Multimodal MT for English to Chinese Translation

Nate Carlson natec18@byu.edu
Department of Computer Science, Brigham Young University, Provo, Utah, 84606

Abstract

The lack of readily available data for multimodal machine translation is a major road block for the advancement of this relatively new area of research. Currently data for this task is only available in 3 language pairs in the Multi30k dataset. This project constructs a dataset for English to Chinese multimodal machine translation by generating synthetic Chinese target sentences using Google's cloud translation API. It then uses a multimodal transformer model to translate the dataset and provides an analysis of the benefits of additional visual context for this language pair. My code is accessible on github at `https://github.com/natbcar/Multimodal-MT` and the data is available upon request.

1 Introduction

Multimodal machine translation refers to the task of translating between two languages with assistance from additional context obtained from different modalities, this project focuses exclusively on images. The hope is that the extra information can be used to resolve ambiguities in the source text and produce higher quality translations. Previous work has shown that incorporating images into a neural translation model can be beneficial to translation quality, especially in the case where there is ambiguity or noise in the source text.

Lack of data is a major road block that researchers in this area face. The process of constructing a dataset to train a MMT model on is non-trivial since each translation pair needs a relevant associated image. The current benchmark dataset for this task is the Multi30k dataset introduced by (2). It currently supports English, German, French, and Czech. Motivated by the desire to further work in this area, this project adds a fifth language to this dataset, Simplified Mandarin Chinese. The Chinese sentences are synthetic, meaning they were obtained by translating English text from Multi30k and other datasets using an off the shelf MT system. I aspire to post edit these translations so that this language pair can be released for public use. More details on the generation of this datset are given in section 2.

Exploring the ability of images to enhance English to Chinese translation is a very interesting problem. Currently only translation between Indo-European languages has been explored for this task. However, there are many ambiguities in English to Chinese translation that current MT systems may be unable to pick up on relying on source text alone. Specifically, () collect Chinese translation norms for a set of 1,429 common English words and find that up to % 71 have at least one correct Chinese translation. MT models will typically be biased to the most common translation of a word in the training dataset, and thus will have a hard time making the correct translation in some fringe cases. This motivates the idea of using another modality, in our case images, to guide the translation system by providing additional context.

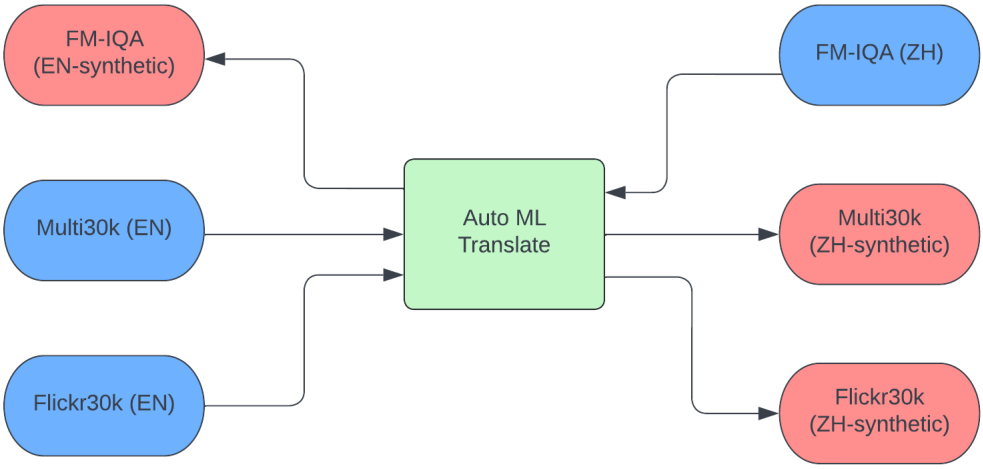

Figure 1: Visual depiction of the synthetic data generation process used to construct a complete EN-ZH MMT data set. The blue boxes represent authentic text and the red boxes represent synthetically generated translations.

2 Data

MMT has only been applied to a small number of language pairs. The current benchmark dataset for MMT is the Multi30k dataset presented by (2). Multi30k was constructed as an extension of the Flickr30k dataset which contains 31,783 images with 5 English descriptions per image totalling 158,915 descriptions. The builders of Multi30k chose one caption for each image and hired professional translators to translate the captions into German. Since it's release the dataset has also been translated to French and Czech. The Freestyle Image Question Answering (FM-IQA) dataset [INCLUDE CITATION!!!] was also of interest for this project. This dataset, currated by Baidu, is a Chinese visual question answering dataset. It contains 158,392 images obtained from the MS COCO dataset [INCLUDE CITATION!!!] along with 316,193 crowd sourced question answer pairs.

Since no English to Chinese MMT datasets are currently available I generated a synthetic dataset to train my models on. I translated sentences using Google's Auto ML translation API. Authentic professionally translated bitext is always preferable to synthetic translations, however, the quality of translations obtained from this pipeline are good and this is probably as close to state of the art that one could get from an off the shelf translation package with no fine tuning. Additionally the models I developed were small so the high quality translations I obtained served as a good substitute for my purposes. Since I had two datasets that had authentic English image descriptions (Multi30k, and Flickr30k) along with one containing authentic Chinese text (FM-IQA) I performed a multiway translation to generate a dataset with X complete sentence pairs and corresponding images. A flowchart of my synthetic data generation is displayed in figure 1.

