Improving German Image Captions using Machine Translation and Transfer Learning

Rajarshi Biswas[1], Michael Barz[1,2], Mareike Hartmann[1], and Daniel Sonntag[1,2]

[1] German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus D3_2, 66123 Saarbruecken, Germany
[2] Applied Artificial Intelligence, Oldenburg University, Marie-Curie Str. 1, 26129 Oldenburg, Germany
{firstname.lastname}@dfki.de

Abstract. Image captioning is a complex artificial intelligence task that involves many fundamental questions of data representation, learning, and natural language processing. In addition, most of the work in this domain addresses the English language because of the high availability of annotated training data compared to other languages. Therefore, we investigate methods for image captioning in German that transfer knowledge from English training data. We explore four different methods for generating image captions in German, two baseline methods and two more advanced ones based on transfer learning. The baseline methods are based on a state-of-the-art model which we train using a translated version of the English MS COCO dataset and the smaller German Multi30K dataset, respectively. Both advanced methods are pre-trained using the translated MS COCO dataset and fine-tuned for German on the Multi30K dataset. One of these methods uses an alternative attention mechanism from the literature that showed a good performance in English image captioning. We compare the performance of all methods for the Multi30K test set in German using common automatic evaluation metrics. We show that our advanced method with the alternative attention mechanism presents a new baseline for German BLEU, ROUGE, CIDEr, and SPICE scores, and achieves a relative improvement of 21.2 % in BLEU-4 score compared to the current state-of-the-art in German image captioning.

Keywords: Natural language understanding and generation · Multimodal technologies · Image Captioning · Natural Language Processing

Image captioning, i.e., the task of automatically describing an image, is an interesting problem of artificial intelligence research. It is multimodal in nature and lies at the intersection of computer vision and natural language processing. The problem has witnessed rapid progress in the last few years owing to the development of novel deep neural architectures, training procedures, rapid advancement in GPU computing power and lastly the availability of large annotated datasets. However, the vast majority of research in this domain concentrates on the English language. The primary reason for this development is the

high availability of annotated image captioning datasets in English compared to other languages. For instance, the English MS COCO dataset [19] contains 164,063 images each with 5 accompanying captions totaling to 820,315 captions. In comparison, the Multi30K [9] dataset, which includes German captions sourced from native speakers, contains 31,014 images with 155,070 accompanying German captions. This is almost one order less in size than the MS COCO dataset. This sparsity of resources is a major obstacle in developing effective neural models for caption generation in German or other non English languages. As a result, there is a gap in research on image caption generation in German. It is studied mostly as a sub-task of multimodal machine translation where the image provides additional information for the translation task. Elliott et al. [9] introduced the first dedicated German image captioning dataset, Multi30K, sourced from native German speakers. Jaffe [15] studied the problem of generating image descriptions in German. For this purpose, they explored different model architectures which use a training corpus containing captions in both English and German. They generate captions for both languages, but discard the English output. Their best approach, which is based on an attention pipeline with random embeddings, is the current state-of-the-art in producing German image captions.

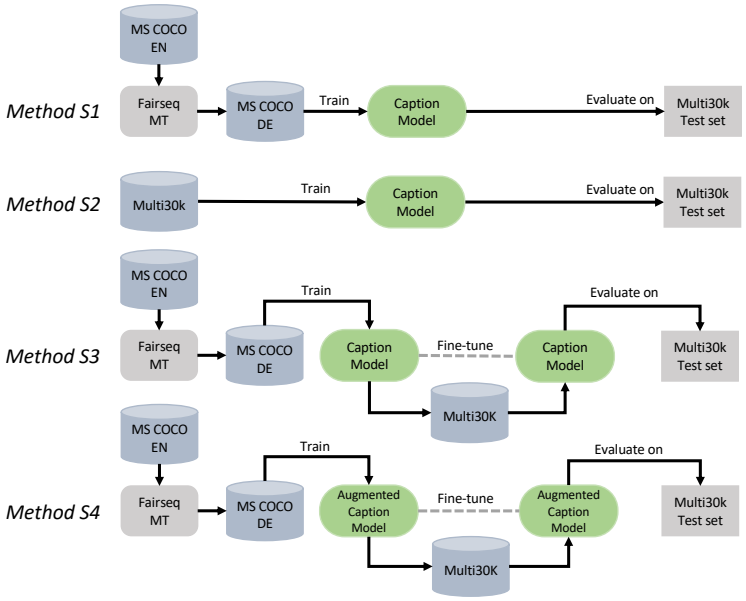


Fig. 1. The methods for German image caption generation that we compare in this work.

