UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Visual question answering per l'italiano: esperimenti con un sistema di traduzione automatica

Candidato: *Chiara De Nigris*

Relatore: *Alessandro Lenci*

Correlatore: *Felice Dell'Orletta*

Anno Accademico 2020-2021

Sommario

1. Introduzione 3

2. Descrizione delle risorse 4

2.1 Il framework LXMERT 4

2.2 Il dataset GQA 6

2.3 Google Translate 10

3. Metodologia 12

3.1 Creazione di un test set per la lingua italiana 12

3.1.1 Immagini 12

3.1.2 Descrizione quantitativa del dataset 12

3.1.3 Domande 12

3.2 Applicazione del modello LXMERT all'input 14

3.2.1 Traduzione dell'input 15

3.2.2 Invocazione del modello 15

3.2.3 Traduzione dell'output 17

4. Analisi dei risultati 20

4.1 Classificazione degli errori 20

4.1.1 Errori del modello 20

4.1.2 Errori nella traduzione della domanda 21

4.1.3 Errori nella traduzione della risposta 22

4.2 Valutazione dell'efficacia del modello 23

4.2.1 Risultati sul test set GQA 24

4.2.2 Confronto risultati tra i due test set 26

4.2.3 Valutazione generale 31

4.2.4 Incidenza del contesto 32

5 Conclusioni 33

6 Bibliografia 34

7 Appendice 35

7.1 Abbreviazioni di uso frequente 35

1. Introduzione

Obiettivo del presente studio è la valutazione dell'efficacia di strumenti pre-addestrati in lingua inglese per risolvere un task sulla lingua italiana, mediante l'utilizzo di sistemi neurali di traduzione automatica. Tale soluzione può rappresentare una valida alternativa all'addestramento ex-novo del modello su una nuova lingua, scelta computazionalmente più costosa ed impegnativa. L'avanzamento tecnologico nel campo della traduzione automatica ha permesso, infatti, di limitare l'elaborazione solo ad input e output per il *porting* da una lingua all'altra. In questo modo vengono sfruttati i risultati raggiunti sull'inglese per risolvere task di *natural language processing* anche in lingue per cui esistono meno risorse.

La metodologia scelta per l'esperimento ha previsto, per prima cosa, la creazione di due campioni, composti da un numero ritenuto esaustivo di immagini e domande. Ogni domanda in inglese è stata correlata alla relativa traduzione in italiano, in modo da ottenere lo stesso numero di elementi in entrambe le lingue. A questo punto, un *framework* di *Visual Question Answering* è stato testato su entrambi gli input. Al contrario dell'input in inglese, che non ha subito alcuna elaborazione, l'input in italiano è stato automaticamente tradotto in inglese, in modo da poter utilizzare il modello nella lingua per cui è stato addestrato. Seguendo lo stesso procedimento, l'output ottenuto come risposta all'input inglese non ha subito modifiche, mentre quello ottenuto dall'input automaticamente tradotto è stato a sua volta tradotto in italiano.

Ai fini delle valutazioni relative all'esperimento in oggetto è stato quantificato il rumore generato dalla traduzione automatica, determinando la validità del processo e definendo possibili miglioramenti della strategia di esecuzione.

2. Descrizione delle risorse

Per il Visual Question Answering, task semantico per le risposte automatiche a domande in linguaggio naturale basate su un'immagine, sono stati principalmente utilizzati il framework LXMERT (Hao Tan, Mohit Bansa, 2019) nella sua versione *pre-trained* e il dataset GQA (Hudson, Drew A. e Christopher D. Manning, 2019).

LXMERT, sviluppato da Hao Tan e Mohit Bansa della UNC Chapel Hil, è un framework per l'individuazione delle connessioni multimodali di input di tipo differente (linguistici e visuali), preaddestrato su dataset di Visual Question Answering.

Uno di questi è GQA, dataset multimodale realizzato da Drew A. Hudson e Christopher D. Manning della Stanford University.

Per la traduzione automatica di input e output si è optato per l'utilizzo dell'API Google Translate.

2.1 Il framework LXMERT

LXMERT (*Learning Cross-Modality Encoder Representations from Transformers*) è un framework pensato per apprendere le connessioni tra informazioni visive e linguistiche.

Come molti dei sistemi *state-of-the-art* in ambito natural language processing, il sistema è basato sull'architettura dei *transformer*. In particolare, è costituito da tre *encoder*: due a modalità singola, che si concentrano rispettivamente su input linguistico e visuale, e uno multimodale.

Dopo essere stato pre-addestrato su grandi quantità di coppie immagine-frase, LXMERT è in grado di apprendere i rapporti sia intermodali che multimodali tra gli oggetti che compongono le immagini.

Per svolgere il task di VQA, LXMERT prende in input un'immagine e la relativa frase (una didascalia o una domanda) e, a partire da esse, mediante la combinazione e l'implementazione di *self-attention* e *cross-attention layers*, è in grado di generare rappresentazioni linguistiche, visuali e multimodali in relazione all'immagine di input.

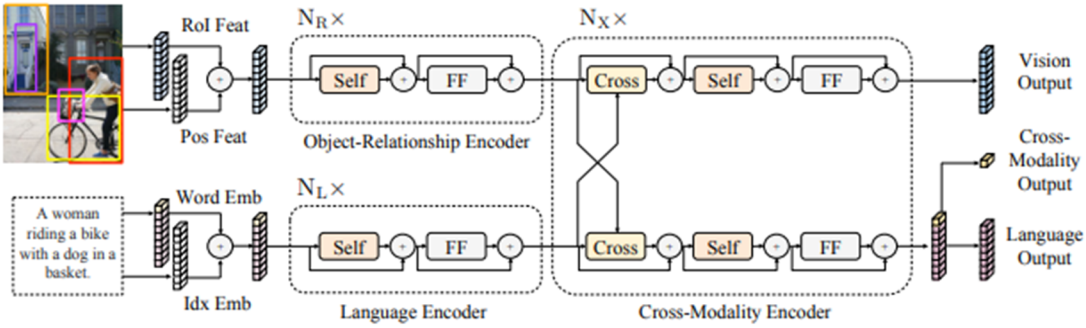


Figura 1 Architettura del framwork LXMERT. 'Self'e 'Cross'sono abbreviazioni per self-attention sub-layers e cross-attention sub-layers. 'FF'indica un feed-forward sub-layer (Hao Tan, Mohit Bansa, 2019, p. 4).

In input, ogni immagine è rappresentata come una sequenza di oggetti ed ogni frase come una sequenza di parole. I *layer* di *input embedding* del modello, infatti, convertono gli input in due modi:

- a livello di oggetti all'interno dell'immagine (*obect-level image embedding*);
- a livello di parole all'interno della frase (*word-level sentence embeddings*).

Nel primo caso, le immagini vengono rappresentate come embedding di oggetti, rilevati da un algoritmo di *object detection* di tipo *Faster R-CNN* (pre-addestrato su Visual Genome, un dataset per la comprensione di immagini) e circoscritti da riquadri, detti *bounding boxes*. Ogni oggetto è rappresentato dalle coordinate della sua bounding box e dalla sua regione di interesse 2048-dimensionale. Per quanto riguarda le frasi, queste vengono rappresentate come un embedding di parole. Come si legge nella sezione dedicata del paper di LXMERT, «ogni frase viene prima divisa in singole parole {w1, . . ., wn} di lunghezza n dallo stesso *WordPiece tokenizer*. Successivamente, (…) la parola "w_i "e il suo indice "i" vengono inseriti in un vettore da un sub-layer di embedding e poi aggiunti al word embedding indicizzato (Hao Tan e Mohit Bansa, 2019, p.2)». Dopo gli embedding layers, vengono applicati prima gli encoder a modalità singola (visuale e linguistico) e poi l'encoder multimodale. Come premesso, i tre encoder sono basati principalmente su due tipi di attention layer: self-attention layers e cross-attention layers, i quali hanno la funzione di recuperare le informazioni da un insieme di vettori di contesto "y_j" relativo a un vettore di query "x", calcolare il punteggio di corrispondenza tra il vettore "x" e ogni vettore di contesto "y_j", normalizzarlo e restituire in output una somma pesata del vettore di contesto e del punteggio normalizzato. Il modello restituisce, quindi**,** tre output: uno a livello linguistico, uno a livello visuale e uno cross-modale.

Citando la sezione relativa agli output del paper: «gli output di linguaggio e visione sono le sequenze di *features* generate dall'encoder cross-modale. Per l'output cross-modale, (...) appendiamo un *token* speciale (…) prima delle parole della frase e il corrispettivo vettore di features di questo token speciale nelle sequenze di features del linguaggio è usato come output cross-modale. (Hao Tan e Mohit Bansa, 2019, p.4)».

I dati per il pre-addestramento sono stati ricavati da cinque dataset visuali e linguistici: MS COCO (Lin et al., 2014), Visual Genome (Krishna et al 2017), VQA (Antol et al., 2015), GQA e VG-QA (Zhu et al., 2016).

