Experiences of Adapting Multimodal Machine Translation Techniques for Hindi

Baban Gain
Indian Institute of Technology Patna
gainbaban@gmail.com

Dibyanayan Bandyopadhyay
Indian Institute of Technology Patna
dibyanayan@gmail.com

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

Abstract

Multimodal Neural Machine Translation (MNMT) is an interesting task in natural language processing (NLP) where we use visual modalities along with a source sentence to aid the source to target translation process. Recently, there has been a lot of works in MNMT frameworks to boost the performance of standalone Machine Translation tasks. Most of the prior works in MNMT tried to perform translation between two widely known languages (e.g. English-to-German, English-to-French). In this paper, We explore the effectiveness of different state-of-the-art MNMT methods, which use various data oriented techniques including multimodal pre-training, for low resource languages. Although the existing methods works well on high resource languages, usability of those methods on low-resource languages is unknown. In this paper, we evaluate the existing methods on Hindi and report our findings.

1 Introduction

Machine Translation (MT) has been a challenging task in natural language processing (NLP). In the last few years, there have been significant progress in MT due to easier accessibility to data, computation and discovery of new deep learning based MT techniques. The argument over using text only dataset for MT is limited when viewed with respect to how human performs translation between two languages. We, as humans comprehend the world and perform action by combining different input modalities (e.g, Text, Image etc). Similarly the same argument can be used to develop machines that can process different input modalities to perform downstream tasks. In this paper we specifically focus on the task of MT combined with visual inputs. We call this task as Multimodal MT (w.r.t the fact that we use visual modalities combined with classical text-only MT framework). We specifically aim to address the translation task between English and a low resource language (Hindi) and illustrate that by combining classical MT methods with visual information and explore how this could improve the task of translation. There are technical motivation too for performing multimodal translation tasks. Text-based systems are unable to translate sentences that are ambiguous, which can have different target translations depending upon the situation. While words surrounding the ambiguous word can clear ambiguity up to some extent, it might be helpful to introduce related information from other modalities. Following these motivation, we describe several baseline approaches well known in the field of multimodal MT and adapt them for a low-resource language. Specifically, we focus on applying the benchmark methodologies for translating between English and Hindi and vice-versa with the help of visual information.

2 Related Work

Several systems attempted to use extra information to translate text. Unimodal systems include document-level NMT, (Wang et al., 2017) which utilises document as context, sentence-level NMT with contextual information (Gain et al., 2021b), etc. Among multimodal systems, (Huang et al., 2016) used an object detection system to extract 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. Su et al. (2018) demonstrated an unsupervised method based on the language translation cycle consistency loss conditioned on

Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021),
in conjuction with RANLP-2021, September 7, 2021. pages 40–44
https://doi.org/10.26615/978-954-452-073-1_007

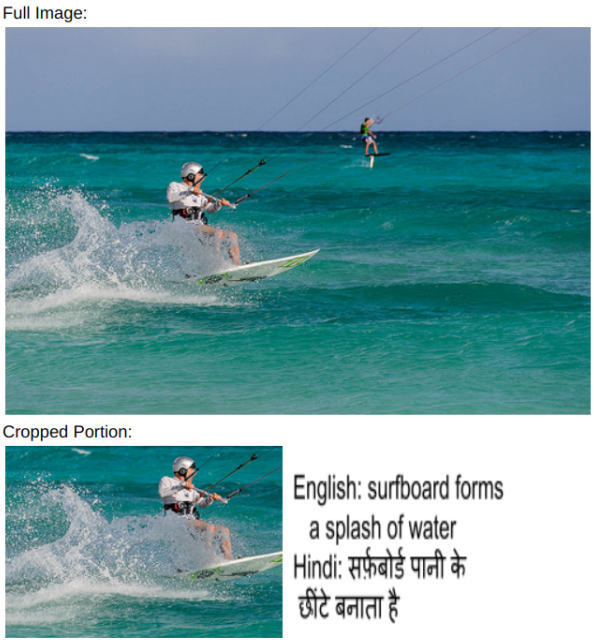

Figure 1: An example of the multimodal dataset

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 attention networks is more useful.

