Generating Image Descriptions using Multilingual Data

Alan Jaffe
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA
apjaffe@andrew.cmu.edu

Abstract

In this paper we explore several neural network architectures for the WMT 2017 multimodal translation sub-task on multilingual image caption generation. The goal of the task is to generate image captions in German, using a training corpus of images with captions in both English and German. We explore several models which attempt to generate captions for both languages, ignoring the English output during evaluation. We compare the results to a baseline implementation which uses only the German captions for training and show significant improvement.

1 Introduction

Neural models have shown great success on a variety of tasks, including machine translation (Sutskever et al., 2014), image caption generation (Xu et al., 2015), and language modeling (Bengio et al., 2003). Recently, huge datasets necessary for training these models have become more widely available, but there are still many limitations. In some cases, the dataset which is available may not match the domain of the task.

In this paper, we attempt to generate image captions in German, using a training corpus of images with captions in both English and German. For each image, we have 5 independently generated captions in each language. Since the training corpus is relatively small (less than 30,000 images), we want to make use of the English language data to improve the German captions. (See figure 1).

It is important to note that since these captions were generated independently in each language rather than translated, they often differ from each other quite a bit. Not only do they often choose to describe different features of an image, but also they sometimes describe contradictory features of the image (one caption describing a man sleeping on a couch while a different caption describes a woman sleeping on a couch). This inconsistency and the relatively small amount of training data makes it very difficult to train a reliable translation system between the languages based on this corpus.

In this paper, we will start by discussing related work in image caption generation. Then we will explain the baseline German image caption generation model, the soft attention model from Xu et al. (2015). Several methods of incorporating the English data to improve the performance will be described. Finally, the experimental setup will be specified and the results will be evaluated.

2 Related Work

The task of multilingual image caption generation has been previously explored by Elliott et al. (2015). Elliott et al. (2015) used an LSTM to generate captions, using features from both a source-language multimodal model and a target-language multimodal model. Other previous work on multilingual images such as Hitschler and Riezler (2016) has focused on image caption translation, where captions are available at test time in a single language, and we wish to use the image as a guide while translating into a different language. The WMT 2016 multimodal machine translation task (Specia et al., 2016) explored precisely this task. Using existing machine translation techniques to translate the given caption provided a very strong baseline. Supplementing these translation with information from the image provided only marginal improvements. For instance Huang et al. (2016) re-ranked the translation output using image features and failed to achieve a higher METEOR score than the baseline.

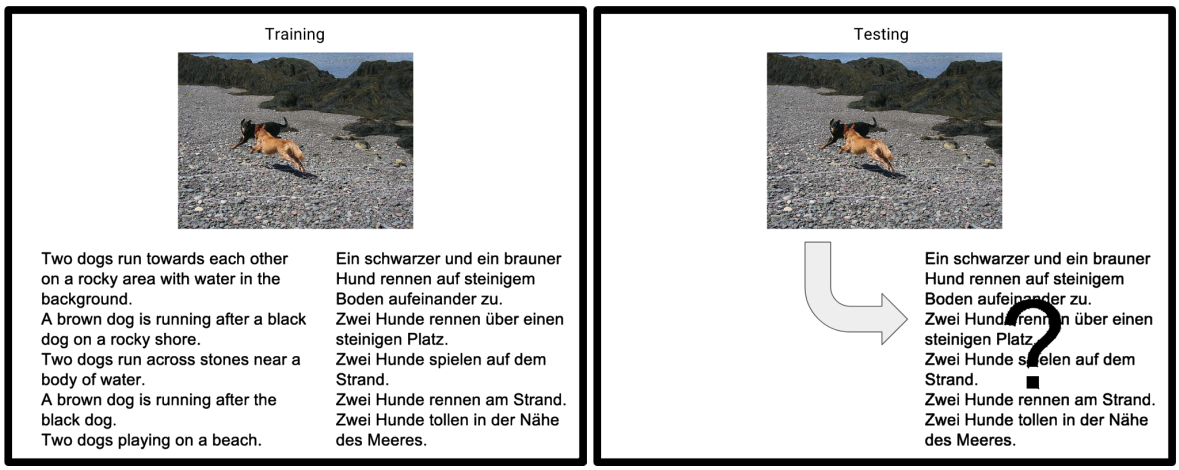

Figure 1: Training data and test data

Similarly, systems developed for the WMT 2016 crosslingual image description multimodal task had access to one or more reference English descriptions of the image (in addition to the image itself) when attempting to generate a German caption, allowing them to use attention-based models that took advantage of both pieces of information. Again though, the image seemed to provide little benefit, and in fact the highest scoring system ignored it altogether.