Di ogni dataset sono stati utilizzati solo il *training* e il *validation set*, evitando i dati dei *test set*. Il dataset finale, allineato linguisticamente e visivamente, consta di 9,18 milioni di coppie immagine-frase e 180.000 immagini distinte. In termini di token, i dati di pre-addestramento contengono circa 100 milioni di parole e 6,5 milioni di oggetti.

LXMERT presenta ottime performance su due dataset di *Image Question-Answering*, GQA e VQA, con un rispettivo miglioramento dell'accuratezza generale del 3,2% e 2,1% rispetto ai migliori risultati ottenuti precedentemente da altri modelli. Inoltre, i dati confermano anche la possibilità di generalizzazione del modello con un miglioramento del 22% sul dataset di *Visual Reasoning* NLVR2 (Suhr et al., 2019).

2.2 Il dataset GQA

GQA è un dataset per il ragionamento visivo e la risposta a domande composizionali su immagini del mondo reale, incentrato sul *Real-world Reasoning* e sul *Question Answering* composizionale.

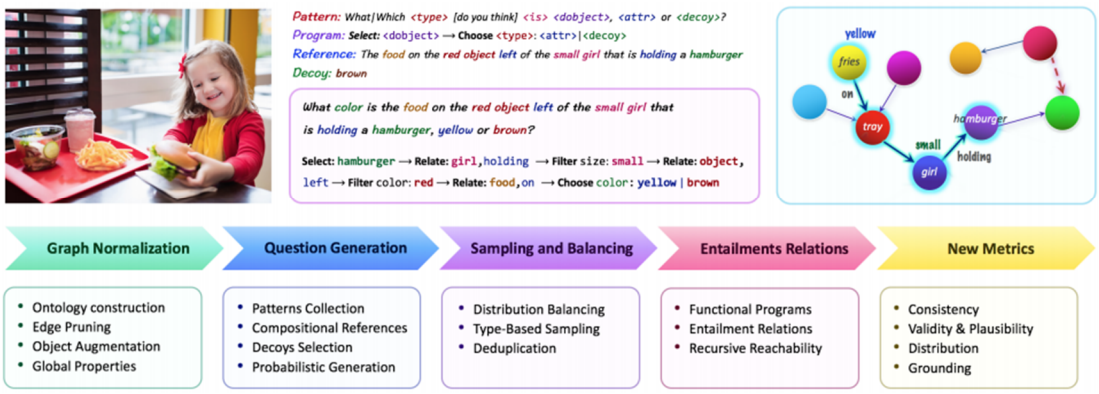


Figura 2 Panoramica del processo di costruzione di GQA. Data un'immagine annotata con uno scene graph dei suoi oggetti attributi e relazioni, vengono prodotte domande composizionali attraversando il grafico. (Hudson, Drew A. e Christopher D. Manning, 2019, p. 3).

È composto da 113.000 immagini e 22 milioni di domande di diversi tipi e vari gradi di composizionalità, che misurano le prestazioni su una serie di abilità di ragionamento.

Come riportato nel paper di GQA, queste includono «il riconoscimento di oggetti e attributi, il tracciamento delle relazioni transitive, il ragionamento spaziale, l'inferenza logica e i confronti. (Hudson, Drew A. e Christopher D. Manning, 2019, p.3)». Alla base del dataset c'è l'idea di sfruttare rappresentazioni semantiche, sia delle scene che delle domande, per consentire una diagnosi dettagliata per diversi tipi di domande.

«Le immagini, le domande e le risposte corrispondenti sono tutte accompagnate dalle rispettive rappresentazioni semantiche:

- ogni immagine è annotata con uno *scene graph* che rappresenta gli oggetti, gli attributi e le relazioni che essa contiene;
- ogni domanda è associata a un programma funzionale che elenca la serie di passaggi di ragionamento necessari per arrivare alla risposta;
- ogni risposta è argomentata da giustificazioni sia testuali che visive, che si riferiscono alla regione pertinente all'interno dell'immagine. (Hudson, Drew A. e Christopher D. Manning, 2019, p.3)».

Per la costruzione del dataset è stato necessario procedere alla ristrutturazione dei scene graph, inizialmente estratti da Visual Genome. Questi sono stati sottoposti a fasi di pulizia, normalizzazione, consolidamento ed espansione e annotati in linguaggio naturale. Si legge nella sezione del paper relativa alla normalizzazione che «il scene graph funge da rappresentazione formalizzata dell'immagine: ogni nodo denota un oggetto, un'entità visiva all'interno dell'immagine, come una persona, una mela, l'erba o una nuvola. È collegato a una bounding box che ne specifica la posizione e le dimensioni ed è contrassegnato con circa 1-3 attributi, proprietà dell'oggetto come, ad esempio, colore, forma o materiale (…). Gli oggetti sono collegati tra loro da relazioni che rappresentano azioni (verbi), relazioni spaziali (preposizioni) e comparativi. (Hudson, Drew A. e Christopher D. Manning, 2019, p.4)».

Figura 3 Esempio di un'immagine annotata con un scene graph. (Hudson, Drew A. e Christopher D. Manning, 2019, p. 1).

Dopo aver proceduto con la normalizzazione dei scene graph, «gli oggetti e le relazioni all'interno delle immagini sono stati associati ai corrispondenti schemi grammaticali (raccolti da VQA 2.0) in modo da procedere con la generazione delle domande (Hudson, Drew A. e Christopher D. Manning, 2019, p.4)».

Responsabile della produzione delle domande è un *question engine* ad hoc che sfrutta due risorse: i scene graph e i *pattern* strutturali, che riassumono il contenuto di ogni immagine adattandolo alle domande. I pattern totali sono 524, 250 costruiti manualmente e 274 estratti da VQA 1.0, ognuno dei quali è associato a una rappresentazione strutturata sotto forma di un programma funzionale. Riprendendo un esempio chiarificatore contenuto nel paper, «la domanda "di che colore è la mela sul tavolo bianco?" è semanticamente equivalente al seguente programma:

```
select: table → filter: white → relate (subject, on): apple → query: color (…).
```

Questi programmi sono composti da operazioni atomiche come la selezione di oggetti, il *traversing* di relazioni o la verifica di attributi, che vengono poi concatenati insieme per creare domande complesse. (Hudson, Drew A. e Christopher D. Manning, 2019, p.5)».

Come si legge nella sezione About della pagina web di GQA, «molte delle domande di GQA coinvolgono più capacità di ragionamento, comprensione spaziale e inferenza, risultando generalmente più impegnative rispetto a quelle contenute nei precedenti dataset di Visual Question Answering. (…). Sebbene le domande siano

state generate automaticamente, esse si basano su scene graphs scritti in linguaggio naturale e quindi sono grammaticali, diverse e idiomatiche.

(GQA, About https://cs.stanford.edu/people/dorarad/gqa/about.html)».

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

Figura 4 Esempi di domande dal dataset GQA
(Hudson, Drew A. e Christopher D. Manning, 2019, p. 4).

Ogni domanda viene classificata come di tipo strutturale o semantico, come mostrato nella tabella 2 del paper.

«Le domande di tipo strutturale derivano dall'operazione finale del programma funzionale delle domande. Possono essere:

1. di verifica: domande che prevedono sì o no come risposta.
   Es. Is it cloudy today?;
2. query: tutte le domande aperte.
   Es. How is the weather in the image?;
3. di scelta: domande che presentano due alternative tra cui scegliere.
   Es. Is the apple green or red?;
4. logiche: domande che coinvolgono un'inferenza logica.
   Es. Is the apple red and shiny?;
5. di confronto: domande che confrontano due o più oggetti.
   Es Who is taller, the boy or the girl?.

Le domande di tipo semantico si riferiscono all'argomento principale della domanda. Possono essere classificate in domande incentrate su:

1. oggetti: domande di esistenza sugli oggetti.

Es. Is there an apple in the picture?;

2. attributi: domande basate sulle proprietà o sulla posizione di un oggetto.
   Es. What color is the apple?;
3. categorie: domande correlate all'identificazione di un oggetto all'interno di una classe.
   Es. What kind of fruit is on the table?;
4. relazioni: domande sul soggetto o sull'oggetto di una relazione descritta.
   Es. What is the small girl wearing?;
5. proprietà globali: domande su caratteristiche della scena nella sua interezza come il clima o il luogo.
   Es. Is it sunny or cloudy?.

(Hudson, Drew A. e Christopher D. Manning, 2019, p.6)».

2.3 Google Translate

Google Translate è un servizio di traduzione automatica creato da Google e lanciato nel 2006 con il fine originario di tradurre dati statistici, utilizzando le trascrizioni delle Nazioni Unite e del Parlamento europeo come dati linguistici. Il traduttore copre più di cento lingue ed è in grado di tradurre, oltre che dati testuali, anche parlato, immagini e video. Ad eccezione dell'inglese, il processo di traduzione da una lingua all'altra non è diretto, ma prevede la traduzione della stringa di input prima in lingua inglese e poi nella lingua di destinazione, seguendo uno schema del tipo: L1 → EN → L2.

