CUNI System for the WMT18 Multimodal Translation Task

Jindřich Helcl and **Jindřich Libovický** and **Dušan Variš**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{helcl, libovicky, varis}@ufal.mff.cuni.cz

Abstract

We present our submission to the WMT18 Multimodal Translation Task. The main feature of our submission is applying a self-attentive network instead of a recurrent neural network. We evaluate two methods of incorporating the visual features in the model: first, we include the image representation as another input to the network; second, we train the model to predict the visual features and use it as an auxiliary objective. For our submission, we acquired both textual and multimodal additional data. Both of the proposed methods yield significant improvements over recurrent networks and self-attentive textual baselines.

1 Introduction

Multimodal Machine Translation (MMT) is one of the tasks that seek ways of capturing the relation of texts in different languages given a shared "grounding" information in a different (e.g. visual) modality.

The goal of the MMT shared task is to generate an image description (caption) in the target language using a caption in the source language and the image itself. The main motivation for this task is the development of models that can exploit the visual information for meaning disambiguation and thus model the denotation of words.

During the last years, MMT was addressed as a subtask of neural machine translation (NMT). It was thoroughly studied within the framework of recurrent neural networks (RNNs) (Specia et al., 2016; Elliott et al., 2017). Recently, the architectures based on self-attention such as the Transformer (Vaswani et al., 2017) became state-of-the-art in NMT.

In this work, we present our submission based on the Transformer model. We propose two ways of extending the model. First, we tweak the architecture such that it is able to process both modalities in a multi-source learning scenario. Second, we leave the model architecture intact, but add another training objective and train the textual encoder to be able to predict the visual features of the image described by the text. This training component has been introduced in RNNs by Elliott and Kádár (2017) and is called the "imagination".

We find that with self-attentive networks, we are able to improve over a strong textual baseline by including the visual information in the model. This has been proven challenging in the previous RNN-based submissions, where there was only a minor difference in performance between textual and multimodal models (Helcl and Libovický, 2017; Caglayan et al., 2017).

This paper is organized as follows. Section 2 summarizes the previous submissions and related work. In Section 3, we describe the proposed methods. The details of the datasets used for the training are given in Section 4. Section 5 describes the conducted experiments. We discuss the results in Section 6 and conclude in Section 7.

2 Related Work

Currently, most of the work has been done within the framework of sequence-to-sequence learning. Although some of the proposed approaches use explicit image analysis (Shah et al., 2016; Huang et al., 2016), most methods use image representation obtained using image classification networks pre-trained on ImageNet (Deng et al., 2009), usually VGG19 (Simonyan and Zisserman, 2014) or ResNet (He et al., 2016a).

In the simplest case, the image can be represented as a single vector from the penultimate layer of the image classification network. This vector can be then plugged in at various places of the sequence-to-sequence architecture (Libovický et al., 2016; Calixto and Liu, 2017).

Several methods compute visual context information as a weighted sum over the image spatial representation using the attention mechanism (Bahdanau et al., 2014; Xu et al., 2015) and combine it with the context vector from the textual encoder in doubly-attentive decoders. Caglayan et al. (2016) use the visual context vector in a gating mechanism applied to the textual context vector. Caglayan et al. (2017) concatenate the context vectors from both modalities. Libovický and Helcl (2017) proposed advanced strategies for computing a joint attention distribution over the text and image. We follow this approach in our first proposed method described in Section 3.1.

The visual information can also be used as an auxiliary objective in a multi-task learning setup. Elliott and Kádár (2017) propose an imagination component that predicts the visual features of an image from the textual encoder representation, effectively regularizing the encoder part of the network. The imagination component is trained using a maximum margin objective. We reuse this approach in our method described in Section 3.2.

3 Architecture

We examine two methods of exploiting the visual information in the Transformer architecture. First, we add another encoder-decoder attention layer to the decoder which operates over the image features directly. Second, we train the network with an auxiliary objective using the imagination component as proposed by Elliott and Kádár (2017).

