CUNI System for the WMT17 Multimodal Translation Task

CUNI System for the WMT17 Multimodal Traslation Task

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

Abstract

In this paper, we describe our submissions to the WMT17 Multimodal Translation Task. For Task 1 (multimodal translation), our best scoring system is a purely textual neural translation of the source image caption to the target language. The main feature of the system is the use of additional data that was acquired by selecting similar sentences from parallel corpora and by data synthesis with back-translation. For Task 2 (cross-lingual image captioning), our best submitted system generates an English caption which is then translated by the best system used in Task 1. We also present negative results, which are based on ideas that we believe have potential of making improvements, but did not prove to be useful in our particular setup.

1 Introduction

Recent advances in deep learning allowed inferring distributed vector representations of both textual and visual data. In models combining text and vision modalities, this representation can be used as a shared data type. Unlike the classical natural language processing tasks where everything happens within one language or across languages, multimodality tackles how the language entities relate to the extra-lingual reality. One of these tasks is multimodal translation whose goal is using cross-lingual information in automatic image captioning.

In this system-description paper, we describe our submission to the WMT17 Multimodal Translation Task. In particular, we discuss the effect of mining additional training data and usability of advanced attention strategies. We report our results on both the 2016 and 2017 test sets and discuss efficiency of tested approaches.

The rest of the paper is organized as follows. Section 2 introduces the tasks we handle in this paper and the datasets that were provided to the task. Section 3 summarizes the state-of-the-art methods applied to the task. In Section 4, we describe our models and the results we have achieved. Section 5 presents the negative results and Section 6 concludes the paper.

2 Task and Dataset Description

The challenge of the WMT Multimodal Translation Task is to exploit cross-lingual information in automatic image caption generation. The state-of-the-art models in both machine translation and automatic image caption generation use similar architectures for generating the target sentence. The simplicity with which we can combine the learned representations of various inputs in a single deep learning model inevitably leads to a question whether combining the modalities can lead to some interesting results. In the shared task, this is explored in two subtasks with different roles of visual and textual modalities.

In the multimodal translation task (Task 1), the input of the model is an image and its caption in English. The system then should output a German or French translation of the caption. The system output is evaluated using the METEOR (Denkowski and Lavie, 2011) and BLEU (Papineni et al., 2002) scores computed against a single reference sentence. The question this task tries to answer is whether and how is it possible to use visual information to disambiguate the translation.

In the cross-lingual captioning task (Task 2), the input to the model at test-time is the image alone. However, additionally to the image, the model is

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 450–457
Copenhagen, Denmark, September 711, 2017.

	en	de	fr
Train. sentences	29,000		
Train. tokens	378k	361k	410k
Avg. # tokens	13.0	12.4	14.1
# tokens range	4–40	2–44	4–55
Val. sentences	1,014		
Val. tokens	13k	13k	14k
Avg. # tokens	13.1	12.7	14.2
# tokens range	4–30	3–33	5–36
OOV rate	1.28%	3.09%	1.20%

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

supplied with the English (source) caption during training. The evaluation method differs from Task 1 in using five reference captions instead of a single one. In Task 2, German is the only target language. The motivation of Task 2 is to explore ways of easily creating an image captioning system in a new language once we have an existing system for another language, assuming that the information transfer is less complex across languages than between visual and textual modalities.

2.1 Data

The participants were provided with the Multi30k dataset (Elliott et al., 2016) – a multilingual extension of Flickr30k dataset (Plummer et al., 2017) – for both training and evaluation of their models.

The data consists of 31,014 images. In Flickr30k, each image is described with five independently acquired captions in English. Images in the Multi30k dataset are enriched with five crowdsourced German captions. Additionally, a single German translation of one of the English captions was added for each image.

The dataset is split into training, validation, and test sets of 29,000, 1,014, and 1,000 instances respectively. The statistics on the training and validation part are tabulated in Table 1.

For the 2017 round of the competition (Elliott et al., 2017), an additional French translation was included for Task 1 and new test sets have been developed. Two test sets were provided for Task 1: The first one consists of 1,000 instances and is similar to the test set used in the previous round of the competition (and to the training and validation data). The second one consists of images, captions, and their translations taken from the MSCOCO image captioning dataset (Lin et al., 2014). A new single test set containing 1,071 images with five reference captions was added for Task 2.

The style and structure of the reference sentences in the Flickr- and MSCOCO-based test sets differs. Most of the sentences in the Multi30k dataset have a similar structure with a relatively simple subject, an active verb in present tense, simple object, and location information (e.g., "*Two dogs are running on a beach.*"). Contrastingly, the captions in the MSCOCO dataset are less formal and capture the annotator's uncertainty about the image content (e.g., "*I don't know, it looks like a lemon.*").

