Attention-based Multimodal Neural Machine Translation

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh[†], Chris Dyer
Language Technologies Institute, Robotics Institute[†]
Carnegie Mellon University
Pittsburgh, PA, USA
{poyaoh|fliu1|sshiang|cdyer}@cs.cmu.edu, jeanoh@nrec.ri.cmu.edu[†]

Abstract

We present a novel neural machine translation (NMT) architecture associating visual and textual features for translation tasks with multiple modalities. Transformed global and regional visual features are concatenated with text to form attendable sequences which are dissipated over parallel long short-term memory (LSTM) threads to assist the encoder generating a representation for attention-based decoding. Experiments show that the proposed NMT outperform the text-only baseline.

1 Introduction

In fields of machine translation, neural network attracts lots of research attention recently that the encoder-decoder framework is widely used. Nevertheless, the main drawback of this neural machine translation (NMT) framework is that the decoder only depends on the last state of the encoder, which may deteriorate the performance when the sentence is long. To overcome this problem, attention based encoder-decoder framework as shown in Figure 1 is proposed. With the attention model, in each time step the decoder depends on both the previous LSTM hidden state and the context vector, which is the weighted sum of the hidden states in the encoder. With attention, the decoder can “refresh” it’s memory to focus on source words that may help to translate the correct words rather than only seeing the last state of the sentences where the words in the sentence and the ordering of words are missing.

Most of the machine translation task only focus textual sentences of the source language and target language; however, in the real world, the sentences may contain information of what people see. Beyond the bilingual translation, in WMT 16’ multimodal translation task, we would like to translate the image captions in English into German. With the additional information from images, we would further resolve the problem of ambiguity in languages. For example, the word “bank” may refer to the financial institution or the land of the river’s edge. It would be confusing if we only look at the language itself. In this task, the image may help to disambiguate the meaning if it shows that there is a river and thus the “bank” means “river bank”.

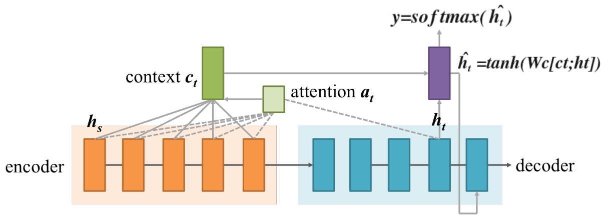

Figure 1: Attention-based neural machine translation framework using a context vector to focus on a subset of the encoding hidden states.

In this paper, we explore approaches to integrating multimodal information (text and image) into the attention-based encoder-decoder architecture. We transform and make the visual features as one of the steps in the encoder as text, and then make it possible to attend to both the text and the image while decoding. The image features we used are (visual) semantic features extracted from the entire images (global) as well as the regional bounding boxes proposed by the region-based convolutional neural networks (R-CNN) (Girshick et al., 2014). In the following section, we first describe the related works, and then we introduce the proposed multimodal attention-based NMT in Section 3, followed by re-scoring of the translation candidates in Section 4. Finally we demonstrate the experiments in Section 5.

2 Related Work

As the advances of deep learning, Neural Machine Translation (NMT) (Kalchbrenner and Blunsom,

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 639–645,
Berlin, Germany, August 11-12, 2016. ©2016 Association for Computational Linguistics

2013; Jean et al., 2014) leveraging encode-decoder architecture attracts research attention. Under the NMT framework, less domain knowledge is required and large training corpora can compensate for it. However, encoder-decoder structure encodes the source sentence into one fixed-length vector, which may deteriorate the translation performance as the length of source sentences increasing. (Bahdanau et al., 2014) extended encoder-decoder structure that the decoder only focuses on parts of source sentence. (Luong et al., 2015) further proposed attention-based model that combine global, attending to all source words, and local, only focusing on a part of source words, attentional mechanism.