Generally, the long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) seems to be quite effective for caption generation and other natural language processing tasks. Dropout has also been shown to reduce overfitting (Srivastava et al., 2014).

Supplementing the basic LSTM model with attention model has been shown to be effective for related tasks as well, such as machine translation (Bahdanau et al., 2014). Multiple methods are possible for determining how the attention is allocated at each step, such as a simple dot-product, linear transformation, or multilayer perceptron. Several of these alternatives were explored by Luong et al. (2015).

Beyond multilingual caption generation, the over-arching task of image caption generation has also been considered before. Vinyals et al. (2015) used a convolutional neural network to encode an image, followed by an LSTM decoder to produce an output sequence. Xu et al. (2015) extended that model by adding an attentional component, using a multilayer perceptron to determine the weight of each part of the image given to the LSTM at each step.

With less than 30,000 images, it is difficult to train a convolutional neural network to identify image features. Caglayan et al. (2016) found that the ResNet (He et al., 2015) trained on ImageNet classification task was quite effective (specifically using layer 'res4fx' which is found at the end of Block-4, after ReLU). Note that this differs from Xu et al. (2015), which used pre-trained features from the Oxford VGGnet (Simonyan and Zisserman, 2014).

3 Image Caption Generation Models

3.1 Baseline

We developed several models, each of which generate both English and German captions. The models were trained on both the English and the German data, but at test time we evaluate the performance only for generating German captions.

Our baseline is implemented as an attentional neural network following the model of Xu et al. (2015). Each image is encoded as 196 vectors, each of which corresponds to a particular section of the image. Each of these vectors consists of 1024 real numbers, derived from layer 'res4fx' of ResNet. (Note that this modifies the original work by Xu et al. (2015), which used Oxford VGGnet with only 512 real numbers for each location in the image.) Xu et al. (2015) considered both a hard and a soft attentional model, but since these performed comparably, we have only re-implemented their soft attentional model.

We generate a caption as a series of words (encoded as 1-hot vectors), terminated by the end of sentence symbol `</s>`. At each timestep, an attention mechanism implemented as a multilayer

perceptron (MLP) predicts how important each part of the image is, based on the previous hidden state h_{t-1}. Softmax is applied over the attention outputs to compute a weighted average of the image vectors. The result is a 1024-dimensional context vector z_t that represents the important parts of the entire image at timestep t.

We use an LSTM as the decoder, which has decoupled input and forget gates and does not use peephole connections. We initialize the LSTM to 0, unlike Xu et al. (2015) which initializes the LSTM using two additional MLP's. Given some previous state (h_{t-1}, c_{t-1}) and input x_t, we compute $(h_t, c_t) = f(h_{t-1}, c_{t-1}, x_t)$ where $x_t = \textit{concat}(\textit{embed}_{t-1}, z_t)$. $\textit{embed}_{t-1}$ is the word embedding of the previous word outputted (or the special token $< s >$ at the start of the sentence), and z_t is the context vector derived from attention over the image. The resulting output h_t is then transformed to $\textit{softmax}(W_{yh}h_t + b_y)$ to compute the probability of each word in the vocabulary. Each timestep $(h_t, c_t) = f(h_{t-1}, c_{t-1}, x_t)$ is computed as follows (Neubig et al., 2017):

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f + 1) \quad (2)$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \quad (3)$$
$$u_t = \tanh(W_{ux}x_t + W_{uh}h_{t-1} + b_u) \quad (4)$$
$$c_t = c_{t-1} \circ f_t + u_t \circ i_t \quad (5)$$
$$h_t = \tanh(c_t) \circ o_t \quad (6)$$

Equation 1 is the input gate, equation 2 is the forget gate, equation 3 is the output gate, and equation 4 computes the update.

Since we re-implemented this baseline and made some changes in the process as detailed above (most notably by omitting the hard attentional model), we wanted to verify that this did not affect performance. The original paper generated English captions only, so we trained a version of our baseline model to generate English captions. Using dropout of 0.02, an English vocabulary size of 12138, and a minibatch size of 32, this achieved a BLEU score of 21.48 (lowercased, ignoring punctation).[1] That result lines up well with the BLEU score of 19.1 reported by Xu et al. (2015) on the Flickr30k dataset, so we are confident that our reimplementation has not weakened the baseline.

[1]Dropout of 0.2 was also tested, with slightly worse results (BLEU = 20.66).

