Deep Learning for Automatic Image Captioning in poor Training Conditions

Caterina Masotti and Danilo Croce and Roberto Basili
Department Of Enterprise Engineering
University of Roma, Tor Vergata
caterinamasotti@yahoo.it
{croce,basili}@info.uniroma2.it

Abstract

English. Recent advancements in Deep Learning show that the combination of Convolutional Neural Networks and Recurrent Neural Networks enables the definition of very effective methods for the automatic captioning of images. Unfortunately, this straightforward result requires the existence of large-scale corpora and they are not available for many languages. This paper describes a simple methodology to automatically acquire a large-scale corpus of 600 thousand image/sentences pairs in Italian. At the best of our knowledge, this corpus has been used to train one of the first neural systems for the same language. The experimental evaluation over a subset of validated image/captions pairs suggests that results comparable with the English counterpart can be achieved.

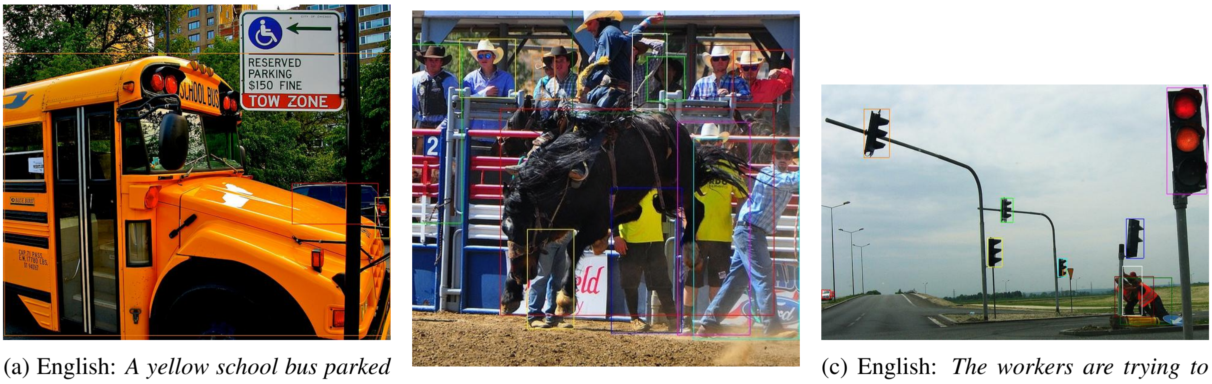
Italiano. *La combinazione di metodi di Deep Learning (come Convolutional Neural Network e Recurrent Neural Network) ha recentemente permesso di realizzare sistemi molto efficaci per la generazione automatica di didascalie a partire da immagini. Purtroppo, l'applicazione di questi metodi richiede l'esistenza di enormi collezioni di immagini annotate e queste risorse non sono disponibili per ogni lingua. Questo articolo presenta un semplice metodo per l'acquisizione automatica di un corpus di 600 mila coppie immagine/frase per l'italiano, che ha permesso di addestrare uno dei primi sistemi neurali per questa lingua. La valutazione su un sottoinsieme del corpus manualmente validato suggerisce che é possibile raggiungere risultati comparabili con i sistemi disponibili per l'inglese.*

1 Introduction

The image captioning task consists in generating a brief description in natural language of a given image that is able to capture the depicted objects and the relations between them, as discussed in (Bernardi et al., 2016). More precisely, given an image I as input, an *image captioner* should be able to generate a well-formed sentence $S(I) = (s_1, ..., s_m)$, where every s_i is a word from a vocabulary $V = \{w_1, ..., w_n\}$ in a given natural language. Some examples of images and corresponding captions are reported in Figure 1. This task is rather complex as it involves non-trivial subtasks to solve, such as object detection, mapping visual features to text and generating text sequences.

Recently, neural methods based on deep neural networks have reached impressive state-of-the-art results in solving this task (Karpathy and Li, 2014; Mao et al., 2014; Xu et al., 2015). One of the most successful architectures implements the so-called *encoder-decoder* end-to-end structure (Goldberg, 2015). Differently by most of the existing encoder-decoder structures, in (Vinyals et al., 2014) the encoding of the input image is performed by a convolutional neural network which transform it in a dense feature vector; then, this vector is "translated" to a descriptive sentence by a Long short-term memory (LSTM) architecture, which takes the vector as the first input and generates a textual sequence starting from it. This neural model is very effective, but also very expensive to train in terms of time and hardware resources[1], because there are many parameters to be learned; not to mention that the model is overfitting-prone, thus it needs to be trained on a training set of annotated images that is as large and heterogeneous

[1] As of now, training a neural encoder-decoder model such as the one presented at http://github.com/tensorflow/models/tree/master/im2txt on a dataset of over $580,000$ image-caption examples takes about two weeks even with a very performing GPU.



