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 second WMT 2017 multimodal task. The goal of the task is to generate image captions in German, using a training corpus with captions in both English and German. We explore several models which attempt to generate captions for both languages simultaneously, 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 translation (Sutskever et al., 2014), caption generation (Xu et al., 2015), and language models (Bengio et al., 2003). Training these models requires vast amounts of data. Recently, these huge datasets have become more widely available, but there are still many limitations. In some cases, the dataset which is available may not exactly match the target application.

In this paper, we attempt to generate image captions in German, using a training corpus 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 translations. (See figure 1).

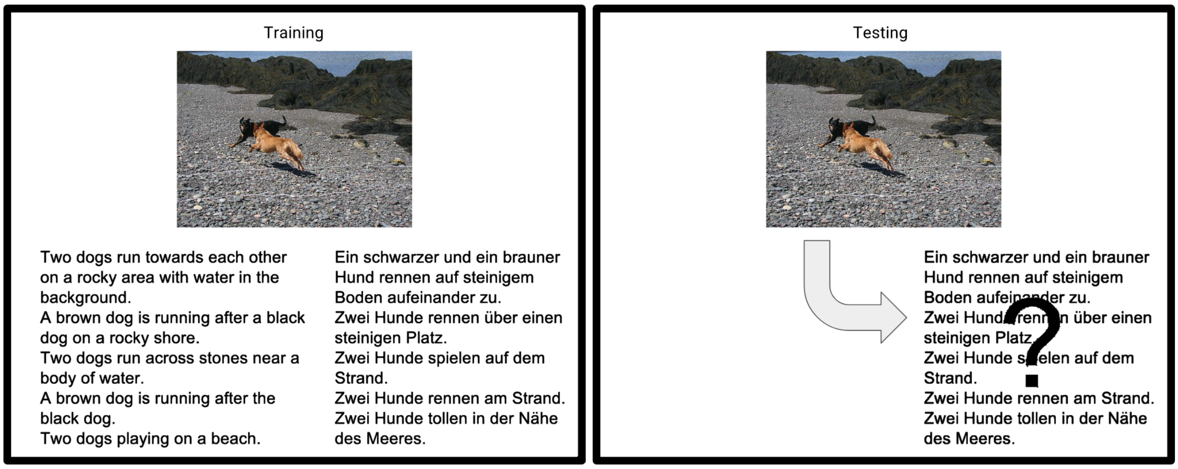
It's important to note that since these captions were generated independently in each language rather than translated, they often differ from each other quite a bit. Not only do they often choose to describe different features of an image, but they sometimes describe contradictory features of the image (such as 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.

We use the soft attention model from (Xu et al., 2015) as our baseline, training it only on the German captions. Then we develop several methods of incorporating the English data to improve the performance.

2 Related Work

Previous work on multilingual images such as (Hitschler and Riezler, 2016) has focused on the case where captions are available at evaluation time in a single language, and we wish to translate into a different language. In that case, it is sufficient to use existing machine translation techniques to translate the given caption, supplemented with information from the image to rerank the translation output. Similarly, systems developed for the WMT 2016 multimodal task such as (Huang et al., 2016) 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.

While there is little work on this exact task, the over-arching task of caption generation has been considered before. (Vinyals et al., 2015) use a convolutional neural network to encode an image, followed by an LSTM decoder to produce an output sequence. (Xu et al., 2015) extends that model by adding an attentional component, using a multilayer perceptron to determine which parts



Training

Two dogs run towards each other on a rocky area with water in the background.
A brown dog is running after a black dog on a rocky shore.
Two dogs run across stones near a body of water.
A brown dog is running after the black dog.
Two dogs playing on a beach.

Ein schwarzer und ein brauner Hund rennen auf steinigem Boden aufeinander zu.
Zwei Hunde rennen über einen steinigen Platz.
Zwei Hunde spielen auf dem Strand.
Zwei Hunde rennen am Strand.
Zwei Hunde tollen in der Nähe des Meeres.

Testing

Ein schwarzer und ein brauner Hund rennen auf steinigem Boden aufeinander zu.
Zwei Hunde rennen über einen steinigen Platz.
Zwei Hunde spielen auf dem Strand.
Zwei Hunde rennen am Strand.
Zwei Hunde tollen in der Nähe des Meeres.

Figure 1: Training data and test data

of the image are given to the LSTM at each step.

The attentional 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), but in this paper we will focus only on the multilayer perceptron method.

