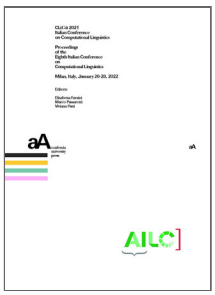

Elisabetta Fersini, Marco Passarotti and Viviana Patti (dir.)

Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it 2021
Milan, Italy, 26-28 January, 2022

Accademia University Press

GQA-it: Italian Question Answering on Image Scene Graphs

Danilo Croce, Lucia C. Passaro, Alessandro Lenci and Roberto Basili

DOI: 10.4000/books.aaccademia.10535
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2022
Published on OpenEdition Books: 20 October 2022
Series: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic EAN: 9791280136947

GQA-it: Italian Question Answering on Image Scene Graphs

Danilo Croce[1], **Lucia C. Passaro**[2], **Alessandro Lenci**[3], **Roberto Basili**[1]
[1] Dept. of Enterprise Engineering, University of Rome "Tor Vergata"
[2] Dept. of Computer Science, University of Pisa
[3] Dept. of Philology, Literature and Linguistics, University of Pisa
croce@info.uniroma2.it, lucia.passaro@unipi.it,
alessandro.lenci@unipi.it, basili@info.uniroma2.it

Abstract

The recent breakthroughs in the field of deep learning have lead to state-of-the-art results in several Computer Vision and Natural Language Processing tasks such as Visual Question Answering (VQA). Nevertheless, the training requirements in cross-linguistic settings are not completely satisfying at the moment. The datasets suitable for training VQA systems for non-English languages are still not available, thus representing a significant barrier for most neural methods. This paper explores the possibility of acquiring in a semi-automatic fashion a large-scale dataset for VQA in Italian. It consists of more than 1 M question-answer pairs over 80k images, with a test set of 3,000 question-answer pairs manually validated. To the best of our knowledge, the models trained on this dataset represent the first attempt to approach VQA in Italian, with experimental results comparable with those obtained on the English original material.

1 Introduction

Multimodal information processing is crucial to deal with a wide array of human actions and real-world computer applications. Notably, when observing a real-world scene, agents – both human and virtual ones – should understand what kinds of objects it depicts and the relations occurring among them. Such understanding allows agents to reason about the scene and the context in which it appears, thus inferring additional information that can be used for different purposes.

In recent years, several Artificial Intelligence (AI) tasks have been proposed in order to challenge systems in drawing inferences from multimodal inputs bringing together both linguistic and visual contents. An important task boosting research in multimodal scenarios is represented by Visual Question Answering (Antol et al., 2015; Srivastava et al., 2020). This task consists of correctly answering natural language questions regarding an input image. This requires the integration of vision, language and commonsense knowledge to answer. In English, several benchmark datasets have been proposed to deal with visual reasoning and question answering (Antol et al., 2015; Hudson and Manning, 2019; Srivastava et al., 2020). However, despite the impressive advances obtained in this context thanks to both new available resources and models, other languages still lack large-scale datasets suitable to learn VQA models.

In this paper, we present the semi-automatic creation of GQA-it, a large-scale Italian dataset based on the balanced version of GQA (Hudson and Manning, 2019). Specifically, we obtained more than 1 million question/answer pairs in Italian over 80K images by applying Neural Machine Translation (NMT) and we manually validated $3,000$ examples to provide a valuable benchmark. Moreover, we adapted to Italian a state-of-the-art VQA neural architecture, namely LXMERT (Tan and Bansal, 2019), and we trained/evaluated it using GQA-it. The experimental evaluation in both languages shows comparable results. This result is particularly significant given the complexity of the task and the adoption of noisy, automatically translated material for training. To the best of our knowledge, this represents one of the first Italian VQA systems. GQA-it will be made available to the research community.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes the new GQA-it dataset. Section 4 presents the experimental evaluation obtained by creating a new model by using GQA-it. Conclusions and future work are drawn in Section 5.

2 Related Work

Available VQA Resources. Pioneering work in VQA has been made by Malinowski and Fritz (2014), collecting a dataset of 2,483 unique English questions about 1,449 real-world images. Then, Antol et al. (2015) introduced the task of Visual Question Answering, defined as follows: *Given an image and a natural language question about the image, the task is to provide an accurate natural language answer.* Both questions and answers are open-ended and can refer to different areas of the image. Indeed, VQA systems require a deep understanding of images and of the objects they depict, as well as reasoning abilities about available (multimodal) information. Along with proposing the new task, the authors also provided the very first large-scale VQA dataset, made of about 600k questions on about 200k images, taken from the Microsoft Objects in Context (MS COCO) dataset (Lin et al., 2014).

