione della risposta, è importante notare che gli errori di genere, di numero e di resa verbale non impediscono di comprendere il significato della risposta, anche se questa non è morfologicamente coerente con la domanda.
Si rivela interessante, quindi, calcolare l'accuratezza al netto degli errori non gravi, ovvero quelli di genere, numero e resa verbale; l'accuratezza per i due dataset è rappresentata nella tabella 3.7.

In questo caso, gli errori di polisemia per il campione di GQA sono 68, implicando un calo dell'accuratezza del 6,8% rispetto all'input in inglese originale.
Nel nuovo dataset gli errori di polisemia e di traduzione errata delle domande sono 58, implicando un calo dell'accuratezza del 5,8% rispetto all'input in inglese.
La differenza di accuratezza ottenuta, in questo caso, si rivela minima.
È interessante notare che Google Translate non è particolarmente utile per la risoluzione

della polisemia nelle risposte. Infatti, il traduttore automatico, per l'italiano, traduce separatamente la domanda e la risposta. Per esempio, traducendo "On which side is the dark bag? Right" si otterrà la traduzione "Da che lato dell'immagine è la borsa scura? Giusto". L'introduzione del contesto, quindi non riesce a risolvere alcun caso di polisemia o di mancata concordanza tra risposta e domanda. Tutti gli errori nella traduzione automatica delle domande sarebbero infatti risolvibili attraverso l'utilizzo di un traduttore automatico in cui il contesto-domanda consente di risolvere gli errori di polisemia e flessione della risposta.

3.3 Confronto tra l'utilizzo di NMT + LXMERT-it e LXMERT-pretrained + NMT

La creazione del dataset GQA-it (Croce, Passaro et al. 2021) dimostra che attraverso la traduzione automatica è possibile ottenere una risorsa utile per il VQA in lingue diverse dall'inglese senza la necessità di una costosa annotazione manuale delle immagini per un'altra lingua; il rumore prodotto dalla traduzione automatica sembra essere tollerato dalla rete utilizzata per l'addestramento (come analizzato nei paragrafi 3.1.2 e 3.1.1).

Entrambi i sistemi analizzati per il VQA in italiano si rivelano soluzioni accettabili. Sia il sistema NMT + LXMERT-it, illustrato nella sezione 2.1, sia il sistema LXMERT-pretrained + NMT, illustrato nella sezione 2.2, ottengono infatti un'accuratezza soddisfacente sui dataset sui quali vengono testati.
Tuttavia, il riaddestramento di un modello compiuto in NMT + LXMERT-it è computazionalmente molto costoso, e i costi di computazione elevati rendono impossibile riprodurre i risultati del modello LXMERT (Tan et al. 2019) pre-addestrato che ottiene l'accuratezza più alta.
L'utilizzo del modello pre-addestrato in inglese attraverso la sola traduzione di domanda in input e risposta in output, compiuto in LXMERT-pretrained + NMT, consente non solo di ridurre i costi computazionali del riaddestramento di un modello, ma anche di utilizzare la versione più performante del *framework* LXMERT; inoltre, questo approccio può essere

facilmente esteso a qualsiasi modello pre-addestrato.
La perdita di accuratezza causata dal rumore introdotto dalla traduzione automatica è bassa per entrambe le metodologie utilizzate, come è possibile osservare dalla tabella 3.8. È bene tenere in considerazione che i traduttori utilizzati per gli esperimenti sono diversi. Google Translate si rivela, tra i traduttori disponibili gratuitamente, la risorsa migliore, ma mette a disposizione un numero limitato di caratteri da tradurre e non può quindi essere utilizzato per la traduzione di tutto il materiale di un dataset delle dimensioni di GQA. Tuttavia, a differenza di Opus NMT, con Google Translate l'utilizzo della domanda-contesto non risolve gli errori di flessione e di polisemia delle risposte e non ottiene quindi traduzione corretta; questo fa sì che negli gli esperimenti con il sistema LXMERT-pretrained + NMT, che utilizza la traduzione automatica di input e output, nessun errore di ambiguità lessicale o di concordanza nelle risposte venga risolto con l'utilizzo del contesto, come accade invece con il sistema NMT + LXMERT-it.