Generally, the long short-term memory (LSTM) model (Hochreiter and Schmidhuber, 1997) seems to be quite effective for this task. Dropout has also been shown to reduce overfitting (Srivastava et al., 2014).

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) pre-trained model was quite effective (specifically using layer 'res4fx' which is found at the end of Block-4, after ReLU), so we will incorporate these pre-trained embeddings into our model. Note that this differs from (Xu et al., 2015), which used pre-trained annotations from the Oxford VGGnet (Simonyan and Zisserman, 2014).

3 Approach

We developed several models, each of which simultaneously 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.

4 Baseline

Our baseline was 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., which used Oxford VGGnet with only 512 real numbers for each location in the image.) Xu et al. 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 model f_{att} implemented as a multilayer perceptron (MLP) predicts how important each part of the image is based on the current hidden state h_{t-1}. We compute the softmax of these attention outputs and use this to compute a weighted average of the image vectors. The result is a fixed-length (1024 long) context vector z_t that represents the important parts of the entire image at timestep t.

We use the `VanillaLSTM` from DyNet (Neubig et al., 2017) as the decoder, which has decoupled input and forget gates and does not use peephole connections. A single timestep $f(h_{t-1}, c_{t-1}, x_t) = (h_t, c_t, o_t)$ takes a previous hidden state (h_{t-1}, c_{t-1}), and context x_t resulting in a new hidden state (h_t, c_t) as well as an output o_t as follows (Neubig et al., 2017):

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

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

We initialize the LSTM to 0, unlike Xu et al. which initializes the LSTM using two additional attentional MLP's. Given some previous state (h_{t-1}, c_{t-1}), we compute $f(h_{t-1}, c_{t-1}, x_t)$ where $x_t = \mathit{concat}(\mathit{embed}_{t-1}, z_t)$. 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 o_t is then transformed to $\mathit{softmax}(W_{yo}o_t + b_y)$ to compute the probability of each word in the vocabulary.

Since we re-implemented this baseline and made some changes in the process, 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). 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.

5 Shared Decoder

The first model tested was the shared decoder model. 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_{yo}, 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.

6 Encoder-decoder Pipeline

The next model tested was the encoder-decoder pipeline (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 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 $(h_t, c_t, o_t) = f_1(h_{t-1}, c_t, \mathit{embed}_{t-1})$. 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.

7 Attentional Pipeline with Averaged Embeddings

Attention has been shown to improve upon simple encoder-decoder models, so we wanted to test adding an additional attentional component. 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.

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^2_{att} implemented as a multilayer perceptron (MLP) predicts how important each averaged word embedding s_w is based on the current 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 fixed-length (256 long) context vector z_t that represents the important parts of the German sentence at timestep t. The next timestep is computed as $(h_t, c_t, o_t) = f_2(h_{t-1}, c_t, x_t)$ 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

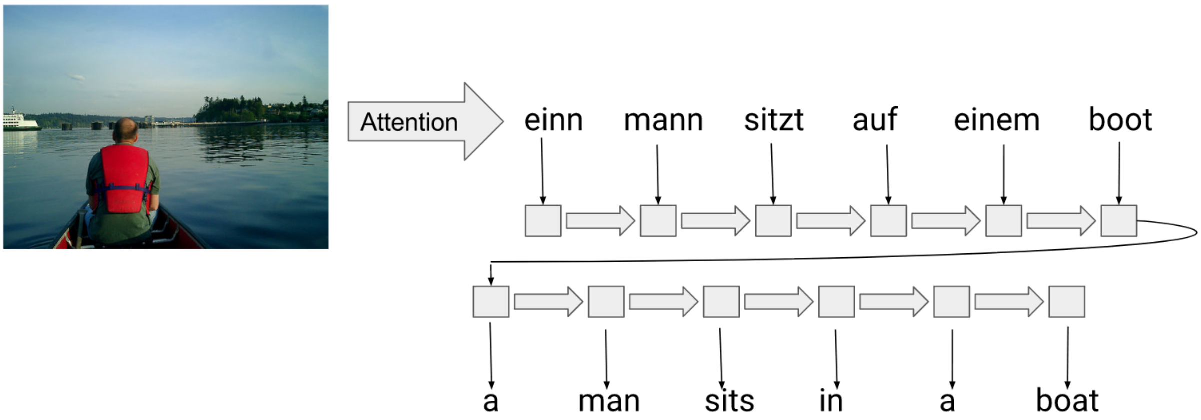


Figure 2: Encoder-decoder Pipeline. The LSTM state after producing the German caption (with attention to the image) is passed along to a new decoder. The new decoder produces an English caption using only the final hidden LSTM state, without referencing the image directly.

embedding size, smaller hidden layer, and smaller vocabulary. To address this issue, we consider a variant using random embeddings.