Afterwards, several other datasets on this topic have been created with the aim to pursue different goals (Goyal et al., 2017; Johnson et al., 2017; Zhu et al., 2016; Krishna et al., 2017). Notably, a common shortcomings of all these datasets is the presence of important real-world biases that are inherited also by neural models exploiting them for learning. Specifically, several studies report on the fact that models are driven by superficial correlations in the training data with the effect of lacking sufficient visual grounding (Agrawal et al., 2018; Goyal et al., 2017; Johnson et al., 2017).

To mitigate these aspects, the GQA dataset (Hudson and Manning, 2019) has been developed starting from Visual Genome (Krishna et al., 2017). The latter resource is valuable for several multimodal tasks, as it contains linguistically and visually more complex annotations. Specifically, images are annotated with the objects they contain and the relationships between them. In addition, Visual Genome contains a wide range of descriptions relative to specific portions of the image. Finally, the resource also comes with a visual question answering layer. However, Visual Genome is very complex from both a linguistic (ambiguity and redundancy) and visual (several regions describe the same objects) perspective, making it difficult to be easily used to train neural VQA models. This is the reason why additional normalization efforts have been performed to create a new resource, GQA (Hudson and Manning, 2019). From an annotation point of view, the resource is similar to Visual Genome, but with a lower linguistic and conceptual variability in terms of objects, relations, and descriptions. Moreover, to deal with the bias present in most of the VQA datasets, the authors created a rich *question engine* by exploiting objects, attributes and relations annotated in Visual Genome (Krishna et al., 2017) along with compositional patterns and lexical resources. In this work, we adopted the GQA dataset because, differring from the other ones, it challenges the reasoning capabilities of the models.

Neural models for VQA. The proliferation of shared tasks on this topic, led to a great technological enhancement in terms of pre-trained end-to-end models to perform visual question answering. A first benchmark is represented by the model proposed by Antol et al. (2015), which uses a CNN for visual feature extraction and a LSTM or Recurrent networks for language processing. The introduction of attention (Chen et al., 2015; Andreas et al., 2016; Yang et al., 2016) improved the results on the VQA benchmark allowing the model to focus on specific portions of the image. Subsequently, Teney et al. (2018) exploited object detection to perform VQA. The model employs R-CNN architecture and achieves good results. The introduction of Transformers and their success in NLP (Devlin et al., 2019) inspired works based on large-scale pre-training and fine-tuning studies on cross-modality. One of the first multimodal models of this generation was proposed by Tan and Bansal (2019) with the development of LXMERT, used in this work. LXMERT has been originally developed to work with GQA and embeds BERT, easily adaptable to Italian through its multilingual counterpart (Pires et al., 2019).

Multilingual approaches for VQA. More recently, new attempts have been devoted to Multilingual Visual Question Answering (Gupta et al., 2020). However, to the best of our knowledge, no gold VQA datasets is available for Italian. Therefore, this work aims to enable the training and evaluation of VQA methods in Italian, regardless of whether they are multilingual or not.

3 GQA-it: the Italian VQA Dataset

In order to build a valuable resource for Italian VQA, we considered the balanced version of GQA, in which the question distribution has been smoothed to obtain a more balanced and repre-

sentative question/answer sample. In particular, we started from the benchmark split provided by Tan and Bansal (2019), namely the `train`[1] and `validation`[2] material. Moreover, the GQA test set is not publicly available. Therefore, we adopted the `test-dev`[3] subset, which represents a subset of the original test material, but it is defined to be highly representative of different linguistic and conceptual phenomena. Moreover, systems evaluated on this smaller dataset are generally in line with respect to the evaluations applied to the larger test set.

