HYBRID MACHINE TRANSLATION USING VISUAL INFORMATION

Zubair Abid
20171076
zubair.abid@research.iiit.ac.in

Souvik Banerjee
20171094
souvik.banerjee@research.iiit.ac.in

Team 11: Jibonedukkho

ABSTRACT

A majority of state of the art machine translation models use the transformer as their central architecture. It provides us with a massively reduced computational time and increased parallel computations, especially compared against its encoder-decoder based equivalents. However, they are still lacking when it comes to the ability to exploit multi-modal information. In this project, we have implemented a multi-modal Neural Machine Translation (NMT) system making use of associated images as described in the paper Neural Machine Translation with Universal Visual Representation (ICLR 2020). We compare the results against baseline Statistical Machine Translation (SMT) and NMT systems, and analyse the performance.

1 INTRODUCTION

A majority of state of the art machine translation models use the transformer as their central architecture. Although incredible in its own right with room for parallel computations and with its vastly reduced computational time compared to attention-based encoder-decoder models, it is still lacking in its ability to exploit multi-modal information. Only a handful of work has been done in this field. A transformer can indeed be modified in a manner that it uses both visual and textual clues to do better in translation related task. There is also room for exploring improvements to the vanilla model by incorporating methods from pre-existing SMT systems.

Visual information has been used to better neural machine translations in some study although the contribution of images is still an open question ((Elliott, 2018)). With our project, we aim to address that question by showing how such information improves translation results and how seamlessly models which uses such information can be integrated seamlessly with other types of model like SMT, thus being a great hybrid model candidate in the process.

In this project, we have implemented such a Hybrid Machine Translation model that uses phrase-based SMT system Moses, and the previously mentioned NMT system which was originally proposed in this paper: Neural Machine Translation with Universal Visual Representation, ICLR 2020. We compare the results against a baseline SMT system and a baseline NMT system.

The language pair used is English-French. For textual data we used WMT14 EN-FR dataset while the Multi-30k dataset (which has caption-image pairs). was used for visual data.

2 LITERATURE SURVEY

Statistical Machine Translation systems, at least back in 2003, were dominantly phrase-table based. We use Moses, developed by Koehn et al. (2007) as our state-of-the-art open source phrase-based

SMT system.

Significant steps to the application of Deep Neural Networks (DNNs) to the machine translation task were made by Sutskever et al., who introduced sequence learning to encoder-decoder architectures that simplified sentence processing. Further improvements were made by the addition of the Attention mechanism (Bahdanau et al., 2016). We will be using this architecture for our baseline NMT system. The Transformer architecture (Vaswani et al.) was introduced, improving results further by not considering sequences and taking advantage of parallel computation; but we skip over that in favour of Universal Visual Representation in our Hybrid MT system.

Using visual representation to better neural machine translation has been attempted in several previous studies (Specia et al., 2016). The goal is to use multi-modal methods to try and return significant improvements over traditional text-only systems. It is not entirely a confirmed result, and studies have explored the effectiveness of actually using images for translation tasks (Elliott, 2018).

A major limitation to this, however, has been the lack of available annotated image data for all pairs of languages. The Multi-30k dataset (Elliott et al., 2016) provides parallel captions in `en-fr`, `en-de`, and `en-ro` pairs for a a few thirty-thousand images, but the data is still rather insufficient for effective use in a deep neural network system, and it also requires annotation for both pairs of languages, which is a tedious and time-consuming task. Zhang et al. (2020) tide over this by requiring only (limited) annotated image data for one of the languages in the language pair. It uses topic-indentification and a topic-image lookup table to use the image features in the pipeline.

3 Research Methods

3.1 Preprocessing

All the systems used a similar preprocessing pipeline (tokenising, cleaning), provided by the Moses decoder. In addition to that, we used `subword-nmt` by Sennrich et al. (2016) for sub-word segmentation and reduce Out-of-vocabulary (OOV) issues. Some sections (like an implementation detail in a system in Section 3.4 required extra preprocessing steps, and is described in detail.

3.2 Data

We primarily used the WMT-14 English to French (EN-FR) translation dataset, and the Multi30K dataset.

1. For the SMT system, we used the provided Europarl V7 as our training set. For testing, we used the newstest-2011 corpus.
2. For the baseline NMT system, we used only the provided Europarl V7 dataset. 1.9M sentences (99.7%) was for used for train, and the remaining was split equally between dev and test.
3. For UVR-NMT, we used the provided Europarl V7 with the same train-test-dev splits, but also used the Multi30k's task 1 captions for training the image lookup table.

