The AFRL-OSU WMT17 Multimodal Translation System: An Image Processing Approach

John Duselis, Michael Hutt, Jeremy Gwinnup
Air Force Research Laboratory
{john.duselis,michael.hutt.ctr, jeremy.gwinnup.1}@us.af.mil

James W. Davis, Joshua Sandvick
Ohio State University
{davis.1719,sandvick.6}@osu.edu

Abstract

This paper introduces the AFRL-OSU Multimodal Machine Translation Task 1 system for submission to the Conference on Machine Translation 2017 (WMT17). This is an atypical MT system in that the image is the catalyst for the MT results, and not the textual content.

1 Introduction

Contemporary scientific meetings have examined the potential benefits of fusing image information with machine translation. For instance, the leading international conference in this area, the Conference on Machine Translation (WMT), is approaching its second year of competition on Multimodal Machine Translation (MMT). First year results in WMT16's Multimodal Task 1 were varied in approaches, informative in their findings, and indicated potential opprtunities for multimodal system improvement. (Specia et al., 2016).

In the WMT16 submissions, the seemingly predominant focal point across the systems was the fact that textual information was the driver for the translation. The image features tended towards being ancilliary inputs or outputs (Libovický et al., 2016; Guasch and Costa-Jussà, 2016; Caglayan et al., 2016) or decision-type functions (Shah et al., 2016) and not the main antagonist for translation (Specia et al., 2016; Elliott et al., 2015). This is sensible as it is an MT competition. However, approaching it from another direction, namely, having the image as the driver for the translation presents a different point of view worth investigating.

This work is sponsored by the Air Force Research Laboratory under AFRL/711 Human Performance Wing Chief Scientist Funding.

The following sections will outline the seemingly novel approach to MMT and give particulars of this unconstrained system.

2 AFRL-OSU System

This section will outline the architecture of the system. This is a first approximation into the process but is expected to undergo further development based on insights from this competition.

2.1 General Overview

Referencing Fig. 1, a generic example taken from (Specia et al., 2016) shows a method where the source caption and image are the drivers for the multimodal translation. In some of WMT16's submissions, the decomposition of the image is incorporated as an additional feature into the MMT system, while others used the features as a function to help pick the best translation.

AFRL-OSU's system is pictorially represented in Figure 2. Currently, there is much work in image captioning systems (Socher et al., 2014; Ghahramani et al., 2014; Mao et al., 2014; Kiros et al., 2014; Vinyals et al., 2015), and WMT17 has even set out a task in its competition for it. Our emphasis is not to try to produce a multilingual image captioning system, rather to use one to accomplish MT as the maturity of the caption engine research progresses.

This system architecture assumes an image caption engine can be trained in a target language to give meaningful output in the form of a set of the most probable n target language candidate captions. A learned mapping function of the encoded source language caption to the corresponding encoded target language candidate captions is thusly employed. Finally, a distance function is applied to retrieve the "nearest" candidate caption to be the translation of the source caption.

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 445–449
Copenhagen, Denmark, September 711, 2017.

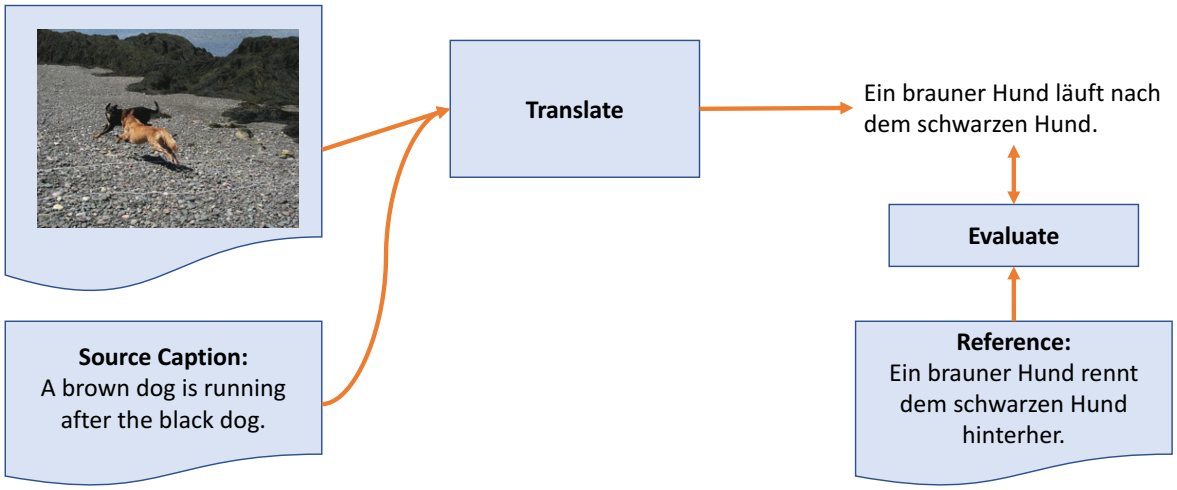

Figure 1: A Text-based Model for Multimodal Machine Translation adapted from (Specia et al., 2016)