3.2 Shared Decoder

The first model tested was the shared decoder model. This is a multitask architecture, with one loss for each language. The idea of this model was to consider English and German as two separate vocabularies, thus each with their own set of word embeddings and word output weights W_{yh}, b_y. Other than that, the remaining parameters were shared, including the LSTM decoder and the attentional MLP. The hope was that by simply using the same parameters for a related task, we would allow data to be shared between the two languages and reduce overfitting.

3.3 Encoder-decoder Pipeline (ENCDEC)

The next model tested was the encoder-decoder pipeline (figure 2). Again, this was a relatively straightforward extension to the baseline. After the baseline model finished producing a German caption, it had some final state (h_t, c_t). We simply resumed decoding to produce an English caption starting from that final state with an independent decoder f_1, separate vocabulary, and this time without any direct access to the image. Each timestep is computed as $(h_t, c_t) = f_1(h_{t-1}, c_{t-1}, \textit{embed}_{t-1})$. This should force the model to keep information about the image in the hidden state throughout the decoding process, hopefully improving the model output.

This is the model that was used as the submission to the WMT multimodal task.

3.4 Attentional Pipeline with Averaged Embeddings (ATTAVG)

Attention has been shown to improve upon simple encoder-decoder models, so we wanted to test adding an additional attentional component. Both the baseline and the previous models mentioned already include attention over the *image*, but here we add attention over the German caption output as well. Once again, the German part of this model is just the baseline. Additionally, for each German word that was actually produced, we want to consider all of the alternatives. Thus at each timestep, we average together the embeddings of every word in the German vocabulary, weighted by the probability of producing each word. The result is one vector s_w (with the same dimension as the word embedding size) for each word w in the German caption.

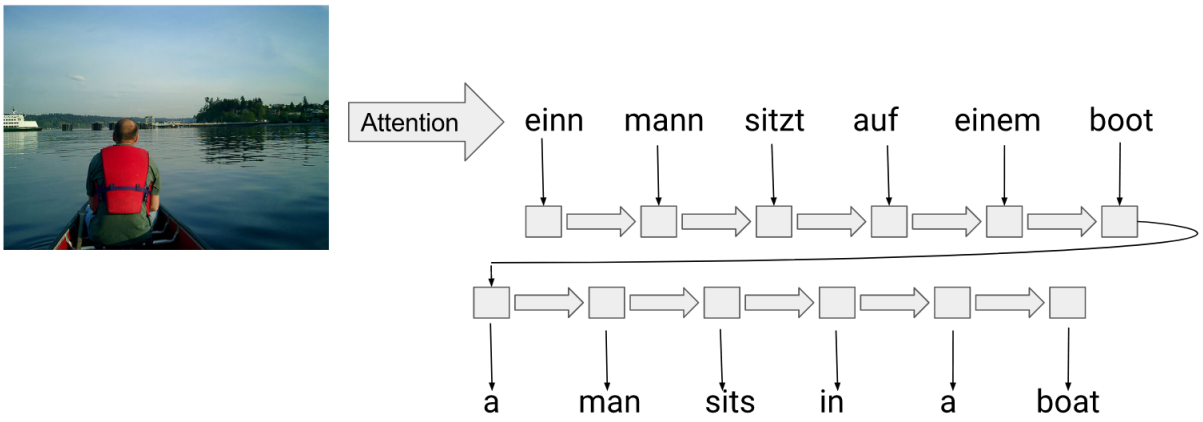

Figure 2: Encoder-decoder Pipeline. The LSTM state after producing the German caption (with attention to the image) is passed along to a new decoder. The new decoder produces an English caption using only the final hidden LSTM state, without referencing the image directly.

Then, we generate the English caption using a separate LSTM with attention over the averaged German word embeddings (and without any access to the underlying image). That is, at each timestep, an attention model f_{att} implemented as a multilayer perceptron (MLP) predicts how important each averaged word embedding s_w is, based on the previous hidden state h_{t-1}. We compute the softmax of these attention outputs and use this to compute a weighted average of the s_w embeddings. The result is a 256-dimensional context vector z_t that represents the important parts of the German sentence at timestep t. The next timestep is computed as $(h_t, c_t) = f_2(h_{t-1}, c_{t-1}, x_t)$ where $x_t = \mathit{concat}(\mathit{embed}_{t-1}, z_t)$. The process is shown in figure 3.

Unfortunately, the implementation of averaged embeddings requires more memory than the other implementations, forcing us to use a smaller word embedding size, smaller hidden layer, and smaller vocabulary. To address this issue, we consider a variant using random embeddings.