3.3 Statistical MT

We used the Moses Statistical Machine Translation toolkit for our baseline system. It was trained on WMT-14 EN-FR data, and tested on Newstest-2011. For preprocessing, regular tokenisation and cleaning as provided by the system scripts was done. We did not truecase the data.

Training took 25 hours on an i5-4210U running on three parallel threads. The generated phrase table was 2.7GiB.

3.4 BASELINE NEURAL MT

In our initial experiments, we wrote a custom implementation of Bahdanau et al. (2016) in PyTorch. However, we were unable to manage the associated memory concerns, which meant that at most we were able to train on 100,000 parallel sentences only, for 7 epochs before it crashed. However, these results showed a ten-fold improvement on the BLEU score in comparison to training for 50 epochs on 20,000 sentence pairs, indicating that better results could be gotten by training on the full dataset.

To achieve this, we used OpenNMT-py as our baseline seq2seq+attention system instead. Performance was comparable to the baseline reference implementation without LSTM Ensembles and Beam search. OOV words were replaced with an UNK token.

We compare the performance in Table 1.

System	Training samples	EN-FR BLEU Score (on 100)
Custom Implementation	20,000	0.076
Custom Implementation	100,000 (interrupted)	0.28
OpenNMT-py	1,914,618	28.27

Table 1: Comparing performance of implementations of Seq2seq

3.5 UVR-NMT

In the UVR-NMT system, the model is trained by associating topics with images by training it on the Multi30K caption-image pairs. Once the model is learnt, we do inference by passing the sentence along with its associated images (as picked by the topic words of the sentence) to the model.

Broadly, the architecture works by retrieving images that are associated with the sentence and using their features (taken from Resnet-50 embeddings (He et al., 2015)) as inputs to an aggregation layer in between encoder and decoder of a transformer architecture.

3.5.1 CREATING THE LOOKUP TABLE

Creating the topic-image lookup table required two major components. This operation was done on the provided dataset that has captions and associated images.

1. **Topic Identification:** In order to filter out stopwords, we use an existing stopword dictionary [1]. This is followed by term frequenct-inverse document frequenct (TF-IDF) on the sentences to identify the "topic" words for each sentence (caption). The top 5 or 10 topic words are chosen for a sentence (caption).

2. **Image Retrieval:** Given the topic words for a caption, we add to each of their entries in an association dictionary the image that the caption was related to. By repeating the process over all sentences (captions) in the dataset, we get a completed association table of topic words mapping to multiple images.

For our project, we used the image-lookup code provided by the authors of the paper. The pseudocode provided is given below in Algorithm 3.5.1. We also replicate the graphic used to illustrate the process in Figure 3.5.1.

[1] `https://github.com/stopwords-iso/stopwords-en`

Algorithm 1 Topic-image Lookup Table Conversion Algorithm, (Zhang et al., 2020)

Require: Input sentences, $S = X_1, X_2, ...X_I$ and paired images $E = e_1, e_2, ..., e_I$
Ensure: Topic-image lookup table Q where each word is associated with a group of images

Obtain the TF-IDF dictionary F= TF-IDF(S)
Transform sentence-image pair to topic-image lookup table Q= LookUp(S, E, F)
procedure TF-IDF(S)
    for each sentence in S **do**
        Filter stop-words in the sentence
        Calculate the TF-IDF weight for each word
    end for
return TF-IDF Dictionary F
end procedure
procedure LOOKUP(S, E, F)
    for each pair $T_i, e_i \in zip(S, E)$ **do**
        Rank and pick out the top-w "topic" words in the sentence according to the TF-IDF score in the dictionary F, and each sentence is reformed as $T = t1, t2, ..., tw$
        Pair the w words with the corresponding image e_i
        for each word t_j in T **do**
            if e_i not in $Q[t_j]$ **then**
                Add e_j to the corresponding image set $Q[t_j]$ for word t_j
            end if
        end for
    end for
return Topic-image lookup table Q
end procedure