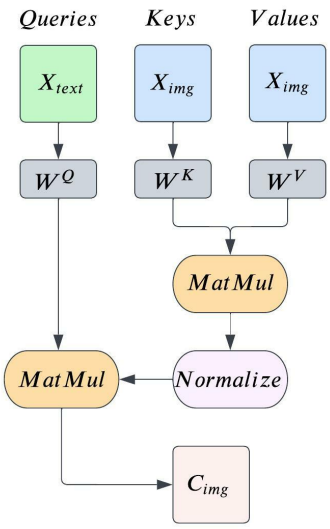

Figure 2: Diagram of the visual cross attention mechanism in the multimodal transformer. Keys, and Values are the projected image features X_{img} while the queries are the outputs of the previous attention block in the encoder or decoder denoted X_{text}

3 Method

For the baseline model I use the Transformer introduced by (5), specifically I used the Annotated Transformer for English to Chinese translation (Cui) coded in pytorch. The implementation outlined in the blog post was used as the baseline model and the code was adapted to incorporate multiple modalities as described below.

3.1 Multimodal Transformer

For the multimodal model I implemented a Transformer with multimodal attention (4). They add an additional attention block in the decoder layers that attends to the visual information. The keys and the values correspond to the visual features while the queries are the output of the first attention block of the decoder. I experiment adding attention in both the encoder and decoder to confirm results of previous work that suggests visual information is most useful at decoding time.

3.2 Spatial Image Features

Image features are extracted from a ResNet-50 model (3) pretrained on the ImageNet dataset. Drawing from past research that indicates the importance of spatial image information in translation, this project extracts spatially aware image feature maps from the penultimate convolutional layer of the ResNet-50. For a single image the feature maps are $7 \times 7 \times 2048$ in dimension. To prepare them to be passed into the multimodal attention block we flatten the first 2 dimensions and project into our model dimension to obtain a tensor of shape $49 \times d_{model}$. To project the image features into the correct dimension we experiment using a simple linear transformation along with a small feed forward network with non-linear activation function.

4 Experimental Setup

4.1 Data

We do not train on data from the FM-IQA dataset restricting our experiments to Flickr30k and Multi30k. There are two main reasons behind this decision. First, I had compute restraints that prohibited me from spending an excessive amout of time training a model. Cutting out FM-IQA nearly halved the training time for a single model. Second, FM-IQA has repetitive sentence structure since I concatenated question answer pairs to form sentences each datapoint is of the form "QUESTION:ANSWER". For example, "How many apples are on the table? There are 5". Using the entire FM-IQA dataset would give us a dataset that is % 50 of the same sentence form. I worried that this repetive simple sentence structure would result in the model over fitting to this type of sentence. The sentences from Flickr30k and Multi30k are slightly more complex and far less repetitive.

I split Flickr30k+Multi30k into train-validation-test sets allocating %80 for train, %20 for validation, and %20 for test. I split by images to ensure that no images showed up in more that one set. In total train has 55,127 instances, and test/validation have 6,891.

4.2 Training

I trained each model for 20 epochs with a batch size of 64. Each model had 6 encoder layers and 6 decoder layers with a model dimension of 256 and feed forward dimension of 2048. I used the NOAM optimizer and label smoothing.

4.3 Human Evaluations

For each pair of translations I displayed the target translation along with translations from both systems in randomized order. I ranked the translations in order of which one I thought was better and counted the total number of times that each system was ranked best. I evaluated 200 sentences under this setup. The results of the evaluations are presented in section 4.3

4.4 Ablation Studies

I ran an ablation study to probe the models ability to leverage correct visual information. I trained the model on the data with the correct images replaced with blank images and random images and evaluated with BLEU and Meteor scores. This should reveal how much of a difference having correct visual features actually makes.

4.5 Source Text Degradation

To further explore the impact of visual context on translation quality I intentionally masked tokens in the source text and trained one multimodal model and one vanilla transformer model. This forces the multimodal model to rely solely on the visual features to infer the correct translation of the masked tokens.

I used two different masking strategies POS masking and deterministic masking. For POS masking I used an off the shelf POS tagger to identify nouns, verbs, and adjectives in the source text. For each token that I identified to belong to one of those three categories randomly masked it with probability 0.3. More specifically, I drew $x \sim Bernoulli(0.3)$ and masked the token if $x = 1$. For deterministic masking I simply masked the last half of each source sentence.

