Large scale datasets for Image and Video Captioning in Italian

Scaiella Antonio*
Università di Roma, Tor Vergata

Danilo Croce**
Università di Roma, Tor Vergata

Roberto Basili†
Università di Roma, Tor Vergata

The application of Attention-based Deep Neural architectures to the automatic captioning of images and videos is enabling the development of increasingly performing systems. Unfortunately, while image processing is language independent, this does not hold for caption generation. Training such architectures requires the availability of (possibly large-scale) language specific resources, which are not available for many languages, such as Italian.

In this paper, we present MSCOCO-it e MSR-VTT-it, two large-scale resources for image and video captioning. They have been derived by applying automatic machine translation to existing resources. Even though this approach is naive and exposed to the gathering of noisy information (depending on the quality of the automatic translator), we experimentally show that robust deep learning is enabled, rather tolerant with respect to such noise. In particular, we improve the state-of-the-art results with respect to image captioning in Italian. Moreover, in the paper we discuss the training of a system that, at the best of our knowledge, is the first video captioning system in Italian.

1. Introduction

Given the massive production of images and videos available from Social Networks and Distributed Sensors, automating the annotation, retrieval and clustering of the corresponding multimedia material is becoming crucial. Even though neural embeddings are growingly adopted to represent multimedia objects, linguistic descriptions also represent a straightforward, and more intuitive, representation of their contents. In fact, captions offer a simple way to summarize, index and search those contents implicit in such different types of data.

In this scenario, the goal of the automatic captioning of images and videos is thus to predict the correct caption(s) given an image or a video, respectively. In other words, a multimedia "captioner" is expected to automatically generate a textual description of a multimedia content, summarizing the depicted entities, the involved actions and those relations holding between them.

* Department of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: scaiellantonio@gmail.com
** Department of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: croce@info.uniroma2.it
† Department of Enterprise Engineering - Via del Politecnico 1, 00133 Rome, Italy.
E-mail: basili@info.uniroma2.it

Slightly more formally, given an image X as input, the output of an image captioner is $S(X) = \left(S_{\{1\}}, \ldots, S_{\{m\}}\right)$ such that $S(X)$ is a meaningful sentence where every $s_{\{i\}}$ is a word belonging to a vocabulary $V = \left(v_{\{1\}}, \ldots, v_{\{n\}}\right)$ of a given language. Similarly, considering the video as a sequence of images (frames), the output of a video captioner is again a meaningful sentence with the same characteristics.

Many recently proposed methods are based on deep neural networks. The results on the task show performances whose quality is sometimes comparable with humans judgments (Hossain et al. 2019). Some approaches directly operate on input multimedia sources, while in some works these are also contextualized within associated texts, i.e. (Feng and Lapata 2013; Batra, He, and Vogiatzis 2018). Most of the existing neural approaches are inspired by the architecture proposed in (Vinyals et al. 2015) where images are first encoded by a Convolutional Neural Network (which transforms them into continuous representations) and then "translated" into descriptive sentences by a recurrent architecture (for example, a Long Short-Term Memory network). At the same time, several approaches proposed for video captioning extend existing approaches for image captioning, while modeling a video as a sequence of images, such as (Venugopalan et al. 2015), or adopting more strict "Sequence to Sequence" approaches (Sutskever, Vinyals, and Le 2014; Yao et al. 2015).

In any case, the training of these neural architectures requires large scale collections of multimedia content paired with one or more captions. Important (and costly) effort led to the production of several datasets, most of which exist only for English. Examples are the MS-COCO dataset (Lin et al. 2014) for image captioning, made of 600,000 captions for about 120,000 images and the MSR-VTT dataset (Xu et al. 2016) for video captioning, made of 200,000 captions for 10,000 videos. Figure 1 reports an example from the MS-COCO dataset, i.e. an image with the corresponding five captions.