Modello	**Dataset**	**Acc.**	**Acc. attesa**	**Perdita**
LXMERT-it	GQA-it (Gold ans.)	51,0%	59,0%	8,0%
	GQA-it (Silver ans.)	52,6%	59,0%	6,4%
LXMERT-pretrained	campione GQA (risposte flesse)	71,4%	80,1%	8,7%
	campione GQA (risposte non flesse)	73,3%	80,1%	6,8%
	nuovo dataset (risposte flesse)	64,8%	75,7%	10,9%
	nuovo dataset (risposte non flesse)	69,9%	75,7%	5,8%

Tabella 3.8: Confronto della perdita di accuratezza utilizzando NMT + LXMERT-it e LXMERT-pretrained + NMT.

Nonostante tali osservazioni, è possibile notare dalla tabella 3.8 che c'è una corrispondenza tra la perdita di accuratezza dei due sistemi NMT + LXMERT-it e LXMERT-pretrained + NMT.
Si presenta una perdita di accuratezza maggiore per entrambi i sistemi considerando le risposte flesse: in NMT + LXMERT-it questa perdita dipende dal fatto che il modello non ha osservato in fase di addestramento forme femminili ma solo forme normalizzate

al maschile, mentre in LXMERT-pretrained + NMT la perdita dipende dal fatto che il traduttore utilizzato, Google Translate, non risolve la flessione utilizzando le domande come contesto. In quest'ultimo caso, tuttavia, è evidente che l'accuratezza del sistema LXMERT-pretrained + NMT otterrebbe risultati migliori disponendo di un traduttore automatico che tenga maggiormente in considerazione il contesto-domanda.
La perdita di accuratezza minore si ottiene con l'utilizzo di LXMERT-pretrained + NMT sul nuovo dataset, senza tenere in considerazione gli errori di flessione.

Questo è un risultato positivo dal momento che il nuovo dataset è quello in cui le domande in italiano sono state composte da due parlanti madrelingua e sono quindi quelle che rispettano maggiormente una forma naturale per la lingua, come illustrato nella sezione 2.2.3.

Conclusioni

Dalle analisi esposte nell'elaborato emerge che la traduzione automatica consente di ottenere risultati molto promettenti per il VQA in italiano.
In primo luogo, la traduzione automatica permette di approssimare la creazione di risorse in italiano senza l'obbligo di creare un nuovo dataset specifico per la lingua a partire da zero. Come dimostrato dalle analisi quantitative e qualitative di GQA-it (Croce, Passaro et al. 2021), la traduzione automatica delle domande e delle risposte presenta un rumore limitato. Per quanto riguarda la traduzione automatica dei *scene graph* di GQA, invece, la traduzione automatica non si rivela sufficiente. La traduzione automatica, infatti, non è efficace per la traduzione fuori contesto di singole parole e deve essere affiancata da una validazione manuale consistente.

Per quanto riguarda, invece, la creazione di sistemi di VQA, il confronto tra l'accuratezza raggiunta in italiano dal modello riaddestrato su GQA-it e dal modello pre-addestrato in inglese con la traduzione automatica di input e output, rivela che i due sistemi ottengono risultati molto simili, con un calo di accuratezza che varia dal 5,8% al 6,8% rispetto ai risultati degli stessi modelli sull'inglese. Tuttavia, l'utilizzo di un traduttore automatico che prenda maggiormente in considerazione il contesto garantirebbe un aumento sostanziale dell'accuratezza del sistema che utilizza il modello pre-addestrato in inglese. Ciò, associato alla considerazione che il riaddestramento di un modello è molto costoso in termini computazionali, porta a favorire il sistema che utilizza la traduzione automatica di input e output. Inoltre, questo approccio può essere esteso a modelli pre-addestrati più efficienti di LXMERT ed è applicabile a tutte le lingue senza costi computazionali aggiuntivi.