Piuttosto che proporre un'analisi tradizionale incentrata su regole grammaticali, gli algoritmi di Google Translate si basano su analisi statistiche cercando, durante la traduzione, di proporre il risultato più probabile sulla base dei dati a disposizione. Prima di restituire la traduzione di output, il *tool* ricerca, all'interno di centinaia di milioni di documenti, schemi già tradotti da traduttori umani, in modo da individuare la traduzione più appropriata. Google mette a disposizione, oltre che un'app e un'interfaccia web, anche un'API per gli sviluppatori, utilizzata per il lavoro nella sua versione *basic*. Per questo esperimento, l'API è stata importata dalla libreria Python Deep-translator 1.4.4 ed utilizzata nel codice in maniera semplice: dopo aver specificato la lingua sorgente e quella di destinazione, sono state passate come input

le domande e le risposte in inglese e restituite in output le corrispettive domande e risposte in italiano.

Nonostante non possa assicurare risultati affidabili quanto quelli di una traduzione eseguita da un parlante nativo, il traduttore automatico di Google fornisce traduzioni coerenti e relativamente accurate. Il motivo principale che ha portato ad optare per questa risorsa è la semplicità, sintattica e lessicale, delle frasi di input, che permette di ottenere delle performance soddisfacenti. Inoltre, il *tool* garantisce velocità di calcolo e autorevolezza dei risultati ottenibili.

3. Metodologia

3.1 Creazione di un test set per la lingua italiana

3.1.1 Immagini

In questo progetto di tesi, sono stati condotti due esperimenti differenti.

Il primo ha riguardato la valutazione della correttezza delle risposte ad alcune domande relative ad immagini contenute in una porzione di GQA, selezionata in modo che non comprendesse immagini presenti nel traning e nel validation set, ma solo nel test set.

Il secondo, invece, ha previsto la creazione di un dataset ad hoc, composto da immagini simili nella composizione e negli oggetti contenuti a quelle di GQA. Le immagini, prive di diritto d'autore, sono state scaricate dal sito Pexels.

La scelta di dividere il test set definitivo secondo quanto descritto ha come fine la valutazione dei risultati, sia su un insieme di dati più familiari a LXMERT, sia su un campione selezionato che non fosse già conosciuto al modello, ma comunque simile. Inoltre, è stato ritenuto necessario comporre i due campioni in modo che contenessero lo stesso numero di immagini per poter, alla fine delle valutazioni distinte, proporre un confronto dei risultati ottenuti.

3.1.2 Descrizione quantitativa del dataset

Dal test set del dataset GQA sono state estratte 100 immagini, ognuna delle quali associata in media a circa 33 domande, per un totale di 3379 domande.

Il nuovo dataset è composto da 100 immagini e 1000 domande, distribuite in blocchi di 10 per ogni immagine.

3.1.3 Domande

Le domande associate alle immagini di GQA sono state estratte direttamente dal dataset e ad ognuna di esse è stata associata anche la relativa traduzione manuale in italiano. Come illustrato nel par. 2.2, le domande di GQA sono state composte da un question engine mediante l'utilizzo di pattern strutturali e risultano, quindi, standardizzate. Sulla base di questi pattern sono state composizionalmente create le 1000 domande per le nuove immagini.

Per garantire coerenza con la struttura delle domande usate come modello, anche le domande del nuovo test set sono state inizialmente composte in inglese. Successivamente, tutte le domande sono state manualmente tradotte dall'inglese all'italiano da parlanti nativi, cercando di preservarne il più possibile la conformazione originaria.

question	domanda
Does the name tag have a different shape than the ball?	La targhetta del nome ha una forma diversa dalla palla?
What animal is it?	Che animale è?
Are there seals or cats?	Ci sono foche o gatti?
What color is the water that is not shallow?	Di che colore è l'acqua che non è superficiale?
Is the water deep?	L'acqua è profonda?

Figura 5 Esempio di questions in inglese tradotte in domande in italiano.

Su 524 pattern, ne sono stati selezionati 44, reperiti dagli esempi contenuti nel paper di GQA.

Rispecchiando la suddivisione in categorie per le domande di tipo strutturale riportate nel par. 2.2, sono di seguito elencati degli esempi di pattern per ognuna delle suddette categorie:

- di verifica: Does the {object} and the {object} have the same color?;
- query: What kind of {object} is {relation} the {object}?;
- di scelta: Is the {object} {attribute} or {attribute}?;
- logiche: Do you see either an {object} or a {object} there?;
- di confronto: Are the {object} and the {object} made of different materials?.

Il grafico seguente riporta la distribuzione da un punto di vista strutturale delle 1000 nuove domande del test set italiano composto per l'esperimento:

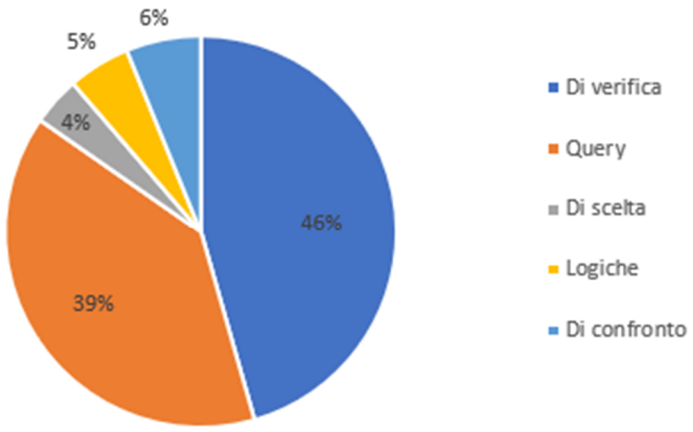


Figura 6 Distribuzione delle domande strutturali.

Rispecchiando la suddivisione in categorie per le domande di tipo semantico riportate nel par. 2.2, sono di seguito elencati degli esempi di pattern per ognuna delle suddette categorie:

- oggetti: Is there a {object} in the picture?;
- attributi: Is the {object} both {attribute} and {attribute}?;
- categoria: What kind of {object} is it, a {object} or a {object}?;
- relazioni: Is the {object} {relation} or {relation} of the {object}?;
- globali: Where is this?.

Il grafico seguente riporta la distribuzione da un punto di vista semantico delle 1000 nuove domande del test set italiano composto per l'esperimento:

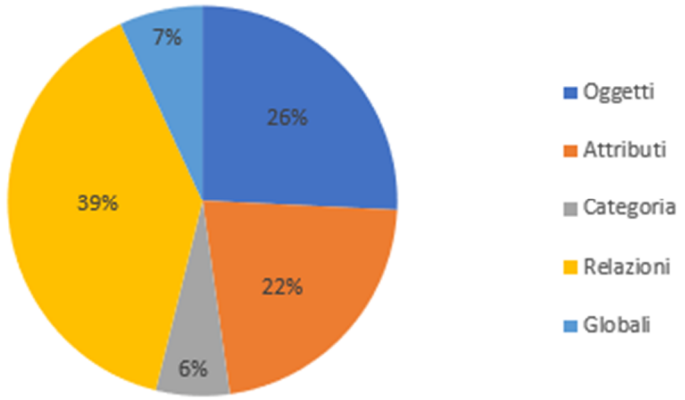


Figura 7 Distribuzione delle domande semantiche.

3.2 Applicazione del modello LXMERT all'input

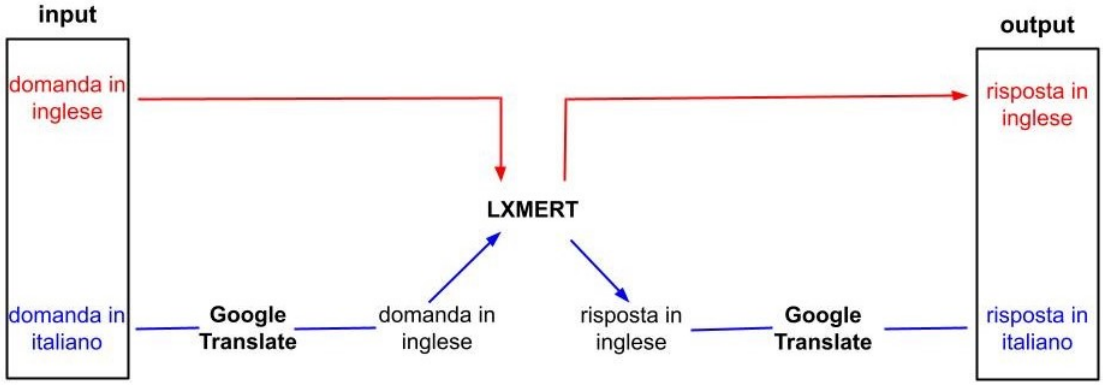


Figura 8 Workflow.

3.2.1 Traduzione dell'input

Tutte le domande, sia quelle in inglese che in italiano, sono state raccolte in un unico file di tipo CSV, composto da tre colonne: una per tenere traccia dell'identificatore dell'immagine di riferimento, una per la *question* in inglese e una per la rispettiva traduzione della domanda in italiano. A questo punto LXMERT è stato valutato su un input inglese originale e su un input inglese generato da una traduzione automatica dall'italiano.