8 Attentional Pipeline with Random Embeddings

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 "averaged" German word embeddings (and without any access to the underlying image).

9 Dual Attention

Finally, we tried one model with the opposite structure from the rest. 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 $(h_t, c_t, o_t) = f_2(h_{t-1}, c_t, x_t)$ where $x_t = \mathit{concat}(\mathit{embed}_{t-1}, z_t, \tilde{z}_t)$.

10 Experimental Setup

All models were implemented using DyNet (Neubig et al., 2017) and trained using the Adam optimizer (Kingma and Ba, 2014) for back propagation. A variant of the Flickr training data was provided for the WMT multimodal task 2 constrained setting (Elliott et al., 2016) and used as the dataset. No external data was used, making this a constrained submission. Each of the models used hidden size 512, embedding size 512, and attention size 256, with the exception of the attentional pipeline with averaged embeddings which used hidden size 256, embedding size 256, and attention size 128. 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. 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. Models were trained until the perplexity on the validation set no longer improved.

11 Experimental Evaluation

The 2016 WMT multimodal task test set was used for evaluation. Results were scored using BLEU (Papineni et al., 2002) and METEOR

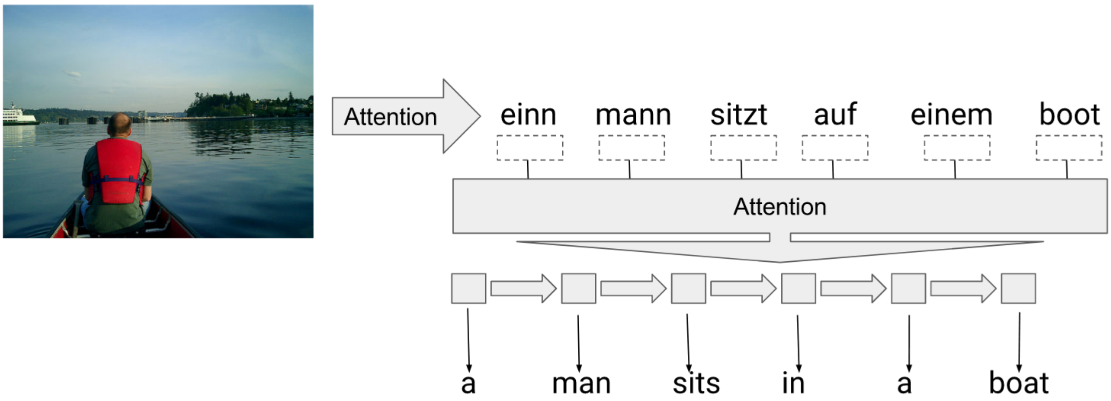


Figure 3: Attention Pipeline. At each timestep as the German caption is being generated, we produce an embedding (box with dashed outline). Depending on whether we are using averaged embeddings or random embeddings, this is either (1) the weighted average of all words in the vocabulary, or (2) the contribution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

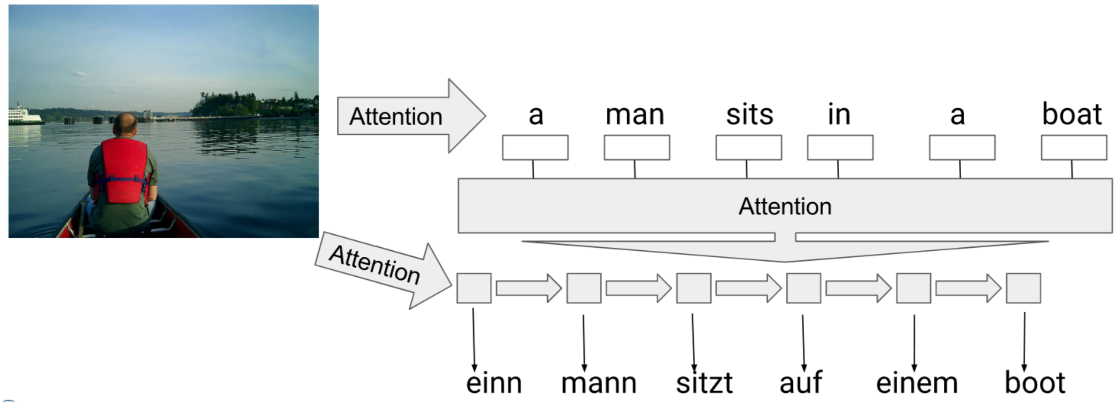


Figure 4: Dual Attention. After generating an English caption, we retrieve the embeddings for the words generated (white box with solid outline). An LSTM with attention over both the English embeddings and the image produces a German caption.