We aim to generate a large-scale dataset in which training and validation material is obtained via automatic neural machine translation and the test material is manually validated. This approach allows us to i.) create a benchmark test set in Italian and ii.) measure how sensitive the system is to the noise introduced by the machine translation. We thus applied Opus-NMT (Tiedemann and Thottingal, 2020), a Transformer-based Neural Machine translation trained on the OPUS parallel corpus, a large scale collection of texts semi-automatically aligned for several language pairs. We selected the model trained on the aligned subset of documents in the English/Italian pairs.[4] The quality of the translated questions is evaluated on a portion of the dataset. Notably, manual validation has been performed on 500 items, consisting of 250 random questions taken from the training set and 250 random questions taken from the test set. Given the characteristics of the texts contained in GQA (simple texts, no sub-sentence level) and the implementation simplicity and reproducibility, we decided to use the BLEU score for the evaluation. Overall, the performance reaches 0.82. This is impressively high, but quite in line with the BLEU obtained by the adopted translation model over the `Tatoeba.it.en` dataset (BLEU=0.72) composed of short sentences with syntactical complexity similar to the GQA dataset.[5]

The translation of answers (here expressed only with one or two tokens) is more problematic. In fact, many answers should be translated differently depending on the context or associated image, e.g., an answer "*bat*" can be translated as the animal "*pipistrello*" or the object "*mazza*". As suggested in Croce et al. (2019), in order to reduce such lexical ambiguity, we translated an answer by pairing it with the corresponding question. This way, we exploit the context sensitive nature of the adopted Transformer-based architecture: the answer "*mouse*" is thus correctly translated when paired with the question "*What's next to the keyboard?*", while generic translations, such as "*topo*", are systematically preferred when no context is made available. Unfortunately, the lexical variability of the automatically translated answers was problematic. In fact, the initial English material was characterized by $1,842$ possible answers types. After the automatic translation, this number increased to $3,306$. This is partially due to the cases in which the context does not improve the translation, e.g., the question "*What's at the top of the photo*?" is not really helpful to disambiguate the answer "*mouse*".

In other cases, multiple ways to translate the same lexical item exist, e.g., "*aircraft*" is translated both as "*aeromobile*" or "*aeroplano*". Finally, while answers involving singular and plural expressions were kept separated in the original dataset, gender is generally not marked in English, differently from Italian. Most of the times a context-sensitive translation inflected the translation in masculine and feminine. For example, "*little*" was translated in "*piccola*", "*piccolo*", "*piccole*" and "*piccoli*" depending on the items involved in the photo. To reduce this lexical variability, we applied a manual normalization to answers associated to more than two questions. We paired each original English answer with the translated ones, in order to manually normalize the translations. While this kind of manual validation is generally ineffective when dealing with machine translation, we considered that, by design, English GQA has a limited amount of polysemy, as questions, answers, and graph annotations have been automatically normalized to reduce the linguistic ambiguity (Hudson and Manning, 2019). In practice, when mentioning a "*sign*", answers (almost) always refer to objects such as a "*signboard*" more than a "*mark*" or a "*gesture*".[6] We preserved singular and plural forms. Actions, e.g.,

[1] `https://nlp.cs.unc.edu/data/lxmert_data/gqa/train.json`

[2] `https://nlp.cs.unc.edu/data/lxmert_data/gqa/valid.json`

[3] `https://nlp.cs.unc.edu/data/lxmert_data/gqa/testdev.json`

[4] `https://github.com/Helsinki-NLP/OPUS-MT-train/tree/master/models/it-en`

[5] The results of the model are available in the Github page.

[6] Only the word "*glass*" was used in both senses of "*bicchiere*" and "*vetro*", while all other words were generally characterized by only one sense.

"*skating*", "*jumping*" or "*sleeping*", were translated as the gerundive forms "*sta facendo skateboard*", "*sta saltando*" e "*sta dormendo*". Unfortunately, the noise introduced when translating adjectives makes the gender of such words problematic, so that we normalized all forms to the masculine gender. After this manual normalization, the number of possible answers across the dataset is $1,701$.

Table 1 shows the 50 most frequent answers in both the English and the Italian dataset, showing that the distribution is generally preserved across languages.