Model Configuration	BLEU	chrF	METEOR
`baseline`	54.27	48.31	76.36
`enc + linear`	53.57	46.93	75.02
`enc + non-linear`	**54.73***	48.23	75.85
`dec + linear`	54.13	47.45	76.12
`dec + non-linear`	**55.07***†	**48.99***†	**76.77***†
`enc + dec + linear`	49.77	43.75	72.83
`enc + dec + non-linear`	**54.50***	**48.53***	76.33

Table 1: Results for different model configurations on the test set. Note `enc` and `dec` denote attention in the encoder and decoder respectively, while `linear` and `non-linear` refer to linear and non-linear image projections. For each metric the best performing model is marked with † and any multimodal model that outperforms the baseline is marked with *****.

5 Results

5.1 Standard Setup

I trained several different model configurations to identify the best way to incorporate visual context into the model. Based off of previous work I expected that visual information added to the decoder would provide the most benefit. I also tested whether a linear or non-linear image feature projection would benefit the model most. In total I trained 7 different model configurations, 1 baseline model and 6 different multimodal variants. The results are presented in table 1.

Multimodal attention in the decoder with a non-linear image projection yielded the best results with a nearly 1 point improvement in BLUE and chrF scores over the baseline. While attention in the decoder did achieve the best scores, it did so at a very small margin. It is unclear if adding multimodal attention in the decoder provides a significant boost in performance. Perhaps the biggest take away from this experiment was the impact of a non-linear image projection on translation quality. Models with non-linear projection heads outperformed their linear counterparts in every case by at least about 1 BLEU point. Noteably, in the models with multimodal attention in the encoder and decoder the non-linear image projection led to a ~ 5 point improvement in BLEU score. This was perhaps the biggest take away from this experiment. In every case having a non-linear projection seems to provide a significant improvement when compared to a non-linear projection.

5.2 Human Evaluations

The results of the human evaluations were inline with the results in table 1. The two systems performed relatively similar with the multimodal system slightly outperforming the baseline. This reinforces the idea that there are few situations where the visual features are really needed to make the correct translation.

	BLEU	METEOR
correct images	**55.07***	76.77
blank images	54.81	76.36
random images	54.14	**76.85***

Table 2: BLEU and Meteor scores on ablation studies. The * denotes the best performing model.

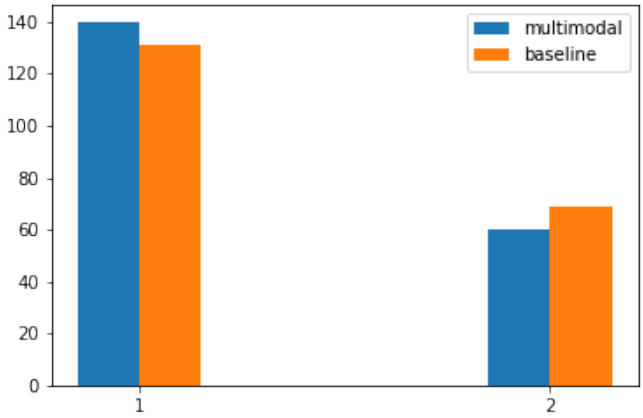

Figure 3: Results of the human evaluation on 200 sentences from `baseline` and `dec + non-linear`

5.3 Ablation Studies

Table 2 includes the results of the ablation studies. Using random images hurt the performance of the model significantly by decreasing the BLEU score by about 4 points and the Meteor score by about X points. Surprisingly replacing all images with blank images did not notably hurt the performance of the model any of the metrics. Using blank images essentially gave the model nothing to learn from the visual features. Since image features for each translation pair were identical the model likely could not learn any mean relationships between specific words and the image features. This model setup was probably more similar to the baseline model with the addition of noise from the blank image features. This along with our original comparison in section 5.1 highlight the fact that there are few cases where the model might need visual context to make a correct translation.

POS Masking	BLEU	METEOR
`baseline`	38.01	60.86
`dec + non-linear`	**39.37***	**62.42***

Deterministic Masking	BLEU	METEOR
`baseline`	28.48	46.96
`dec + non-linear`	**30.66***	**50.67***

Table 3: Results for degradation of source text using probabilistic POS masking and deterministic masking. Note the * denotes the best performing model.

5.4 Source Text Degradation

The results of the source text degradation experiments are presented in table 3. We chose the best performing multimodal model `dec + non-linear` to compare against the baseline model. The multimodal model outperforms the baseline in both probabilistic POS masking and deterministic masking with an average increase in BLEU of $\Delta = 1.77$ and chrF of $\Delta = 1.5$. This difference in performance highlights the multimodal models ability to use visual information to correctly infer missing information in the source text.

6 Conclusion

This project has introduced a synthetically generated multimodal dataset for English to Chinese machine translation. I built a multimodal transformer inspired by (4) and applied it to the new synthetic dataset. Applying MMT to this specific language pair has not previously been done. The results were consistent with previous work that has been done in this area, indicating that while visual information can provide a small boost in translation quality it is not significant unless there is missing information or noise in the source text. In future work I aspire to investigate possibly post editing the synthetic Chinese translations with hopes of publicly releasing this data to the research community to advance further research in multimodal machine translation.

References

[Cui] Cui, C. The annotated transformer: English-to-chinese translator.

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

[3] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[4] Helcl, J., Libovický, J., and Variš, D. (2018). CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.