(Denkowski and Lavie, 2014), with all sentences lower-cased and punctation removed. Scores on the test set are shown in table 1, along with output examples in figure 5.

Oddly, the evaluation results did not show very good correlation between BLEU and METEOR. I tested 52 different output samples which were produced during tests of the various models and 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. My informal human evaluation of the outputs tended to agree more with the BLEU evaluations than the METEOR evaluations.

12 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 caption 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.

Google. 2017. Google translate. Accessed: 2017-05-16. https://translate.google.com.

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* .

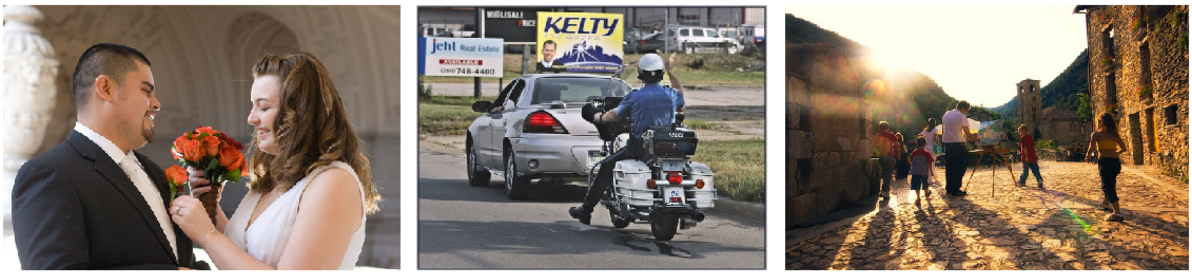
Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '02, pages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

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks

	Dropout	Vocabulary size (German/English)	Minibatch	BLEU-4	METEOR
Single-language attentional	0.02	17855/12138	32	10.35	20.73
Single-language attentional	0.2	9996/8368	8	10.20	18.97
Shared decoder*	0.2	17855/12138	24	11.51	20.87
Encoder-decoder Pipeline*	0.2	9996/8368	32	11.53	21.90
Attentional Pipeline with Random Embeddings*	0.2	9996/8368	32	11.84	20.53
Attentional Pipeline with Averaged Embeddings	0.2	6729/6310	8	9.18	19.67
Dual attention	0.2	17855/12138	24	10.51	19.68

Table 1: Model evaluation results. * indicates statistically significant improvement relative to the baseline ($p < 0.05$) with paired bootstrap resampling, based on BLEU-4 score on the 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.



ein mann und eine frau unterhalten sich
eine frau sitzt auf einer bank
ein mann und eine frau unterhalten sich
ein mann und eine frau stehen an einem tisch
ein mann und eine frau unterhalten sich
eine frau hält ein baby
eine frau und eine frau unterhalten sich

A man and a woman are talking
A woman sitting on a bench
A man and a woman are talking
A man and a woman are standing at a table
A man and a woman are talking
A woman is holding a baby
A woman and a woman are talking

ein mann auf einem motorroller
ein mann fährt auf einem fahrrad
ein mann sitzt auf einer bank
ein mann mit helm fährt auf einer straße
ein mann sitzt auf einem motorroller
ein mann sitzt auf einer straße
ein mann fährt auf einem motorrad

A man on a motorroller
A man is driving on a bicycle
A man sitting on a bench
A man with helmet driving on a road
A man sitting on a motorroller
A man is sitting on a street
A man is driving on a motorcycle

ein paar personen die auf einem weg gehen
menschen auf einem
mehrere personen gehen über einen weg
mehrere personen gehen auf einem weg
ein mann und eine frau gehen auf einem gehweg
ein mann steht auf einem dach
mehrere personen stehen auf einem weg

A few people going on a walk
People on one
Several people go over a path
Several people go on a walk
A man and a woman walking on a walkway
A man is standing on a roof
Several people stand on one path

Figure 5: Sample output on three randomly selected images from the validation set for each of the systems tested, along with English translations provided by Google Translate (Google, 2017). The captions for each image are listed in the order of the systems in table 1.

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.