3.1 Doubly Attentive Transformer

The Transformer network follows the encoder-decoder scheme. Both parts consist of a number of layers. Each encoder layer first attends to the previous layer using self-attention, and then applies a single-hidden-layer feed-forward network to the outputs. All layers are interconnected with residual connections and their outputs are normalized by layer normalization (Ba et al., 2016). A decoder layer differs from an encoder layer in two aspects. First, as the decoder operates autoregressively, the self-attention has to be masked to prevent the decoder to attend to the "future" states. Second, there is an additional attention sub-layer applied after self-attention which attends to the final states of the encoder (called *encoder-decoder*, or *cross* attention).

The key feature of the Transformer model is the use of attention mechanism instead of recurrence relation in RNNs. The attention can be conceptualized as a soft-lookup function that operates on an associative array. For a given set of queries Q, the attention uses a similarity function to compare each query with a set of keys K. The resulting similarities are normalized and used as weights to compute a context vector which is a weighted sum over a set of values V associated with the keys. In self-attention, all the queries, keys and values correspond to the set of states of the previous layer. In the following cross-attention sub-layer, the set of resulting context vectors from the self-attention sub-layer is used as queries, and keys and values are the states of the final layer of the encoder.

The Transformer uses scaled dot-product as a similarity metric for both self-attention and cross-attention. For a query matrix Q, key matrix K and value matrix V, and the model dimension d, we have:

$$\mathcal{A}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \quad (1)$$

The attention is used in a multi-head setup. This means that we first linearly project the queries, keys, and values into a number of smaller matrices and then apply the attention function $\mathcal{A}$ independently on these projections. The set of resulting context vectors C is computed as a sum of the outputs of each attention head, linearly projected to the original dimension:

$$C = \sum_{i=1}^{h} \mathcal{A}(QW_i^Q, KW_i^K, VW_i^V)W_i^O \quad (2)$$

where $(W_i^O)^\top$, W_i^Q, W_i^K, and $W_i^V \in \mathbb{R}^{d \times d_h}$ are trainable parameters, d is the dimension of the model, h is the number of heads, and d_h is a dimension of a single head. Note that despite K and V being identical matrices, the projections are trained independently.

In this method, we introduce the visual information to the model as another encoder via an additional cross-attention sub-layer. The keys and values of this cross-attention correspond to the vectors in the last convolutional layer of a pre-trained image processing network applied on the input image. This sub-layer is inserted between the textual cross-attention and the feed-forward network, as illustrated in Figure 1. The set of the context vectors from the textual cross-attention is used as
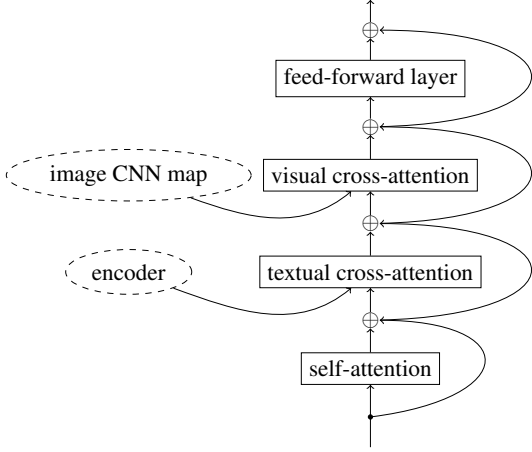
Figure 1: One layer of the doubly-attentive Transformer decoder with 4 sub-layers connected with residual connections.

	en	de	fr	cs
Training	29,000 sentences			
Tokens	378k	361k	410k	297k
Average length	13.0	12.4	14.1	10.2
# tokens range	4–40	2–44	4–55	2–39
Validation	1,014 sentences			
Tokens	13k	13k	14k	10k
Average length	13.1	12.7	14.2	10.2
# tokens range	4–30	3–33	5–36	4–27
OOV rate	1.28%	3.09%	1.20%	3.95%

Table 1: Multi30k statistics on training and validation data – total number of tokens, average number of tokens per sentence, and lengths of the shortest and the longest sentence.

queries, and the context vectors of the visual cross-attention are used as inputs to the feed-forward sub-layer. Similarly to the other sub-layers, the input is linked to the output by a residual connection. Equation 3 shows the computation of the visual context vectors given trainable matrices Z_i^Q, Z_i^K, Z_i^V, and Z_i^O for $i = 1, \ldots, h$; the set of textual context vectors is denoted by C_{txt} and the extracted set of image features as F:

$$C_{img} = \sum_{i=1}^{h} \mathcal{A}(C_{txt}Z_i^Q, FZ_i^K, FZ_i^V)Z_i^O. \quad (3)$$

3.2 Imagination

We use the imagination component of Elliott and Kádár (2017) originally proposed for training multimodal translation models using RNNs. We adapt it in a straightforward way in our Transformer-based models.