Figure 1
An example of image-caption from MSCOCO dataset

Other similar but smaller corpora exist for different languages, most of all composed of sentences that are manually annotated or translated in different languages:

IAPR-TC12, with 20,000 English, German and Spanish described images (Escalante et al. 2010), the Pascal Sentences Dataset, made of of 1,000 Japanese/English described images (Funaki and Nakayama 2015) and Multi30K (Elliott et al. 2016) made of 30,000 German/English described images.

In (Masotti, Croce, and Basili 2017, 2018), a possible alternative to the manual construction of such datasets is explored. The authors propose automatic machine translation as a way to derive annotated data through the direct translation of the original (English) material. In principle, the result is a large scale set of (images,captions) whose texts are in Italian and directly applicable to the training stage of a Neural architecture. Most noticeably, the work in (Masotti, Croce, and Basili 2018) empirically demonstrates that captions produced in Italian by neural models trained over the noisy dataset are of a better quality than the ones obtained by direct translations of English captions. In other words, the pairing of an automatic captioner with a translation system (both trained on manual annotations) is subject to a stronger performance drop than compared to the standard neural architecture trained over automatically translated input material.

In this paper, we thus propose two large scale resources to train neural architectures for image and video captioning in Italian[1]. They are derived by automatically translating the textual descriptions of images from MS-COCO and video from MSR-VTT. While the latter represents a brand new resource made of 200,000 video/caption pairs, the former dataset is generated by re-translating the original MS-COCO. Although this may seem redundant, we assume that the general improvements obtained in Automatic Translation since (Masotti, Croce, and Basili 2017), especially since the introduction of Trasformer-based architectures (Vaswani et al. 2017) will positively impact on the quality of derived neural captioners. In addition, we created a realistic test set as two sets of manually validated portions of both datasets: in fact, while model generation should be robust to noise in training material, representative performance measures strictly require validated material. We also investigate the performance of an Image Captioning model based on the Attention Mechanism (Xu et al. 2015) showing how the use of the new dataset instead of the old one, with the same model, improves the result. Finally, in parallel, we discuss the design and evaluation of the first neural system for Video Captioning in Italian, still based on Attention mechanisms.

In the rest of the paper, Section 2 introduces the resources developed in this work. In Section 3 the experimental evaluation of two neural architectures trained on these resources is discussed, while Section 4 derives the conclusions.

2. The corpus

In this section we present the two large-scale datasets for image and video captioning in Italian. These are obtained by automatic translation of the corresponding English versions[2]. It is worth noting that a subset of each corpus has been manually validated, in order to guarantee the sound evaluation of systems trained on possibly noisy annotated captions. Even though this is not the main focus of this work, this validated material also enables the evaluation of the automatic translation system, that obviously impacts

1 We publicly released both resources at the following GitHub links:
mscoco-it: https://github.com/crux82/mscoco-it
msr-vtt-it: https://github.com/crux82/msr-vtt-it

2 Captions have been translated by using Microsoft Azure Translator (https://azure.microsoft.com/it-it/services/cognitive-services/translator-text-api/) between July 2019 and August 2019.

the overall process: the higher the quality of the produced translations, the higher is the expected quality of the neural captioning system[3]. We thus compared the validated sentences with the automatic translations by using the `sacrebleu`[4] library, obtaining a BLEU score of 0.70 (Papineni et al. 2002). This score is very high, especially if compared with the traditional evaluations of modern translations systems: however, it must be said that our setting is easier if compared with standard machine translation ones. Here annotators are not asked to write the translations without knowing the output of the system, but they are asked to fix the produced translations. As a consequence, a higher number of common sub-sequences between the input sentences and the validated ones are expected, resulting in a higher BLUE score. In any case, this BLEU score suggests that the Italian material is characterized by a low level of noise due to the automatic translation process and it bodes well for the final quality of the captioning system.