3 Related Work

Several promising neural architectures for multimodal translation task have been introduced since the first competition in 2016.

In our last year's submission (Libovický et al., 2016), we employed a neural system that combined multiple inputs – the image, the source caption and an SMT-generated caption. We used the attention mechanism over the textual sequences and concatenated the context vectors in each decoder step.

The overall results of the WMT16 multimodal translation task did not prove the visual features to be particularly useful (Specia et al., 2016; Caglayan et al., 2016).

To our knowledge, Huang et al. (2016) were the first who showed an improvement over a textual-only neural system with model utilizing distributed features explicit object recognition. Calixto et al. (2017) improved state of the art using a model initializing the decoder state with the image vector, while maintaining the rest of the neural architecture unchanged. Promising results were also shown by Delbrouck and Dupont (2017) who made a small improvement using bilinear pooling.

Elliott and Kádár (2017) brought further improvements by introducing the "imagination" component to the neural network architecture. Given the source sentence, the network is trained to output the target sentence jointly with predicting the image vector. The model uses the visual information only as a regularization and thus is able to use additional parallel data without accompanying images.

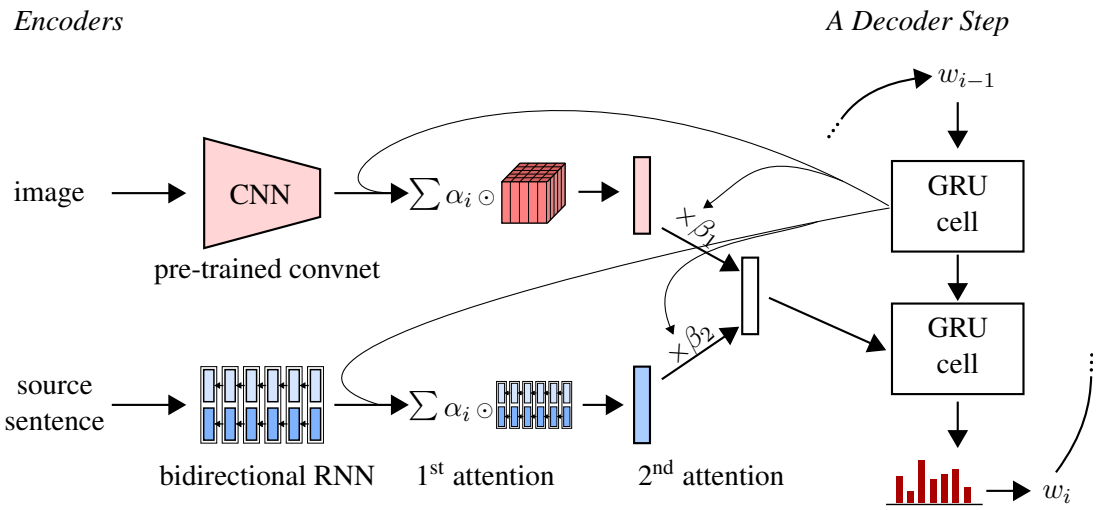

Figure 1: An overall picture of the multimodal model using hierarchical attention combination on the input. Here, α and β are normalized coefficients computed by the attention models, w_i is the i-th input to the decoder.

4 Experiments

All models are based on the encoder-decoder architecture with attention mechanism (Bahdanau et al., 2014) as implemented in Neural Monkey (Helcl and Libovický, 2017).[1] The decoder uses conditional GRUs (Firat and Cho, 2016) with 500 hidden units and word embeddings with dimension of 300. The target sentences are decoded using beam search with beam size 10, and with exponentially weighted length penalty (Wu et al., 2016) with α parameter empirically estimated as 1.5 for German and 1.0 for French. Because of the low OOV rate (see Table 1), we used vocabularies of maximum 30,000 tokens and we did not use sub-word units. The textual encoder is a bidirectional GRU network with 500 units in each direction and word embeddings with dimension of 300. We use the last convolutional layer VGG-16 network (Simonyan and Zisserman, 2014) of dimensionality $14 \times 14 \times 512$ for image processing. The model is optimized using the Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-4} with early stopping based on validation BLEU score.

4.1 Task 1: Multimodal Translation

We tested the following architectures with different datasets (see Section 4.3 for details):

- purely textual (disregarding the visual modality);
- multimodal with context vector concatenation in the decoder (Libovický et al., 2016);
- multimodal with hierarchical attention combination (Libovický and Helcl, 2017) – context vectors are computed independently for each modality and then they are combined together using another attention mechanism as depicted in Figure 1.

