Does Multimodality Help Human and Machine for Translation and Image Captioning?

Ozan Caglayan[1,3], Walid Aransa[1], Yaxing Wang[2], Marc Masana[2], Mercedes García-Martínez[1], Fethi Bougares[1], Loïc Barrault[1] and Joost van de Weijer[2]
[1] LIUM, University of Le Mans, [2] CVC, Universitat Autonoma de Barcelona,
[3] Galatasaray University
[1]FirstName.LastName@lium.univ-lemans.fr
[2]{joost,mmasana,yaxing}@cvc.uab.es
[3]ocaglayan@gsu.edu.tr

Abstract

This paper presents the systems developed by LIUM and CVC for the WMT16 Multimodal Machine Translation challenge. We explored various comparative methods, namely phrase-based systems and attentional recurrent neural networks models trained using monomodal or multimodal data. We also performed a human evaluation in order to estimate the usefulness of multimodal data for human machine translation and image description generation. Our systems obtained the best results for both tasks according to the automatic evaluation metrics BLEU and METEOR.

1 Introduction

Recently, deep learning has greatly impacted the natural language processing field as well as computer vision. Machine translation (MT) with deep neural networks (DNN), proposed by (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) and (Bahdanau et al., 2014) competed successfully in the last year's WMT evaluation campaign (Bojar et al., 2015).

In the same trend, generating descriptions from images using DNNs has been proposed by (Elliott et al., 2015). Several attempts have been made to incorporate features from different modalities in order to help the automatic system to better model the task at hand (Elliott et al., 2015; Kiros et al., 2014b; Kiros et al., 2014a).

This paper describes the systems developed by LIUM and CVC who participated in the two proposed tasks for the WMT 2016 Multimodal Machine Translation evaluation campaign: Multimodal machine translation (Task 1) and multimodal image description (Task 2).

The remainder of this paper is structured in two parts: The first part (section 2) describes the architecture of the four systems (two monomodal and two multimodal) submitted for Task 1. The standard phrase-based SMT systems based on Moses are described in section 2.1 while the neural MT systems are described in section 2.2 (monomodal) and section 3.2 (multimodal). The second part (section 3) contains the description of the two systems submitted for Task 2: The first one is a monomodal neural MT system similar to the one presented in section 2.2, and the second one is a multimodal neural machine translation (MNMT) with shared attention mechanism.

In order to evaluate the feasibility of the multimodal approach, we also asked humans to perform the two tasks of this evaluation campaign. Results show that the additional English description sentences improved performance while the straightforward translation of the sentence without the image did not provide as good results. The results of these experiments are presented in section 4.

2 Multimodal Machine Translation

This task consists in translating an English sentence that describes an image into German, given the English sentence itself and the image that it describes.

2.1 Phrase-based System

Our baseline system for task 1 is developed following the standard phrase-based Moses pipeline as described in (Koehn et al., 2007), SRILM (Stolcke, 2002), KenLM (Heafield, 2011), and GIZA++ (Och and Ney, 2003). This system is trained using the data provided by the organizers and tuned using MERT (Och, 2003) to maximize BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) scores on the validation set.

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 627–633,
Berlin, Germany, August 11-12, 2016.

We also used Continuous Space Language Model[1] (CSLM) (Schwenk, 2010) with the auxiliary features support as proposed by (Aransa et al., 2015). This CSLM architecture allows us to use sentence-level features for each line in the training data (i.e. all n-grams in the same sentence will have the same auxiliary features). By this means, better context specific LM estimations can be obtained.

We used four additional scores to rerank 1000-best outputs of our baseline system: The first two scores are obtained from two separate CSLM(s) trained on the target side (i.e. German) of the parallel training corpus and each one of the following auxiliary features:

- **VGG19-FC7 image features:** The auxiliary feature used in the first CSLM are the image features provided by the organizers which are extracted from the FC7 layer (relu7) of the VGG-19 network (Simonyan and Zisserman, 2014). This allows us to train a multimodal CSLM that uses additional context learned from the image features.

- **Source side sentence representation vectors:** We used the method described in (Le and Mikolov, 2014) to compute continuous space representation vector for each source (i.e. English) sentence that will be provided to the second CSLM as auxiliary feature. The idea behind this is to condition our target language model on the source side as additional context.

The two other scores used for n-best reranking are the log probability computed by our NMT system that will be described in the following section and the score obtained by a Recurrent Neural Network Language Model (RNNLM) (Mikolov et al., 2010). The weights of the original moses features and our additional features were optimized to maximize the BLEU score on the validation set.