Per testare le prestazioni del modello sulle domande in inglese originale, infatti, queste gli sono state date in input, senza apportare alcuna modifica. Le domande in italiano, invece, sono state singolarmente tradotte in inglese in modo automatico da Google Translate.

```
for i,domanda_ita in enumerate(domande_ita[name]):
domanda_ing=
GoogleTranslator(source='it',target='en').translate(domanda_ita)[1]
```

3.2.2 Invocazione del modello

Una volta costituito l'input definitivo, si è passato all'esecuzione del modello vera e propria.

Dopo aver scaricato oggetti, attributi e risposte di GQA, ogni immagine contenuta nel test set di riferimento è stata processata e mappata da una *Faster R-CNN* nei suoi oggetti e nei suoi attributi. Ad ogni oggetto, circoscritto all'interno di una regione delimitata da una bounding box, è stata associata la propria probabilità e un attributo di riferimento, anch'esso legato ad un valore probabilistico. Ogni bounding box è correlata da una *label*, che etichetta il nome dell'oggetto che l'area contiene e l'attributo più probabile ad esso associato.

Per maggiore chiarezza, di seguito si riporta il codice per l'implementazione di quanto appena descritto:

```
#download di oggetti attributi e risposte di GQA
objids = utils.get_data(OBJ_URL)
attrids = utils.get_data(ATTR_URL)
gqa_answers = utils.get_data(GQA_URL)
```

[1] Processo di traduzione automatica di ogni domanda contenuta all'interno del dizionario domande_ita dall'italiano all'inglese.

```
for immagine in lista_immagini:
    #visualizzazione dell'immagine
    frcnn_visualizer=SingleImageViz(URL_loc,id2obj=objids,
    id2attr=attrids)
    #frcnn
    images, sizes, scales_yx = image_preprocess(URL_loc)
    output_dict = frcnn(
        images,
        sizes,
        scales_yx=scales_yx,
        padding="max_detections",
        max_detections=frcnn_cfg.max_detections,
        return_tensors="pt"
    )
    #aggiunta di bounding boxes e labels all'immagine
    frcnn_visualizer.draw_boxes(
        output_dict.get("boxes"),
        output_dict.pop("obj_ids"),
        output_dict.pop("obj_probs"),
        output_dict.pop("attr_ids"),
        output_dict.pop("attr_probs"),
    )
    #normalizzazione delle bounding boxes
    normalized_boxes = output_dict.get("normalized_boxes")
    features = output_dict.get("roi_features")
    #visualizzazione dell'immagine
    display(Markdown(f'\n'))
    display(Markdown(f'\n'))
    showarray(frcnn_visualizer._get_buffer())[2]
```

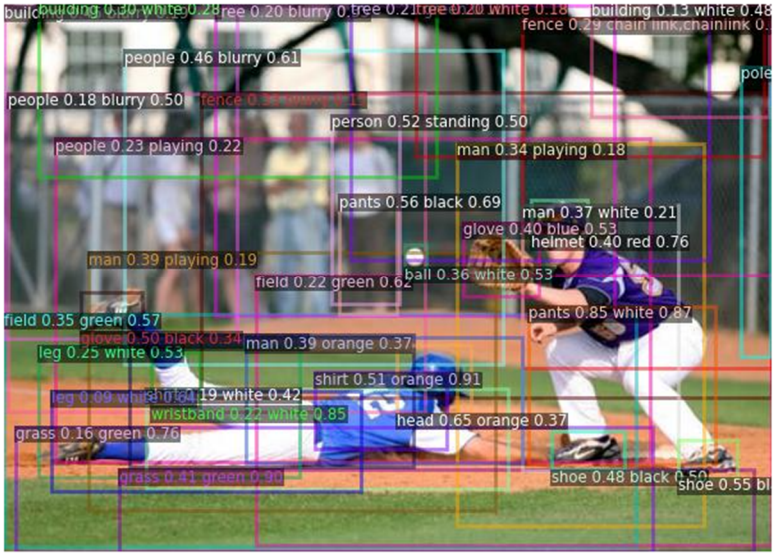


Figura 9 Esempio di immagine mappata nei suoi oggetti e nei suoi attributi.

[2] Mapping di un'immagine.

Successivamente, ogni domanda associata all'immagine di riferimento è stata *tokenizzata* ed associata alla risposta più probabile da un attention layer, mediante l'utilizzo di componenti per il question answering addestrate su GQA, come comprensibile leggendo il seguente frammento di codice:

```
for domanda in domande_ing[name]:
        #tokenizzazione della domanda
        inputs = lxmert_tokenizer(
            test_question,
            padding="max_length",
            max_length=20,
            truncation=True,
            return_token_type_ids=True,
            return_attention_mask=True,
            add_special_tokens=True,
            return_tensors="pt"
        )
        #calcolo della risposta
        output_gqa = lxmert_gqa(
            input_ids=inputs.input_ids,
            attention_mask=inputs.attention_mask,
            visual_feats=features,
            visual_pos=normalized_boxes,
            token_type_ids=inputs.token_type_ids,
            output_attentions=False,
        )
        #predizione della risposta
        pred_gqa=output_gqa["question_answering_score"].argmax(-1)
```
[3]

3.2.3 Traduzione dell'output

Dall'esecuzione del modello sono stati ottenuti due output differenti, entrambi in lingua inglese: le risposte alle domande originali in inglese e le risposte alle domande tradotte in inglese dall'italiano.

Le risposte alle domande in inglese originale sono state semplicemente stampate in output, senza subire alcuna modifica.

Per quanto riguarda le seconde, invece, queste sono state automaticamente tradotte in italiano da Google Translate.

Per valutare l'incidenza del contesto, sono stati proposti due tipi diversi di traduzioni:

[3] Predizione della risposta per ogni domanda. La variabile *lxmert_tokenizer* è associata al tokenizzatore, mentre *lxmert_gqa* è il componente per il question answering addestrato su GQA.

- quella della risposta isolata, ossia fuori contesto:

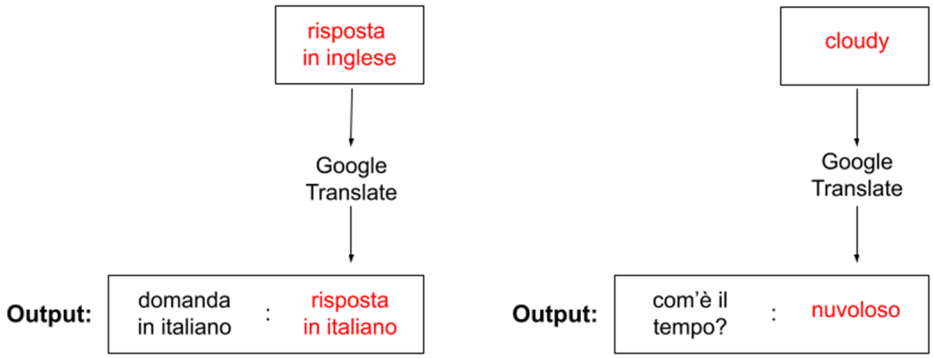


Figura 10 Esempio di traduzione di una domanda fuori contesto.

Di seguito il codice relativo:

```
for domanda_ita in domande_ita[name]:
       #predizione della risposta
        pred_gqa = output_gqa["question_answering_score"].argmax(-1)
        #traduzione risposta in italiano (senza contesto)
        risp_ita=GoogleTranslator(source='en',
target='it').translate(gqa_answers[pred_gqa])[4]
```

- quella della risposta insieme alla relativa domanda, cioè in contesto;

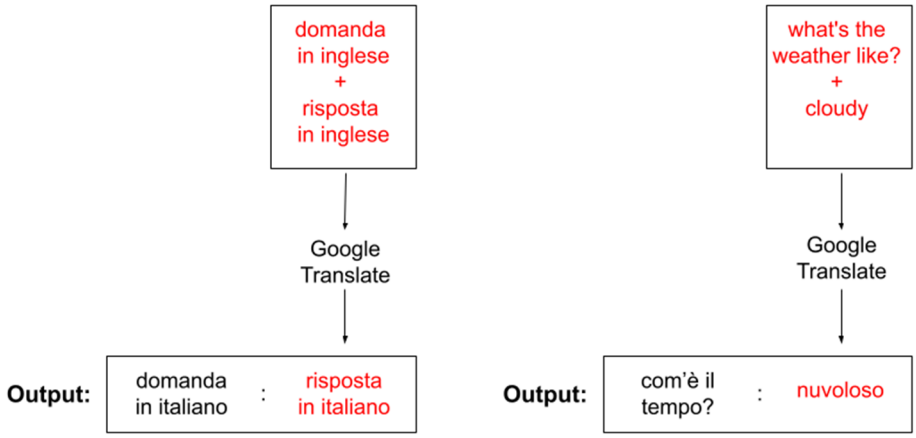


Figura 11 Esempio di traduzione di una domanda in contesto.

