The AFRL-Ohio State WMT18 Multimodal System: Combining Visual with Traditional

Jeremy Gwinnup
AFRL
jeremy.gwinnup.1
@us.af.mil

Joshua Sandvick
Ohio State University
sandvick.6
@osu.edu

Michael Hutt
AFRL
michael.hutt.ctr
@us.af.mil

Grant Erdmann
AFRL
grant.erdmann
@us.af.mil

John U. Duselis
AFRL
john.duselis@us.af.mil

James W. Davis
Ohio State University
davis.1719@osu.edu

Abstract

AFRL-Ohio State extends its usage of visual domain-driven machine translation for use as a peer with traditional machine translation systems. As a peer, it is enveloped into a system combination of neural and statistical MT systems to present a composite translation.

1 Introduction

Most of the submissions to the Second Conference on Machine Translation (WMT17) Multimodal submissions for Task 1 (Elliott et al., 2017) used the visual domain to enhance machine translation of the image+caption pair. The exception was a Visual Machine Translation (VMT) system where the image is the driver for the translation (Duselis et al., 2017). While the scores for this submission did not approach baseline, except by human scoring, it did introduce the concept that the visual domain can approach parity with the traditional text based MT systems.

The AFRL-Ohio State Third Conference on Machine Translation (WMT18) submission also explores viability of a VMT system enhancing current techniques. Previous work by Calixto et al. (2017) ensembled different multimodal machine translation (MMT) systems, with the visual domain used in conjunction with the text domain. Similarly, we incorporate the VMT system with a small sampling of neural and statistical MT systems in order to give indicators on how the performance is affected by mutual inclusion.

2 The AFRL-Ohio State 2018 Multimodal System Submission

A Visual Machine Translation system is one that utilizes the visual domain, whether it is a video or picture, as the driver for MT. This assumes that there is a visual analogue for the relevant source text. This is a specialized form of Multimodal Machine Translation (MT) in which the image is producing candidate target language sentences.

Current trends in MT use system combinations or ensembles of various MT systems (statistical, neural, rule-based, etc.) to create a consensual final answer. A key ingredient to this method is introducing variability of MT outputs to reach the conclusion (Freitag et al., 2014). We posit that adding the VMT to the system will enhance the overall results.

AFRL-Ohio State submitted three systems for official scoring. The focus of explanation will be on the 4Combo system because it underwent human evaluation, but the other two will be revisited in the analysis portion. No post-editing was performed for any of the submission systems.

2.1 The AFRL-Ohio State WMT17 Submission

Here is an overview of the VMT system submitted to the WMT17 submission (Duselis et al., 2017). This system architecture assumes a captionator 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 thus employed. Finally, a distance function is applied, and the nearest candidate caption is selected to be the translation of the source caption.

2.2 Captionator

The current instantiation of our VMT system uses the Google Show and Tell captionator (Vinyals et al., 2015) trained on the training set from Flickr30k, augmented with data from ImageCLEF 2008 (Grubinger et al., 2006).

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 612–615

https://doi.org/10.18653/v1/W18-64067

The captionator was trained on the 29,000 training image+German caption pairs, plus 20,000 image+German captions from ImageCLEF 2008. This was slightly fewer than the number used on the WMT17 submission. Additional models were trained on the constrained set of the 29,000 WMT pairs, one with a single caption per image and another with five captions per image. However, the Show and Tell system generated a high number of 'unknown word' tokens. Filtering out the sentences with unknown tokens produced a bias towards short, generic captions. Augmenting with the ImageCLEF data produced noticeably better results. This was the only change for the captionator. Consistent with the prior year's submission, no accommodations were made for out of vocabulary words.

2.3 Caption Selection

Stemming from critique and results from WMT17, the simple neural network was revised to center around a two sided Long Short Term Memory (LSTM) encoder. One side of the LSTM was trained to encode English sentences, while the other was trained to encode German sentences. Each of the LSTMs has a state size of 256 nodes. The multiclass hinge loss function was used to evaluate the encodings, penalizing the loss by the highest-scoring incorrect match between the English and German sentences in a training batch.

The training data comprised the WMT18 Multimodal Task 1 English and German training sentences from the 2018 Multi30k dataset. The words were tokenized and transformed to lower case, and punctuation was removed. Words were then embedded using the FastText pretrained word embedding vectors (Bojanowski et al., 2017), with dimension 300. The Adam optimizer (Kingma and Ba, 2014) was employed to train the network parameters with a batch size of 32. The network was trained for approximately 100 epochs using TensorFlow on a GeForce GTX 1080.

We tested the caption selection mechanism on the 2018 Multi30k datasets, encoding both the given English captions and the given German captions. Each English caption was matched with the German caption in the set with minimum hinge loss. On the 29,000-image training set, each English caption was correctly matched with its corresponding German caption 99.4% of the time. On the 1,014-image test set, the matching accuracy was 92.4%.