Rather than using the embedding of each modality independently, Some works (Hardoon et al., 2004; Andrew et al., 2013; Ngiam et al., 2011; Srivastava and Salakhutdinov, 2014) focus on learning joint space of different modalities. In machine translation fields, (Zhang et al., 2014; Su et al., 2015) learned phrase-level bilingual representation using recursive auto-encoder. Beyond textual embedding, (Kiros et al., 2014) proposed CNN-LSTM encoder to project two modalities into the same space. Based on the jointly learning of multiple modalities or languages, we find it possible to evaluate the quality of the translations that if the space of the translated sentence is similar to the source sentence or the image, it may imply that the translated sentence is good.

3 Attention-based Multimodal Machine Translation

Based on the encoder-decoder framework, the attention-based model aim to handle the missing order and source information problems in the basic encoder-decoder framework. At each time step t in the decoding phrase, the attention-based model attends to subsets of words in the source sentences that can form up the context which can help the decoder to predict the next word. This model infers a variable-length alignment weight vector $\mathbf{a_t}$ based on the current target state $\mathbf{h}_t$ and all source states $\mathbf{h}_s$. The context feature vector $\mathbf{c}_t = \mathbf{a}_t \cdot \mathbf{h}_s$ is the weighted sum of the source states $\mathbf{h}_s$ according to $\mathbf{a_t}$, which is defined as:

$$\mathbf{a}_t(s) = \frac{e^{score(\mathbf{h}_t, \mathbf{h}_s)}}{\sum_s' e^{score(\mathbf{h}_t, \mathbf{h}_s')}} \tag{1}$$

The scoring function $score(\mathbf{h}_t, \mathbf{h}_s)$ can be referred as a content-based measurement of the similarity between the currently translating target and the source words. We utilize a transformation matrix $\mathbf{W}_a$ which associates source and target hidden state to learn the general similarity measure by:

$$score(\mathbf{h}_t, \mathbf{h}_s) = \mathbf{h}_t \mathbf{W}_a \mathbf{h}_s \tag{2}$$

We produce an attentional hidden state $\hat{\mathbf{h}}_t$ by learning $\mathbf{W}_c$ of a single layer perceptron activated by $tanh$. The input is simply the concatenation of the target hidden state $\mathbf{h}_t$ and the source-side context vector $\mathbf{c}_t$:

$$\hat{\mathbf{h}}_t = tanh(\mathbf{W}_c[\mathbf{c}_t; \mathbf{h}_t]) \tag{3}$$

After generating the context feature vector and the attentional hidden state, the target word is predicted through the softmax layer with the attentional hidden state $\mathbf{h}_t$ vector by $p(y_t|\mathbf{x}) = softmax(\mathbf{W}_s \hat{\mathbf{h}}_t)$. The following we will introduce how we incorporate images features based on the attention models.

3.1 Model 1: LSTM with global visual feature

Visual features from convolution neural network (CNN) may provide additional information to textual features in machine translation with multiple modalities. As depicted in Figure 2, we propose to append visual features at the head/tail to the original text sequence in the encoding phase. Note that for simplicity, we omit the attention part in the following figures.

Global (i.e., whole image) visual feature are extracted from the last fully connected layer known as $fc7$, a 4096-dimensional semantic layer in the 19-layered VGG (Simonyan and Zisserman, 2014). With the dimension mismatch and the inherent difference in content between the visual and textual embedding, a transformation matrix $\mathbf{W}_{img}$ is proposed to learn the mapping. The encoder then encode both textual and visual feature sequences to generate the representation for decoding. In the decoding phase, the attention model weights all the possible hidden states in the encoding phase and produce the context vector $\mathbf{c}_t$ with Eq. 1 and Eq. 2 for NMT decoding.

3.2 Model 2: LSTM with multiple regional visual features

In addition to adding only one global visual feature, we extend the original NMT model by incorporating multiple regional features in the hope

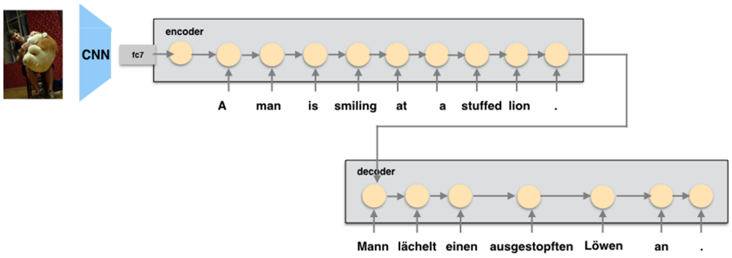

Figure 2: Model 1: Attention-based NMT with single additional global visual feature. Decoder may attend to both text and image steps of encoding. For clarity, the possible attention path is hidden here.