Di seguito il codice relativo:

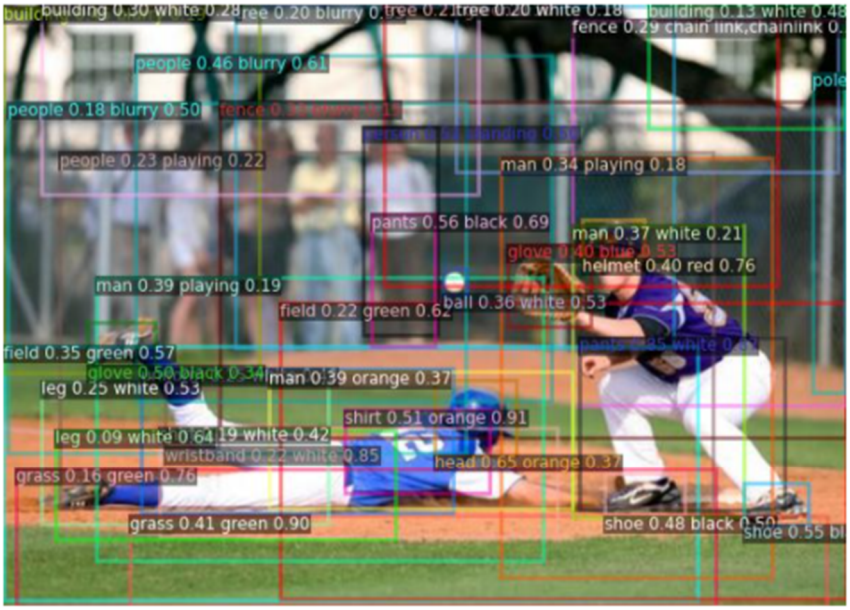
```
for domanda_ita in domande_ita[name]:
        #predizione della risposta
```

[4] Traduzione automatica della domanda fuori contesto.

```
        pred_gqa = output_gqa["question_answering_score"].argmax(-1)
        #traduzione risposta in italiano (in contesto)
        DomandaRisposta = domanda_ita + ``+gqa_answers[pred_gqa]
        risp_in_contesto=GoogleTranslator(source='en',
target='it').translate(DomandaRisposta)
```
[5]

Ogni immagine è stata quindi associata a tre coppie di output diversi:

- **Inglese originale**: *question-answer* in inglese originale;
- **Italiano con contesto**: coppia domanda-risposta in italiano, composta dalla domanda in italiano originale e la risposta tradotta automaticamente in italiano dall'inglese, insieme alla domanda;
- **Italiano fuori contesto**: coppia domanda-risposta in italiano, composta dalla domanda in italiano originale e la risposta tradotta automaticamente in italiano dall'inglese, in maniera distinta rispetto alla domanda.



Inglese (originale)

Question: *What is behind of the grey fence?* --> Answer: **trees**

Italiano (con contesto)

Domanda: *Cosa c'è dietro la staccionata grigia?* --> Risposta: **alberi**

Italiano (senza contesto)

Domanda: *Cosa c'è dietro la staccionata grigia?* --> Risposta: **alberi**

Figura 12 Esempio di output.

[5] Traduzione automatica della domanda con contesto.

4. Analisi dei risultati

L'efficacia del modello è stata valutata prima sulla coppia input-output in inglese originale e poi sui risultati tradotti automaticamente.

Al termine del lavoro, sono state proposte due valutazioni differenti:

- una relativa alle prestazioni del modello sulle cento immagini del test set GQA;
- una ricavata dalla comparazione fra i risultati sul test set estratto da GQA e quelli sul nuovo test set.

  Come esposto nel par. 3.1.2, i due campioni sono composti dallo stesso numero di immagini, ma il test set composto dai dati estratti da GQA presenta 2379 domande in più. Per procedere con un confronto bilanciato dei risultati è stata proposta una versione *balanced* di quest'ultimo, isolando casualmente solo dieci domande per ognuna delle immagini, per un totale di 1000.

4.1 Classificazione degli errori

Le prestazioni del modello sono state valutate da due annotatori, classificando gli errori in tre gruppi: errori del modello, errori nella traduzione della domanda ed errori nella traduzione della risposta.

4.1.1 Errori del modello

Vengono classificati come errori del modello quei casi in cui la risposta alla domanda è sbagliata già rispetto all'input originale in inglese: l'errore è, quindi, dipendente dalla predizione del framework di VQA.

Riprendendo la classificazione semantica delle domande descritta nel par. 3.1.3, gli errori sono stati distinti nelle seguenti sottocategorie:

- **errori globali:** riguardano le risposte errate a domande globali, relative, ad esempio, alle condizioni atmosferiche o all'identificazione della scena;
- **errori sugli attributi:** riguardano le risposte errate a domande sul riconoscimento di attributi quali, ad esempio, colori o materiali;
- **errori sugli oggetti:** riguardano le risposte errate a domande sul riconoscimento degli oggetti;

- **errori sulle relazioni:** riguardano le risposte errate a domande relative alla posizione di un oggetto, o a livello assoluto all'interno dell'immagine o rispetto ad un altro.

4.1.2 Errori nella traduzione della domanda

Vengono classificati come errori nella traduzione della domanda i casi in cui la risposta all'input in inglese originale risulta corretta, mentre la domanda viene tradotta in modo errato dall'italiano all'inglese da Google Translate.

Gli errori appartenenti a questa classe sono stati distinti nelle seguenti sottocategorie:

- **Errori di polisemia:**

  Gli errori di polisemia riguardano la traduzione in italiano di un vocabolo polisemico, ovvero portatore di più significati, in un'accezione diversa da quella con cui esso è utilizzata in inglese. Questi tipi di errori stravolgono la domanda, comportando una risposta errata.

  Un esempio di questo tipo di errori è la domanda in inglese *"*What is tied to the pants*?"*, resa con il corrispettivo italiano "Cosa è legato ai pantaloni?". La frase in italiano è stata tradotta da Google Translate come *"*What is related to the pants?", traducendo, quindi, il termine italiano *legato* con l'accezione di *relazionato* e non di *allacciato*. Il modello risponde correttamente all'input inglese originale ma non a quello tradotto dall'italiano, che si presenta come una domanda del tutto differente.

- **Errori generali di traduzione:**

  Gli errori generali di traduzione riguardano quei casi in cui la domanda è stata tradotta in modo completamente errato, senza possibilità di circoscrivere l'errore in modo più dettagliato.

  Un esempio di questo tipo di errori è la domanda in inglese "Which kind is the food?"*,* resa con il corrispettivo italiano *"*Di che tipo è il cibo?". La frase in italiano è stata tradotta da Google Translate come *"*What is the food like?", travisando il contenuto della domanda. Il modello, infatti, risponde correttamente all'input inglese originale ma non a quello tradotto dall'italiano, che si presenta come una domanda del tutto differente e viene addirittura scambiata per una domanda sul clima (probabilmente a causa della somiglianza con la formula standard "What is the weather like?") ottenendo come risposta *cloudy*.

4.1.3 Errori nella traduzione della risposta

Vengono classificati come errori nella traduzione della risposta i casi in cui la risposta alla domanda non è sbagliata in inglese, ma è tradotta erroneamente in italiano.

Gli errori appartenenti a questa classe sono stati distinti nelle seguenti sottocategorie:

- **Errori di polisemia**:

  Gli errori di polisemia riguardano la traduzione in italiano di una parola polisemica in un'accezione diversa da quella con cui essa è utilizzata in inglese. Questi comportano, quindi, una risposta che ha senso in inglese ma non più in italiano.

  L'esempio più frequente di questo tipo di errori è la traduzione della risposta *right* alle domande di localizzazione che ha, intuitivamente, in inglese il significato di destra o destro. Tuttavia, in italiano la parola viene sempre tradotta con il termine *giusto* che è sì una delle possibili traduzioni corrette, ma non assume alcun significato in relazione al contesto in cui viene usata.

- **Errori di accordo di genere:**

  Gli errori di genere riguardano la mancata concordanza del genere della risposta con la domanda nella traduzione dall'inglese all'italiano. È bene precisare che questo genere di errori abbassa sicuramente il grado di precisione del sistema ma non rende la risposta incomprensibile.

  Esempi di questo tipo di errori sono le risposte relative al colore degli oggetti. Senza che vi sia alcuna connessione con la domanda, alcuni colori vengono tradotti sempre al femminile dall'italiano all'inglese, altri sempre al maschile. Ad esempio, il colore *white* viene sempre reso come *bianca*, anche in casi in cui è richiesto il maschile.

- **Errori di accordo di numero:**

  Gli errori di numero riguardano il mancato accordo del numero della risposta con la domanda nella traduzione dall'inglese all'italiano. Anche in questo caso, a fronte di un minor grado di precisione del sistema, la risposta è comunque comprensibile.