In this work, we aim at improving caption generation in German by utilizing the large-scale MS COCO dataset in English. This is different from Jaffe [15] who used the Multi30K dataset for model training only. We transfer the English resources by translating all captions to German using the state-of-the-art neural machine translator Fairseq [21]. This way, we distantly leverage the higher availability of resources in the English-to-German translation domain. In total, we compare four different methods for generating German image captions on the Multi30K test set (see Figure 1). We include two baseline methods and two more advanced methods based on fine-tuning. All methods are based on an adapted version of the encoder-decoder based neural architecture described in [3, 5]. The baseline models are trained on the translated MS COCO dataset (*S1*) and the train set of the Multi30K dataset (*S2)*, respectively. For both advanced methods, we pre-train the model on the translated MS COCO dataset and fine-tune it using the train set of the Multi30K dataset. We use the same model as for the baseline methods (*S3*) and a model with an alternative attention mechanism as described in [4] (*S4*). We hypothesize that both fine-tuning methods perform better than the baseline methods in terms of common evaluation metrics. Also, we expect that the alternative attention mechanism *S4* further improves the image caption quality and beats the current state-of-the-art by Jaffe [15].

The rest of the paper is structured as follows: We discuss the progress in multilingual caption generation in section 1. Then we discuss the technical details of our approach in section 2 followed by a detailed report of our evaluation and its results in section 3. We discuss the results in section 4 and conclude our paper in section 5.

1 Related Work

Approaches for multilingual image captioning can be divided into two broad categories: translation-based approaches and alignment-based approaches. Translation-based approaches rely on machine translation models to either translate generated captions to the target language or to create an image captioning dataset in the target language for training language-specific models. Elliott et al. [7] are one of the first to study the task of multilingual image caption generation. They use features from both source and target language model and generate the captions using an LSTM [13] based decoder. Hitschler et al. [12] translate image captions from one language to the other using the image as additional input. This image guided translation is the focus of the WMT 2016 multimodal machine translation task [23] and the WMT 2017 task [8] with some variations, such as, unavailability of the source language at test time. These WMT tasks on multimodal machine translation find that purely text-based machine translation techniques provide a strong baseline when translating captions from one language to another. Additionally, they found that supplementing machine translation techniques with information from the image only results in a marginal improvement. For example, the work in Huang et al. [14] re-ranked the translation output using image features, but could not improve the METEOR score compared to their base-

line. This trend is also observed for the task of generating cross-lingual image descriptions in WMT 2016. In spite of using attention based models, the image does not provide much benefit towards generating captions in German and all the highest scoring systems in WMT 2016 for the cross-lingual image description multimodal task ignored the image. In the WMT 2017 task this observation is repeated, that is, text-only systems perform better and obtain higher scores compared to multimodal systems that use images as context. Jaffe [15] generate image captions in German as part of the WMT 2017 multimodal translation sub task on multilingual image caption generation. They use the Multi30K dataset for this purpose and explore different neural architectures. They use the images with both English and German captions for training their models. In fact, they generate captions in both languages, but discard the English output during evaluation. Also, they experiment with textual attention for caption generation. Their architecture that uses attention over the German caption output achieves the highest scores in terms of the BLEU-4 and METEOR metrics.

Alignment-based approaches rely on a joint embedding space. These methods aim to first learn an alignment between the given image and corresponding English captions in a common latent space. This alignment is then used to relate to the target language. They assume better alignment leads to better captions generated in the target language. Through this process they try to make up for the lack of annotated training data in the target language. For instance, Miyazaki et al. [20] pre-train a captioning model on the MS COCO dataset. Later they modify this model and train the modified model on Japanese data for generating captions. Wu et al. [26] combine merits from both, alignment-based and translation-based approaches, for multilingual image captioning in a unified architecture. In their work, given an input image, they generate English captions and, then, the caption in the target language. Similarly, Thapliyal et al. [24] propose a system that uses existing English annotations and their translations at training time. At run time their system generates an English caption and then a corresponding caption in the target language. Lan et al. [17] propose a fluency guided framework where they aim to learn a cross-lingual captioning model from machine translated sentences. Their proposed framework automatically estimates the fluency of the sentences and uses the estimated fluency scores as part of the cost function to train an image captioning model for the target language. The work of Gu et al. [10] first uses a pivot language for capturing the characteristics of the image captioner and then uses a pivot-target language parallel corpus to align the image captioner to the target language.

