IITP at WAT 2021: System description for English-Hindi Multimodal Translation Task

Baban Gain[*]
Indian Institute of Technology Patna
Patna, India
gainbaban@gmail.com

Dibyanayan Bandyopadhyay[*]
Indian Institute of Technology Patna
Patna, India
dibyanayan@gmail.com

Asif Ekbal
Indian Institute of Technology Patna
Patna, India
asif@iitp.ac.in

Abstract

Neural Machine Translation (NMT) is a predominant machine translation technology nowadays because of its end-to-end trainable flexibility. However, NMT still struggles to translate properly in low-resource settings specifically on distant language pairs. One way to overcome this is to use the information from other modalities if available. The idea is that despite differences in languages, both the source and target language speakers see the same thing and the visual representation of both the source and target is the same, which can positively assist the system. Multimodal information can help the NMT system to improve the translation by removing ambiguity on some phrases or words. We participate in the 8th Workshop on Asian Translation (WAT - 2021) for English-Hindi multimodal translation task and achieve 42.47 and 37.50 BLEU points for Evaluation and Challenge subset, respectively.

1 Introduction

Recent progress in neural machine translation (NMT) focuses on translating a source language into a particular target language. Various methods have been proposed for this task and most of them deal with the textual data. There are certain drawbacks while performing machine translation using only textual datasets.

Human performs translation which is based upon language grounding: our sense of meaning emerges from interacting with the world. NMT methods do not have any mechanism to perform language grounding; thus they are devoid of capturing the true meaning of sentences or phrases while translating them into the other languages. For example, it needs to translate the word "cricket", it can get confused if it is the game cricket or the insect cricket. But the visual information can clear the ambiguity. Multi-modal translation aims to alleviate this issue by training an NMT model on textual data along with associated images to perform language grounding.

This shared task deals with developing multi-modal NMT models for English-Hindi translation. The choice of languages depends on the following issues: *i).* Hindi is the most spoken language in India and the fourth most spoken language in the world with 600 million speakers[1]. Despite the huge amount of speakers, suitable resources in Hindi is limited due to the various factors. *ii)* Automatic translation of texts from one language to the another is a difficult task. Specifically, when one or both of them are resource-poor and distant from each other.

In Multimodal NMT (MNMT), information from the other modalities like audio, image, video, etc. are used along with text to generate the translation. In low-resource languages, this is particularly used to improve the low-quality translations as even though vocabularies, grammar of two languages are different but their visual representation is the same. There are several proposed multi-modal methods for translations that exploit the features of the associated image for better translation. State-of-the-art methods might achieve better accuracy than the models we used. Our main motivation for using simplistic models is to demonstrate a proof-of-concept to be used for multi-modal translation among the resource-poor language pairs. We achieved good results on both Challenge and Evaluation set in different evaluation metrics including BLEU, RIBES, AMFM. In subsequent modifications, we aim to develop our models incorporating

---

[*]Equal contribution

several state-of-the-art features. The following sections describe our processes in greater details.

2 Related Works

There have been many attempts to use information other than the source for better translation. Uni-modal systems include document-level NMT (Wang et al., 2017), sentence-level NMT with contextual information (Gain et al., 2021), etc. Among multimodal systems, (Huang et al., 2016) used an object detection system and extracted local and global image features. Thereafter, they used those image features as additional inputs to encoder and decoder. (Delbrouck and Dupont, 2017) used attention mechanism on visual inputs for the source hidden states. (Lin et al., 2020) used Dynamic Context-guided Capsule Network (Sabour et al., 2017) (DCCN) for iterative extraction of related visual features.