GQA
yes (17.6%) - *no* (17.6%) - *left* (5.2%) - *right* (5.1%) - *man* (1.2%) - *white* (1.2%) - *black* (1.1%) - *bottom* (0.9%) - *woman* (0.9%) - *chair* (0.9%) - *blue* (0.9%) - *top* (0.8%) - *table* (0.8%) - *brown* (0.8%) - *boy* (0.7%) - *gray* (0.6%) - *dog* (0.6%) - *green* (0.6%) - *bed* (0.6%) - *cat* (0.6%) - *girl* (0.6%) - *red* (0.5%) - *car* (0.5%) - *horse* (0.5%) - *color* (0.4%) - *bus* (0.4%) - *desk* (0.4%) - *large* (0.4%) - *orange* (0.4%) - *couch* (0.4%) - *small* (0.4%) - *yellow* (0.4%) - *shelf* (0.4%) - *elephant* (0.4%) - *people* (0.4%) - *shirt* (0.3%) - *train* (0.3%) - *wood* (0.3%) - *metal* (0.3%) - *truck* (0.3%) - *child* (0.3%) - *laptop* (0.3%) - *jacket* (0.3%) - *giraffe* (0.3%) - *player* (0.3%) - *field* (0.3%) - *cabinet* (0.3%) - *lady* (0.3%) - *guy* (0.3%) - *pink* (0.2%) -
GQA-it
sì (17.6%) - *no* (17.6%) - *sinistra* (5.2%) - *destra* (5.1%) - *uomo* (1.2%) - *bianco* (1.2%) - *nero* (1.1%) - *ragazzo* (1.0%) - *inferiore* (0.9%) - *donna* (0.9%) - *sedia* (0.9%) - *blu* (0.9%) - *in alto* (0.8%) - *marrone* (0.8%) - *tavola* (0.8%) - *auto* (0.6%) - *grigio* (0.6%) - *cane* (0.6%) - *verde* (0.6%) - *letto* (0.6%) - *divano* (0.6%) - *gatto* (0.6%) - *ragazza* (0.6%) - *rosso* (0.5%) - *cavallo* (0.5%) - *autobus* (0.4%) - *colore* (0.4%) - *piccolo* (0.4%) - *scrivania* (0.4%) - *grande* (0.4%) - *arancione* (0.4%) - *giallo* (0.4%) - *ripiano* (0.4%) - *elefante* (0.4%) - *persone* (0.4%) - *cappello* (0.4%) - *camicia* (0.3%) - *armadio* (0.3%) - *strada* (0.3%) - *bambino* (0.3%) - *treno* (0.3%) - *camion* (0.3%) - *legno* (0.3%) - *campo* (0.3%) - *metallo* (0.3%) - *laptop* (0.3%) - *giacca* (0.3%) - *giraffa* (0.3%) - *giocatore* (0.3%) - *signora* (0.3%)

Table 1: The 50 most frequent answers in the datasets. For each word the percentage of associated questions is reported.

Finally, to provide a valuable resource for real-scale evaluation of NLP systems, we manually validated a subset of the test material, by correcting $3,000$ question/answer pairs, randomly selected to preserve data balance. In particular, we also restored the gender inflection, lost during the previous normalization process.

The resulting dataset, namely **GQA-it**[7] is a large scale (possibly noisy) dataset made of more than 1.08 M of question/answers insisting on more than 80k images, with a test set partially validated. Specific statistics about GQA-it are reported in Table 2. Note that "silver" refers to non-validated material, while "gold" refers to manually validated ones. Each question/answer pair is connected to an image and the identifiers are aligned to the original GQA resource, thus enabling the reuse of further levels of valuable information, such as the knowledge graph associated with each image. Figure 1 shows both English and Italian Question Answer pairs for an example image taken from GQA-it.

[7]The resource is publicly available at `https://github.com/crux82/gqa-it`.

Dataset	**#images**	**#quest./ans. pairs**
`train`	72,140	943,000
`valid`	10,234	132,062
`test-dev` (**silver**)	398	12,578
`test-dev` (**gold**)	398	3,000

Table 2: Statistics of The GQA-it dataset. The gold `test-dev` is a subset of the silver one.

Figure 1: Examples from the GQA-it dataset (image id n90294):
Q(A)$_{en}$: Is the remote to the right or to the left of the book? (right). **Q(A)**$_{it}$: *Il telecomando è a destra o a sinistra del libro? (destra)*
Q(A)$_{en}$: How thick is the book to the left of the remote? (thick). **Q(A)**$_{it}$: *Quanto è spesso il libro a sinistra del telecomando? (spesso)*
Q(A)$_{en}$: What device is to the left of the calculator made of plastic? (charger). **Q(A)**$_{it}$: *Quale dispositivo si trova a sinistra della calcolatrice di plastica? (caricabatterie)*
Q(A)$_{en}$: What's the charger made of? (platic). **Q(A)**$_{it}$: *Di cosa è fatto il caricabatterie? (plastica)*
Q(A)$_{en}$: Are there any phones? (no). **Q(A)**$_{it}$: *Ci sono dei telefoni? (no).*