Our advanced methods can be classified as translation-based, because we use the Fairseq neural machine translator to translate the MS COCO dataset into German. However, our approach differs from previous works on translation based methods in two key aspects. First, we use a fine-tuning process where we pre-train our captioning model on the translated dataset and subsequently fine tune the model on the German captioning dataset, Multi30K, sourced from native speakers of the language. We assume this process can help in learning language specific nuances through this process. And, it is much simpler compared

to approaches using a pivot language. Second, we apply a modified attention scheme [4] that has been shown to improve caption generation for English.

2 Method

We implement four methods for generating image captions in German based on the neural image captioning model presented in [5, 4]. We include two baseline models and two advanced models based on fine-tuning (see Figure 1). The baseline models use the MS COCO dataset translated into German and the Multi30K dataset respectively for training the captioning model. For the advanced models, we pre-train the caption model using the translated MS COCO dataset and then fine-tune it on the Multi30K dataset. Also, the baseline methods *S1, S2* and the method *S3* use the caption model from [5, 27]. In contrast, in the advanced method *S4* we fine-tune the image captioning model with the more effective augmented attention mechanism proposed in [4].

2.1 Image Captioning Datasets

We translate the original MS COCO dataset [19] from English into German using the Fairseq neural machine translator. The translated MS COCO dataset contains 82,783 images with 5 corresponding captions in the training set while the validation set contain 5,000 images each with 5 groundtruth captions per image. We refer to this data split as the *COCO_Split*. We also use the Multi30K German image captioning dataset for training the captioning models in our methods. For Multi30K, the training set contains 29,000, the validation set contains 1014 and the test set contains 1000 images respectively with 5 corresponding captions per image. We denote this break up as the *M30k_Split*.

2.2 Image Captioning Model

For all methods, we use the neural encoder-decoder model with visual attention mechanism adapted from [3, 5, 27]. The image encoder part of this model is based on the ResNet-101 model with 101 layers [11]. We do not perform any pre-processing on the images. We apply spatially adaptive max-pooling which results in a fixed size output of $14 \times 14 \times 2048$ for each image. Thus, each image is encoded as 196 vectors with a dimension of 2048. The decoder in our caption generation model is an LSTM [13], and we build our vocabulary by dropping word types with a frequency < 5. We set the dimensions for the LSTM hidden state, image, word and attention embeddings to 512 and train the model under the cross entropy objective, using the ADAM [16] optimizer. All models are trained for 30 epochs, followed by 30 epochs of fine-tuning for methods *S3* and *S4*.

2.3 Caption Generation Methods

To build our baseline method *S1*, we train the caption generation model as described above using the translated MS COCO dataset. Our baseline method *S2* is trained using the *M30k_ Split*. For *S3*, we pre-train the model on the translated German MS COCO dataset using *COCO_ Split*. This allows the model to learn the initial mapping from images to the German language from the translated corpus. Subsequently, we fine-tune this model on the Multi30K dataset, sourced from native speakers, using the *M30k_ Split*. For *S4*, we use the attention mechanism presented in [4], which has been shown to improve English captioning systems. This mechanism incorporates object-specific localized maps from a region proposal network for this purpose. Specifically, we represent an input image I as a set of feature vectors, $I = \{f_1, f_2, ..., f_n\}$ where $f_i \in \mathbb{R}^d$. Each element in this set represents the encoding of a bounding box detected by a region proposal network that is encoded using the ResNet-101 model [11]. We extract the image regions inside the final bounding boxes obtained after non-maxima suppression and embed them into the feature space learned by ResNet-101 pre-trained on the ImageNet [6] dataset. We set a high threshold (0.8) for the classification probability for the regions to be selected. Subsequently, we compute visual attention on the joint embedding space formed by the union of high-level features obtained from the encoder of the caption generator and the low-level features obtained from the object specific local regions of the input image. We use 10 additional feature vectors for every image to represent the local regions. So, our attention mechanism at every time-step produces a mask over 206 spatial locations. This mask is applied to a set of image features and the result is spatially averaged to produce a 2048 dimensional representation of the attended portion of the image. We pre-train this caption model with the augmented attention mechanism first on the translated MS COCO dataset using *COCO_ Split* and then fine-tune the trained model on the Multi30K dataset with *M30k_ Split* for learning language specific nuances. For a quick reference all considered methods are listed below.