The imagination component serves effectively as a regularizer to the encoder, making it consider the visual meaning together with the words in the source sentence. This is achieved by training the model to predict the image representations that correspond to those computed by a pre-trained image classification network. Given a set of encoder states h_j, the model computes the predicted image representation as follows:

$$\hat{y}_{img} = W_2^R \max(0, W_1^R \sum_j h_j) \quad (4)$$

where $W_1^R \in \mathbb{R}^{r \times d}$ and $W_2^R \in \mathbb{R}^{n \times r}$ are trainable parameter matrices, d is the Transformer model dimension, r is a hidden layer dimension of the

imagination component, and n is the dimension of the image feature vector. Note that Equation 4 corresponds to a single-hidden-layer feed-forward network with a ReLU activation function applied on the sum of the encoder states.

We train the visual feature predictor using an auxiliary objective. Since the encoder part of the model is shared, additional weight updates are propagated to the encoder during the model optimization w.r.t. this additional loss. For the generated image representation $\hat{y}$ and the reference representation y, the error is estimated as margin-based loss with margin parameter α:

$$\mathcal{L}_{imag} = \max(0, \alpha + d(\hat{y}, y) - d(\hat{y}, y_c)) \quad (5)$$

where y_c is a contrastive example randomly drawn from the training batch and d is a distance function between the representation vectors, in our case the cosine distance.

Unlike Elliott and Kádár (2017), we sum both translation and imagination losses within the training batches rather than alternating between training of each component separately.

4 Data

The participants were provided with the Multi30k dataset (Elliott et al., 2016), an extension of the Flickr30k dataset (Plummer et al., 2017) which contains 29,000 train images, 1,014 validation images and 1,000 test images. The images are accompanied with six captions which were independently obtained through crowd-sourcing. In

Multi30k, each image is accompanied also with German, French, and Czech translations of a single English caption. Table 1 shows statistics of the captions contained in the Multi30k dataset.

Since the Multi30k dataset is relatively small, we acquired additional data, similarly to our last year submission (Helcl and Libovický, 2017). The overview of the dataset structure is given in Table 2.

First, for German only, we prepared synthetic data out of the WMT16 MMT Task 2 training dataset using back-translation to English (Sennrich et al., 2016). This data consists of five additional German descriptions of each image. Along with the data for Task 1 which is the same as the training data this year, the back-translated part of the dataset contains 174k sentences.

Second, for Czech and German, we selected pseudo in-domain data by filtering the available general domain corpora. For both languages, we trained a character-level RNN language model on the corresponding language parts of the Multi30k training data. We use a single layer bidirectional LSTM (Hochreiter and Schmidhuber, 1997) network with 512 hidden units and character embeddings with dimension of 128. For Czech, we compute perplexities of the Czech sentences in the CzEng corpus (Bojar et al., 2016b). We selected 15k low-perplexity sentence pairs out of 64M sentence pairs in total by setting the perplexity threshold to 2.5. For German, we used the additional data from the last year (Helcl and Libovický, 2017), which was selected out of several parallel corpora (EU Bookshop (Skadiņš et al., 2014), News Commentary (Tiedemann, 2012) and CommonCrawl (Smith et al., 2013)).

Third, also for Czech and German, we applied the same criterion on monolingual corpora and used back-translation to create synthetic parallel data. For Czech, we took 333M sentences of CommonCrawl and 66M sentences of News Crawl (which is used in the WMT News Translation Task; Bojar et al., 2016a) and extracted 18k and 11k sentences from these datasets respectively.