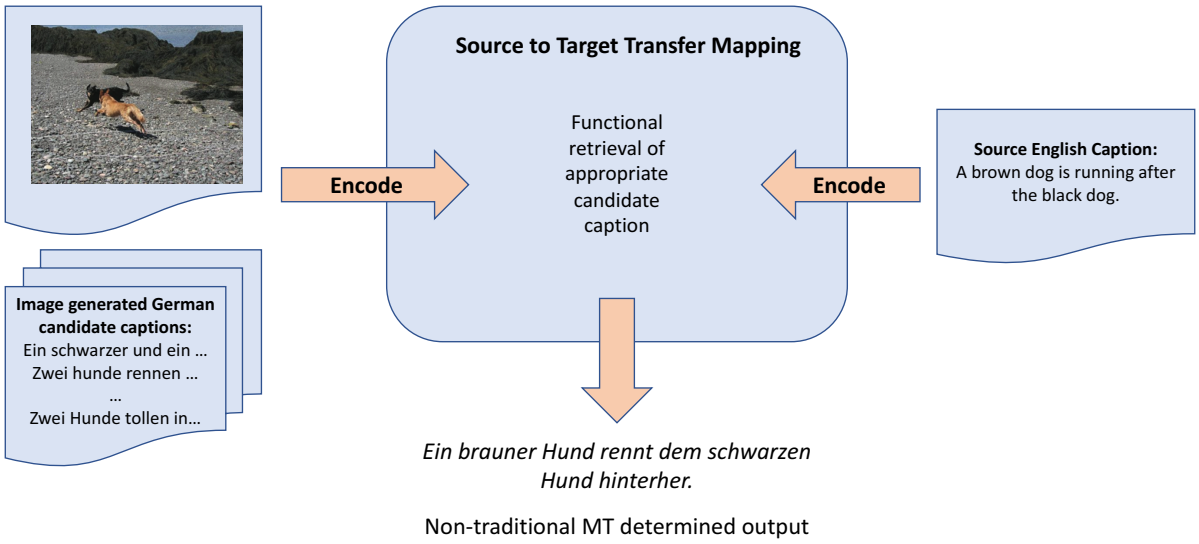

Figure 2: An Image-based Model for Multimodal Machine Translation.

2.2 Theoretical Overview

Details of the system architecture are illustrated in Figure 3. Given an image i (top left), using an image captioning engine trained in the target language t, we produce n candidate captions: C^t_{ij} for $j = 1, ..., n$.

After obtaining the candidate sentences, we transform them into a fixed vector length encoding with

$$v^t_{ij} = G^t(C^t_{ij}) \quad (1)$$

where $G^t(\cdot)$ is the target encoder.

Similarly (from the top right of Figure 3), the source language caption C^s_i is encoded using

$$v^s_i = G^s(C^s_i) \quad (2)$$

where $G^s(\cdot)$ is the source encoder.

At this point, both the target captions and the source caption are encoded in separate monolingual, monomodal subspaces. In order to execute the retrieval process, we develop a transfer mapping of the source language encodings to the space of target language encodings. We learn this source-to-target mapping using training pairs of source language encodings and target language encodings provided by traditional MT of the source language examples (Sennrich et al., 2016). Hence the mapping attempts to learn MT translation from the encoding representations themselves. The architecture employed is a multi-layer neural network.