2.1 Image Captioning

The image captioning task requires a large number of training examples and among existing datasets (Hossain et al. 2019), one of the largest one is MSCOCO (Lin et al. 2014). It was released in its first version in the 2014 and is composed approximately of 122,000 annotated images for training and validation, plus 40,000 more for testing. As shown in Fig. 1, each image is paired with 5 or 6 human-validated descriptions, for a total of 600k (image,caption) pairs fully available for the training and validation stages.

In particular, the original MSCOCO split consists in 82,783 captions composing the training dataset, 40,504 composing the validation set and 40,775 composing the test set. Unfortunately, captions in the test dataset are not publicly available, as they are only used in competitions. To overcome this issue, some works apply alternative splits. For example, the neural architecture proposed in (Vinyals et al. 2015) is trained on all the MSCOCO training set plus 85% of the validation set (approximately 116,000 training images, for a total of 580,000 training image-caption pair); 6,000 images from the validation set are left out and split in a development set and a test set of 2,000 and 4,000 images, respectively.

In Italian, the first version of MSCOCO-it (Masotti, Croce, and Basili 2017) follows the same specifications. A subset of captions from the original development set was manually validated (noted by v. in Table 1), thus resulting into 308 images as the development and 596 as the test set. Some (few) images were associated with captions that are only partially validated by annotators (denoted by p.). All the others, denoted by n., are left not analyzed. Overall, the statistics about the Italian dataset are shown, in terms of numbers of represented images and captions, together with the size of the resulting dataset as number of different tokens.

This work proposes a second version of MSCOCO-it where all training set plus an 85% of the validation set was fully re-translated. Here, we maintained the original validated translation, but also accomplish the validation for all the partially validated images. Validations were carried out by six annotators, not expert in Deep Learning or Natural Language Processing, but native Italian speakers.

Given the limited average length of input captions, (i.e. 10 words for caption) translations are of a good quality. For example: "*a man in shorts gets ready to hit a tennis*

3 Even though this score is measured only on the test datasets, we can speculate it reflects the quality of the translations also in the training/development subsets, since no bias is applied on this splitting.

4 `https://github.com/mjpost/sacreBLEU`

ball" is translated into "*un uomo in pantaloncini si prepara a colpire una palla da tennis*" or "*A group of three people standing on top of a snow covered slope*" into "*un gruppo di tre persone in piedi sulla cima di un pendio coperto di neve*". In some texts, word senses are mistakenly assigned, such as "*Three computer monitors sitting on top of a wooden table*" translated in "*Tre monitor per computer seduto sulla cima di un tavolo di legno*" or "*A vase of freshly cut flowers on a table*" into "*Un vaso di fiori freschi su una tabella*". In other cases, the translation is grossly incorrect, such as "*Man in body suit surfing on a large wave*" translated into "*Uomo nel vestito del corpo surf su un'onda di grandi dimensioni*" or "*a couple of kids are holding up umbrellas*" into '*Un paio di ragazzi sono holding up ombrelloni*". This is more common when jargon expressions (such as *"body suit"*) or informal expressions (e.g. not so common phrasal verbs) are employed in captions.

Table 1
Statistics for the MSCOCO-it dataset.

		#images	#captions	#words
training	n.	116,195	581,286	∼6,900,000
development	v.	308	1,516	∼18,000
	p.	(14)	25	∼300
	n.	1,696	8,486	∼102,000
test	v.	596	2,941	∼34,600
	p.	(23)	41	∼500
	n.	3,422	17,120	∼202,000

Table 2
Statistics for the second version MSCOCO-it-v2 of the MSCOCO-it dataset.

		#images	#captions	#words
training	n.	116,195	581,286	∼6,900,000
development	v.	308	1,541	∼18,000
	n.	1,696	8,486	∼102,000
test	v.	596	2,982	∼35,000
	n.	3,422	17,120	∼202,000

We are interested in evaluating the potential good impact of the novel resource in the neural training of the image captioner. The experimental results reported in the following sections will in fact connect the quality of the training material to the quality of an image captioner, which is trained over the two different dataset and compared on the same test set. Overall, it is worth noting the size of this (possible noisy) dataset made of hundred thousands of examples for a language (Italian) for which such resource has never been available.