  Un esempio di questo tipo di errori è la risposta alla domanda "Are the gray pants short or long?" resa in italiano come "I pantaloni grigi sono corti o lunghi?". Alla domanda in inglese, la risposta, corretta, fornita da LXMERT è

long che viene però tradotto in italiano come *lungo* e non come *lunghi*, determinando il mancato accordo con il plurale di *pantaloni*.

- **Errori di accordo di genere e numero:**

  Gli errori di genere e numero riguardano quei casi in cui è assente sia la concordanza di genere che di numero con la domanda di riferimento. Anche questi errori, che in effetti sono simili ai precedenti, influenzano il grado di precisione del sistema senza rendere la risposta incomprensibile.

  Un esempio di questo tipo di errori è la risposta alla domanda "How big are the windows?" resa in italiano come "Quanto sono grandi le finestre?". Alla domanda in inglese, la risposta, corretta, fornita da LXMERT è *small* che viene però tradotto in italiano come *piccolo* e non come *piccole*, determinando non solo la mancata concordanza con il plurale ma anche con il femminile di *finestre*.

- **Errori di resa verbale:**

  Gli errori di resa verbale riguardano l'errata traduzione di un verbo dall'inglese all'italiano. Anche in questo caso, si rileva che si tratta di errori che non inficiano la comprensibilità della risposta, pur abbassando il grado di precisione del sistema.

  Un esempio di questo tipo di errori è la traduzione di verbi al gerundio in inglese con forme non verbali in italiano. Ad esempio, il termine *standing* come risposta a domande del tipo "Cosa sta facendo la persona?" viene tradotto come *in piedi* mancando, quindi, di una vera e propria componente verbale.

4.2 Valutazione dell'efficacia del modello

La valutazione proposta consiste in un paragone tra il grado di accuratezza delle risposte prodotte dall'input originale e quello delle risposte ottenute per l'italiano, prima relativamente al solo test set GQA, poi in un confronto tra una parte di quest'ultimo e il nuovo test set composto ad hoc per l'esperimento. Inoltre, è stata valutata l'incidenza della presenza del contesto nella traduzione delle risposte, quantificando il numero di casi in cui la traduzione della domanda insieme alla risposta ha determinato un miglioramento delle prestazioni.

4.2.1 Risultati sul test set GQA

I risultati sulle risposte alle domande in inglese originale che compongono il test set GQA confermano un'ottima performance da parte di LXMERT: 2613 risposte su 3379 sono state valutate come corrette (77,3% del totale).

Per quanto riguarda le valutazioni sull'input tradotto automaticamente dall'italiano, le risposte corrette ammontano a 2344 su 3379 (69,4% del totale).

La differenza tra le risposte corrette alle domande scritte originariamente in inglese e quelle alle domande tradotte in inglese dall'italiano conta, quindi, 269 casi (7,9% del totale).

	Occorrenze	**Percentuale**
Risposte corrette all'input in inglese originale	2613	77,3%
Risposte corrette all'input in inglese tradotto	2344	69,4%

Tabella 1 Quantificazione delle risposte corrette nel test set GQA.

Una differenza bassa come quella riscontrata tra il numero di risposte corrette all'input originale e quello relativo all'input tradotto conferma la validità del traduttore automatico scelto.

Per quanto riguarda gli errori, sono stati differenziati gli errori del modello, comuni ad entrambi i tipi di input, e quelli di traduzione, relativi solo all'input tradotto. Gli errori del modello sono distribuiti come illustrato dalla tabella seguente:

	Tipo di errore	**Occorrenze**	**Percentuale**
Errori del modello	Riconoscimento ambiente	7	0,91%
Occorrenze: 766	Riconoscimento relazione	91	11,9%
(22,5%)	Riconoscimento oggetto	358	46,7%
	Riconoscimento attributi	310	40,4%

Tabella 2 Quantificazione e classificazione degli errori del modello nel test set GQA.

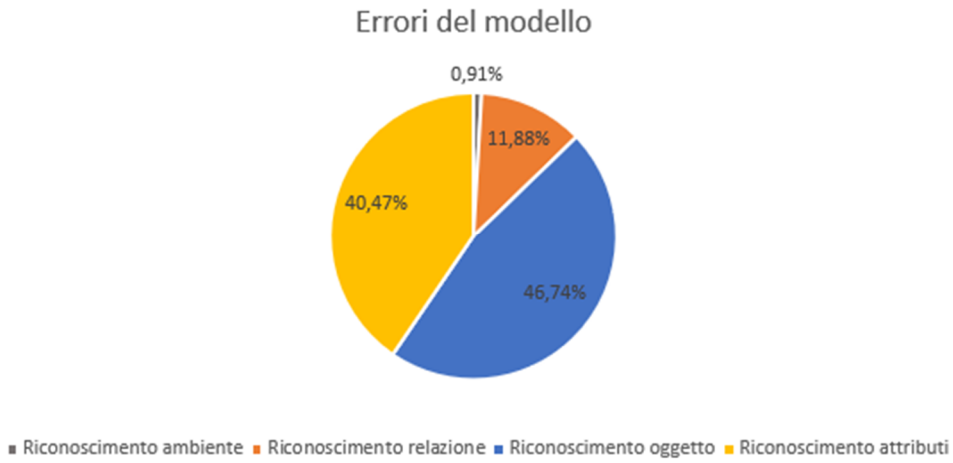


Figura 13 Classificazione degli errori del modello nel test set GQA.

Gli errori di traduzione, invece, sono distribuiti come illustrato dalla tabella seguente:

	Tipo di errore	**Occorrenze**	**Percentuale**
Traduzione domanda **Occorrenze: 16** **(0,47%)**	Polisemia	6	37,5%
	Traduzione errata	10	62,5%
Traduzione risposta **Occorrenze: 253** **(7,5%)**	Genere	42	16,6%
	Polisemia	179	70,8%
	Resa verbale	18	7,1%
	Numero	12	4,7%
	Genere e Numero	2	0,8%

Tabella 3 Quantificazione e classificazione degli errori di traduzione nel test set GQA.

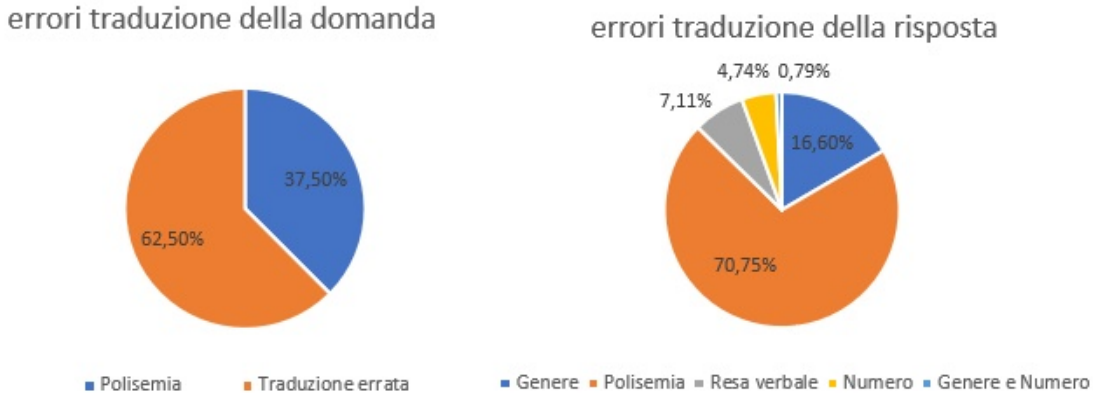


Figura 14 Classificazione degli errori di traduzione nel test set GQA della domanda e della risposta.

È possibile notare come gli errori di traduzione della domanda costituiscano solo lo 0,47% dei casi, mentre gli errori di traduzione della risposta il 7,5%. Focalizzandosi su questi ultimi, è bene sottolineare, come già spiegato nel par. 4.1.3, che gli errori di genere, numero e resa verbale minano sì la resa grammaticale della risposta ma non la privano di significato, rendendola ugualmente comprensibile.

Gli errori veramente rilevanti, quindi, sono gli errori di traduzione della domanda, che contano solo di 16 occorrenze, e quelli di polisemia in relazione alla traduzione della risposta, che sono 179. Proponendo una distinzione tra errori gravi ed errori trascurabili, è possibile circoscrivere alla prima categoria solo le due classi appena commentate. Gli errori gravi sull'input italiano sono, quindi, 195, determinando un aumento solo del 5,7% (e non del 7,9%) rispetto agli errori comuni anche all'input in inglese.

4.2.2 Confronto risultati tra i due test set

Il confronto dei risultati tra i due test set offre, tra le altre cose, la possibilità di valutare quanto sia generalizzabile il framework LXMERT.

Come esposto nel par. 2.1, LXMERT è stato addestrato su più dataset multimodali, tra cui GQA. Sebbene il campione composto per l'esperimento sia costituito da immagini che il modello non ha mai visto prima, poiché appartenenti al test set, è comunque importante considerare che la tipologia di immagini e di domande contenute in GQA gli sono sicuramente più familiari rispetto ad un input anche solo in parte differente.