2.2 Neural MT System

The fundamental model that we experimented[2] is an attention based encoder-decoder approach (Bahdanau et al., 2014) except some notable changes in the recurrent decoder called Conditional GRU.

[1] github.com/hschwenk/cslm-toolkit
[2] github.com/nyu-dl/dl4mt-tutorial

We define by X and Y, a source sentence of length N and a target sentence of length M respectively. Each source and target word is represented with an embedding vector of dimension E_X and E_Y respectively:

$$X = (x_1, x_2, ..., x_N), x_i \in \mathbb{R}^{E_X} \quad (1)$$

$$Y = (y_1, y_2, ..., y_M), y_j \in \mathbb{R}^{E_Y} \quad (2)$$

A bidirectional recurrent encoder reads an input sequence X in forwards and backwards to produce two sets of hidden states based on the current input and the previous hidden state. An annotation vector h_i for *each* position i is then obtained by concatenating the produced hidden states.

An attention mechanism, implemented as a simple fully-connected feed-forward neural network, accepts the hidden state h_t of the decoder's recurrent layer and one input annotation at a time, to produce the attention coefficients. A softmax activation is applied on those attention coefficients to obtain the attention weights used to generate the weighted annotation vector for time t. The initial hidden state h_0 of the decoder is determined by a feed-forward layer receiving the mean annotation vector.

We use Gated Recurrent Unit (GRU) (Chung et al., 2014) activation function for both recurrent encoders and decoders.

2.2.1 Training

We picked the following hyperparameters for all NMT systems both for Task 1 and Task 2. All embedding and recurrent layers have a dimensionality of 620 and 1000 respectively. We used Adam as the stochastic optimizer with a minibatch size of 32, Xavier weight initialization (Glorot and Bengio, 2010) and L2 regularization with $\lambda = 0.0001$ except the monomodal Task 1 system for which the choices were Adadelta, sampling from $\mathcal{N}(0, 0.01)$ and L2 regularization with $\lambda = 0.0005$ respectively.

The performance of the network is evaluated on the validation split using BLEU after each 1000 minibatch updates and the training is stopped if BLEU does not improve for 20 evaluation periods. The training times were 16 and 26 hours respectively for monomodal and multimodal systems on a Tesla K40 GPU.

Finally, we used a classical left to right beam-search with a beam size of 12 for sentence generation during test time.

2.3 Data

Phrase-based and NMT systems for Task 1 are trained using the dataset provided by the organizers and described in Table 1. This dataset consists of 29K parallel sentences (direct translations of image descriptions from English to German) for training, 1014 for validation and finally 1000 for the test set. We preprocessed the dataset using the punctuation normalization, tokenization and lowercasing scripts from Moses. In order to generalize better over the compound structs in German, we trained and applied a compound splitter[3] (Sennrich and Haddow, 2015) over the German vocabulary of training and validation sets. This reduces the target vocabulary from 18670 to 15820 unique tokens. During translation generation, the splitted compounds are stitched back together.

Side	Vocabulary	Words
English	10211	377K
German	15820	369K

Table 1: Training Data for Task 1.

2.4 Results and Analysis

The results of our phrase-based baseline and the four submitted systems are presented in Table 2. The **BL+4Features** system is the rescoring of the baseline 1000-best output using all the features described in 2.1 while **BL+3Features** is the same but excluding FC7 image features. Overall, we were able to improve test set scores by around 0.4 and 0.8 on METEOR and BLEU respectively over a strong phrase-based baseline using auxiliary features.

Regarding the NMT systems, the monomodal NMT achieved a comparative BLEU score of 32.50 on the test set compared to 33.45 of the phrase-based baseline. The multimodal NMT system that will be described in section 3.2, obtained relatively lower scores when trained using Task 1's data.

3 Multimodal Image Description Generation

The objective of Task 2 is to produce German descriptions of images given the image itself and one or more English descriptions as input.

[3] github.com/rsennrich/wmt2014-scripts

3.1 Visual Data Representation

To describe the image content we make use of Convolutional Neural Networks (CNN). In a breakthrough work, Krizhevsky et al. (Krizhevsky et al., 2012) convincingly show that CNNs yield a far superior image representation compared to previously used hand-crafted image features. Based on this success an intensified research effort started to further improve the representations based on CNNs. The work of Simonyan and Zisserman (Simonyan and Zisserman, 2014) improved the network by breaking up large convolutional features into multiple layers of small convolutional features, which allowed to train a much deeper network. The organizers provide these features to all participants. More precisely they provide the features from the fifth convolutional layer, and the features from the second fully connected layer of VGG-19. Recently, Residual Networks (ResNet) have been proposed (He et al., 2015). These networks learn residual functions which are constructed by adding skip layers (or projection layers) to the network. These skip layers prevent the vanishing gradient problem, and allow for much deeper networks (over hundred layers) to be trained.