4 Experimental Evaluation

To assess the quality of the produced GQA-it dataset, we trained and evaluated a state-of-the-

art VQA system over the automatically generated material and evaluated over the 3,000 manually validated test set. In particular, we evaluated LXMERT (Learning Cross-Modality Encoder Representations from Transformers) presented by Tan and Bansal (2019).[8] This neural architecture models the VQA problem by stacking three neural encoders: an object/relationship encoder encoding (which encodes the input images), a language encoder (which encodes the input questions) and a cross-modality encoder (that combines the above multimodal embeddings). In a nutshell, LXMERT extracts visual and linguistic information, combines them in the cross-modal encoder and applies a (linear) classifier that associates each image/question pair to one of the n possible answers considered in the dataset.

The object detector uses a Faster R-CNN model (Ren et al., 2015) built over the ResNet-101 backbone (He et al., 2015) and pre-trained on the Visual Genome dataset (Krishna et al., 2017) to encode salient area of the input images. The language encoder is implemented as a BERT based model (Devlin et al., 2019). In Tan and Bansal (2019) best results are obtained without using existing pre-trained BERT models: the weights of this encoder are randomly initialized and pre-trained (together with the weights of cross-modality encoder) using a dedicated large scale dataset. This is composed of image captions and related questions of about 9 millions sentences. This pre-training stage is implemented by defining 5 auxiliary tasks, e.g., the cross-modal alignment task ("does the sentence describes the image?"). Nonetheless, experimental results showed that good performances can be also obtained by adopting a pre-trained BERT model. In order to effectively train LXMERT over GQA-it, we replaced the specialized English model with a standard pre-trained BERT model, in particular, multilingual BERT (Pires et al., 2019), which is also available for Italian. We preserved the original object/relationship encoder (which is language independent) and randomly initialized the cross-modality encoder.

Performances are measured in terms of Accuracy, i.e., the percentage of questions that exactly received the correct answer. All experiments were conducted using the same parameters used in Tan and Bansal (2019) but we investigated up to 15 epochs in the fine-tuning. Results are reported in Table 3. To compare the effectiveness of LXMERT on English and Italian data, we selected the common subset of $3,000$ question/answer pairs in both languages. The task is extremely challenging: A system assigning random answers would achieve an accuracy of 0.05%. Considering that the dataset is quite imbalanced, a baseline system assigning the most frequent answer (here, "*yes*"/ "*si*") achieves 17.6%. First, we applied the best model from Tan and Bansal (2019) (namely `en-pretrain`) that is pre-trained over the dedicated corpus: while it achieves 60.0% (almost the state-of-the-art) on the entire English `test-dev` dataset, it achieves 59.0% on this subset. Tan and Bansal (2019) show that performances drop to 56.2% when using the original pre-trained BERT, and the English multilingual counterpart here achieves 55.3%. This drop in performances confirms the findings of Tan and Bansal (2019) and represents a sort of upper-bound for the experiments in Italian, as all the above setups are not affected by the noise introduced in the training material of GQA-it.

	Model	Accur.
-	**baseline (most freq. answer)**	17.6%
en	**`LXMERT en-pretrain`**	59.0%
	`LXMERT bert-multi.`	55.3%
it	**`LXMERT en-pretrain` + MT**	47.1%
	`LXMERT bert multi.` + MT	44.8%
	`LXMERT-it` (gold ans.)	51.0%
	`LXMERT-it` (silver ans.)	52.6%

Table 3: Results of LXMERT and LXMERT-it on 3,000 questions of GQA and GQA-it.

In order to assess the value of the new Italian resource, we first evaluated a trivial workflow that re-used the above English models in an Italian setting (first two rows in the Italian section of Table 3). First, we automatically translated the Italian questions using Opus-NMT in English ($\text{mt}_{it \rightarrow en}$). Second, we applied the English LXMERT models (`en-pretrain` and `bert-multilingual`) to derive the English answers. Finally, we applied Opus-NMT to translate back answers to Italian ($\text{mt}_{en \rightarrow it}$), after pairing them with the questions, as discussed in the previous section (cf. Table 3, rows `LXMERT en-pretrain` + MT and `LXMERT bert multi.` + MT). Indeed, this trivial workaround achieved significant results, i.e., 47.1% and 44.8%. This drop is par-

[8] `https://github.com/airsplay/lxmert`

tially due to the $it \rightarrow en$ translation, as the performances of the `en-pretrain` model drops from 59.0% to 54.5% when applied to English questions derived via machine translation, while the `bert-multilingual` from 55.3% to 51.3%. We suppose that the language model of LXMERT is not robust to the noise induced by the NMT. The remaining performance drop is clearly due to the translation $en \rightarrow it$, mainly due to polysemy and the other phenomena discussed in the previous section.