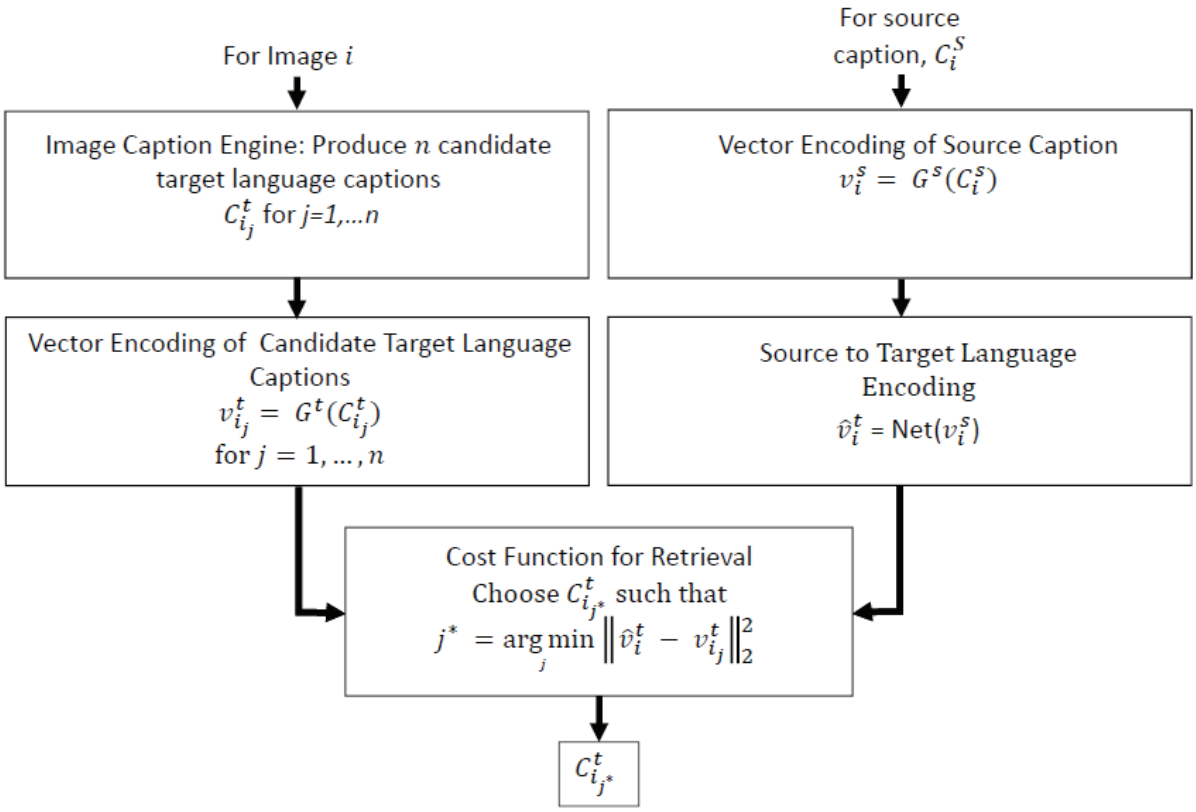

Figure 3: Architectural Diagram of the Processing Chain

2.3 Implementation

The actual AFRL-OSU unconstrained implementation went through many iterations of tool sets before settling. The captioning engine selected for this competition was Google's Show and Tell system (Vinyals et al., 2015). It was trained on the WMT16 training and validation splits using the MultiFlickr30K images and German (Elliott et al., 2016) and ImageClef 2006-2008 multilingual datasets (Rorissa et al., 2006). For testing, 1000 captions ($C_{i_j}^t$ for $j = 1, \ldots, 1000$) per image were produced. Any caption with sentence length less than five words was not considered, but was not replaced. Captions were put into all lowercase without punctuation.

The monolingual word encodings, G^t and G^s, used to vector encode the source language caption and target language captions employed the word encodings compiled and published by Facebook (Bojanowski et al., 2016). Because Facebook's data was chosen over any word encodings produced internally, vector length was fixed at 300. This dataset was produced by Facebook by crawling and cleaning up data from Wikipedia pages using their fastText software and encoding algorithm outlined in (Bojanowski et al., 2016). Sentence encodings used in the AFRL-OSU system were derived from averaging of in-vocabulary constituent word encodings.

To transform source encoded data into the target language encoded subspace, a multi-layer neural network was constructed. The WMT16 training/validation splits were used for the training English source captions (5 captions per image with a total of 29000 images). These English captions were encoded into 300x1 vectors, each L2-normalized. The training target outputs were generated using Edinburgh's WMT16 Neural MT System (Sennrich et al., 2016) to translate captions from English to German in the same 300x1 vector format, and again L2-normalized. The neural network was configured with 1 hidden layer (500 nodes) and a mean squared error loss-function. To test the approach 10% of the training data was kept for evaluation. During training, 25% of the remaining training data was withheld for validation with a maximum of 10000 epochs. The resulting network provides a source-to-target mapping of the source caption encoding

$$\hat{v}_i^t = Net(v_i^s) \tag{3}$$

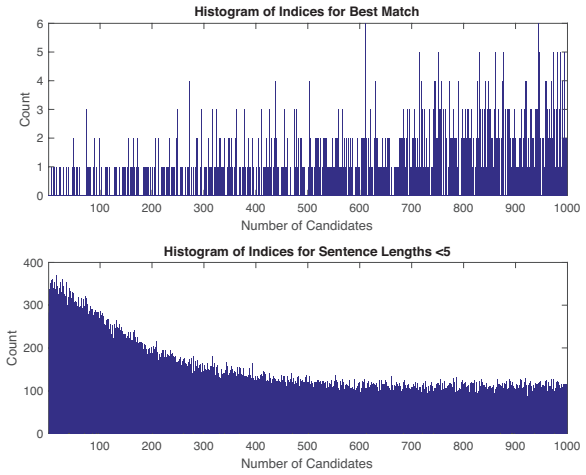

Figure 4: Histogram analysis. Top: Histogram of indices for the best match in the candidates. Bottom: Histogram of indices for candidate sentences with invalid (<5) length.