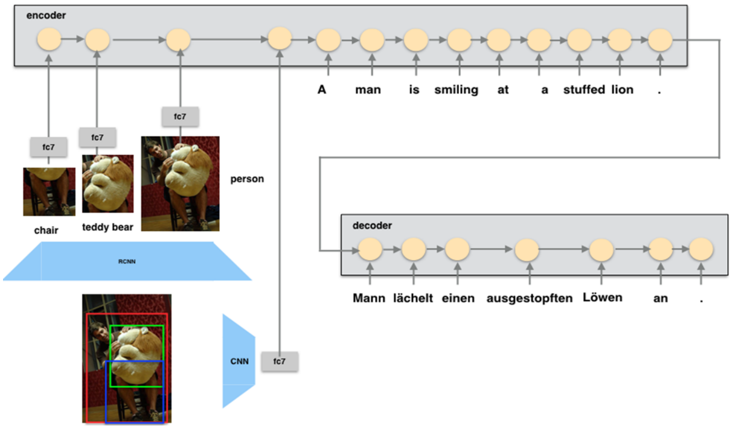

Figure 3: Model 2: Attention-based NMT with multiple additional regional visual features.

that those regional visual attributes would assist LSTM to generate better and more accurate representations. The illustration of the proposed model is depicted in 3. We will first explain how to determine multiple regions from one image and explain how these visual features are extracted and sorted.

Intuitively, objects in an image are most likely to appear in both source and target sentences. Therefore. we utilize the region proposal network (RPN) in the region-based convolutional neural network (Ren et al., 2015) (R-CNN) to identify objects and their bounding boxes in an image and then extract visual feature from those regions. In order to integrate these images to the original sequence in the LSTM model, we design a heuristic approach to sort those visual features. The regional features are fed in the ascending order of the size of the bounding boxes; followed by the original global visual feature and the text sequence. Visual features are sequentially fed in such order since important features are designed to be closer to the encoded representation. Heuristically, larger objects may be more noticeable and essential in an image described by both the source and target language contexts.

In the implementation, we choose top 4 regional objects plus the whole image and then extracted their $fc7$ with VGG-19 to form the visual sequence followed by the text sequence. If there are less than 4 objects recognized in the original image, zero vectors are padded instead for the batch process during training.

3.3 Model 3: Parallel LSTM threads

To further alleviate the assumption that regional objects share some pre-defined order, we further propose a parallel structure as shown in Figure 4. The encoder of NMT is composed of multiple encoding threads where all the LSTM parameters are shared. In each thread, a (regional) visual feature is followed by the text sequence. This parallel structure would associate the text to the most relevant objects in the encoding phase and distinguish them when computing attention during decoding. Intuitively, the text sequence follows a regional object would be interpreted as encoding the visual information with the textual description (i.e., encoding captions as well as visual features for that object). An encoder hidden state for attention can be interpreted as the “word” imprinted by the semantics features of some regional object. The decoder can therefore distinctively attend to