Studi futuri potrebbero estendere le analisi al VQA multilingue: tutti i dataset utilizzati nell'elaborato, infatti, dispongono di annotazioni inglesi e italiane allineate e viene anche presentata una nuova risorsa annotata manualmente da annotatori madrelingua italiani.

Bibliografia

Agrawal, Aishwarya, Dhruv Batra e Devi Parikh (2016). «Analyzing the behavior of visual question answering models». In: *arXiv preprint arXiv:1606.07356*.

Agrawal, Aishwarya, Dhruv Batra, Devi Parikh e Aniruddha Kembhavi (2018). «Don't just assume; look and answer: Overcoming priors for visual question answering». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980.

Agrawal, Aishwarya, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra e Devi Parikh (2016). *VQA: Visual Question Answering*. arXiv: `1505.00468 [cs.CL]`.

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould e Lei Zhang (2018). «Bottom-up and top-down attention for image captioning and visual question answering». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086.

Andreas, Jacob, Marcus Rohrbach, Trevor Darrell e Dan Klein (2016). «Neural module networks». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 39–48.

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick e Devi Parikh (2015). «Vqa: Visual question answering». In: *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives (2007). «Dbpedia: A nucleus for a web of open data». In: *The semantic web*. Springer, pp. 722–735.

Bahdanau, Dzmitry, Kyunghyun Cho e Yoshua Bengio (2014). «Neural machine translation by jointly learning to align and translate». In: *arXiv preprint arXiv:1409.0473*.

Bigham, Jeffrey P, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White

et al. (2010). «Vizwiz: nearly real-time answers to visual questions». In: *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André FT Martins et al. (2019). «Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)». In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz et al. (2016). «Findings of the 2016 conference on machine translation». In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131–198.

Callison-Burch, Chris, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz e Josh Schroeder (2008). «Further meta-evaluation of machine translation». In: *Proceedings of the third workshop on statistical machine translation*, pp. 70–106.

Callison-Burch, Chris, Miles Osborne e Philipp Koehn (2006). «Re-evaluating the role of BLEU in machine translation research». In: *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Cer, Daniel, Christopher D Manning e Dan Jurafsky (2010). «The best lexical metric for phrase-based statistical MT system optimization». In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 555–563.

Cheng, Yong (2019). «Neural Machine Translation». In: *Joint Training for Neural Machine Translation*. Springer, pp. 1–10.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau e Yoshua Bengio (ott. 2014). «On the Properties of Neural Machine Translation: Encoder–Decoder Approaches». In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Doha, Qatar: Association for Computational

Linguistics, pp. 103–111. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012.

Chowdhury, Gobinda G (2003). «Natural language processing». In: *Annual review of information science and technology* 37.1, pp. 51–89.

Cores, Daniel, Manuel Mucientes e Vıéctor M Brea (2020). «RoI feature propagation for video object detection». In: *ECAI 2020*. IOS Press, pp. 2680–2687.

Coughlin, Deborah (2003). «Correlating automated and human assessments of machine translation quality». In: *Proceedings of MT summit IX*. Citeseer, pp. 63–70.

Croce, Danilo, Lucia C. Passaro, Alessandro Lenci e Roberto Basili (2021). «GQA-IT: Italian Question Answering on Image Scene Graphs». In.

Croce, Danilo, Alexandra Zelenanska e Roberto Basili (2019). «Enabling deep learning for large scale question answering in Italian». In: *Intelligenza Artificiale* 13.1, pp. 49–61.

Dasiopoulou, Stamatia, Vasileios Mezaris, Ioannis Kompatsiaris, V-K Papastathis e Michael G Strintzis (2005). «Knowledge-assisted semantic video object detection». In: *IEEE Transactions on Circuits and Systems for Video Technology* 15.10, pp. 1210–1224.