2.2 Video Captioning

As for image captioning, several English benchmarks exist for Video Captioning. Examples of such datasets are MSVD, YouCook, M-VAD, TACoS, and MPII-MD (Aafaq et al. 2020). The first large-scale video benchmark for video understanding was MSR-VTT (Xu et al. 2016). In its current version (2017), MSR-VTT provides 10,000 web video clips

with 41.2 hours and 200,000 clip-sentence pairs. Each clip is annotated with about 20 natural sentences written by human annotators. This corpus is one of the largest open-domain video captioning datasets with a wide variety of video topics. In fact, the videos generically cover a comprehensive list of 20 categories (or topics), such as music, movie, cooking or sports. Table 3 shows the statistics of MSR-VTT dataset.

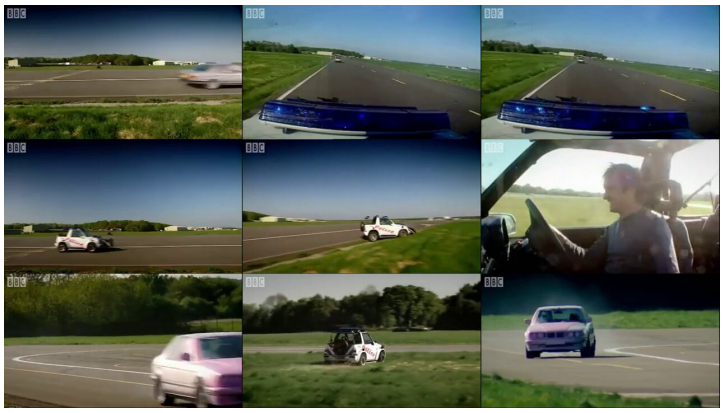

Figure 2
Example of a video included in the MSR-VTT dataset. One of the available captions for this video is *"a man is driving a small police car on a track"*.

By translating this dataset we obtained the first resource for the training of data-driven video captioning systems in Italian: MSR-VTT-it. The resource have the following video split: 6,513 video (and the corresponding captions) for training, 497 for validation and 2,990 for tests as summarized in Table 4. It is natural, like in the captioning task over MSCOCO-it, that some captions are not properly translated. The original captions of the video whose frames are shown in Figure 2 are the following:

1. *"a man is driving a small police car on a track"*
2. *"a british guy rides a police car through a grassy field"*
3. *"a man with a blue visored helmet is driving a car"*
4. *"there is a man driving a car into the grass"*
5. *"a car race is organized and displayed between three vehicles of vastly different performance"*

The translated captions are hereafter reported where wrong lexical choices or grammatical errors are underlined:

1. *"un uomo sta guidando una piccola auto della polizia su una pista"*
2. *"un ragazzo britannico cavalca una macchina della polizia attraverso un campo erboso"*
3. *"un uomo con un casco blu con visiera sta guidando una macchina"*
4. *"c'è un uomo alla guida di una macchina in erba"*
5. *"una gara automobilistica è organizzata ed esposta tra tre veicoli di prestazioni molto diverse"*

As for image captioning data, we developed a manually validated testset from a randomly selected set of 100 test videos, made thus of 2,000 validated images-caption pairs, reported in Table 4.

Table 3
MSR-VTT general statistics

#Video	7,180
#Clip	10,000
#Sentence	200,000
#Word	∼1,850,000
Vocabulary	29,316
Duration(hr)	41.1

The validation of the test set was carried out by six annotators which were asked only to check and correct the translations after watching the original video. Annotators are not expert in the field of Deep Learning or Natural Language Processing, but are Italian native speakers.