Sul balanced test set composto da 1000 domande di GQA sono state valutate come corrette 801 risposte all'input originale in inglese (80,1% del totale) e 714 risposte alle domande tradotte automaticamente in inglese dall'italiano (71,4% del totale).

Sul campione contenente le nuove immagini, invece, le risposte corrette sono state 757 alle domande in inglese originale (75,7% del totale) e 648 all'input tradotto automaticamente in inglese dall'italiano (64,8%).

	Balanced test set GQA	**Nuovo test set**
Risposte corrette all'input in inglese originale	801 (80,1%)	757 (75,7%)
Risposte corrette all'input in inglese tradotto	714 (71,4%)	648 (64,8%)

Tabella 4 Confronto tra le risposte corrette per il balanced test set GQA e per il nuovo test set composto.

Come atteso, il Framework ha migliori prestazioni sul balanced test set GQA. Ai fini dell'esperimento, non è particolarmente rilevante la disuguaglianza tra le risposte in inglese sui due test set o tra i risultati in italiano, quanto invece il divario tra le risposte corrette in inglese e in italiano in ogni campione distinto. Sul balanced test set GQA, infatti, la differenza tra le risposte corrette in inglese (80,1%) e quelle corrette in italiano (71,4%) è dell'8,7% mentre questo stesso valore raggiunge il 10,9% nel nuovo test set. Questo calo di prestazioni sui nuovi dati, che è comunque minimo, conferma il legame tra GQA e LXMERT.

Riprendendo la differenziazione proposta nel par. 4.2.1 per il test set GQA, si riporta la classificazione degli errori del modello e degli errori di traduzione per entrambi i test set.

Gli errori del modello sono distribuiti come illustrato dalle tabelle seguenti:

Balanced test set GQA			
	Tipo di errore	Occorrenze	Percentuale
Errori del modello **Occorrenze: 199** **(19,8%)**	Riconoscimento ambiente	3	1,5%
	Riconoscimento relazione	25	12,6%
	Riconoscimento oggetto	87	43,7%
	Riconoscimento attributi	84	42,2%

Tabella 5 Quantificazione e classificazione degli errori del modello nel balanced test set GQA.

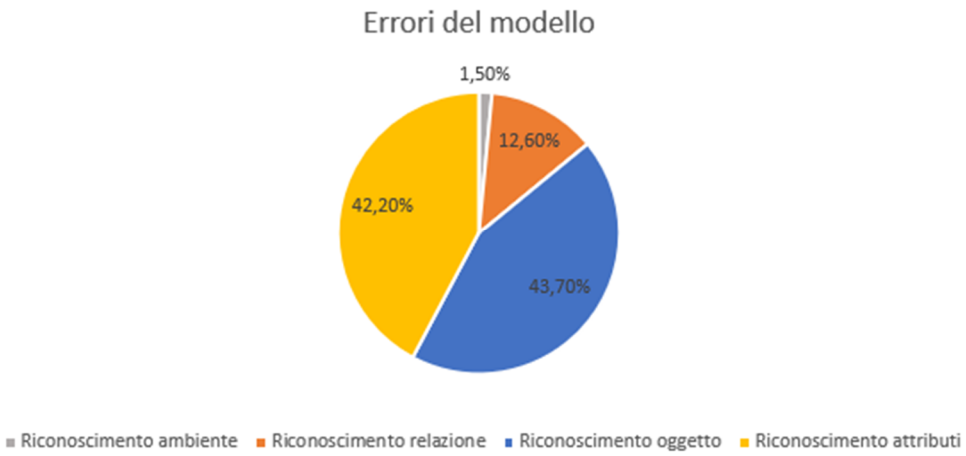


Figura 15 Classificazione degli errori del modello nel balanced test set GQA.

Nuovo test set			
	Tipo di errore	Occorrenze	Percentuale
Errori del modello **Occorrenze: 243** **(24,3%)**	Riconoscimento ambiente	19	7,8%
	Riconoscimento relazione	23	9,5%
	Riconoscimento oggetto	112	36,6%
	Riconoscimento attributi	89	46,1%

Tabella 6 Quantificazione e classificazione degli errori del modello nel nuovo test set.

Figura 16 Classificazione degli errori del modello nel nuovo test set.

Gli errori di traduzione, invece, sono distribuiti come illustrato dalle tabelle seguenti:

Balanced test set GQA			
Errore	Tipo di errore	Occorrenze	Percentuale
Traduzione domanda **Occorrenze: 4** **(0,4%)**	Polisemia	2	50%
	Traduzione errata	2	50%
Traduzione risposta **Occorrenze: 83** **(8,3%)**	Genere	8	9,6%
	Polisemia	64	77,2%
	Resa verbale	5	6%
	Numero	5	6%
	Genere e Numero	1	1,2%

Tabella 7 Quantificazione e classificazione degli errori di traduzione nel balanced test set GQA.

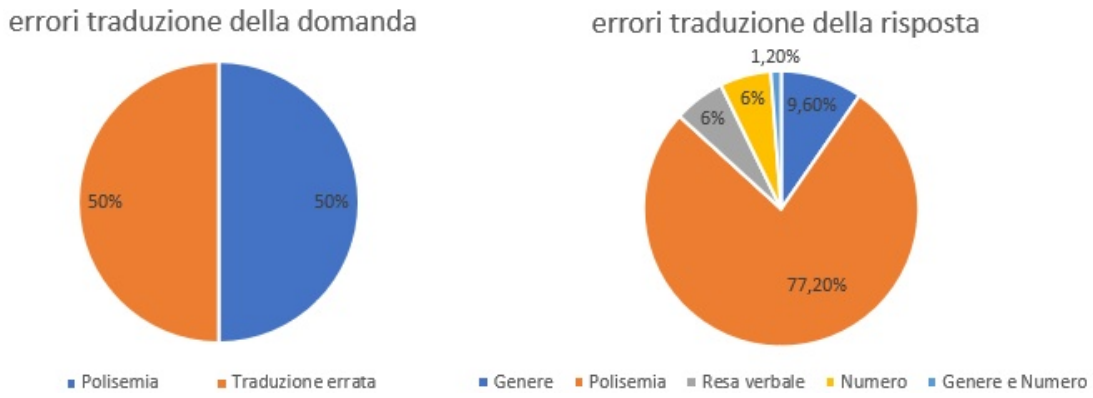


Figura 17 Classificazione degli errori di traduzione sul balanced test set GQA della domanda e della risposta

Nuovo test set			
Errore	Tipo di errore	Occorrenze	Percentuale
Traduzione domanda **Occorrenze: 21** **(2,1%)**	Polisemia	8	38,1%
	Traduzione errata	13	61,9%
Traduzione risposta **Occorrenze: 88** **(8,8%)**	Genere	29	32,9%
	Polisemia	37	42,1%
	Resa verbale	22	25%
	Numero	0	0%
	Genere e Numero	0	0%

Tabella 8 Quantificazione e classificazione degli errori di traduzione nel nuovo test set.

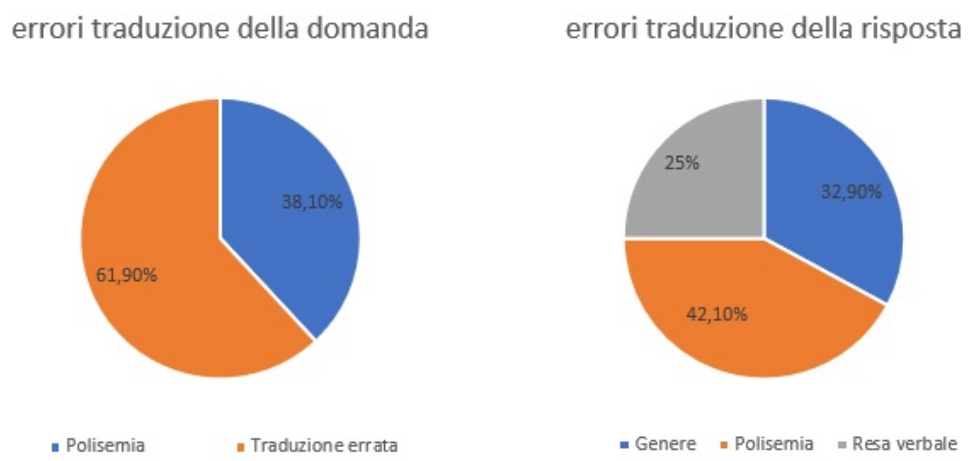


Figura 18 Classificazione degli errori di traduzione sul nuovo test set della domanda e della risposta.

È possibile notare che la differenza più ingente tra i due campioni è negli errori di traduzione della domanda: sul balanced test set GQA, questi ammontano solo a 4, costituendo lo 0,4% del totale mentre sul nuovo test set sono 21, il 2,1% del totale. In entrambi i casi, questo tipo di errori non sono in numero elevato ma, se sul balanced test set GQA sono praticamente trascurabili, nel nuovo campione aumentano dell'1,7%. Questo dipende, ancora una volta, dalla natura dell'input che nel primo caso è più familiare al framework, mentre nel secondo è risultato, in alcuni casi, più ostico.