Datta, Debajit, Preetha Evangeline David, Dhruv Mittal e Anukriti Jain (2020). «Neural machine translation using recurrent neural network». In: *International Journal of Engineering and Advanced Technology* 9.4, pp. 1395–1400.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li e Li Fei-Fei (2009). «Imagenet: A large-scale hierarchical image database». In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee e Kristina Toutanova (2018). «Bert: Pre-training of deep bidirectional transformers for language understanding». In: *arXiv preprint arXiv:1810.04805*.

Dollár, Piotr, Christian Wojek, Bernt Schiele e Pietro Perona (2012). «Pedestrian Detection: An Evaluation of the State of the Art». In: *PAMI* 34.

Fayek, Haytham M e Justin Johnson (2020). «Temporal reasoning via audio question answering». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28, pp. 2283–2294.

Ferraro, Francis, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell et al. (2015). «A survey of current datasets for vision and language research». In: *arXiv preprint arXiv:1506.06833*.

Firat, Orhan, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural e Yoshua Bengio (2017). «Multi-way, multilingual neural machine translation». In: *Computer Speech & Language* 45, pp. 236–252.

Forsyth, David e Jean Ponce (2011). *Computer vision: A modern approach.* Prentice hall.

Fukui, Akira, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell e Marcus Rohrbach (2016). «Multimodal compact bilinear pooling for visual question answering and visual grounding». In: *arXiv preprint arXiv:1606.01847*.

Gao, Haoyuan, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang e Wei Xu (2015). «Are you talking to a machine? dataset and methods for multilingual image question answering». In: *arXiv preprint arXiv:1505.05612*.

Girshick, Ross (2015). «Fast r-cnn». In: *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448.

Girshick, Ross, Jeff Donahue, Trevor Darrell e Jitendra Malik (2014). «Rich feature hierarchies for accurate object detection and semantic segmentation». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

Google, Sito (2017). *Translation API Language Support*.

Goyal, Yash, Tejas Khot, Douglas Summers-Stay, Dhruv Batra e Devi Parikh (2017). «Making the v in vqa matter: Elevating the role of image understanding in visual question answering». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913.

Gridach, Mourad (2020). «A framework based on (probabilistic) soft logic and neural network for NLP». In: *Applied Soft Computing* 93, p. 106232.

Gupta, Akshay Kumar (2017). «Survey of visual question answering: Datasets and techniques». In: *arXiv preprint arXiv:1705.03865*.

Gupta, Deepak, Pabitra Lenka, Asif Ekbal e Pushpak Bhattacharyya (2020). «A Unified Framework for Multilingual and Code-Mixed Visual Question Answering». In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 900–913.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li et al. (2018). «Achieving human parity on automatic chinese to english news translation». In: *arXiv preprint arXiv:1803.05567*.

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie e Ross Girshick (giu. 2020). «Momentum Contrast for Unsupervised Visual Representation Learning». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren e Jian Sun (2016). «Deep residual learning for image recognition». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hirschman, L. e R. Gaizauskas (2001). «Natural language question answering: the view from here». In: *Natural Language Engineering* 7.4, pp. 275–300. DOI: 10.1017/S1351324901002807.

Hjelmås, Erik e Boon Kee Low (2001). «Face detection: A survey». In: *Computer vision and image understanding* 83.3, pp. 236–274.

Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin e Hamid Laga (2019). «A comprehensive survey of deep learning for image captioning». In: *ACM Computing Surveys (CsUR)* 51.6, pp. 1–36.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten e Kilian Q Weinberger (2017). «Densely connected convolutional networks». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Hudson, Drew A e Christopher D Manning (2019). «Gqa: A new dataset for real-world visual reasoning and compositional question answering». In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709.

Johnson, Justin, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick e Ross Girshick (2017). «Clevr: A diagnostic dataset for compositional language and elementary visual reasoning». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910.

Johnson, Mark (2009). «How the statistical revolution changes (computational) linguistics». In: *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pp. 3–11.