To select the optimal layer for image representation we performed an image classification task on a subsection of images from SUN scenes (Xiao et al., 2010). We extract the features from the various layers of ResNet-50 and evaluate the classification performance (Figure 1). The results increase during the first layers but stabilize from Block-4 on. Based on these results and considering that a higher spatial resolution is better, we have selected layer 'res4fx' (end of Block-4, after ReLU) for the experiments on multimodal MT. We also compared the features from different networks on the task of image description generation with the system of Xu et al. (Xu et al., 2015). The results for generating English descriptions (Table 3) show a clear performance improvement from VGG-19 to ResNet-50, but comparable results are obtained when going to ResNet-152. Therefore, given the increase in computational cost, we have decided to use ResNet-50 features for our submission.

3.2 Multimodal NMT System

The multimodal NMT system is an extension of (Xu et al., 2015) and the monomodal NMT system described in Section 2.2.

System Description	Validation Set		Test Set	
	METEOR (norm)	BLEU	METEOR (norm)	BLEU
Phrase-based Baseline (BL)	53.71 (58.43)	35.61	52.83 (57.37)	33.45
BL+3Features	54.29 (58.99)	36.52	53.19 (57.76)	34.31
BL+4Features	54.40 (59.08)	36.63	53.18 (57.76)	34.28
Monomodal NMT	51.07 (54.87)	35.93	49.20 (53.10)	32.50
Multimodal NMT	44.55 (47.97)	28.06	45.04 (48.52)	27.82

Table 2: BLEU and METEOR scores on detokenized outputs of baseline and submitted Task 1 systems. The METEOR scores in parenthesis are computed with `-norm` parameter.

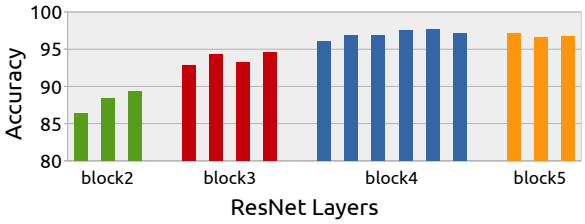

Figure 1: Classification accuracy on a subset of SUN scenes (Xiao et al., 2010) for ResNet-50: The colored groups represent the building blocks while the bars inside are the stacked blocks (He et al., 2015).

Network	BLEU-1	BLEU-2	BLEU-3	BLEU-4
VGG-19	58.2	31.4	18.5	11.3
ResNet-50	68.4	45.2	30.9	21.1
ResNet-152	68.3	44.9	30.7	21.1

Table 3: BLEU scores for various deep features on the image description generation task using the system of Xu et al. (Xu et al., 2015).

The model involves two GRU layers and an attention mechanism. The first GRU layer computes an intermediate representation $s_j^{'}$ as follows:

$$s_j^{'} = (1 - z_j^{'}) \odot \underline{s}_j^{'} + z_j^{'} \odot s_{j-1} \quad (3)$$

$$\underline{s}_j^{'} = \tanh(W^{'} E[y_{j-1}] + r_j^{'} \odot (U^{'} s_{j-1})) \quad (4)$$

$$r_j^{'} = \sigma(W_r^{'} E[y_{j-1}] + U_r^{'} s_{j-1}) \quad (5)$$

$$z_j^{'} = \sigma(W_z^{'} E[y_{j-1}] + U_z^{'} s_{j-1}) \quad (6)$$

where E is the target word embedding, $\underline{s}_j^{'}$ is the hidden state, $r_j^{'}$ and $z_j^{'}$ are the reset and update gate activations. $W^{'}$, $U_r^{'}$, $W_r^{'}$, $U_r^{'}$, $W_z^{'}$ and $U_z^{'}$ are the parameters to be learned.