(a) English: *A yellow school bus parked in a handicap spot*, Italian: *Uno scuolabus giallo parcheggiato in un posto per disabili*.

(b) English: *A cowboy rides a bucking horse at a rodeo*, Italian: *Un cowboy cavalca un cavallo da corsa a un rodeo*.

(c) English: *The workers are trying to pry up the damaged traffic light*, Italian: *I lavoratori stanno cercando di tirare su il semaforo danneggiato*.

Figure 1: Three images from the MSCOCO dataset, along with two human-validated descriptions.

as possible, in order to achieve a good generalization capability. Hardware and time constraints do not always allow to train a model in an optimal setting, and, for example, cutting down on the dataset size could be necessary: in this case we have *poor training conditions*. Of course, this reduces the model's ability to generalize on new images at captioning time. Another cause of poor training conditions is the lack of a good quality dataset, for example in terms of annotations: the manual captioning of large collections of images requires a lot of effort and, as of now, human-annotated datasets only exist for a restricted set of languages, such as in English. As a consequence, training such a neural model to produce captions in another language (e.g. in Italian) is an interesting problem to explore, but also challenging due to the lack of data resources.

A viable approach is building a resource by *automatically translating the annotations from an existing dataset*: much less expensive than manually annotating images, but of course it leads to a loss of human-like quality in the language model. This approach has been considered in this work to perform one of the first neural-based image captioning in Italian: more precisely, the annotations of the images from the MSCOCO dataset, one of the largest datasets in English of image/caption pairs, have been automatically translated to Italian in order to obtain a first resource for this language: this has been exploited to train a neural captioner and whose quality can be improved over time (e.g., by manually validating the translations). Then, a subset of this Italian dataset has been used as training data for the neural captioning system defined in (Vinyals et al., 2014), while a subset of the test set has been manually validated for evaluation purposes.

In particular, prior to the experimentations in Italian, some early experiments have been performed with the same training data originally annotated in English, to get a reference benchmark about convergence time and evaluation metrics on a dataset of smaller size. These results in English will suggest if the Italian image captioner shows similar performance when trained over a reduced set of examples or the noise induced in the automatic translation process compromises the neural training phase. Moreover, these experiments have also been performed with the introduction of a pre-trained word embedding, (derived using the method presented in (Mikolov et al., 2013)), in order to measure how it affects the quality of the language model learned by the captioner, with respect to a randomly initialized word embedding that is learned together with the other model parameters.

Overall the contributions of this work are threefold: (*i*) the investigation of a simple, automatized way to acquire (possibly noisy) large-scale corpora for the training of neural image captioning methods in poor training conditions; (*ii*) the manual validation of a first set of human-annotated resources in Italian; (*iii*) the implementation of one of the first automatic neural-based Italian image captioners.

In the rest of the paper, the adopted neural architecture is outlined in Section 2. The description of a brand new resource for Italian is presented in Section 3. Section 4 reports the results of the early preparatory experimentations for the English language and then the ones for Italian. Finally, Section 5 derives the conclusions.

2 The Show and Tell Architecture

The Deep Architecture considered in this paper is the *Show and Tell* architecture, described in (Vinyals et al., 2014) and sketched in Figure 2. It follows an encoder-decoder structure where the image is encoded in a dense vector by a state-of-the-art deep CNN, in this case *InceptionV3* (Szegedy et al., 2015), followed by a fully connected layer; the resulting feature vector is fed to a LSTM, used to generate a text sequence, i.e. the caption. As the CNN encoder has been trained over an object recognition task, it allows encoding the image in a dense vector that is strictly connected to the entities observed in the image. At the same time, the LSTM implements a language model, in line with the idea introduced in (Mikolov et al., 2010): it captures the probability of generating a given word in a string, given the words generated so far. In the overall training process, the main objective is to train a LSTM to generate the next word given not only the string produced so far, but also a set of image features. As the first CNN encoder is (mostly) language independent, it can be totally re-used even in the captioning of images in other languages, such as Italian. On the contrary, the language model underlying the LSTM needs new examples to be trained.

