

SHEF-Multimodal: Grounding Machine Translation on Images

Kashif Shah, Josiah Wang, Lucia Specia

University of Sheffield

211 Portobello Street, Sheffield, UK

{kashif.shah, j.k.wang, l.specia}
@sheffield.ac.uk

Abstract

This paper describes the University of Sheffield’s submission for the WMT16 Multimodal Machine Translation shared task, where we participated in Task 1 to develop German-to-English and English-to-German statistical machine translation (SMT) systems in the domain of image descriptions. Our proposed systems are standard phrase-based SMT systems based on the Moses decoder, trained only on the provided data. We investigate how image features can be used to re-rank the n -best list produced by the SMT model, with the aim of improving performance by grounding the translations on images. Our submissions are able to outperform the strong, text-only baseline system for both directions.

1 Introduction

This paper describes the University of Sheffield’s submission for a new WMT16 Multimodal Machine Translation shared task. The task is aimed at the generation of image descriptions in a target language, given an image and one or more descriptions in a different (source) language. We participated in Task 1, which takes a source language description and translates it into the target language, supported by information from images. We submitted systems for the translation between English and German in both directions.

Multimodal approaches for various applications related to language processing have been gaining wider attention from the research community in recent years. The main motivation is to investigate whether contextual information from various sources can be helpful in improving system performance. Multimodal approaches have been ex-

plored in various tasks such as image and video description, as well as question answering about images (see Section 4). However, not much work has been done to explore multimodality in the context of machine translation. Whilst a large number of approaches have been developed to improve translation quality, they concern solely textual information. The use of non-textual sources such as images and speech has been largely ignored partially because of the lack of datasets and resources. This shared task provides an interesting opportunity to investigate the effectiveness of information from images in improving the performance of machine translation systems.

The main objective of our proposed system is to explore how image features can be used to re-rank an n -best list of translations from a standard phrase-based Statistical Machine Translation (SMT) system. This is in contrast to existing work (Elliott et al., 2015) that uses image features jointly with image descriptions to train a Neural Network-based translation model. The dataset provided for this shared task contains short segments with simple grammar and repetitive vocabulary. Therefore, it is expected that a standard phrase-based SMT system can already produce reasonably good quality translations.

The intuition behind our approach is that image features may help further improve the translation of image descriptions, for example disambiguating words with multiple senses, when these alternatives are available in the n -best list produced by the SMT model. This approach also has the advantage over joint visual-textual alternatives in that the translation model itself is learnt independently from images, and thus does not require dataset-specific images at training time to generate candidate translations. In fact, images are only used at test time for n -best list re-ranking, and the visual classifier is pre-trained on a generic image dataset.

We use image features from a Convolutional Neural Network (CNN) along with standard Moses features to re-rank the n -best list. We also propose an alternative scheme for the German-to-English direction, where terms in the English image descriptions are matched against 1,000 WordNet synsets, and the probability of these synsets occurring in the image estimated using CNN predictions on the images. The aggregated probabilities are then used to re-rank the n -best list, with the intuition that the best translations will contain words representing these entities. Our submissions that re-rank the n -best translations with image vectors are able to marginally outperform the strong, text-only baseline system for both directions.

In Section 2 we describe the procedure to extract image features. In Section 3 we explain the experiments along with their results. We finally give a brief overview of related work in Section 4, before presenting some conclusion and future directions (Section 5).

2 Image features

Image features were extracted using the 16-layer version of VGGNet (VGG-16) (Simonyan and Zisserman, 2014), which is a Deep Convolutional Neural Network (CNN) pre-trained on 1,000 object categories of the classification/localisation task of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). More specifically, we used MatConvNet (Vedaldi and Lenc, 2015) to extract the final fully-connected layer (FC8) of VGG-16 after applying the softmax function. The 1,000-dimensional vector from this layer provides class posterior probability estimates for 1,000 object categories, each corresponding to a distinct WordNet concept (synset).

The 1,000 dimensional vector were used as features in our systems to re-rank the top- n output translations from the SMT model (Section 3.2). Each feature is an estimate of the probability that a given object category is depicted in the image. Note that the posterior probability estimates for VGGNet are not perfect (the top-5 error rate was 7.3% in the ILSVRC2014 challenge, where a prediction is considered correct if the correct category is within the top 5 guesses), and we expect such errors to propagate downstream to the translation task. Moreover, the classifiers are tuned

to the 1,000 categories of ILSVRC, and many categories may not be relevant to the Flickr30K dataset (Young et al., 2014) that is used for this task, and vice-versa, that is, many of the objects in the Flickr30K dataset may not exist in the ILSVRC dataset. This implies that the classification error in our dataset is probably much higher.

3 Experiments

3.1 Data

The data used for the shared task is a version of the Flickr30K dataset. For the translation task, the Flickr30K dataset was extended in the following way: for each image, one of the English descriptions was selected and manually translated into German by a professional translator. The resulting parallel data and corresponding images for training are divided into training, development and test sets. As training and development data, 29,000 and 1,014 triples were provided, respectively, each containing an English source sentence, its German human translation and corresponding image. As test data, set of 1,000 tuples containing an English description and its corresponding image was provided. More details about the shared task data can be found in (Elliott et al., 2016).