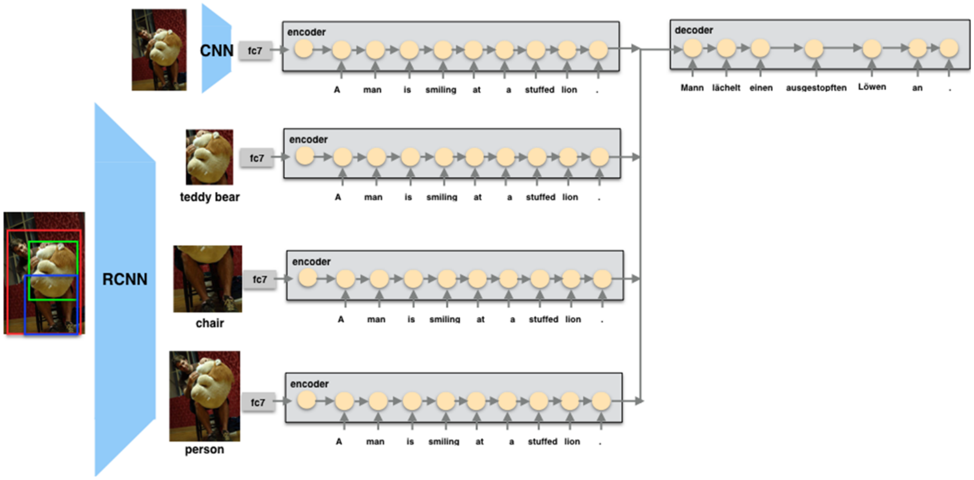
CNN
fc7
encoder
A man is smiling at a stuffed lion .
decoder
Mann lächelt einen ausgestopften Löwen an .
RCNN
teddy bear
chair
person

Figure 4: Model 3: Parallel LSTM threads with multiple additional regional visual features.

words that describe different visual objects in multiple threads.

In the encoding phase, parameters in LSTM are shared over threads. All possible hidden states over multiple threads are recorded for attention. At the end of encoding phase, the outputs of different encoding threads are fused together to generate the final embedding of the whole sentence as well as all the image objects. In the decoding phase, candidates of global attention are all the text hidden states over multiple threads. For example, at time t, the decoder may choose to attend to ‘bear’ at the second thread (which sees a teddy bear image at the beginning) as well as the ’bear’ in the global image thread. At time $t+1$, the decoder may switch to another thread and focus on “the man” with the person image.

For implementation simplicity for batch training, we limit the number of regional objects to 4 and add one global image thread. We also choose an average pooling in the encoder fusion process and back-propagate accordingly.

4 Re-scoring of Translation Candidates

In the neural machine translation, the easiest way to decode is to greedily get the words with highest probability step-by-step. To achieve better performance, ensemble of models are required. Translation candidates are generated from multiple models, and we aim to figure out which candidate should be the best one. The following we describe the approaches we investigated to re-score the translation candidates using monolingual and bilingual information.

4.1 Monolingual Re-scoring

To evaluate the quality of the translation, the most simple approach is to check whether the translated sentences are readable. To achieve this, using language model is an effective way to check whether the sentences fit into the model that trained on a large corpus. If the language model score is high, it implies that the sentence holds the high probability to be generated from the corpus. We trained a single layer LSTM with 300 hidden state to predicting the next word. Image caption datasets MSCOCO and IAPR TC-12 (overall 56,968 sentences) are used as training data.

4.1.1 Bilingual autoencoder

A good translation would also recognize the sentence in the source language. We utilize bilingual autoencoder (Ngiam et al., 2011) depicted as in Fig.5 to reconstruct source language given the source language. Bilingual autoencoder only uses single modality (here we used source language or target language) and re-constructs the both modalities. We project bilingual information into the joint space (the bottleneck layer); if the two target and source sentences have similar representation, the model is able to reconstruct both sentences. Moreover, if the similarity of values of bottleneck layer is high, it may indicate that the source sentence and the translated sentence are similar in concepts; therefore, the quality of the translation would be better. The inputs of the autoencoder are the last LSTM encoder states trained on monolingual image captions dataset. The dimension of the input layer is 256, and 200 for the middle, and 128 for the joint layer.

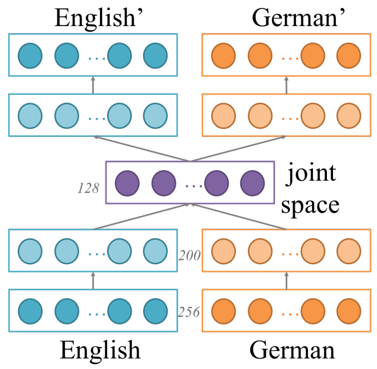

Figure 5: Bilingual auto-encoder to re-construct both English and German using only one of them.