3. First evaluation

In this section we report the experimental evaluation of two different captioning systems enabled by the two resources presented in this work. In both cases we adopted an open source implementation of a deep architecture for image and video captioning, with the aim of maximizing the reproducibility of the obtained results. Systems were trained on a NVIDIA Tesla T4 GPU and evaluated using traditional metrics, i.e., BLEU (Papineni et al. 2002), METEOR (Lavie and Agarwal 2007), ROUGE (Lin 2004), Cider (Vedantam, Zitnick, and Parikh 2015).

Table 4
MSR-VTT-it statistics as the numbers of available video and Italian captions.

		#videos	#captions
training	n.	6,513	130,260
development	n.	497	9,940
test	v.	100	2,000
	n.	2,990	59,800

3.1 Image Captioning

The results reported in (Masotti, Croce, and Basili 2017) were obtained by adopting the architecture presented in (Vinyals et al. 2015) trained over a subset of the 20% of the training material. In this work we improved that evaluation in two directions. First, we trained a different architecture, based on the approach presented in (Xu et al. 2015), which exploits Attention Mechanisms[5]: these are in fact demonstrated to improve

5 We used the architecture implemented using PyTorch and available at the following link: `http://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning`

the quality of the generated captions since the architecture focuses on specific areas of the input images when generating each word. Second, the adoption of GPU-based hardware allowed to scale to the size of the entire training set.

Table 5
Italian Image captioning results

Model	Bleu_4	Cider	Rouge_L
(Masotti, Croce, and Basili 2018)	0.26	0.79	/
This work (old data)	0.28	0.93	0.48
This work (new data)	0.29	0.96	0.48

In a nutshell, the architecture combines a CNN, based on ResNet (He et al. 2016) which encodes the images into low-dimensional embeddings. Then a long short-term memory network produces the caption by generating one word at each time step, conditioned onto a context vector (which implements the Attention), the previous hidden state and the previously generated word. the context vector allows the LSTM, at each time step, to focus more carefully on some portions of the image rather than on all visual aspects by fostering a more modular learning of visual and lexical correlations.

To select the best network parameters, a validation was carried out over the development set by selecting those configuration achieving on average the best score on all the metrics we considered. The learning rate was set at standard $4e^{-4}$ with an initial random initialization of network weights and we used a batch size of 32 image-caption pairs. The dimension of word embeddings, attention linear layers and decoder RNN have been all set to 512. To avoid network overfitting a dropout at 0.5 was applied. In addition to dropout, the only other regularization strategy we used was early stopping on BLEU score. Since 20th epoch onwards we used "Fine Tuning" of the ResNet based encoder to evaluate possible improvements in the captions generation.

Results are reported in Table 5. In the first row, the results from (Masotti, Croce, and Basili 2018) are reported. In the second row, we report the results of our architecture trained on the same dataset from (Masotti, Croce, and Basili 2018), but considering all available training captions. Then, we evaluated the same architecture on the new dataset. Results confirm the beneficial impact of the new architecture trained over the entire dataset, with a significant improvement especially in term of Cider. Most importantly the beneficial impact of the new available dataset is confirmed by the improved results in terms of BLUE4 and Cider. Figure 3 shows an image which the system associated to the caption "*Un uomo in sella ad una moto su una strada sterrata*" (in English, "*A man riding a motorcycle on a dirt road*"). Moreover, in the same figure, the different areas where the network focused when generating each word are shown.