3.5 Attentional Pipeline with Random Embeddings (ATTRND)

This model is a slight variant on the attentional pipeline with averaged embeddings. At each timestep, instead of averaging together the embeddings of every word, we sample one random word from the distribution of predicted probabilities. The embedding of that word is multiplied by its probability, giving us a value that represents the contribution of that word to the weighted average. This again yields one vector for each word in the German caption. And again we generate the English caption using an LSTM with attention over the sampled German word embeddings (and without any access to the underlying image), as shown in figure 3.

3.6 Dual Attention (DUALATT)

Finally, we tried one model with the opposite structure from the rest (figure 4). We first generate the *English* caption using the baseline method, and then train an LSTM with attention over both the English caption and the image (using two separate MLPs).

That is, after we've generated an English caption using the baseline model, we consider it as a pseudo-reference. When generating the German sentence, we take attention over the image vectors as usual to get z_t, and we take attention over the word embeddings for the actual English caption generated to get $\tilde{z}_t$, both conditioned on the hidden state h_{t-1}. That allows us to compute the next timestep as $(h_t, c_t) = f_2(h_{t-1}, c_{t-1}, x_t)$ where $x_t = \mathit{concat}(\mathit{embed}_{t-1}, z_t, \tilde{z}_t)$.

4 Experimental Setup

All models were implemented using DyNet (Neubig et al., 2017), specifically using the `VanillaLSTM` class. Models were trained using the Adam optimizer (Kingma and Ba, 2014). Multi30k, an expanded of the Flickr 30k training data, was provided for the WMT multimodal task 2 constrained setting (Elliott et al., 2016) and

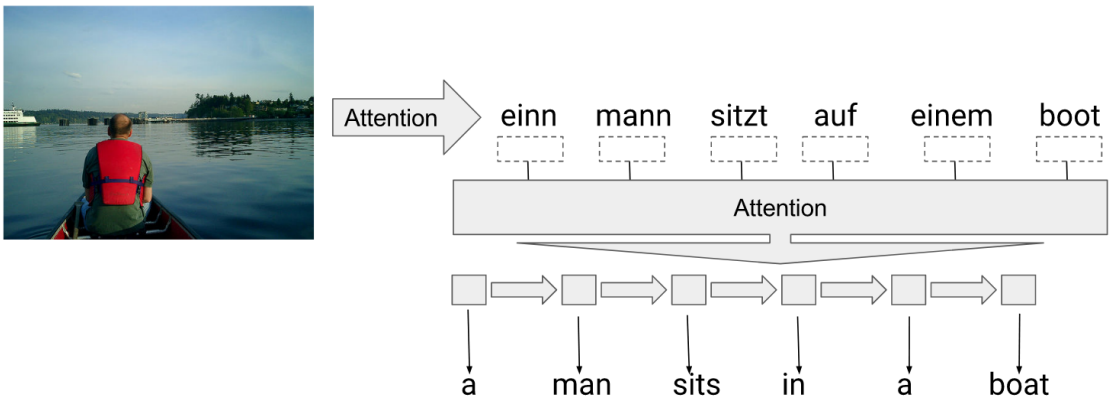

Figure 3: Attention Pipeline. At each timestep as the German caption is being generated, we produce an embedding (box with dashed outline). Depending on whether we are using averaged embeddings or random embeddings, this is either (1) the weighted average of all words in the vocabulary, or (2) the contribution of one randomly selected word to that weighted average. An LSTM with attention produces an English caption using these embeddings.