Multimodal Machine Translation (MMT) for English-Hindi has not been well explored yet. (Dutta Chowdhury et al., 2018) used synthetic data for training. Furthermore they used multi-modal, attention-based MMT which incorporate visual features into different parts of both the encoder and the decoder (Calixto and Liu, 2017). (Sanayai Meetei et al., 2019) used a Recurrent Neural Network (RNN) based approach achieving BLEU score of 28.45 on Evaluation set and 12.58 on Challenge set. (Laskar et al., 2020) exploited monolingual data for better translation. Recent works tried to focus on developing unsupervised model for multi-modal NMT. Su et al. (2018) demonstrated an unsupervised method based on the language translation cycle consistency loss conditional on the image. This is done to learn the bidirectional multi-modal translation simultaneously. Moreover, Su et al. (2021) showed that jointly learning text-image interaction instead of modeling them separately using attentional networks is more useful. This result is in line with several state-of-the-art visual transformer related models, such as VisualBERT (Li et al., 2019), UNITER (Chen et al., 2019) etc.

3 Methodology

3.1 Dataset Description

We use Hindi Visual Genome 1.1 dataset (Parida et al., 2019)(Nakazawa et al., 2020)(Nakazawa et al., 2021) for our experiments. This dataset consists of 28,929 parallel English-Hindi sentence

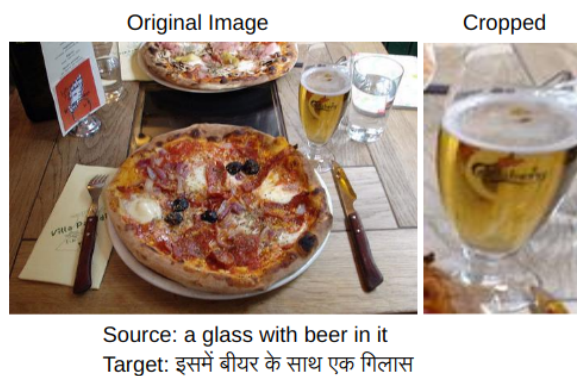Source: a glass with beer in it
Target: इसमें बीयर के साथ एक गिलास

Figure 1: An example of multimodal dataset

pairs along with the associated images. Furthermore, we use HindEnCorp dataset for pre-training containing 273K English-Hindi sentence pairs without images. Statistics of the datasets are shown in Table 1.

Multimodal dataset consists of an image along with a description of certain rectangular portion of the image. We are given the coordinates of the portion. We aim to translate the description with help of the image. An example of multimodal dataset is given in Figure 1.

3.2 Pre-processing

For text data, we lowercase all the utterances. Then, we jointly learn byte-pair-encoding (Sennrich et al., 2016) combining both source and target with a vocabulary of 10,000. We treat the images by cropping a specified rectangular portions. This operation is used to discard the portions that do not contribute much to the translation performance. After we get those cropped-out images, we use the pre-trained VGG19-bn (Simonyan and Zisserman, 2015) to obtain the image representations. We use OpenNMT-py (Klein et al., 2017) framework to perform this step.

3.3 Training

We use OpenNMT-py (Klein et al., 2017) for our NMT systems. We use Bidirectional RNN encoder and doubly attentive RNN decoder (Calixto et al., 2017) for our experiments. We train our system in two ways *viz*. With pre-training, and Without pre-training.:

1. **With pre-training** We pre-train one of our models on HindEnCorp dataset. This step does not use any visual features as the dataset used for pre-training is devoid of any visual

Dataset	Type	Sentences	Avg length source	Avg length target
Pre-training	Parallel	273,885	12.33	13.36
Train	parallel + multimodal	28929	4.95	5.02
Valid	parallel + multimodal	998	4.93	4.99
Evaluation	parallel + multimodal	1595	4.92	4.92
Challenge	parallel + multimodal	1400	5.85	6.17

Table 1: Descriptions of datasets used for our experiments

Figure 2: An example of translation generated by the system. Here, the target is *Ek vyakti railgari mein chad raha hai (A man climbing into train.)* The translation by Google NMT system is *Train mein chadta Aadmi (Man climbs into train)*; whereas our NMT system translates it as: *Ek aadmi ek train mein chadta hai (A man climbs into a train.)*

features. After pre-training, we fine-tune the pre-trained model with VisualGenome dataset containing textual and visual features.