Finally, we use the whole EU Bookshop as an additional out-of-domain parallel data. Since the size of this dataset is large relative to the sizes of the other parts, we oversample the rest of the data to balance the in-domain and out-of-domain portions of the training dataset. The oversampling factors are shown in Table 2.

	de	fr	cs
Multi30k		29k	
– oversampling factor	273×	366×	9×
Task 2 BT	145k	—	—
in-domain parallel	3k	—	15k
in-domain BT	30k	—	29k
– oversampling factor	39×	—	7×
EU Bookshop	9.3M	10.6M	445k
COCO (English only)		414k	

Table 2: Overview of the data used for training our models with oversampling factors. The EU Bookshop data was not oversampled. BT stands for back-translation.

For the unconstrained training of the imagination component, we used the MSCOCO (Lin et al., 2014) dataset which consists of 414k images along with English captions.

5 Experiments

In this year's round, two variants of the MMT tasks were announced. As in the previous years, the goal of Task 1 is to translate an English caption into the target language given the image. The target languages are German, French and Czech. In Task 1a, the model receives the image and its captions in English, German, and French and is trained to produce the Czech translation. In our submission, we focus only on Task 1.

In our submission, we experiment with three distinct architectures. First, in *textual* architectures, we leave out the images from the training altogether. We use this as a strong baseline for the multimodal experiments. Second, *multimodal* experiments use the doubly attentive Transformer decoder described in Section 3.1. Third, the experiments referred to as *imagination* employ the imagination component as described in Section 3.2.

We train the models in constrained and unconstrained setups. In the constrained setup, only the Multi30k dataset is used for training. In the unconstrained setup, we train the model using the additional data described in Section 4. We run the multimodal experiments only in the constrained setup.

In the unconstrained variant of the imagination experiments, the dataset consists of examples that can miss either the textual target values (MSCOCO extension), or the image (additional

		en-cs		en-fr		en-de	
		single	averaged	single	averaged	single	averaged
Caglayan et al. (2017)		N/A		54.7/71.3	56.7/73.0	37.8/57.7	41.0/**60.5**
Cons.	Textual	29.6/28.9	30.9/29.5	59.2/73.7	59.7/74.4	38.1/56.2	38.3/56.0
	Imagniation	29.8/29.4	30.5/29.6	59.4/74.2	59.7/74.4	38.8/56.4	39.2/56.8
	Multimodal	30.5/29.7	31.0/29.9	60.6/75.0	60.8/75.1	38.4/53.1	38.7/57.2
Unc.	Textual	31.2/30.1	32.3/30.7	62.0/76.7	62.5/76.7	39.6/58.7	40.4/59.0
	Imagination	**36.3/32.8**	35.9/32.7	**62.8/77.0**	**62.8/77.0**	**42.7**/59.1	42.6/59.4

Table 3: Results on the 2016 test set in terms of BLEU score and METEOR score. We compare our results with the last year's best system (Caglayan et al., 2017) which used model ensembling instead of weight averaging.

parallel data). In these cases, we train only the decoding component with specified target value (i.e. imagination component on visual features, or the Transformer decoder on the textual data). As said in Section 3.2, we train both components by summing the losses when both the image and the target sentence are available in a training example.

In all experiments, we use the Transformer network with 6 layers with model dimension of 512 and feed-forward hidden layer dimension of 4096 units. The embedding matrix is shared between the encoder and decoder and its transposition is reused as the output projection matrix (Press and Wolf, 2017). For each language pair, we use a vocabulary of approximately 15k wordpieces (Wu et al., 2016). We extract the vocabulary and train the model on lower-cased text without any further pre-processing steps applied. We tokenize the text using the algorithm bundled with the tensor2tensor library (Vaswani et al., 2018). The tokenization algorithm splits the sentence to groups of alphanumeric and non-alphanumeric groups, throwing away single spaces that occur inside the sentence. We conduct the experiments using the Neural Monkey toolkit (Helcl and Libovický, 2017).[1]

For image pre-processing, we use ResNet-50 (He et al., 2016a) with identity mappings (He et al., 2016b). In the doubly-attentive model, we use the outputs of the last convolutional layer before applying the activation function with dimensionality of $8 \times 8 \times 2048$. We apply a trainable linear projection to the maps into 512 dimensions to fit the Transformer model dimension. In the imagination experiments, we use average-pooled maps with 2048 dimensions. Following Elliott and Kádár (2017), we set the margin parameter α from Equation 5 to 0.1.