1. (*S1*) we train the caption generation model using only the translated MS COCO dataset.
2. (*S2*) we train the caption generation model using only the Multi30K dataset.
3. (*S3*) we train the image caption generation model on the translated MS COCO dataset and then fine tune the model on the Multi30K dataset.
4. (*S4*) we train the image caption model with augmented attention on the translated MS COCO dataset and then fine tune the model on the Multi30K dataset.

3 Evaluation

We test all methods, explained above, using the Multi30K test set and compare the generated captions using automated metrics commonly used in the image captioning research community (see Section 3.1). Our goal is to ascertain the most effective method for generating image captions in German. In this regard,

we investigate the effect of pre-training a model on the translated MS COCO dataset and the impact of using the alternative attention mechanism on the quality of generated image captions. Also, we compare the scores of all four methods to the results reported in [15] as they achieved the highest metric scores for caption generation in German.

3.1 Metrics

We compute a group of automated metrics commonly used in the image captioning research community: BLEU [22], METEOR [2], ROUGE [18], CIDEr [25] and SPICE [1]. These metrics primarily focus on the n-gram overlap between the generated and ground-truth captions. For convenience, we provide a short description for each metric. BLEU scores are computed by directly matching n-grams between individual machine generations and a corresponding set of ground-truth references. It is always between 0 and 1 where 0 indicates no overlap and 1 indicates a perfect overlap. Depending on the size of the n-grams you get different BLEU scores, i.e, BLEU-1, BLEU-2, BLEU-3, BLEU-4. METEOR evaluates outputs from a machine translation system. It computes the harmonic mean of unigram precision and recall. Recall is weighted higher than precision. ROUGE measures the longest matching sequence of words. An advantage of it is that it does not require consecutive matches but in-sequence matches that reflects sentence level order. CIDEr measures the similarity of a generated sentence against a set of ground truth sentences composed by humans and shows high agreement with consensus as assessed by humans. SPICE denotes semantic propositional image caption evaluation. It uses semantic information in the form of a scene graph to measure the similarity between the ground-truth and machine generated captions.

3.2 Hypothesis

We hypothesize that the advanced methods, *S3* and *S4*, yield a better performance than the baseline methods *S1, S2* that do not use fine-tuning in terms of the metrics mentioned above. We expect that the method (*S4*) generates the best German captions compared to our other methods, but also to the current state-of-the-art performance in German image captioning as reported in [15] which provides baseline scores for BLEU-4 and METEOR.

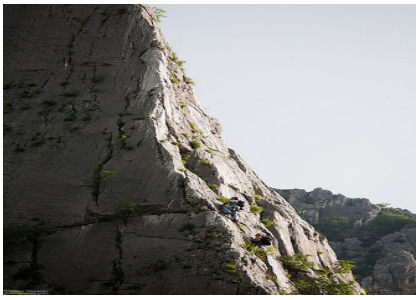
3.3 Results

The scores for all German image caption generation methods are summarized in Table 1. The scores are computed using the standard metric computation package which ensures comparability with Jaffe [15]. Among our methods, *S4* yields the best scores: it achieves higher scores for BLEU-1,2,3,4, ROUGE, CIDEr, and SPICE metrics. Only the METEOR score obtained with *S4* is lower by a small margin of 0.006 compared to *S3* and by 0.003 than *S2*. We observe that

(a) **S4:** ein mann mit hut und sonnenbrille sitzt auf einem felsen und schaut auf das wasser (*a man in a hat and sunglasses is sitting on a rock and looking at the water*); **S3:** ein mann mit mütze sitzt auf einem felsen und schaut auf sein handy (*a man in a hat is sitting on a rock and looking at his mobile phone*); **S2:** ein mann mit hut sitzt auf einem skateboard (*a man in a hat is sitting on a skateboard*); **S1:** ein mann sitzt auf einem UNK (*a man is sitting on a UNK*)

(b) **S4:** eine gruppe von menschen sitzt an einem tisch mit essen (*a group of people is sitting at a table with food*); **S3:** eine frau und ein mann sitzen an einem tisch und essen kuchen (*a woman and a man are sitting at a table and eating cake*); **S2:** eine gruppe von menschen sitzt an einem tisch mit einem tisch (*a group of people is sitting at a table with a table*); **S1:** zwei frauen sitzen an einem tisch und essen (*two women are sitting at a table and eating*)