4.2 Bilingual dictionary

In the WMT 16' multimodal task, captions are structured with simple grammars; therefore, only considering language model may be insufficient to distinguish good translations. In order to directly consider whether the concepts mentioned in the source sentences are all well-translated, we utilize the bilingual dictionary Glosbe[1], in which we use the words in one language extracting the corresponding words in the other language. We directly count the number of words in the source language that the synonyms in target language are also in the translated results as the re-ranking score.

5 Experiments

5.1 Experimental Setup

In the official WMT 2016 multimodal translation task dataset (Elliott et al., 2016), there are 29,000 parallel sentences from English to German for training, 1014 for validation and 1000 for testing. Each sentence describes an image from Flickr30k dataset (Young et al., 2014). We preprocessed all the descriptions into lower case with tokenization and German compound word splitting.

Global visual features (*fc7*) are extracted with VGG-19 (Simonyan and Zisserman, 2014). For regional visual features, the region proposal network in RCNN (Girshick et al., 2014) first recognizes bounding boxes of objects in an image and then we computed 4096-dimensional *fc7* features from these regions with VGG-19. The RPN of RCNN is pre-trained on ImageNet dataset [2] and then fine-tuned on MSCOCO dataset [3] with 80 object classes.

Table 1: BLEU and METEOR of the proposed multimodal NMT

	BLEU	METEOR
Text baseline	34.5 (0.7)	51.8 (0.7)
m1:image at tail	34.8 (0.6)	51.6 (0.7)
m1:image at head	35.1 (0.8)	52.2 (0.7)
m2:5 sequential RCNNs	36.2 (0.8)	53.4 (0.6)
m3:5 parallel RCNNs	**36.5** (0.8)	**54.1** (0.7)

We use a single-layered LSTM with 256 cells and 128 batch size for training. The dimension of word embedding is 256. $\mathbf{W}_{img}$ is a 4096×256 matrix transforming visual features into the same embedding space as words. When training NMT, we follow (Luong et al., 2015) with similar settings: (a) we uniformly initialized all parameters between -0.1 and 0.1, (b) we trained the LSTM for 20 epochs using simple SGD, (c) the learning rate was initialized as 1.0, multiplied by 0.7 after 12 epochs, (d) dropout rate was 0.8. Note that the same dropout mask and NMT parameters are shared by all LSTM threads in model 3.

5.2 Results of Adding Visual Information

The quantitative performance of the proposed models can be seen in Table 1. We evaluate BLEU and METEOR scores with tokenization under the official settings of WMT 2016 multimodal machine translation challenge. The text-only baseline is the NMT implementation with global attention. Adding single global visual feature from an image at the head of a text sequence improves BLEU by 0.6% and METEOR by 0.4% respectively.

The results show that the additional visual information improves the translations in this dataset. However, the lukewarm improvement is not as significant as we expected. One possible explanation is that the information required for the multimodal translation task is mostly self-contained in the source text transcript. Adding global features from whole images do not provide extra supplementary information and thus results in a subtle improvement.

Detailed regional visual features provide extra attributes and information that may help the NMT translates better. In our experiment, the proposed model2 with multiple regional and one global visual features showed an improvement of 1.7% in BLEU and 1.6% in METEOR while model3

[1] https://glosbe.com/en/de/
[2] http://image-net.org/
[3] http://mscoco.org/

showed an improvement of 2.0% in BLEU and 2.3% in METEOR. The results correspond to our observation that most sentences would describe important objects which could be identified by R-CNN. The most commonly mentioned object is "person". It's likely that the additional attributes provided by the visual features about the person in an image help to encode more detailed context and thus benefit NMT decoding. Other high frequency objects are "car", "baseball", "cellphone", etc.