In this work, we will train this architecture over a corpus that has been automatically translated from the MSCOCO dataset. We thus speculate that the LSTM will learn a sort of simplified language model, more inherent to the automatic translator than to an Italian speaker. However, we are also convinced that the quality achievable by modern translation systems (Bahdanau et al., 2014; Luong et al., 2015), combined with the generalization that can be obtained by a LSTM trained over thousands of (possibly noisy) translations will be able to generate reasonable and intelligible captions.

3 Automatic acquisition of a Corpus of Captions in Italian

In this section we present the first release of the MSCOCO-it, a new resource for the training of data-driven image captioning systems in Italian. It has been built starting from the MSCOCO dataset for English (Lin et al., 2014): in particular we considered the training and validation subsets, made respectively of $82,783$ and $40,504$ images, where every image has 5 human-written annotations in English. The Italian version of the dataset has been acquired with an approach that automatizes the translation task: for each image, all its five annotations have been translated with Bing[2]. The result is a big amount of data whose annotations are fully translated, but not of the best quality with respect to the Italian fluent language. This automatically translated data can be used to train a model, but for the evaluation a test set of human-validated examples is needed: so, the translations of a subset of the MSCOCO-it have been manually validated. In (Vinyals et al., 2014), two subsets of $2,024$ and $4,051$ images from the MSCOCO validation set have been held out from the rest of the data and have been used for development and testing of the model, respectively. A subset of these images has been manually validated: 308 images from the development set and 596 from the test set. In Table 1, statistics about this brand new corpus are reported, where the specific amount of unvalidated (*u.*) and validated (*v.*) data is made explicit[3].

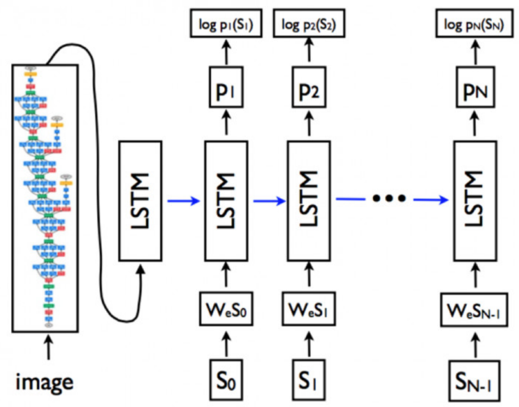


Figure 2: The Deep Architecture presented in (Vinyals et al., 2014). LSTM model combined with a CNN image embedder and word embeddings. The unrolled connections between the LSTM memories are in blue.

4 Experimental Evaluation

In order to be consistent with a scenario characterized by *poor training conditions* (limited hardware resources and time constraints) all the experimentations in this paper have been made by training

[2] Sentences have been translated between December 2016 and January 2017.

[3] Although Italian annotations are available for all the images of the original dataset, in the table some images were not counted because they are corrupted and therefore have not been used.

		#images	#sent	#words
training	*u.*	116,195	581,286	6,900,546
valid.	*v.*	308	1,516	17,913
	u.	1,696	8,486	101,448
	p.	(14)	25	304
test	*v.*	596	2,941	34,657
	u.	3,422	17,120	202,533
	p.	(23)	41	479
total		**122,217**	**611,415**	**7,257,880**

Table 1: Statistics about the MSCOCO-it corpus. *p.* stands for *partially validated*, since some images have only some validated captions out of five. The partially validated images are between parentheses because they are already counted in the validated ones.

the model on significantly smaller samples of data with respect to the whole MSCOCO dataset (made of more than 583,000 image-caption examples).

First of all, some early experimentations have been performed on smaller samples of data from MSCOCO in English, in order to measure the loss of performance caused by the reduced size of the training set[4]. Each training example is a image-caption pair and they have been grouped in data *shards* during the training phase: each shard contains about 2,300 image-caption examples. The model has been trained on datasets of 23,000, 34,500 and 46,000 image-caption pairs (less than 10% of the entire dataset).

In order to balance the reduced size of the training material and provide some kind of linguistic generalization, we evaluated the adoption of pre-trained word embedding in the training/tagging process. In fact, in (Vinyals et al., 2014) the LSTM architecture initializes randomly all vectors representing input words; these are later trained together with the other parameters of the network. We wondered if a word embedding already pre-trained on a large corpus could help the model to generalize better on brand new images at test time. We introduce a word embedding learned through a Skip-gram model (Mikolov et al., 2013) from an English dump of Wikipedia. The LSTM architecture has been trained on the same shards but initializing the word vectors with this pretrained word embedding.