Kafle, Kushal e Christopher Kanan (2016). «Answer-type prediction for visual question answering». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4976–4984.

— (2017). «Visual question answering: Datasets, algorithms, and future challenges». In: *Computer Vision and Image Understanding* 163, pp. 3–20.

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma et al. (2017). «Visual genome: Connecting language and vision using crowdsourced dense image annotations». In: *International journal of computer vision* 123.1, pp. 32–73.

Kumar, Ankit, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus e Richard Socher (2016). «Ask me anything: Dynamic memory networks for natural language processing». In: *International conference on machine learning*. PMLR, pp. 1378–1387.

Lenci, Alessandro, Simonetta Montemagni e Vito Pirrelli (2005). *Testo e computer. Introduzione alla linguistica computazionale*. Carocci editore.

Lewis, Michael B e Hadyn D Ellis (2003). «How we detect a face: A survey of psychological evidence». In: *International Journal of Imaging Systems and Technology* 13.1, pp. 3–7.

Lu, Dengsheng e Qihao Weng (2007). «A survey of image classification methods and techniques for improving classification performance». In: *International journal of Remote sensing* 28.5, pp. 823–870.

Lu, Jiasen, Jianwei Yang, Dhruv Batra e Devi Parikh (2016). «Hierarchical question-image co-attention for visual question answering». In: *Advances in neural information processing systems* 29, pp. 289–297.

Malinowski, Mateusz e Mario Fritz (2014). «A multi-world approach to question answering about real-world scenes based on uncertain input». In: *Advances in neural information processing systems* 27, pp. 1682–1690.

Masotti, Caterina, Danilo Croce e Roberto Basili (2018). «Deep learning for automatic image captioning in poor training conditions». In: *IJCoL. Italian Journal of Computational Linguistics* 4.4-1, pp. 43–55.

McEnery, Tony e Andrew Wilson (2003). «Corpus linguistics». In: *The Oxford handbook of computational linguistics*, pp. 448–463.

Noh, Hyeonwoo e Bohyung Han (2016). «Training recurrent answering units with joint loss minimization for vqa». In: *arXiv preprint arXiv:1606.03647*.

Pan, Jia-Yu, Hyung-Jeong Yang, Pinar Duygulu e Christos Faloutsos (2004). «Automatic image captioning». In: *2004 IEEE International Conference on Multimedia and Expo (ICME)(IEEE Cat. No. 04TH8763)*. Vol. 3. IEEE, pp. 1987–1990.

Papineni, Kishore, Salim Roukos, Todd Ward e Wei-Jing Zhu (2002). «Bleu: a method for automatic evaluation of machine translation». In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee e Luke Zettlemoyer (2018). «Deep contextualized word representations». In: *arXiv preprint arXiv:1802.05365*.

Pires, Telmo, Eva Schlinger e Dan Garrette (2019). «How multilingual is multilingual BERT?» In: *arXiv preprint arXiv:1906.01502*.

Radford, Alec, Karthik Narasimhan, Tim Salimans e Ilya Sutskever (2018). «Improving language understanding by generative pre-training». In.

Ren, Mengye, Ryan Kiros e Richard Zemel (2015). «Image question answering: A visual semantic embedding model and a new dataset». In: *Proc. Advances in Neural Inf. Process. Syst* 1.2, p. 5.

Ren, Shaoqing, Kaiming He, Ross Girshick e Jian Sun (2015). «Faster r-cnn: Towards real-time object detection with region proposal networks». In: *Advances in neural information processing systems* 28, pp. 91–99.

Rogers, Anna, Matt Gardner e Isabelle Augenstein (2021). «QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension». In: *arXiv preprint arXiv:2107.12708*.

Scaiella, Antonio, Danilo Croce e Roberto Basili (2019). «Large scale datasets for Image and Video Captioning in Italian». In: *IJCoL. Italian Journal of Computational Linguistics* 5.5-2, pp. 49–60.