[1] https://github.com/ufal/neuralmonkey

For each model, we keep 10 sets of parameters that achieve the best BLEU scores (Papineni et al., 2002) on the validation set. We experiment with weight averaging and model ensembling. However, these methods performed similarly and we thus report only the results of the weight averaging, which is computationally less demanding.

In all experiments, we use the Adam optimizer (Kingma and Ba, 2014) with initial learning rate 0.2, and Noam learning rate decay scheme (Vaswani et al., 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ and 4,000 warm-up steps.

6 Results

We report the quantitative results of measured on the Multi30k 2016 test set in Table 3.

The Transformer architecture achieves generally comparable or better results than the RNN-based architecture. Adding the visual information has a significant positive effect on the system performance, both when explicitly provided as a model input and when used as an auxiliary objective. In the constrained setup which used only the data from the Multi30k dataset, the doubly-attentive decoder performed best.

The biggest gain in performance was achieved by training on the additional parallel data. The imagination architecture outperforms the purely textual models.

As the performance of single models increases, the positive effect of weight averaging diminishes. The effect of checkpoint averaging is smaller than the results reported by Caglayan et al. (2017) who use ensembles of multiple models trained with a different initialization – we use only checkpoints from a single training run.

During the qualitative analysis, we noticed that mostly for Czech target language, the systems are

often incapable of capturing morphology. In order to quantify this, we also measured the BLEU scores using the lemmatized system outputs and references. The difference was around 4 BLEU points for Czech, less than 3 BLEU points for French, and around 2 BLEU points for German. These differences were consistent among different types of models.

We hypothesize that in the imagination experiments, the visual information is used to learn a better representation of the textual input, which eventually leads to improvements in the translation quality. In the multimodal experiments, the improvements can come from the refining of the textual representation rather than from explicitly using the image as an input.

In order to determine whether the visual information is used also at the inference time, we performed an adversarial evaluation by providing the trained multimodal model with randomly selected "fake" images. In French and Czech, BLEU scores dropped by more than 1 BLEU point. This suggests that the multimodal models utilize the visual information at the inference time as well. The German models seem to be virtually unaffected. We hypothesize this might be due to a different methodology of acquiring the training data for German and the other two target languages (Elliott et al., 2016).

7 Conclusions

In our submission for the WMT18 Multimodal Translation Task, we experimented with the Transformer architecture for MMT. The experiments show that the Transformer architecture outperforms the RNN-based models.

Experiments with a doubly-attentive decoder showed that explicit incorporation of visual information improves the model performance. The adversarial evaluation confirms that the models also take into account the visual information.

The best translation quality was achieved by extending the training data by additional image captioning data and parallel textual data. It this unconstrained setup, the best scoring model employs the imagination component that was previously introduced in RNN-based sequence-to-sequence models.

References

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

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

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016a. Findings of the 2016 conference on machine translation (WMT16). In *Proceedings of the First Conference on Machine Translation (WMT). Volume 2: Shared Task Papers*, volume 2, pages 131–198, Stroudsburg, PA, USA. Association for Computational Linguistics, Association for Computational Linguistics.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. CzEng 1.6: Enlarged Czech-English parallel corpus with processing tools dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

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*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, Miami, FL, USA. IEEE.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *CoRR*, abs/1705.04350.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645.

Jindřich Helcl and Jindřich Libovický. 2017. CUNI system for the WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 450–457. Association for Computational Linguistics.

Jindřich Helcl and Jindřich Libovický. 2017. Neural Monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics*, 107:5–17.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9:1735–1780.

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*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

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

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation*, pages 646–654, Berlin, Germany. Association for Computational Linguistics.

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

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

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vision*, 123(1):74–93.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation*, pages 660–665, Berlin, Germany. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In

Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland. European Language Resources Association (ELRA).

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich 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-15)*, pages 2048–2057, Lille, France. JMLR Workshop and Conference Proceedings.