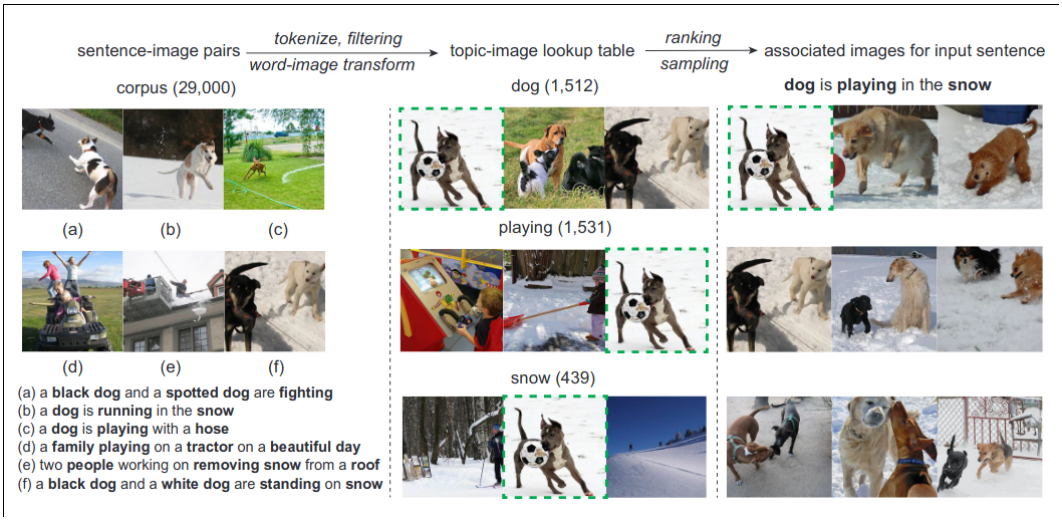

Figure 1: Lookup table process. (Zhang et al., 2020)

3.5.2 Description of the Architecture

The general architecture can be described as a transformer model based on Vaswani et al. with an added layer in between the encoder and decoder layers, referred to as the aggregation layer. The aggregation layer is added specifically for considering the image features. Given the lookup table, the preprocessed source(English, in our case) sentences are stripped of their stop word and weights are assigned to each of the remaining words. From the words that are available in the topic-image lookup table, we can get for each sentence input the associated images, and also the image features that are used as input to a Feedforward Network. The output of this is then split into K and V along

with H taken from the output of the Encoder, and attention is applied. The output is then normalised and sent as input to the Decoder stack.

Refer to Figure 7 for an architecture diagram.

3.5.3 USING THE LOOKUP TABLE FOR NEW SENTENCES

Given a new sentence (in a group of sentences), the process of getting the associated image features is roughly as follows:

1. **Extract topics in a sentence** using TF-IDF
2. **Look up the topics** in the lookup table to get the associated images for each topic, and concatenate them.
3. **Get image features** using the pretrained Resnet-50.

4 RESULTS

System	Architecture	EN-FR BLEU Score (on 100)
Moses	Phrase-table SMT	25.08
OpenNMT-py	Seq2Seq+Attention	28.27
This work	UVR-NMT base transformer	35.17
Vaswani et al.	Transformer (base)	38.1
Zhang et al. (2020) (reference)	UVR-NMT base transformer	39.64

Table 2: Comparing our work with existing systems

5 ANALYSIS AND DISCUSSION

1. For large amounts of data, the validation loss was higher than for smaller amounts of data, even though the final BLEU score was better for the large sample. We believe that there is an optimal data size for which the results are maximised
2. The Vanilla transformer and our implementation of the UVT model do not differ by much in performance, despite the computational overhead of getting image features for each sentence.

6 LIMITATIONS

1. The Moses decoder takes a very long time to train (well over a day) on the given data in order to achieve any sort of substantive results. Working with lower counts of data inputs are also not very effective, as experiments conducted showed their performance to be well below the par.
2. Computational limitations: A vanilla implementation of Seq2Seq with Attention consumes a lot more video memory than normally available, necessitating massive reductions in the amount of input data usable by the system, which severely impacts performance. Google Colab was unusable for over 10,000 samples of training data.
3. The images are feature extracted using a resnet50 neural network along with the preprocessing of 1M sentences. These preparations are tedious and thus a great deal of time is spent before the data is even passed as an input to the transformer model. Each train epoch also takes more time owing to the extra layer between the encoder and decoder.
4. The phrase table was not binarised, and as such it takes a long time to initialise the translation engine

5. We didn't use fairseq, which the reference paper authors did. We believe that this has led to some performance loss, due to non-optimal code.
6. We didn't rank the images retrieved from each sentence, unlike the actual paper. This may have had an impact on our performance.