Somers, Harold (1996). «Machine translation». In: *The Oxford handbook of translation studies*.

Suhr, Alane, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai e Yoav Artzi (2018). «A corpus for reasoning about natural language grounded in photographs». In: *arXiv preprint arXiv:1811.00491*.

Sultana, Farhana, Abu Sufian e Paramartha Dutta (2018). «Advancements in image classification using convolutional neural network». In: *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. IEEE, pp. 122–129.

Szeliski, Richard (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Tan, Hao e Mohit Bansal (2019). «Lxmert: Learning cross-modality encoder representations from transformers». In: *arXiv preprint arXiv:1908.07490*.

Tavosanis, Mirko Luigi Aurelio (2019). «Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano». In.

Thomee, Bart, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth e Li-Jia Li (2015). «The New Data and New Challenges in Multimedia Research». In: *CoRR* abs/1503.01817. arXiv: 1503.01817. URL: http://arxiv.org/abs/1503.01817.

Tiedemann, Jörg e Santhosh Thottingal (2020). «OPUS-MT–building open translation services for the world». In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480.

Toral, Antonio, Sheila Castilho, Ke Hu e Andy Way (2018). «Attaining the unattainable? reassessing claims of human parity in neural machine translation». In: *arXiv preprint arXiv:1808.10432*.

Tu, Zhaopeng, Zhengdong Lu, Yang Liu, Xiaohua Liu e Hang Li (2016). «Modeling coverage for neural machine translation». In: *arXiv preprint arXiv:1601.04811*.

Uijlings, Jasper RR, Koen EA Van De Sande, Theo Gevers e Arnold WM Smeulders (2013). «Selective search for object recognition». In: *International journal of computer vision* 104.2, pp. 154–171.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin (2017). «Attention is all you need». In: *Advances in neural information processing systems*, pp. 5998–6008.

Veit, Andreas, Tomas Matera, Lukas Neumann, Jiri Matas e Serge Belongie (2016). «Coco-text: Dataset and benchmark for text detection and recognition in natural images». In: *arXiv preprint arXiv:1601.07140*.

Way, Andy (2018). «Quality expectations of machine translation». In: *Translation quality assessment*. Springer, pp. 159–178.

Wu, Qi, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick e Anton van den Hengel (2016). *Visual Question Answering: A Survey of Methods and Datasets*. arXiv: `1607.05910 [cs.CV]`.

Wu, Qi, Peng Wang, Chunhua Shen, Anthony Dick e Anton Van Den Hengel (2016). «Ask me anything: Free-form visual question answering based on knowledge from external sources». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4622–4630.

Xu, Huijuan e Kate Saenko (2016). «Ask, attend and answer: Exploring question-guided spatial attention for visual question answering». In: *European Conference on Computer Vision*. Springer, pp. 451–466.

Yang, Shuo, Ping Luo, Chen-Change Loy e Xiaoou Tang (giu. 2016). «WIDER FACE: A Face Detection Benchmark». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Peng, Yash Goyal, Douglas Summers-Stay, Dhruv Batra e Devi Parikh (2016). «Yin and yang: Balancing and answering binary visual questions». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5014–5022.

Zhao, Bingchen e Xin Wen (2020). «Distilling visual priors from self-supervised learning». In: *European Conference on Computer Vision*. Springer, pp. 422–429.

Zhao, Zhou, Qifan Yang, Deng Cai, Xiaofei He e Yueting Zhuang (2017). «Video Question Answering via Hierarchical Spatio-Temporal Attention Networks.» In: *IJCAI*, pp. 3518–3524.

Zhou, Bolei, Yuandong Tian, Sainbayar Sukhbaatar, Arthur Szlam e Rob Fergus (2015). «Simple baseline for visual question answering». In: *arXiv preprint arXiv:1512.02167*.

Zhu, Yuke, Oliver Groth, Michael Bernstein e Li Fei-Fei (2016). «Visual7w: Grounded question answering in images». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004.