3.2 Video Captioning

In this evaluation we adopted the model presented in (Laokulrat et al. 2016), which also exploits Attention Mechanisms[6]. This architecture extends the one adopted for Image Captioning used in the previous evaluations. Since a video can be considered as a sequence of images (i.e., the frames), this approach essentially implements a sequence-

6 We used the architecture implemented using PyTorch and available at the following link: `http://github.com/xiadingZ/video-caption.pytorch`

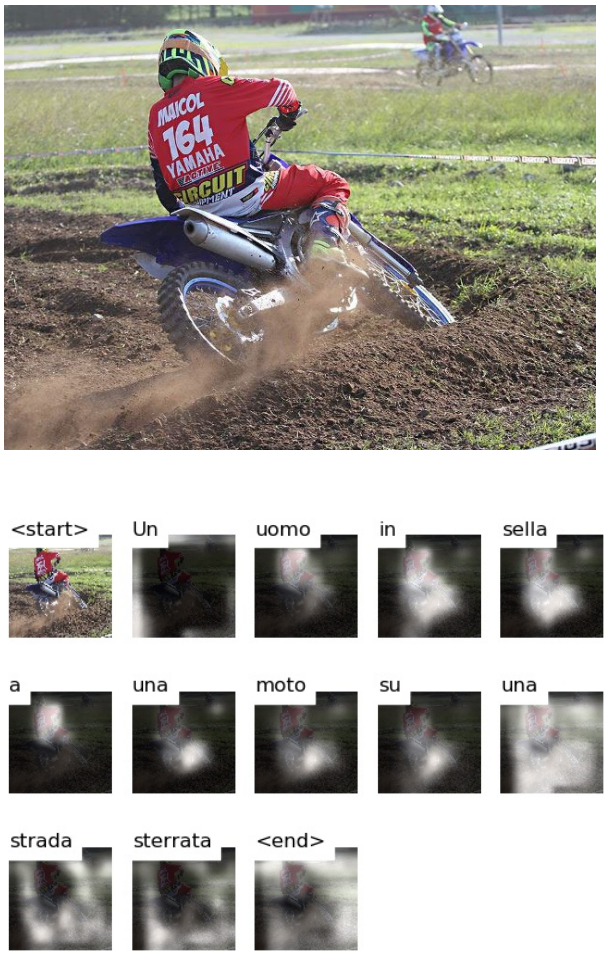

Figure 3
Example of caption generated by the model trained on MSCOCO-it-v2

to-sequence model. In the first encoding stage, a sample of k images are first extracted from the video and encoded by a Convolutional Neural Network (again the ResNet implementation, (He et al. 2016)) to be used in input to a recurrent neural network (again a LSTM network).

Then, in the decoding phase, the LSTM generates word by word the caption by taking into consideration at each time step the hidden state of the network, the hidden state of the previous time step, the word generated at the previous time step and a context vector used to represent a sort of temporal attention (Bahdanau, Cho, and Bengio 2015). This last element allows to weight the contribution of each image in input to focus on those one more important to generate the caption. In the evaluations, the learning rate was set at standard $4e^{-4}$ with an initial random initialization of network weights. The dimension of the word embeddings, the attention linear layers and the decoder RNN has been set at 512 and a recurrent dropout (set to 0.5) is used. The dimension of features encoding the frames is set to 2048 and a a batch size of 128 video-

caption pairs is used. A sample of $k = 30$ was imposed. The network parameters were chosen by selecting those maximizing on average the various measures (BLEU4, CIDER, METEOR and ROUGE-L) on the validation set.

As far as the video captioning task is concerned, we have no reference being the first experimental work done. So we will limit ourselves to make comparisons between the trained network with and without Attention mechanism (Bahdanau, Cho, and Bengio 2015) and in particular focusing on the Attention mechanism. Moreover, we focused on the reliability of the results that can be obtained by the available test material, which is also partially validated.

Table 6
Performances on the Italian Video captioning task.