We lastly used the squared Euclidean Distance between the source transformed English caption encoding $\hat{v}_i^t$ to the collection of candidate target caption encodings to select the best candidate sentence index j^*

$$j^* = \arg\min_j ||\hat{v}_i^t - v_{i_j}^t||_2^2 \quad (4)$$

The "best" match (to the source language caption) produced from the captioning engine is the sentence $C_{i_{j^*}}^t$. From the test data (with ground truth source-to-target labels), we received a top-1 of 77% and top-5 of 87%.

3 Results

The final submission consisted of generating 1000 captions per image with the top score being selected. The minimum of 5 words per sentence discounted 150963 candidate captions.

The top caption satisfying Eq.4 as the minimal value was scored against the output from the Edinburgh WMT16 Neural MT system and had a METEOR (Denkowski and Lavie, 2014) score of 19.8 (Sennrich et al., 2016). Figure 4 provides some trends for locations of zero vectors and top scoring vectors.

4 Conclusion

Assuming sufficient baseline results from an image-centric MMT system evaluated in this competition, there exist several opportunities for understanding the implications of such a system and also to improve its capabilities.

The captioning engine used is employed as a black box and assumed meaningful output for processing. Knowing the inner workings of the caption engine should allow tuning to produce more meaningful results. The authors also look forward to the results of this Multimodal Competition's Task 2 to obtain a better captioning engine (either improvements on the current system, or a different method altogether).

The monolingual word encodings attained from the Facebook models were constrained to 300 elemental vector length. Exploration into not only the size, but also construction of the data is warranted.

The cost function used, squared Euclidean Distance, is a first attempt. Looking at a variety of functions may harvest better results.

The authors only submitted the top ranked caption for scoring in this competition. However, 33 candidate submissions received a 0.0 sentence level METEOR score. Therefore, approaching a selection from the Top m captions that would maximize the METEOR is worth investigating.

This paper outlined the AFRL-OSU WMT17 Multimodal Translation system where the image is the focal point for MT. The authors hope that it spurs some alternative thinking and research in the area of multimodal MT.

Acknowlegements

The authors wish to thank Rebecca Young for her involvement in the human evaluation portion of the WMT17 Multimodal Translation task. The authors also wish to thank Rico Sennrich for making models and data available from the Edinburgh WMT16 Neural MT system, saving valuable time and effort during development.

References

2016. *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*. The Association for Computer Linguistics. http://aclweb.org/anthology/W/W16/.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vec-

tors with subword information. *arXiv preprint arXiv:1607.04606* .

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? *CoRR* abs/1605.09186. http://arxiv.org/abs/1605.09186.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions pages 70–74.

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

Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors. 2014. *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. http://papers.nips.cc/book/advances-in-neural-information-processing-systems-27-2014.

Sergio Rodríguez Guasch and Marta R. Costa-Jussà. 2016. WMT 2016 multimodal translation system description based on bidirectional recurrent neural networks with double-embeddings. In (DBL, 2016), pages 655–659. http://aclweb.org/anthology/W/W16/W16-2362.pdf.

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

Jindrich Libovický, Jindrich Helcl, Marek Tlustý, Pavel Pecina, and Ondrej Bojar. 2016. CUNI system for WMT16 automatic post-editing and multimodal translation tasks. *CoRR* abs/1606.07481. http://arxiv.org/abs/1606.07481.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *NIPS Deep Learning Workshop* .

Abebe Rorissa, Paul Clough, William Hersh, Abebe Rorissa, and Miguel Ruiz. 2006. Imageclef and imageclefmed: Toward standard test collections for image storage and retrieval research. *Proceedings of the American Society for Information Science and Technology* 43(1):1–6. https://doi.org/10.1002/meet.14504301130.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. *CoRR* abs/1606.02891. http://arxiv.org/abs/1606.02891.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. Shef-multimodal: Grounding machine translation on images. In (DBL, 2016), pages 660–665. http://aclweb.org/anthology/W/W16/W16-2363.pdf.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL* 2:207–218.

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/W/W16/W16-2346.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.