(c) **S4:** ein mann sitzt an einem tisch und schreibt etwas auf ein papier (*a man is sitting at a table and writing something on a paper*); **S3:** ein mann sitzt an einem tisch und schreibt in ein heft (*a man is sitting at a table and writing something in a notebook*); **S2:** ein mann sitzt an einem tisch mit einem laptop (*a man is sitting at a table with a laptop*); **S1:** zwei männer sitzen an einem tisch und spielen (*two men are sitting at a table and playing a game*)



(d) **S4:** ein mann klettert an einem seil gesichert eine felswand hinauf (*a man is climbing up a rock face secured by a rope*); **S3:** ein mann klettert an einem seil gesichert an einem seil (*a man is climbing on a rope secured by a rope*); **S2:** ein mann klettert an einem felsen (*a man is climbing up a rock*); **S1:** ein mann fährt auf einem UNK durch eine UNK (*a man is driving on a UNK through a UNK*)

Fig. 2. Example German image captions generated with the methods explored in our work. Italics in brackets provide English translations of the generated captions.

all metric scores, apart from METEOR, gradually increase from *S1* to *S4*. This trend also extends to the method (*S3*) and the method (*S4*). Also, the BLEU-4 score of *S4* is better than the corresponding score reported in the current state-of-the-art approach by Jaffe [15] by an absolute margin of 0.025 that is a relative improvement of 21.19 percent. However, our METEOR score is lower by 0.048. We use the same technique as Jaffe to compute the metrics and believe this inconsistency could be due to the low correlation between BLEU and METEOR as observed by Jaffe [15]. Unfortunately, the authors did not report other metrics for image captioning like CIDEr and SPICE for which our approach *S4* obtains highest scores among our methods.

Table 1. Performance scores of different methods used for generating German image captions on the Multi30K test set.

Strategy	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE	CIDEr	SPICE
Jaffe [15]	–	–	–	0.118	**0.205**	–	–	–
S4	**0.527**	**0.352**	**0.227**	**0.143**	0.157	**0.369**	**0.307**	**0.035**
S3	0.508	0.317	0.191	0.107	0.163	0.358	0.250	0.029
S2	0.482	0.297	0.178	0.101	0.160	0.351	0.227	0.027
S1	0.456	0.270	0.151	0.081	0.151	0.326	0.177	0.023

4 Discussion

The results obtained in our work (see Table 1) show that the method (*S4*) with the alternative attention mechanism results in higher BLEU-4 score compared to the value reported in the state-of-the-art approach [15] for German image captioning, indicating that our hypothesis could be confirmed in terms of the BLEU-4 metric. The comparison of *S4* with *S3, S2, S1* establishes the merit in using the augmented attention mechanism. This is also observed in the examples shown in Figure 2 which shows that the captions generated using *S4* are comparatively better than the other methods. Also, the captions generated using *S3, S2, S1* do not capture the relevant details in the image compared to *S4*. Our results also show the benefit of pre-training the caption generation model on the translated MS COCO dataset followed by fine-tuning it on the smaller Multi30K dataset. Importantly, there is a consistent gradual increase in the BLEU, ROUGE, CIDEr, and SPICE scores as we transition from *S1* to *S4*. A comparison of the scores from *S1* and *S2* shows that the Multi30K training data, sourced from native German speakers, is more influential to the caption generation model compared to the machine translated German MS COCO dataset in our test setting. Using *S3*, we show that pre-training followed by fine tuning could be one of the possible ways to overcome the requirement of large amount of annotated data for training an image captioning model in German as it achieves better scores compared to both *S1* and *S2* across all the metrics. Finally, we

show that training the caption generation model with the augmented attention mechanism using fine-tuning in *S4* results in highest improvement relative to all the strategies we used in our work. This is evidenced through higher BLEU-1,2,3,4, ROUGE, CIDEr and SPICE scores compared to those obtained by *S1, S2, S3*. Moreover, *S4* even obtains higher BLEU-4 score compared to the current state-of-the-art in German image captioning.

5 Conclusion

In this work, we implemented and evaluated four methods for caption generation in German with the goal of achieving state-of-the-art performance. We showed that our methods could serve as possible ways of overcoming the problem of sparse availability of training data for image captioning in the German language. Our best performing method uses an alternative attention mechanism from the literature [4] and leverages the vast resources available in English, i.e., the MS COCO dataset for cross-lingual information transfer in the context of image captioning via the Fairseq neural machine translator. The model is pre-trained on the translated MS COCO dataset and fine-tuned on the German Multi30K dataset sourced from native speakers. This model achieves the best BLEU-1,2,3,4, ROUGE, CIDEr, and SPICE scores compared to our baseline methods. Moreover, the model with alternative attention mechanism obtained a higher BLEU-4 score than the state-of-the-art approach by Jaffe [15] by an absolute margin of 0.025 that is a relative improvement of 21.19 percent.