A shared attention layer similar to (Firat et al., 2016) that consists of a fully-connected feed-forward network is used to compute a set of modality specific attention coefficients e_{ij}^{mod} at each timestep j:

$$e_{ij}^{mod} = U_{att} \tanh(W_{catt} h_i^{mod} + W_{att} s_j^{'}) \quad (7)$$

The attention weight between source modality context i and target word j is computed by applying a softmax on e_{ij}^{mod} :

$$\alpha_{ij} = \frac{\exp(e_{ij}^{txt})}{\sum_{k=1}^{N} \exp(e_{kj}^{txt})} \quad (8)$$

$$\beta_{ij} = \frac{\exp(e_{ij}^{img})}{\sum_{k=1}^{196} \exp(e_{kj}^{img})} \quad (9)$$

The final multimodal context vector c_j is obtained as follows:

$$c_j = \tanh(\sum_{i=1}^{N} \alpha_{ij}\, h_i^{txt} + \sum_{i=1}^{196} \beta_{ij}\, h_i^{img}) \quad (10)$$

The second GRU generates s_j from the intermediate representation $s_j^{'}$ and the context vector c_j as follows:

$$s_j = (1 - z_j) \odot \underline{s}_j + z_j \odot s_j^{'} \quad (11)$$

$$\underline{s}_j = \tanh(W c_j + r_j \odot (U s_j^{'})) \quad (12)$$

$$r_j = \sigma(W_r c_j + U_r s_j^{'}) \quad (13)$$

$$z_j = \sigma(W_z c_j + U_z s_j^{'}) \quad (14)$$

where $\underline{s}_j^{'}$ is the hidden state, r_j and z_j are the reset and update gate activations. W, U_r, W_r, U_r, W_z and U_z are the parameters to be learned.

Finally, in order to compute the target word, the following formulations are applied:

$$o_j = L_o \tanh(E[y_{j-1}] + L_s s_j + L_c c_j) \quad (15)$$

$$P(y_j | y_{j-1}, s_j, c_j) = \text{Softmax}(o_j) \quad (16)$$

where L_o, L_s and L_c are trained parameters.

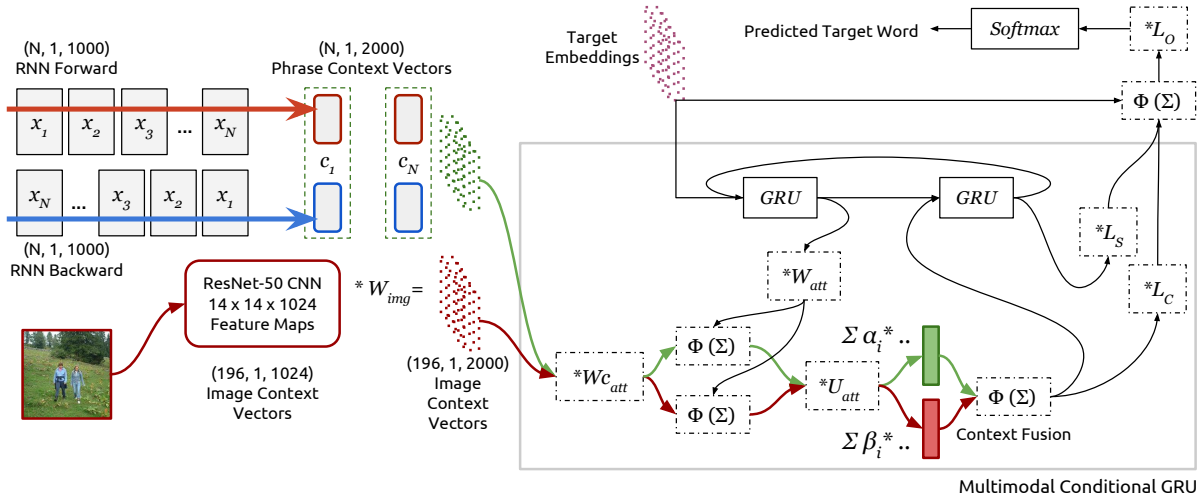

Figure 2: The architecture of the multimodal NMT system. The boxes with $*$ refers to a linear transformation while $\Phi(\Sigma)$ means a $tanh$ applied over the sum of the inputs. The figure depicts a running instance of the network over a single example.

3.2.1 Generation

Since we are provided 5 source descriptions for each image in order to generate a single German description, we let the NMT generate a German description for each source and pick the one with the highest probability and preferably without an UNK token.

3.3 Data

The organizers provided an extended version of the Flickr30K Entities dataset (Elliott et al., 2016) which contains 5 *independently* crowd-sourced German descriptions for each image in addition to the 5 English descriptions originally found in the dataset. It is possible to use this dataset either by considering the cross product of 5 source and 5 target descriptions (a total of 25 description pairs for each image) or by only taking the 5 pairwise descriptions leading to 725K and 145K training pairs respectively. We decided to use the smaller subset of 145K sentences.

Side	Vocabulary	Words
English	16802	1.5M
German	10000	1.3M

Table 4: Training Data for Task 2.