4.2 Task 2: Cross-lingual Captioning

We conducted two sets of experiments for this subtask. In both of them, we used an attentive image captioning model (Xu et al., 2015) for the cross-lingual captioning with the same decoder as for the first subtask.

The first idea we experimented with was using a multilingual decoder provided with the image and a language identifier. Based on the identifier, the decoder generates the caption either in English or in German. We speculated that the information transfer from the visual to the language modality is the most difficult part of the task and might be similar for both English and German.

The second approach we tried has two steps. First, we trained an English image captioning system, for which we can use larger datasets. Second, we translated the generated captions with the multimodal translation system from the first subtask.

4.3 Acquiring Additional Data

In order to improve the textual translation, we acquired additional data. We used the following

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

technique to select in-domain sentences from both parallel and monolingual data.

We trained a neural character-level language model on the German sentences available in the training part of the Multi30k dataset. We used a GRU network with 512 hidden units and character embedding size of 128.

Using the language model, we selected 30,000 best-scoring German sentences from the SDEWAC corpus (Faaß and Eckart, 2013) which were both semantically and structurally similar to the sentences in the Multi30k dataset.

We tried to use the language model to select sentence pairs also from parallel data. By scoring the German part of several parallel corpora (EU Bookshop (Skadiņš et al., 2014), News Commentary (Tiedemann, 2012) and CommonCrawl (Smith et al., 2013)), we were only able to retrieve a few hundreds of in-domain sentences. For that reason we also included sentences with lower scores which we filtered using the following rules: sentences must have between 2 and 30 tokens, must be in the present tense, must not contain non-standard punctuation, numbers of multiple digits, acronyms, or named entities, and must have at most 15 % OOV rate w.r.t. Multi30k training vocabulary. We extracted additional 3,000 in-domain parallel sentences using these rules. Examples of the additional data are given in Table 2.

By applying the same approach on the French versions of the corpora, we were pable to extract only few additional in-domain sentences. We thus trained the English-to-French models in the constrained setup only.

Following Calixto et al. (2017), we back-translated (Sennrich et al., 2016) the German captions from the German side of the Multi30k dataset (i.e. 5+1 captions for each image), and sentences retrieved from the SDEWAC corpus. We included these back-translated sentence pairs as additional training data for the textual and multimodal systems for Task 1. The back-translation system used the same architecture as the textual systems and was trained on the Multi30k dataset only. The additional parallel data and data from the SDEWAC corpus (denoted as additional in Table 3) were used only for the text-only systems because they were not accompanied by images.

For Task 2, we also used the MSCOCO (Lin et al., 2014) dataset which consists of 300,000 images with 5 English captions for each of them.

SDEWAC Corpus (with back-translation)
zwei Männer unterhalten sich . . . . . *two men are talking to each other* .
ein kleines Mädchen sitzt auf einer Schaukel . . . . . *a little girl is sitting on a swing* .
eine Katze braucht Unterhaltung . . . . . *a cat is having a discussion* .
dieser Knabe streichelt das Schlagzeug . . . . . *this professional is petting the drums* .
Parallel Corpora
Menschen bei der Arbeit . . . . . *People at work*
Männer und Frauen . . . . . *Men and women*
Sicherheit bei der Arbeit . . . . . *Safety at work*
Personen in der Öffentlichkeit . . . . . *Members of the public*

Table 2: Examples of the collected additional training data.

4.4 Results

In Task 1, our best performing system was the text-only system trained with additional data. These were acquired both by the data selection method described above and by back-translation. Results of all setups for Task 1 are given in Table 3.

Surprisingly, including the data for Task 2 to the training set decreased the METEOR score on both of the 2017 test sets. This might have been caused by domain mismatch. However, in case of the additional parallel and SDEWAC data, this problem was likely outweighed by the advantage of having more training data.

In case of multimodal systems, adding approximately the same amount of data increased the performance more than in case of the text-only system. This suggests, that with sufficient amount of data (which is a rather unrealistic assumption), the multimodal system would eventually outperform the textual one.

The hierarchical attention combination brought major improvements over the concatenation approach on the 2017 test sets. On the 2016 test set, the concatenation approach yielded better results, which can be considered a somewhat strange result, given the similarity of the Flickr test sets.

The baseline system was Nematus (Sennrich et al., 2017) trained on the textual part of Multi30k only. However, due to its low score, we suspect the model was trained with suboptimal parameters because it is in principle a model identical to our constrained textual submission.