Conversely, the model trained over GQA-it, namely LXMERT-it, achieves 51.0% accuracy, which improves the previous results and it is more in line with the results obtained with `bert-multilingual` in English. Evaluating LXMERT-it w.r.t. the answers generated with the proposed methodology (namely *silver* answers) raises the accuracy to 52.6%. A manual analysis of the differences reveals that they are mainly due to gender inflections (e.g., "*alto*" vs "*alta*", in English "*tall*"). Unfortunately, these cases will inevitably be misclassified by LXMERT-it since it only observed masculine forms during training (which were introduced during the initial normalization phase.

We performed a qualitative error analysis on a random sample of the test set (10%). We identified 6 main error classes. Overall, 44% of the questions produced a wrong answer. First of all, we can make some considerations on these errors. On the one hand, specific errors are due to the wrong identification of objects in the images. In this paper, we did not modify the visual component of the architecture, and therefore the corresponding errors could not be avoided. Many other errors may be attributed to issues related to the machine translation, and in general with the creation of a noisy system for visual question answering. In particular, some errors are critical for the correct comprehension of questions and answers, and in general for using the Italian VQA model. In fact, some errors compromise the correct understanding of the answers (e.g., "*right*" translated in Italian as "*corretto*" instead of "*destra*"), while others allow the correct (albeit noisy) use of the system, such as the use of synonyms and hypernyms of the gold class.

Error Type	Example(s)	Perc.
Object	*tavola* ('table') vs *sedia* ('chair')	31%
Synonyms or hypernyms	*persona* ('person') vs *donna* ('woman')	17%
Attributes	*blu* ('blue') vs *nero* ('black'); *chiuso* ('closed') vs *aperto* ('open')	14%
Morph. feat.	*bella* ('beautiful') vs *bello* ('beautiful'); *persona* ('person') vs *persone* ('people')	3%
Actions	*sta dormendo* ('sleeping') vs *sta sdraiato* ('is lying down)'	3%
Spatial feat.	*destra* ('right') vs *sinistra* ('left')	2%
Residual	*si* ('yes') vs *no* ('no')	31%

Table 4: Classification of errors in `LXMERT-it`.

5 Conclusions

This paper presents GQA-it, a collection of more than 1 M question/answer pairs in Italian associated to 80k images in support of research in VQA in Italian. GQA-it has been obtained with machine translation, and the quality of the resulting resource is demonstrated through both direct evaluation of the translation and indirect evaluation of a state-of-the-art model trained on this material.

This work represents a first step to leverage a large-scale VQA resource like GQA for Italian, a resource whose quality can still largely been improved. In particular, the knowledge graphs behind each image will be extremely valuable to improve the final resource (e.g., using a generation process as in (Hudson and Manning, 2019)) or the VQA process. Finally, the available alignment between GQA and GQA-it will foster research in cross-lingual VQA.

The aim of this paper was to explore the possibility of semi-automatically inducing large-scale Italian dataset for VQA. Obviously, we are aware that there is plenty of room for improvement in many respects. First, a wide range of approaches could be tested, aimed at reducing the noise due to the adaptation of English resources to Italian ones. Specifically, a viable option could be to leverage the question and the image together with each other in order to provide a more consistent translation. Finally, a multimodal masked language modeling step on text-image pairs could enrich the Italian BERT model and make it comparable with the English counterpart. We plan to probe these research avenues in the near future.

Acknowledgments

We would like to thank the "Istituto di Analisi dei Sistemi ed Informatica - Antonio Ruberti" (IASI) for supporting the experimentations through access to dedicated computing resources.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and 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*, pages 4971–4980.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

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

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and 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*, pages 6904–6913.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and 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*, pages 900–913, Suzhou, China, December. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Drew A Hudson and 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*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and 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*, pages 2901–2910.

Ranjay Krishna, 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. *International journal of computer vision*, 123(1):32–73.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2020. Visual question answering using deep learning: A survey and performance analysis.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November. Association for Computational Linguistics.

Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

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