The preprocessing is exactly the same as Task 1 except that we only kept sentence pairs with sentence lengths $\in [3, 50]$ and with a ratio of at most 3. This results in a final training dataset of 131K sentences (Table 4). We picked the most frequent 10K German words and replaced the rest with an UNK token for the target side. Note that compound splitting was not done for this task.

3.4 Results and Analysis

System	Validation		Test	
	METEOR	BLEU	METEOR	BLEU
Monomodal	36.3	24.0	35.1	23.8
Multimodal	34.4	19.3	32.3	19.2

Table 5: BLEU and METEOR scores of our NMT based submissions for Task 2.

As we can see in Table 5, the multimodal system does not surpass monomodal NMT system. Several explanations can clarify this behavior. First, the architecture is not well suited for integrating image and text representations. This is possible as we did not explore all the possibilities to benefit from both modalities. Another explanation is that the image context contain too much irrelevant information which cannot be discriminated by the lone attention mechanism. This would need a deeper analysis of the attention weights in order to be answered.

4 Human multimodal translation and/or description

To evaluate the importance of the different modalities for the image description generation and translation task, we have performed an experiment

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Image + sentences	54.30	35.95	23.28	15.06	39.16
Image only	51.26	34.74	22.63	15.01	38.06
Sentence only	39.37	23.27	13.73	8.40	32.98
Our system	60.61	44.35	31.65	21.95	33.59

Table 6: BLEU and METEOR scores for human translation/description generation experiments.

where we replace the computer algorithm with human participants. The two modalities are the five English description sentences, and the image. The output is a single description sentence in German. The experiment asks the participants for the following tasks:

- Given both the image and the English descriptions: '*Describe the image in one sentence in German. You can get help from the English sentences provided.*'
- Given only the image: '*Describe the image in one sentence in German.*'
- Given only one English sentence: '*Translate the English sentence into German.*'

The experiment was performed by 16 native German speakers proficient in English with age ranging from 23 to 54 (coming from Austria, Germany and Switzerland, of which 10 are female and 6 male). The experiment is performed on the first 80 sentences of the validation set. Participants performed 10 repetitions for each task, and not repeating the same image across tasks. The results of the experiments are presented in Table 6. For humans, the English description sentences help to obtain better performance. Removing the image altogether and providing only a single English description sentence results in a significant drop. We were surprised to observe such a drop, whereas we expected good translations to obtain competitive results. In addition, we have provided the results of our submission on the same subset of images; humans clearly obtain better performance using METEOR metrics, but our approach is clearly outperforming on the BLEU metrics. The participants were not trained on the train set before performing the tasks, which could be one of the reasons for the difference. Furthermore, given the lower performance of only translating one of the English description sentences on both metrics, it could possibly be caused by existing biases in the data set.

5 Conclusion and Discussion

We have presented the systems developed by LIUM and CVC for the WMT16 Multimodal Machine Translation challenge. Results show that integrating image features into a multimodal neural MT system with shared attention mechanism does not yet surpass the performance obtained with a monomodal system using only text input. However, our multimodal systems do improve upon an image captioning system (which was expected). The phrase-based system can benefit from rescoring with multimodal neural language model as well as rescoring with a neural MT system.

We have also presented the results of a human evaluation performing the same tasks as proposed in the challenge. The results are rather clear: image captioning can benefit from multimodality.

Acknowledgments

This work was supported by the Chist-ERA project M2CR[4]. We kindly thank KyungHyun Cho and Orhan Firat for providing the DL4MT tutorial as open source and Kelvin Xu for the arctic-captions[5] system.

References

Walid Aransa, Holger Schwenk, and Loic Barrault. 2015. Improving continuous space language models using auxiliary features. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, pages 151–158, Da Nang, Vietnam, December.

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, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

[4] m2cr.univ-lemans.fr

[5] github.com/kelvinxu/arctic-captions

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

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

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv preprint arXiv:1601.01073*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

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

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014a. Multimodal neural language models. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 595–603. JMLR Workshop and Conference Proceedings.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Meeting of the Association for Computational Linguistics*, pages 177–180.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, volume 2, page 3.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51, March.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA.

Holger Schwenk. 2010. Continuous space language models for statistical machine translation. In *The Prague Bulletin of Mathematical Linguistics, (93):137–146.*

Rico Sennrich and Barry Haddow. 2015. A joint dependency model of morphological and syntactic structure for statistical machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 114–121. Association for Computational Linguistics.

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

Andreas Stolcke. 2002. Srilm - an extensible language modeling toolkit. *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

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

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE.

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*, pages 2048–2057.