3.2 Training

Both our submissions are based on the Moses SMT toolkit (Koehn et al., 2007) to build phrase-based SMT models. They are constructed as follows: First, word alignments in both directions are calculated using GIZA++ (Och and Ney, 2000). The phrases and reordering tables are extracted using the default settings of the Moses toolkit. The parameters of Moses were tuned on the provided development set, using the MERT (Och, 2003) algorithm. 4-gram back-off language models were built using the target side of the parallel corpus. Training was performed using only the data provided by the task organisers, and so systems for both directions were built in the constrained setting.

We extracted the 100 best translations with our SMT model and re-ranked them using the image features described in Section 2, along with the standard Moses features. We used an off-the-shelf tool ¹ to re-rank the n -best translations. More

¹<https://github.com/moses-smt/mosesdecoder/tree/master/scripts/nbest-rescore>

specifically, we performed following steps:

- We ran our Moses decoder to generate 100-best lists for each translation in the development set.
- We extracted and added following image feature scores to the already existing feature values for each translation in the n -best list:
 - *prob*: Aggregated probability estimates of entities being depicted in the image and also being mentioned in the candidate translations. Here, we match terms occurring in the English candidate translations to the 1,000 synsets of ILSVRC, and estimate the probability of these synsets occurring in the image using the CNN predictions. In cases where more than one entity is matched, we average the probabilities of all matched synsets. The intuition is that the top translations should mention the entities depicted in the image, while lower ranked translations will have fewer entities mentioned, and thus a lower probability score overall. This feature is used only in the de-en direction since we only have access to the English version of WordNet.
 - *vec*: 1,000-dimensional FC8 vector from the CNN for both en-de and de-en directions. As mentioned in Section 2, each element in the vector corresponds to the posterior probability estimate of a WordNet synset, with the vector summing to 1 after applying the softmax function. Note that each element in the vector is considered as an independent score, with its weight learnt during re-ranking.
- We ran the optimiser K-best MIRA (Cherry and Foster, 2012) to learn new weights for all features in the n -best list. The optimiser creates a new config file that contains new weights for each feature. The choice of MIRA to learn new weights over MERT is based on the fact that MIRA is known to perform better than MERT for larger feature sets in terms of efficiency and performance.
- We used the original config file to generate 100-best lists for the test set.

Lang.	Train	Dev	Test	BLEU	Meteor
en-de	29000	1014	1000	0.383	0.576
de-en	29000	1014	1000	0.434	0.363

Table 1: Datasets size and results of a baseline system on the development set.

Lang.	Score	Re-Rank _{prob}	Re-Rank _{vec}
en-de	BLEU	-	0.386
	Meteor	-	0.580
de-en	BLEU	0.431	0.437
	Meteor	0.360	0.366

Table 2: Results on the development set after re-ranking.

Lang.	System	Meteor	Meteor-norm
en-de	Baseline	0.525	0.573
	Re-Rank _{vec}	0.526	0.574
de-en	Baseline	0.363	0.398
	Re-Rank _{vec}	0.365	0.401

Table 3: Results on the test set: Baseline Moses vs Re-ranking approach.

- We added the above mentioned image features to the test n -best list.
- Finally, we re-scored the 100-best list using the re-scoring weights file and extracted the top best translation for each source segment.

For our experiments, we used the same tuning set to train the re-ranker that was used to optimise the Moses decoder original features. We note that it could be better to use a distinct tuning set than the one on which the decoder weights were optimised.

3.3 Results

The results of our submissions for the German-English and English-German tasks are summarised in Tables 1, 2 and 3. Table 1 shows our baseline Moses systems (text-only) along with training, development and test data sizes. Table 2 presents our results with re-ranking on the development set. The system *Re-Rank_{prob}* uses the decoder features with additional aggregated probability estimates features, while the system

Re-Rank_{vec} uses decoder features along with the 1,000-dimensional vector produced by VGGNet.

It can be observed that re-ranking with a 1,000-dimension image vector improves over the baseline for both directions, whereas posterior probability feature degrades the result. Note that although all (*n*-best) translation hypotheses for a given source description get the same image feature values (1,000 dimension image vector), the combination of the decoder features with these image vectors make the optimiser produce different discriminative weights, which may lead to better translation choices.

We submitted a system for each translation direction with vector features as the official submissions. It can be seen in Table 3 that our systems were able to improve over the baseline in the official metrics in both directions, although only marginally. Moreover, both systems are among the top three systems in the official ranking that outperform the strong Moses SMT baseline. The output of our systems is significantly different from that of the baseline: 260 out of the 1,000 segments differ between the baseline and the re-ranking approach. Figure 1 shows some examples of English-to-German translations for the test set from our proposed system using VGGNet FC8 features for re-ranking (*Re-Rank_{vec}*), in comparison to translations by the Moses baseline. In all cases, the translations produced by the two systems are different. In the first example, the Moses baseline translation, although not entirely correct, can be considered more accurate. In the second example, both translations are accurate, but that produced by the re-ranking approach matches exactly the reference. Finally, in the third example, the translation by the re-ranking approach is significantly better, and also much closer to the reference. An interesting observation is the fact that while the baseline system does not produce any translation that is exactly the same as the reference, the re-ranking approach produces 37 translations that are exactly the same as the reference translations. A better understanding on the differences between the baseline and re-ranking approaches would require more systematic human evaluation, which we plan to do in the future.