		Task 1: en → de			Task 1: en → fr	
		2016	Flickr	MSCOCO	Flickr	MSCOCO
Baseline	C	—	19.3 / 41.9	18.7 / 37.6	44.3 / 63.1	35.1 / 55.8
Textual	C	34.6 / 51.7	28.5 / 49.2	23.2 / 43.8	**50.3** / 67.0	**43.0 / 62.5**
Textual (+ Task2)	U	36.6 / 53.0	28.5 / 45.7	24.1 / 40.7	—	—
Textual (+ additional)	U	**36.8 / 53.1**	**31.1 / 51.0**	**26.6 / 46.0**	—	—
Multimodal (concat. attn)	C	32.3 / 50.0	23.6 / 41.8	20.0 / 37.1	40.3 / 56.3	32.8 / 52.1
Multimodal (hier. attn.)	C	31.9 / 49.4	25.8 / 47.1	22.4 / 42.7	49.9 / **67.2**	42.9 / **62.5**
Multimodal (concat. attn.)	U	**36.0 / 52.1**	26.3 / 43.9	23.3 / 39.8	—	—
Multimodal (hier. attn.)	U	34.4 / 51.7	**29.5 / 50.2**	**25.7 / 45.6**	—	—
Task 1 winner (LIUM-CVC)	C	—	33.4 / 54.0	28.7 / 48.9	55.9 / 72.1	45.9 / 65.9

Table 3: Results of Task 1 in BLEU / METEOR points. 'C' denotes constrained configuration, 'U' unconstrained, '2016' is the 2016 test set, 'Flickr' and 'MSCOCO' denote the 2017 test sets. The two unconstrained textual models differ in using the additional textual data, which was not used for the training of the multimodal systems.

		Task 2
Baseline	C	**9.1 / 23.4**
Bilingual captioning	C	2.3 / 17.6
en captioning + translation	C	4.2 / 22.1
en captioning + translation	U	6.5 / 20.6
other participant	C	**9.1** / 19.8

Table 4: Results of Task 2 in BLEU / METEOR points.

	Flickr30k
Xu et al. (2015)	**19.1** / 18.5
ours: Flickr30k	15.3 / **18.7**
ours: Flickr30k + MSCOCO	17.9 / 16.6

Table 5: Results of the English image captioning systems on Flickr30k test set in BLEU / METEOR points

In Task 2, none of the submitted systems outperformed the baseline which was a captioning system (Xu et al., 2015) trained directly on the German captions in the Multi30k dataset. The results of our systems on Task 2 are shown in Table 4.

For the English captioning, we trained two models. First one was trained on the Flickr30k data only. In the second one, we included also the MSCOCO dataset. Although the captioning system trained on more data achieved better performance on the English side (Table 5), it led to extremely low performance while plugged into our multimodal translation systems (Table 4, rows labeled "en captioning + translation"). We hypothesize this is caused by the different styles of the sentences in the training datasets.

Our hypothesis about sharing information between the languages in a single decoder was not confirmed in this setup and the experiments led to relatively poor results.

Interestingly, our systems for Task 2 scored poorly in the BLEU score and relatively well in the METEOR score. We can attribute this to the fact that unlike BLEU which puts more emphasis on precision, METEOR considers strongly also recall.

5 Negative Results

In addition to our submitted systems, we tried a number of techniques without success. We describe these techniques since we believe it might be relevant for future developments in the field, despite the current negative result.

5.1 Beam Rescoring

Similarly to Lala et al. (2017), our oracle experiments on the validation data showed that rescoring of the decoded beam of width 100 has the potential of improvement of up to 3 METEOR points. In the oracle experiment, we always chose a sentence with the highest sentence-level BLEU score. Motivated by this observation, we conducted several experiments with beam rescoring.

We trained a classifier predicting whether a given sentence is a suitable caption for a given image. The classifier had one hidden layer with 300 units and had two inputs: the last layer of the VGG-16 network processing the image, and

the last state of a bidirectional GRU network processing the text. We used the same hyperparameters for the bidirectional GRU network as we did for the textual encoders in other experiments. Training data were taken from both parts of the Multi30k dataset with negative examples randomly sampled from the dataset, so the classes were represented equally. The classifier achieved validation accuracy of 87% for German and 74% for French. During the rescoring of the 100 hypotheses in the beam, we selected the one which had the highest predicted probability of being the image's caption.