Gli errori di traduzione della risposta si presentano, invece, come quasi uguali: nel primo test set rappresentano l'8,8% dei casi, nel nuovo l'8,3%.

Le differenze principali all'interno di questa classe sono:

- il numero di errori di genere: solo il 9,6% sul balanced test set GQA ma il 32,9% sul nuovo test set;
- di polisemia: il 76,1% nel primo caso ma solo il 42,1% nel secondo;.
- di resa verbale: solo il 6% nel primo caso ma il 25% nel secondo.

Inoltre, nel nuovo test set non è presente neanche un'occorrenza di errori di numero o di genere e numero, valori che in ogni caso restano molto bassi anche sul balanced test set GQA dove ammontano, rispettivamente, al 2,5% e allo 0,5%.

Seguendo la stessa classificazione proposta nella valutazione del test set GQA integrale nel par. 4.2.1, gli errori gravi sono 68 sul balanced test set GQA, il 6,8%, e 58 sul nuovo test set italiano, il 5,8%. Questo ultimo risultato ci porta a riconsiderare le differenze di performance sui due test set che, se prima era già considerabile minima, ora risulta quasi azzerata.

4.2.3 Valutazione generale

Riassumendo i risultati ottenuti ed illustrati dettagliatamente nei par. 4.2.1 e 4.2.2, è possibile notare che la differenza di performance di LXMERT sull'input inglese originale e su quello in inglese tradotto non è mai particolarmente rilevante.

Di seguito, i grafici che riassumono i dati relativi alle percentuali di errori nelle risposte alle domande in inglese originale:

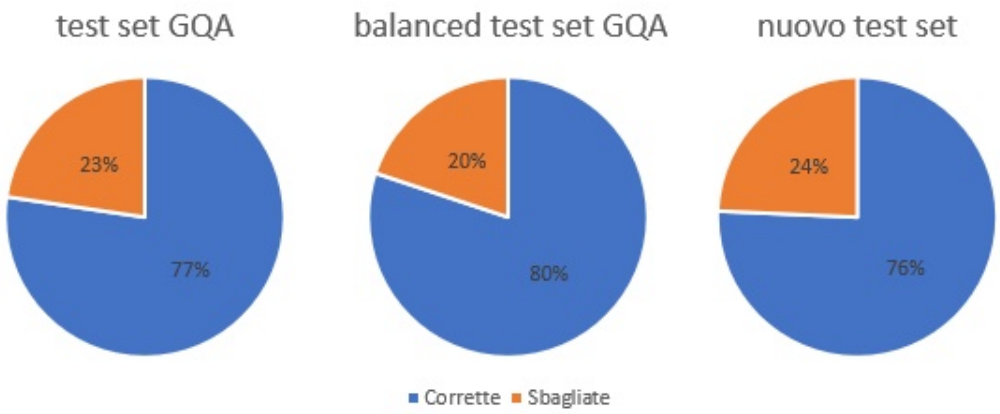


Figura 19 Valutazione delle risposte all'input inglese originale sul test set di GQA, sul balanced test set di GQA e sul nuovo test set.

Si riportano per confronto anche i grafici che riassumono i dati relativi alle percentuali di errori nelle risposte alle domande in inglese tradotto dall'italiano:

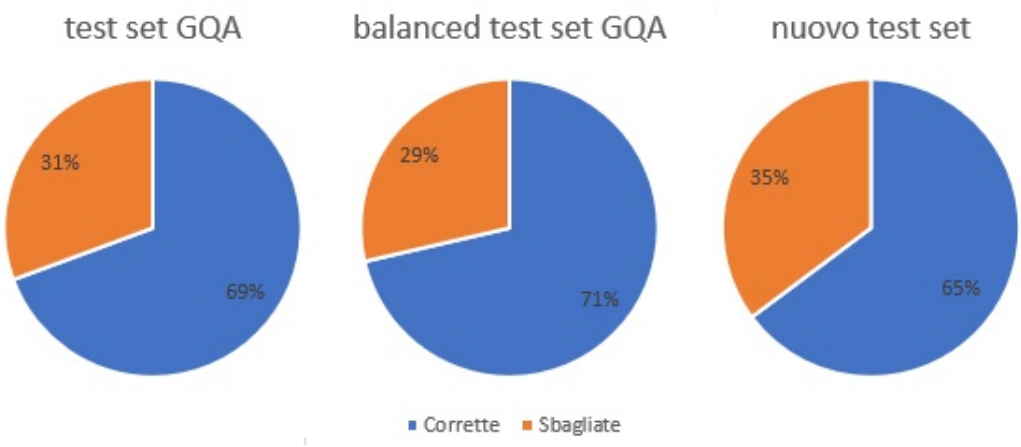


Figura 20 Valutazione delle risposte all'input italiano, poi tradotto automaticamente in inglese, rispettivamente (da sinistra a destra): sul test set di GQA, sul balanced test set di GQA e sul nuovo test set.

Comparando i grafici, è visibile che la percentuale di risposte corrette ha un calo sull'input italiano ma è altrettanto chiaro che la decrescita non è ingente in nessuno dei tre casi.

La differenza più significativa è quella tra i risultati sul nuovo test set che però, come illustrato nel par. 4.2.2, può essere notevolmente ridotta prendendo in considerazione solo gli errori più gravi.

4.2.4 Incidenza del contesto

Nonostante ci si aspettassero risultati differenti, dalla valutazione è emerso che il contesto non ha alcuna incidenza con le prestazioni.

L'introduzione del contesto ha prodotto leggeri miglioramenti della traduzione della domanda, tranne che nel balanced test set GQA, in cui non è stato riscontrato alcun caso. Nel dettaglio, sulle 3379 domande del test set GQA ci sono stati 5 casi di miglioramento; sulle 1000 domande del nuovo test set i casi sono stati 8.

Sebbene valori così bassi dimostrino la mancata correlazione tra una traduzione corretta e l'introduzione del contesto, in considerazione anche del suo ridotto costo computazionale, mantenere la traduzione congiunta di domanda e risposta non può che apportare dei miglioramenti al sistema, per quanto parziali e limitati.

5 Conclusioni

Tenendo conto dei risultati ottenuti dalle valutazioni proposte, l'idea di lavorare solo sulla manipolazione di input e output per il *porting* da una lingua all'altra si conferma una strategia efficace.

La differenza tra le performance sull'input inglese e quelle sull'input italiano non supera mai l'11%, percentuale di rumore accettabile non solo nell'ottica di un esperimento da un costo di realizzazione basso in termini computazionali e temporali, ma anche in rapporto ai possibili miglioramenti applicabili al lavoro. Un punto di partenza per futuri lavori di raffinamento è, sicuramente, l'individuazione di tecniche per aumentare l'incidenza del contesto. Questo porterebbe ad una cospicua riduzione di errori di polisemia, genere, numero e resa verbale, limitando gli output errati alle traduzioni integralmente sbagliate delle domande, che sono già in numero molto esiguo.

Un'importante opportunità di miglioramento è offerta, inoltre, dalla costante e veloce evoluzione dei sistemi di traduzione automatica. Grazie ai progressi in questo campo, sarà possibile raffinare anche le traduzioni, ottenendo un azzeramento quasi totale degli errori.

Inoltre, sebbene LXMERT si confermi meno efficace con dati a lui meno familiari, è molto interessante notare la validità della metodologia proposta non solo sui dati estratti da GQA, ma anche su input nuovi, confermando la possibilità di generalizzazione del processo.

L'applicazione di sistemi di traduzione automatica ad input e output permette di estendere l'utilizzo di modelli computazionali sofisticati anche a sistemi linguistici meno diffusi, evitando scelte computazionalmente costose e complesse. Tenendo anche in considerazione gli spunti di ottimizzazione proposti, la strategia applicata per l'utilizzo di un framework come LXMERT in lingua italiana può quindi, facilmente essere estesa anche ad altri modelli con risultati più che soddisfacenti.

6 Bibliografia

[1] Hudson, Drew A., and Christopher D. Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. . 3, 4, 5, 6.

[2] Hao Tan, Mohit Bansa. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers.” - arXiv preprint arXiv:1908.07490, 2019 - arxiv.org. p 2, 4.

[3] GQA, About
https://cs.stanford.edu/people/dorarad/gqa/about.html (visitato il 3 luglio 2021).

[4] PyPi, deep-translator 1.4.4
https://pypi.org/project/deep-translator/#id1 (visitato il 3 luglio 2021)

[5] Wikipedia en, voce Google Translate
https://en.wikipedia.org/wiki/Google_Translate (visitato il 3 luglio 2021)

7 Appendice

7.1 Abbreviazioni di uso frequente

Application programming interface	API
Pagina	p.
Paragrafo	par.
Region-based convolutional neural network	R-CNN
Visual Question Answering	VQA

Ringraziamenti

Ringrazio la dottoressa Lucia Passaro per il prezioso contributo nella realizzazione del lavoro e le colleghe Elisa Barisani e Simona Sette per aver condiviso il processo di definizione dell'esperimento e la valutazione dei risultati.