4 Related work

In computer vision, considerable progress has been made in the field of visual object recogni-

tion in recent years, especially since the CNN-based AlexNet (Krizhevsky et al., 2012) convincingly won the ILSVRC2012 challenge by a large margin compared to its closest rival. Progress in image classification (“what does this picture depict?”) has since improved from strength to strength, from an error rate of 16.4% (correct label in top 5 guesses) by AlexNet down to 3.6% by ResNet (He et al., 2015) in the 2015 challenge. Despite the high success rate, there is still much work to be done in the object classification and localisation challenge (“what object category does this picture depict and where?”) and the object detection challenge (“find all instances of this object category in all images, if any”), although the performance for these has also improved tremendously in recent years.

With the improved performance of object classifiers/detectors, there has also been increased interest in applying these classifiers/detectors to various downstream tasks, especially those that involve multiple modalities. For example, CNNs has been used in conjunction with Recurrent Neural Networks (RNN) (Mikolov et al., 2010) to generate image descriptions, e.g. (Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Vinyals et al., 2015). Other multimodal tasks that have been explored include video description generation (Chen and Dolan, 2011; Yu and Siskind, 2013), visual question answering (Antol et al., 2015; Ren et al., 2015; Malinowski et al., 2015), multilingual image question answering (Gao et al., 2015), and multimodal translation of image descriptions (Elliott et al., 2015). Whilst the work of Elliott et al. (2015) focuses on learning multimodal image description translation in a joint fashion using CNNs and RNNs, our work uses a conventional phrase-based SMT decoder combined with features extracted from a CNN for re-ranking.

5 Conclusions

We presented the development of our SMT systems that incorporate image features for the first German-English and English-German WMT Multimodal Machine Translation shared task. In the official evaluation, the English-German system was ranked third according to the Meteor score, while the German-English system was ranked first, although there were only two other systems for this direction. Small but consistent improvements over than a strong text-only SMT baseline




	EN	A young brunette woman eating and drinking something.
	DE (Baseline)	Eine junge Frau mit braunen Haaren und isst und trinkt etwas .
	DE (Re-Rank _{vec})	Ein junger brünette Frau isst und trinkt etwas .
	Reference	Eine junge brünette Frau isst und trinkt etwas.
<hr/>		
	EN	A black boy is sitting in the sand.
	DE (Baseline)	Ein dunkelhäutiger Junge sitzt im Sand .
	DE (Re-Rank _{vec})	Ein schwarzer Junge sitzt im Sand .
	Reference	Ein schwarzer Junge sitzt im Sand.
<hr/>		
	EN	A man with a black vest holding a model airplane
	DE (Baseline)	Ein Mann in einer schwarzen Weste und einem Modellflugzeug
	DE (Re-Rank _{vec})	Ein Mann mit einer schwarzen Weste hält einem Modellflugzeug
	Reference	Ein Mann mit einer schwarzen Weste hält ein Modellflugzeug

Figure 1: Example English-to-German (EN-DE) output translations for Re-Rank_{vec} on the test set, compared against the Moses baseline (before re-ranking).

system were found in both directions.

Our initial set of experiments can be improved in many directions. For instance, it would be interesting to explore incorporating image features directly into the decoding step and tuning the weights along with Moses parameters. It is also worth investigating other layers of image models instead of the final fully-connected layer to be used with textual features. Finally, increasing the size of n -best to re-rank translations could increase the chances of achieving better results by providing the re-ranker with more variety in terms of alternative translations.

Acknowledgments

This work was supported by the QT21 (H2020 No. 645452, Lucia Specia), Cracker (H2020 No. 645357, Kashif Shah) and the ERA-NET CHIST-ERA D2K 2011 VisualSense (ViSen) project (UK EPSRC Grant EP/K019082/1, Josiah Wang).

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*.
- David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 190–200, Portland, Oregon, USA, June.
- Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 427–436, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.
- Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. *CoRR*, abs/1605.00459.
- Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Advances in Neural Information Processing Systems*, pages 2287–2295.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*.
- Tomas Mikolov, Martin Karafit, Luks Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 1045–1048. ISCA.
- Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the Association for Computational Linguistics*, pages 440–447, Hongkong, China, October.
- 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*, pages 160–167. Association for Computational Linguistics.
- Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2953–2961.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.
- Andrea Vedaldi and Karel Lenc. 2015. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.
- 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. *Transactions of the Association for Computational Linguistics*, 2:67–78, February.
- Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the Association for Computational Linguistics (ACL)*.