2.4 "Standard" Machine Translation

Inspired by Gwinnup et al. (2017), we trained multiple MT systems with differing toolkits and characteristics for use in system combination with our VMT efforts. These toolkits include: OpenNMT (Klein et al., 2017), Marian (Junczys-Dowmunt et al., 2018), and Moses (Koehn et al., 2007).

All systems were trained with the approximately 41 million parallel lines of preprocessed German–English data provided by the WMT18 organizers.

2.4.1 OpenNMT

The OpenNMT system was trained using the German-English Parallel Data from the WMT18 organizers for the News Task, but excluding the ParaCrawl Data. It incorporates case features and a vocabulary from 2000 byte-pair encoding merges. This small vocabulary was chosen to reduce the number of out-of-vocabulary tokens resulting from morphology and compounding.

2.4.2 Marian

The Marian toolkit was used to train a baseline system using the pre-BPE'd data provided by the WMT18 news task organizers. This system employed a deep bi-directional (or "BiDeep") architecture as outlined in Miceli Barone et al. (2017) and Sennrich et al. (2017). Further details of the exact settings used to train this system are available in the wmt2017-uedin example shown in the marian-examples GitHub repository[1].

2.4.3 Moses

For variety, a phrase-based Moses system was trained using the same BPE'd data as the above Marian system. This system employed a hierarchical reordering model (Galley and Manning, 2008), 5-gram operation sequence model (Durrani et al., 2011) and a 5-gram BPE'd KenLM (Heafield, 2011) language model trained on the target side of the provided parallel data.

2.5 System Combination

RWTH's Jane System combination (Freitag et al., 2014) was used to combine the outputs of the three traditional MT systems with the output of our VMT approach.

[1] https://github.com/marian-nmt/marian-examples/tree/master/wmt2017-uedin

3 Analysis

3.1 Results

The AFRL-Ohio State WMT18 `4Combo` submission, although a better showing than our WMT17 submission, failed to meet baseline. Comparing the VMT component to last year's system showed the expected improvement in results. The official results are presented in Table 3.1, mirroring the results presented in Specia et al. (2018). `VMT` is the visually driven MT system. `2Combo` is the VMT+Marian, `3Combo` is the Marian+Moses+OpenNMT system. `4Combo` is all four systems.

System	BLEU ↑	Meteor ↑	TER ↓
VMT	5.0	17.7	80.1
2Combo	10.0	25.4	79.0
3Combo	23.8	44.5	59.7
4Combo	24.3	45.4	58.6

Table 1: Systems Scoring

Examining the `3Combo` and `4Combo` outputs, we note a positive performance trend when adding the VMT system to combinations of traditional MT systems.

3.1.1 Captionator Output - Oracle Scoring

To gain more insight, a document level Meteor and BLEU Oracle scoring for the captionator output was applied.

The three observables were the most probable sentence from the captionator, the AFRL-Ohio State caption selection mechanism, and the best scoring caption output. This analysis is based on the WMT17 multimodal validation set.

We performed *a posteriori* analysis, to determine how well our caption selector compares with other possibilities. We considered two options. First, the one-best is the caption the captionator considers the most likely, without regard to the source-side text. Second, we found an oracle caption for each image, based on Meteor score. The oracle captions determine an upper-bound on the Meteor score the caption selector can achieve. Results are shown in Table 2.

Method	BLEU ↑	METEOR ↑
1-best	1.53	10.69
LSTM	5.74	18.59
Oracle	18.78	36.74

Table 2: Oracle scoring for the VMT system.

4 Future Work

Our purpose in developing the visual domain is to include it as an equal to the text or as a driver for the MT at a higher level of abstraction than the neural layer. Using the captionator to produce sentences limits the VMT to the the captionator's abilities. Instead, we next plan to employ a more generalized approach to estimate objects or concepts that are particularly difficult to translate directly from the image (or video clip, if available) rather than attempting to estimate an actual sentence structure. We expect the use of such information from the visual content to be more amenable to bias or influence other MT systems.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, pages 1045–1054, Portland, Oregon.

John Duselis, Michael Hutt, Jeremy Gwinnup, James Davis, and Joshua Sandvick. 2017. The afrl-osu wmt17 multimodal translation system: An image processing approach. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 445–449, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Cleared for public release on 9 Aug 2018. Originator reference number RH-18-118708. Case number 88ABW-2018-3967.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open source machine translation system combination. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856.

Michael Grubinger, Paul D. Clough, Henning Müller, , and Thomas Deselaers. 2006. The iapr bencmark: A new evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*.

Jeremy Gwinnup, Timothy Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, Brian Thompson, and Jonathan Taylor. 2017. The AFRL-MITLL WMT17 systems: Old, New, Borrowed, BLEU. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 303–309, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

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

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Loïc Barrault, Fethi Bougares, and Desmond Elliot. 2018. WMT18 shared task: Multimodal machine translation.

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)*.