Name	Test	Bleu_4	Cider	Meteor	Rouge_L
No Attention	v.	0.28	0.34	0.24	0.51
	n.	0.33	0.31	0.25	0.52
With Attention	v.	0.36	0.39	0.26	0.54
	n.	0.35	0.32	0.25	0.54

Figure 4
Example of caption generated by the model trained on MSR-VTT-it. (In English: A man is cooking food in a frying pan)

In table 6 the results of two different systems against two test sets are reported. Results obtained over the validated portion of the test set are denoted by (v.): not validated material is reported in rows (n.). The outcomes confirm the beneficial impact of temporal attention, reported in (Laokulrat et al. 2016): from the first two rows (where the context vector was neglected) to the last two rows, a systematic improvement across

different metrics is reported. An example of captioning obtained by using attention is shown in Figure 3.2. This sounds much interesting as the generalization capability of networks trained on noisy linguistic input is remarkable. Overall, we confirmed the beneficial impact of these resources that, although noisy, trigger the training of large scale networks for Italian, with results comparable with the systems existing for other resource rich languages.

4. Conclusions

In this paper, we proposed two new large scale corpora for Image and Video captioning aimed at enabling the training of effective neural architectures for the Italian language. The work improves the performance on the Image Captioning task for the Italian language and, at the same time, lay the ground-work for future work on Video Captioning in Italian. This last task remains much more difficult than the previous one given the need to capture many more features in the frame sequence than are simply absent over individual images. With our experiments, using models that are not too complex, we hope to support the advancement of the state of the art for the Image and Video caption tasks in Italian. The availability of the two corpora as publicly available resources is expected to trigger more research work on the improvement of the corpus quality as well as on the development of newer neural models through possible language specific architecture.

Acknowledgments

We would like to thank Carlo Gaibisso, Bruno Luigi Martino and Francis Farrelly of the Istituto di Analisi dei Sistemi ed Informatica “Antonio Ruberti" (IASI) for supporting the experimentations through access to dedicated computing resources made available by the Artificial Intelligence & High Performance Computing laboratory.

References

Aafaq, Nayyer, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2020. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52(6):115:1–115:37.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, (ICLR 2015)*, San Diego, CA, USA, May.

Batra, Vishwash, Yulan He, and George Vogiatzis. 2018. Neural caption generation for news images. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Elliott, Desmond, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August. ACL.

Escalante, Hugo Jair, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes-y-Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Feng, Yansong and Mirella Lapata. 2013. Automatic caption generation for news images. *Transactions on Pattern Analysis and Machine Intelligence.*, 35(4):797–812, April.

Funaki, Ruka and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590, Lisbon, Portugal, September. ACL.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, Las Vegas, NV, USA, June. IEEE Computer Society.

Hossain, MD. Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6), February.

Laokulrat, Natsuda, Sang Phan, Noriki Nishida, Raphael Shu, Yo Ehara, Naoaki Okazaki, Yusuke Miyao, and Hideki Nakayama. 2016. Generating video description using sequence-to-sequence model with temporal attention. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers*, pages 44–52, Osaka, Japan, December.

Lavie, Alon and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Prague, Czech Republic, June. ACL.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Masotti, Caterina, Danilo Croce, and Roberto Basili. 2017. Deep learning for automatic image captioning in poor training conditions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, Italy, December.

Masotti, Caterina, Danilo Croce, and Roberto Basili. 2018. Deep learning for automatic image captioning in poor training conditions. *Italian Journal of Computational Linguistics*, 4(1):43–56.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania, July. ACL.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems*, pages 3104–3112, Montreal, Quebec, Canada, December.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, Long Beach, CA, USA, December.

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4566–4575, Boston, MA, USA, June. IEEE Computer Society.

Venugopalan, Subhashini, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado, May–June. ACL.

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3156–3164, Boston, MA, USA, June. IEEE Computer Society.

Xu, Jun, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 5288–5296, Las Vegas, NV, USA, June. IEEE Computer Society.

Xu, Kelvin, 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. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 2048–2057, Lille, France, July.

Yao, Li, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Describing videos by exploiting temporal structure. In *2015 International Conference on Computer Vision, ICCV 2015*, pages 4507–4515, Santiago, Chile, December. IEEE Computer Society.