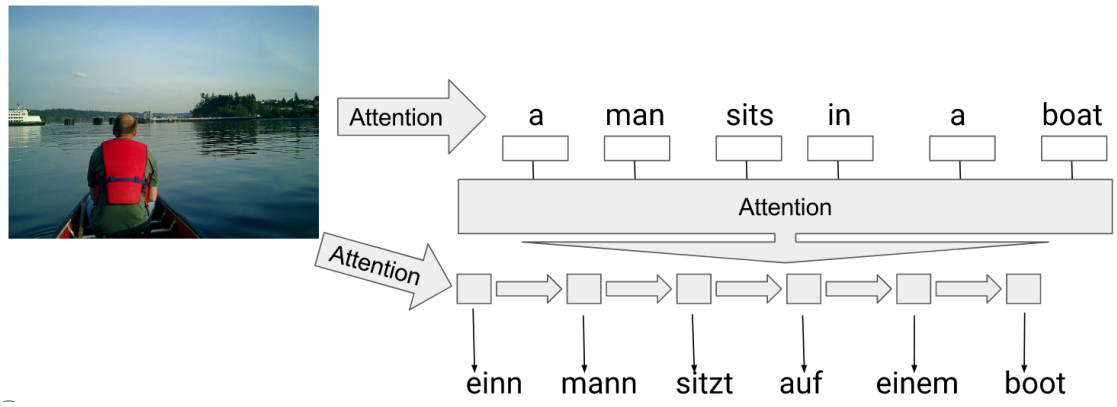

Figure 4: Dual Attention. After generating an English caption, we retrieve the embeddings for the words generated (white box with solid outline). An LSTM with attention over both the English embeddings and the image produces a German caption.

	Dropout	Vocabulary size (German/English)	Minibatch	BLEU-4	METEOR
Baseline 1	0.02	17855/12138	32	10.35	20.73
Baseline 2	0.2	9996/8368	8	10.20	18.97
Shared decoder*	0.2	17855/12138	24	11.51	20.87
ENCDEC*	0.2	9996/8368	32	11.53	21.90
ATTRND*	0.2	9996/8368	32	11.84	20.53
ATTAVG	0.2	6729/6310	8	9.18	19.67
DUALATT	0.2	17855/12138	24	10.51	19.68

Table 1: Model evaluation results. * indicates statistically significant improvement relative to baseline 1 ($p < 0.05$) with paired bootstrap resampling, based on BLEU-4 score on the 2016 test set. Multiple combinations of vocabulary size, minibatch size, and dropout were tested for each model, but only the best combination (by BLEU score on the validation set) is reported here.

used as the dataset. This dataset consists of 29000 images for training, 1014 images for validation, 1000 images for test 2016, and 1000 images for test 2017. Each image had 5 independently generated English and German captions. Since the English and German captions were generated independently, the pairing between English and German captions within each set of 5 was randomized on each epoch, for a total of 25 pairs per image. No external data was used, making this a constrained submission.

Each of the models used LSTM hidden size 512, embedding size 512, and hidden dimension 256 for the Attention MLP. The one exception was ATTAVG which due to memory limits used LSTM hidden size 256, embedding size 256, and hidden dimension 256 for the attention MLP. Minibatching was used, with each batch formed by grouping together similar length captions to improve efficiency. Minibatch sizes, vocabulary sizes, and dropout settings are noted in table 1. The order of the batches was randomized on each epoch. Models were trained until the perplexity on the validation set no longer improved.

5 Results

The WMT 2016 multimodal task test set was used for evaluation. Results were scored using BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014), with all sentences lower-cased and punctuation removed. Scores on the 2016 test set are shown in table 1.

The system submitted to the WMT multimodal task was ENCDEC. On the 2017 test set, it achieved a BLEU score of 9.1 (matching the official baseline and exceeding all other systems submitted). It also achieved a Meteor score of 19.8 (worse than the official baseline of 23.4) and a TER score of 63.3 (better than the official baseline of 91.4 and all other systems submitted). The fact that each of these three scoring methods shows a different result relative to the baseline is somewhat concerning.

In general, the evaluation results did not show very good correlation between BLEU and METEOR. We tested output samples derived from 52 experiments conducted with varying configurations during the course of the study. We found that the correlation between BLEU and METEOR was approximately 0.18. Strikingly, the top-ranked output according to METEOR scored more than 3 BLEU points lower than the baseline. Our informal human evaluation of the outputs tended to agree more with the BLEU evaluations than the METEOR evaluations.

6 Conclusion

We tested five alternative methods for supplementing a German caption dataset with English captions to improve performance, and in three cases achieved statistically significant improvements. This indicates that multilingual image captioning data is a valuable resource, even when learning only a single language. The best performing model measured by BLEU was the attentional pipeline with random embeddings, which improved on the baseline by 1.5 BLEU points. The best performing model measured by METEOR was the encoder-decoder pipeline, which improved on the baseline by 1.2 METEOR points.

References

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 627–633. http://www.aclweb.org/anthology/W/W16/W16-2358.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions pages 70–74.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709* .

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

Julian Hitschler and Stefan Riezler. 2016. Multimodal pivots for image caption translation. *CoRR* abs/1601.03916. http://arxiv.org/abs/1601.03916.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

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

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Kishore Papineni, 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*. pages 543–553.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR* abs/1409.3215. http://arxiv.org/abs/1409.3215.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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. http://arxiv.org/abs/1502.03044.