7 Future Scope

The model can be used for Indian language translation tasks(English to Indian language translations or Indian to Indian languages). The fact that Indian languages have considerably different syntactic properties than English could drastically alter the SMT results(i.e the pre-training weights can turn out be different). It would also be interesting to see how much visual annotations affect morphologically complex Indian languages.

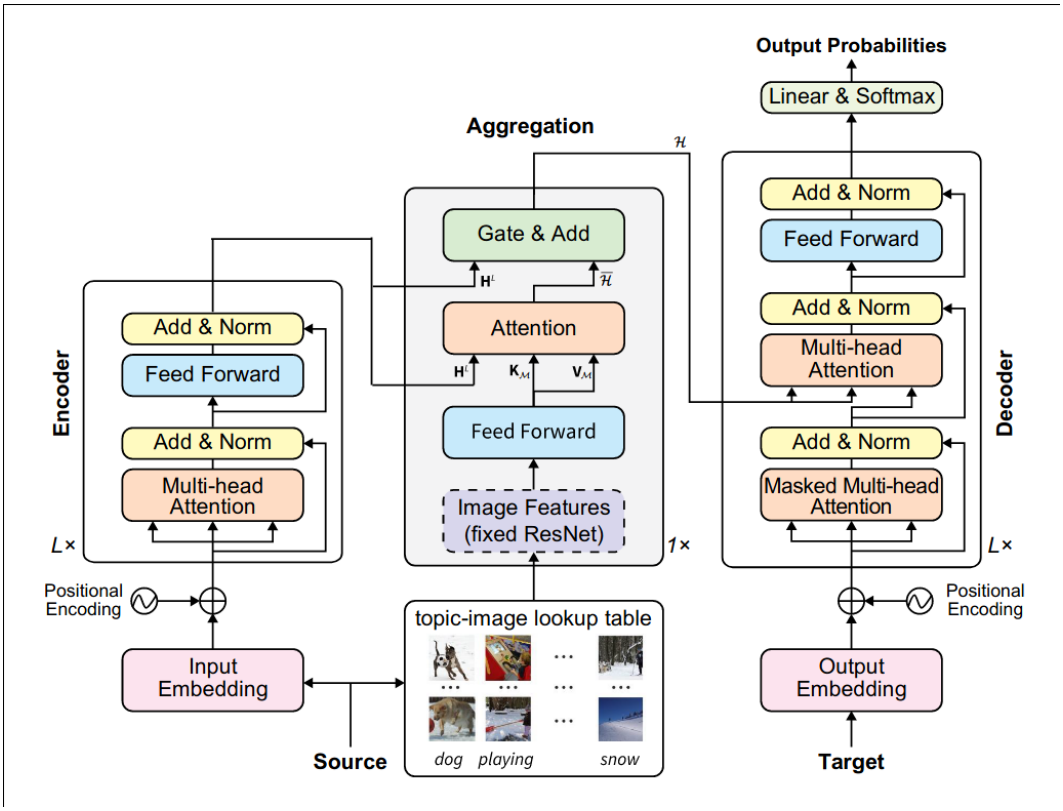

Figure 2: The architecture of UVR-NMT. Provided by Zhang et al. (2020)

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473 [cs, stat]*, May 2016. URL `http://arxiv.org/abs/1409.0473`. arXiv: 1409.0473.

Desmond Elliott. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2974–2978. Association for Computational Linguistics, Oct 2018. doi: 10.18653/v1/D18-1329. URL `https://www.aclweb.org/anthology/D18-1329`.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pp.

70–74. Association for Computational Linguistics, Aug 2016. doi: 10.18653/v1/W16-3210. URL https://www.aclweb.org/anthology/W16-3210.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385 [cs]*, Dec 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

Philipp Koehn, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, and et al. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, pp. 177. Association for Computational Linguistics, 2007. doi: 10.3115/1557769.1557821. URL http://portal.acm.org/citation.cfm?doid=1557769.1557821.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1162. URL http://aclweb.org/anthology/P16-1162.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 543–553. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-2346. URL http://aclweb.org/anthology/W16-2346.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. pp. 9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. pp. 11.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. Neural machine translation with universal visual representation. pp. 14, 2020.