For the proposed LSTM with multiple regional visual features (model 2), the semantic features in $fc7$ of the regions-of-interest in an image provide additional regional visual information to form a better sentence representation. We also experimented other sorting methods including descending size, random, and categorical order to generate the visual sequences. However, ascending-ordered sequences achieve the best result.

For the proposed parallel LSTM architecture with regional visual features (model 3), the regional visual features further help the NMT decoder to attend more accurately and accordingly to focus on the right thread where the hidden states are twiddle by the local visual attributes. The best result of our models achieve 36.5% in BLEU and 54.1% in METEOR, which is comparable to the state-of-the-art Moses results in this challenge.

5.3 Results of Re-Scoring

The experimental results of re-scoring are shown in table 2, we compare our re-scoring methods based on the candidates generated by our best multimodal NMT modal (model 3). The second row is the results using LSTM monolingual language model with hidden size as 300. The reason why we can barely achieve improvement might be that the grammar in the caption task is much easier compared to other translation tasks such as dialog or News; therefore, the candidate sentences with low score of evaluation (BLEU or METEOR) may also looks like a sentence, but without relevance to the source sentence.

The third row shows the re-scoring results with the bi-lingual autoencoder. This approach results in drops in both BLEU and METEOR. The reason might be that the quality and quantity of our Bi-lingual corpus is insufficient for the purpose of learning a good autoencoder. Furthermore, we observe the test perplexity is higher than the training and validation perplexity, showing the over-fitting in language modeling and the effects of unknown words. It's clear that more investigation is required for designing a better bilingual autoencoder for re-scoring.

Table 2: Results of re-scoring using monolingual LSTM, Bi-lingual auto-encoder, and dictionary based on multimodal NMT results.

	BLEU	METEOR
Original Model 3	**36.5** (0.8)	54.1 (0.7)
Language model	36.3 (0.8)	53.3 (0.6)
Bilingual autoencoder	35.9 (0.8)	53.4 (0.7)
Bilingual dictionary	35.7 (0.8)	**55.2** (0.6)

The last row shows the results using the bilingual dictionary. For each word in the source sentence and the target candidates, we retrieve the term and the translation in the other language, and count the number of matching. We can achieve much more improvement on METEOR compared to other methods. This is because that the quality of the translation of captions depends on how much we correctly translate the objects and their modifiers. The bad translation can still achieve fair performance without re-scoring because the sentence structure is similar to good translation. For example, a lot of sentences start with "A man" and both good and bad translation can also translate the sentences start with "Ein Mann". The bilingual dictionary is proved to be an efficient re-scoring approach to distinguish these cases.

6 Conclusions

We enhanced the attention-based neural machine translation by incorporating information in multiple modalities. We explored different encoder-decoder architectures including the LSTM with multiple sequential global/regional visual and textual features as states for attention and the parallel LSTM threads approach. Our best model achieved 2.0% improvement in BLEU score and 2.3% in METEOR using the visual features of an entire image and interesting regional objects within. For re-scoring translation candidates, we investigated monolingual LSTM language model, bilingual autoencoder, and bilingual dictionary re-scoring. We further achieved an additional 1.1% improvements in METEOR using a bilingual dictionary. Integration of more modalities such as audio would be a challenging but interesting next step.

Acknowledgments

This material is based in part on work supported by the National Science Foundation (NSF) under grant number IIS-1638429. Also, this work was conducted in part through collaborative participation in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016, and in part by ONR under MURI grant "Reasoning in Reduced Information Spaces" (no. N00014-09-1-1052). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Galen Andrew, Raman Arora, Karen Livescu, and Jeff Bilmes. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, Atlanta, Georgia.

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

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. *CoRR*, abs/1605.00459.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*.

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, December.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. *CoRR*, abs/1412.2007.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. Seattle, October. Association for Computational Linguistics.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

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

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 689–696. Omnipress.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*.

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

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980.

Jinsong Su, Deyi Xiong, Biao Zhang, Yang Liu, Junfeng Yao, and Min Zhang. 2015. Bilingual correspondence recursive autoencoder for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1248–1258.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 111–121, Baltimore, Maryland, June. Association for Computational Linguistics.