2. **Without pre-training** We do not pre-train the model. We directly fine-tune the models on VisualGenome dataset which contains both text and associated image. Consequently, both textual and visual features are used.

Following step is taken into account while doing inference step:
We take the best hypothesis from both the models and filter out any hypothesis containing <unk>token. Then, we pick the hypothesis with best log-likelihood during generation.

3.4 Hyper-parameters

We set the word embedding size and size of RNN hidden states to 500. We set the batch size to 40 and train for a maximum 25 epochs. We restrict maximum source and target sequence length to 50. We use the Adam optimizer (Kingma and Ba, 2017) for optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. During training, we use 0.3 as dropout rate to avoid over-fitting. During generation of translation, we use 5 as the beam width.

4 Experimental Results

We obtain impressive results on our submissions. There are two sets designed for evaluating our model, *i) Evaluation set*, *ii) Challenge set*. We evaluate our model on both of these test set and tabulate our results in Table 2. We use different evaluation metrics (BLEU, RIBES, AMFM) to test our model. The results shown in the table are sorted according to the obtained BLEU scores. As it can be seen from Table 2, we obtain 42.47 BLEU points and achieve second position in terms of BLEU on Evaluation set on multimodal task. Please refer to Figure 2 for example of translation by our system. We obtain 37.50 BLEU points on Challenge set. One reason for not so good results on Challenge set could be:

- The challenge test set was created by searching for (particularly) ambiguous English words based on the embedding similarity and manually selecting those where the image helps to resolve the ambiguity. Hence, it is difficult to translate compared to the Evaluation set, which was randomly selected.

- Difference between utterance length during training and testing, i.e. while average length

Team	Evaluation			Challenge		
	BLEU	RIBES	AMFM	BLEU	RIBES	AMFM
Volta	44.21	0.818689	0.835480	52.02	0.854139	0.874220
iitp (Ours)	42.47	0.807123	0.819720	37.50	0.790809	0.830230
CNLP-NITS	40.51	0.803208	0.820980	39.28	0.792097	0.812360
CNLP-NITS	39.46	0.802055	0.823270	33.57	0.754141	0.787320
Organizer	38.63	0.767422	0.772870	20.34	0.644230	0.669760

Table 2: Details of obtained results by different submissions

of Train, Evaluation and Validation set is 5 but average length of Challenge set is 6.

5 Conclusion

We participate in WAT-2021 Multimodal Translation Task for English to Hindi. We achieve good results on both the Challenge and Evaluation sets achieving 42.47 and 37.50 BLEU points, respectively. We rank second place on Evaluation set and third place on Challenge set on WAT-2021 Multimodal Translation Task for English to Hindi. In future, we would like to extend our work by training with additional monolingual data and better ways to incorporate multimodal features.

References

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. Modulating and attending the source image during encoding improves multimodal translation.

Koel Dutta Chowdhury, Mohammed Hasanuzzaman, and Qun Liu. 2018. Multimodal neural machine translation for low-resource language pairs using synthetic data. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 33–42, Melbourne. Association for Computational Linguistics.

Baban Gain, Rejwanul Haque, and Asif Ekbal. 2021. Not all contexts are important: The impact of effective context in conversational neural machine translation. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

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

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Multimodal neural machine translation for English to Hindi. In *Proceedings of the 7th Workshop on Asian Translation*, pages 109–113, Suzhou, China. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. *Proceedings of the 28th ACM International Conference on Multimedia*.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui

Chu, Akiko Eriguchi, Kaori Abe, and Sadao Oda, Yusuke Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation*, Bangkok, Thailand. Association for Computational Linguistics.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Suzhou, China. Association for Computational Linguistics.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*, 23(4):1499–1505. Presented at CICLing 2019, La Rochelle, France.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, 554:47–60.

Yuanhang Su, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, and Fei Huang. 2018. Unsupervised multi-modal neural machine translation. *CoRR*, abs/1811.11365.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.