# Shards	BLEU-4	METEOR	CIDEr
1	10,1 / 11,5	13,4 / 13,1	18,8 / 24,4
2	15,7 / 18,9	18,2 / 16,3	36,1 / 51,9
5	22,0 / 22,7	20,2 / 20,4	64,1 / 65,0
10	22,4 / 24,7	22,0 / 21,7	73,2 / 73,7
20	26,5 / 26,2	21,9 / 22,3	79,3 / 79,1
NIC	27,7	23,7	85,5
NICv2	32,1	25,7	99,8
`im2txt`	31,2	25,5	98,1

Table 2: Results on `im2txt` for the English language with a training set of reduced size, without / with and the use of a pre-trained word embedding. Moreover benchmark results are reported.

Table 2 reports results on the English dataset in terms of BLEU-4, CIDEr and METEOR, the same used in (Vinyals et al., 2014): in the first five rows, results are reported both in the case of randomly initialized word embedding and pre-trained ones. We compare these results with the ones achieved by the original NIC and NICv2 networks presented in (Vinyals et al., 2014), and the ones measured by testing a model available in the web[5], trained on the original whole training set.

Results obtained by the network when trained on a reduced dataset are clearly lower w.r.t. the NIC results, but it is straightforward that similar result are obtained, especially considering the reduced size of the training material. The contribution of pre-trained word embeddings is not significant, in line with the findings from (Vinyals et al., 2014). However, it is still interesting noting that the lexical generalization of this unsupervised word embeddings is beneficial, especially when the size of the training material is minimal (e.g. when 1 shard is used, especially if considering the CIDEr metrics). As the amount of training data grows, its impact on the model decreases, until it is not significant anymore.

# Shards	BLEU-4	METEOR	CIDEr
1	11.7 / 12.9	16.4 / 16.9	27.4 / 29.4
2	16.9 / 17.1	18.8 / 18.7	45.7 / 45.6
5	22.0 / 21.4	21.2 / 20.9	62.5 / 60.8
10	22.4 / 22.9	22.0 / 21.5	71.9 / 68.8
20	23.7 / 23.8	22.2 / 22.0	73.0 / 73.2

Table 3: Metrics for the experimentations on `im2txt` for the Italian language with a training set of reduced size, without / with and the use of a pre-trained word embedding.

For what concerns the results on Italian, the experiments have been performed by training the model on samples of 23,000, 34,500 and 46,000 examples, where the captions are automatically

[4] A proper tuning phase was too expensive so we adopted the parameters provided in `https://github.com/tensorflow/models/tree/master/im2txt`

[5] `http://github.com/tensorflow/models/issues/466`

translated with Bing. The model has been evaluated against the validated sentences, and results are reported in Table 3. Results are impressive as they are in line with the English counterpart. It supports the robustness of the adopted architecture, as it seems to learn even from a noisy dataset of automatically translated material. Most importantly, it confirms the applicability of the proposed simple methodology for the acquisition of datasets for image captioning.

When trained with 20 shards, the Italian captioner generates the following description of the images shown in Figure 1: Image 1a: "*Un autobus a due piani guida lungo una strada.*", Image 1b: "*Un uomo che cavalca una carrozza trainata da cavalli.*", Image 1c: "*Una persona che cammina lungo una strada con un segnale di stop.*"

An attempt to use a word embedding that has been pre-trained on a large corpus (more precisely, on a dump of Wikipedia in Italian) has also been made, but the empirical results reported in Table 3 show that its contribution is not relevant but still significant when fewer examples are adopted.

5 Conclusions

In this paper a simple methodology for the training of neural models for the automatic captioning of images is presented. We generated a large scale of about $600,000$ image captions in Italian by using an automatic machine translator. Although the noise introduced in this step, it allows to train one of the first neural-based image captioning systems for Italian. Most importantly, the quality of this system seems comparable with the English counterpart, if trained over a comparable set of data. These results are impressive and confirm the robustness of the adopted Neural Architecture. We believe that the obtained resource paves the way to the definition and evaluation of Neural Models for Image captioning in Italian, and we hope to contribute to the Italian Community, hopefully using the validated dataset in a future Evalita[6] champaign.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.*, 55(1):409–442, January.

Yoav Goldberg. 2015. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632.

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

[6] http://www.evalita.it/