Acknowledgments

This work was funded by the German Federal Ministry of Research (BMBF) under grant number 01IW20005 (XAINES).

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. ECCV (2016)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization pp. 65 – 72 (2005)
3. Biswas, R.: Diverse Image Caption Generation And Automated Human Judgement through Active Learning. Master's thesis, Saarland University (2019)
4. Biswas, R., Barz, M., Sonntag, D.: Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. KI - Künstliche Intelligenz, German Journal on Artificial Intelligence - Organ des Fachbereiches "Künstliche Intelligenz" der Gesellschaft für Informatik e.V. (KI) **34**(4), 571–584 (Jul 2020). https://doi.org/10.1007/s13218-020-00679-2, https://doi.org/10.1007/s13218-020-00679-2

5. Biswas, R., Mogadala, A., Barz, M., Sonntag, D., Klakow, D.: Automatic Judgement of Neural Network-Generated Image Captions, vol. 11816, pp. 261–272. Springer (09 2019). https://doi.org/10.1007/978-3-030-31372-2_22
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Elliott, D., Frank, S., Hasler, E.: Multilingual image description with neural sequence models. arXiv: Computation and Language (2015)
8. Elliott, D., Frank, S., Barrault, L., Bougares, F., Specia, L.: Findings of the second shared task on multimodal machine translation and multilingual image description. In: Proceedings of the Second Conference on Machine Translation. pp. 215–233. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4718, https://www.aclweb.org/anthology/W17-4718
9. Elliott, D., Frank, S., Sima'an, K., Specia, L.: Multi30K: Multilingual English-German image descriptions. In: Proceedings of the 5th Workshop on Vision and Language. pp. 70–74. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-3210, https://www.aclweb.org/anthology/W16-3210
10. Gu, J., Joty, S., Cai, J., Wang, G.: Unpaired Image Captioning by Language Pivoting: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I, pp. 519–535. Springer International Publishing (09 2018). https://doi.org/10.1007/978-3-030-01246-5_31
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
12. Hitschler, J., Schamoni, S., Riezler, S.: Multimodal pivots for image caption translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2399–2409. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1227, https://www.aclweb.org/anthology/P16-1227
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
14. Huang, P.Y., Liu, F., Shiang, S.R., Oh, J., Dyer, C.: Attention-based multimodal neural machine translation. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp. 639–645. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2360, https://www.aclweb.org/anthology/W16-2360
15. Jaffe, A.: Generating image descriptions using multilingual data. In: Proceedings of the Second Conference on Machine Translation. pp. 458–464. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/W17-4750, https://www.aclweb.org/anthology/W17-4750
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980

17. Lan, W., Li, X., Dong, J.: Fluency-guided cross-lingual image captioning. In: Proceedings of the 25th ACM International Conference on Multimedia. p. 1549–1557. MM '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3123266.3123366, https://doi.org/10.1145/3123266.3123366
18. Lin, C.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
20. Miyazaki, T., Shimizu, N.: Cross-lingual image caption generation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1780–1790. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1168, https://www.aclweb.org/anthology/P16-1168
21. Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M.: fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations). pp. 48–53. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-4009, https://www.aclweb.org/anthology/N19-4009
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. Association for Computational Linguistics pp. 311 – 318 (2002)
23. Specia, L., Frank, S., Sima'an, K., Elliott, D.: A shared task on multimodal machine translation and crosslingual image description. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. pp. 543–553. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/W16-2346, https://www.aclweb.org/anthology/W16-2346
24. Thapliyal, A.V., Soricut, R.: Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 160–170. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.16, https://www.aclweb.org/anthology/2020.acl-main.16
25. Vedantam, R., Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. Computer Vision and Pattern Recognition pp. 4566 – 4575 (2015)
26. Wu, Y., Zhao, S., Chen, J., Zhang, Y., Yuan, X., Su, Z.: Improving captioning for low-resource languages by cycle consistency. In: 2019 IEEE International Conference on Multimedia and Expo (ICME). pp. 362–367 (2019). https://doi.org/10.1109/ICME.2019.00070
27. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2048–2057. PMLR, Lille, France (07–09 Jul 2015), http://proceedings.mlr.press/v37/xuc15.html