In other experiments, we tried to train a regression predicting the score of a given output sentence. Unlike the previous experiment, we built the training data from scored hypotheses from output beams obtained by translating the training part of the Multi30k dataset. We tested two architectures: the first one concatenates the terminal states of bidirectional GRU networks encoding the source and hypothesis sentences and an image vector; the second performs an attentive average pooling over hidden states of the RNNs and the image CNN using the other encoders terminal states as queries and concatenates the context vectors. The regression was estimating either the sentence-level BLEU score (Chen and Cherry, 2014) or the chrF3 score (Popović, 2015).

Contrary to our expectations, all the rescoring techniques decreased the performance by 2 METEOR points.

5.2 Reinforcement Learning

Another technique we tried without any success was self-critical sequence training (Rennie et al., 2016). This modification of the REINFORCE algorithm (Williams, 1992) for sequence-to-sequence learning uses the reward of the training-time decoded sentence as the baseline. The systems were pre-trained with the word-level cross-entropy objective and we hoped to fine-tune the systems using the REINFORCE towards sentence-level BLEU score and GLEU score (Wu et al., 2016).

It appeared to be difficult to find the right moment when the optimization criterion should be switched and to find an optimal mixing factor of the cross-entropy loss and REINFORCE loss. We hypothesize that a more complex objective mixing strategy (like MIXER (Ranzato et al., 2015)) could lead to better results than simple objective weighting.

6 Conclusions

In our submission to the 2017 Multimodal Task, we tested the advanced attention combination strategies (Libovický and Helcl, 2017) in a more challenging context and achieved competitive results compared to other submissions. We explored ways of acquiring additional data for the task and tested two promising techniques that did not bring any improvement to the system performance.

Acknowledgments

This research has been funded by the Czech Science Foundation grant no. P103/12/G084, the EU grant no. H2020-ICT-2014-1-645452 (QT21), and Charles University grant no. 52315/2014 and SVV project no. 260 453.

This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

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.

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/W16-2358.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Incorporating global visual features into attention-based neural machine translation. *CoRR* abs/1701.06521. http://arxiv.org/abs/1701.06521.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA, pages 362–367. http://www.aclweb.org/anthology/W14-3346.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. Multimodal compact bilinear pooling for multimodal neural machine translation. *CoRR* abs/1703.08084. http://arxiv.org/abs/1703.08084.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Edinburgh, United Kingdom, pages 85–91. http://www.aclweb.org/anthology/W11-2107.

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. http://arxiv.org/abs/1605.00459.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *CoRR* abs/1705.04350. http://arxiv.org/abs/1705.04350.

Gertrud Faaß and Kerstin Eckart. 2013. Sdewac–a corpus of parsable sentences from the web. In *Language processing and knowledge in the Web*, Springer, pages 61–68.

Orhan Firat and Kyunghyun Cho. 2016. Conditional gated recurrent unit with attention mechanism. https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf. Published online, version `adbaeea`.

Jindřich Helcl and Jindřich Libovický. 2017. Neural monkey: An open-source tool for sequence learning. *The Prague Bulletin of Mathematical Linguistics* (107):5–17. https://doi.org/10.1515/pralin-2017-0001.

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*. Association for Computational Linguistics, Berlin, Germany, pages 639–645. http://www.aclweb.org/anthology/W/W16/W16-2360.

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

Chiraag Lala, Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2017. Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. *The Prague Bulletin of Mathematical Linguistics* (108):197–208. https://doi.org/doi: 10.1515/pralin-2017-0020.

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)*. Association for Computational Linguistics, Vancouver, Canada.

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*. Association for Computational Linguistics, Berlin, Germany, pages 646–654. http://www.aclweb.org/anthology/W/W16/W16-2361.

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. http://arxiv.org/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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

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. https://doi.org/10.1007/s11263-016-0965-7.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Lisbon, Portugal, pages 392–395. http://aclweb.org/anthology/W15-3049.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *CoRR* abs/1511.06732. http://arxiv.org/abs/1511.06732.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. *CoRR* abs/1612.00563. http://arxiv.org/abs/1612.00563.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain, pages 65–68. http://aclweb.org/anthology/E17-3017.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 86–96. http://www.aclweb.org/anthology/P16-1009.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556. http://arxiv.org/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)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

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)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 1374–1383. http://www.aclweb.org/anthology/P13-1135.

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*. Association for Computational Linguistics, Berlin, Germany, pages 543–553. http://www.aclweb.org/anthology/W16-2346.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

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. http://arxiv.org/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 David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. JMLR Workshop and Conference Proceedings, Lille, France, pages 2048–2057. http://jmlr.org/proceedings/papers/v37/xuc15.pdf.