There exists a few multimodal Machine Translation (MMT) methods for English-Hindi language pair. (Gupta et al., 2021) proposed to enhance the textual input by bringing the visual information to a textual domain by extracting object tags from the image. For pre-training, they used IIT corpus (Kunchukuttan et al., 2018). Other methods include usage of doubly attentive decoders as mentioned in Section 4.2.

3 Dataset Description & Pre-processing

We use Hindi Visual Genome 1.1 dataset (Nakazawa et al., 2021) introduced in The 8-th Workshop on Asian Translation (WAT) consisting of a total of 32,922 sentences splited between 28929 train, 998 valid, 1595 test and 1400 challenge sentences. Each sentence represents a caption/description of a rectangular portion of an image, associated with the sentence. Coordinates of the rectangular portion is available with the dataset.

We convert all the datasets into lowercase. Then, we combine HindEnCorp (Bojar et al., 2014) and Visual Genome 1.1 training set and learn byte-pair-encoding (Sennrich et al., 2016) with 20,000 operations using subword-nmt [1].

For image processing, we use the following two methods:

- We crop the rectangular portion of the image, which represents the caption and discard the remaining part of the image as they may not contribute much to the translation performance and can introduce noise. In our experiments in Section 5, we represent experiments with cropped image with an *crop* identifier.
- Sometimes, the cropped images are too small and can miss out important information. Also, in some cases, translation system might utilise background image information. Therefore, we perform another set of experiments with no cropping. In Section 5, absence of *crop* identifier indicate usage of full image.

We extract pre-trained ResNet50 features from the images as ResNet50 is used for image features in most of the existing methods.

4 Methods

4.1 Multimodal Transformer

We use a transformer based Multimodal Machine Translation approach proposed by (Yao and Wan, 2020), where they suggested that if every word is considered as a node, then Transformer can be regarded as a variant of Graph Neural Network (GNN) (Yao et al., 2020) which treats each sentence as a fully-connected graph with words as node.

We adapt this technique for our language pair. As described in the paper using their default configurations, we initialize word embedding by 300 dimension pre-trained GloVe word embeddings. We do not pre-train the systems. We train the systems with Visual Genome dataset (Nakazawa et al., 2021).

4.2 Doubly-Attentive Decoder

A multimodal architecture introduced by (Calixto and Liu, 2017)(Calixto et al., 2017) consists of bidirectional Recurrent Neural Network (BRNN) as encoder type, and doubly-attentive RNN as decoder type which incorporates two independent attention mechanisms, one over source language words and the other over different areas of an image. The decoder incorporates spatial visual features obtained

[1] https://github.com/rsennrich/subword-nmt

Table 1: Results obtained by different systems. BLEU is calculated using multi-bleu.perl. sBLEU_intl represents sacreBLEU with intl tokenizer

Method	Test Set				Challenge Set			
	BLEU	RIBES	sacreBLEU	sBLEU_intl	BLEU	RIBES	sacreBLEU	sBLEU_intl
Volta(Gupta et al., 2021)	44.21	0.8186	-	-	52.02	0.8541	-	-
iitp(Gain et al., 2021a)	42.47	0.8071	-	-	37.50	0.7908	-	-
CNLP-NITS(Laskar et al., 2021)	40.51	0.8032	-	-	39.28	0.7920	-	-
CNLP-NITS	39.46	0.8020	-	-	33.57	0.7541	-	-
WAT 2021 Organizer(Nakazawa et al., 2021)	38.63	0.7674	-	-	20.34	0.6442	-	-
Multimodal_Transformer(Yao and Wan, 2020)	38.30	0.7596	37.6	38.1	24.29	0.6708	23.7	24.2
Multimodal_Transformer_crop	38.53	0.7703	37.9	38.3	27.91	0.6882	27.4	27.8
NMT_src+img(Calixto et al., 2017)	39.83	0.7968	39.6	39.6	31.41	0.7387	30.8	31.3
NMT_src+img_crop	39.72	0.7910	39.6	39.6	31.81	0.7348	31.4	31.6
Transformer_text_only(Vaswani et al., 2017)	41.97	0.8091	41.9	41.9	28.53	0.6933	28.3	28.5

using pre-trained convolutional neural networks (CNNs).

There are several ways to adapt this method. (Dutta Chowdhury et al., 2018) used synthetic data for training. (Sanayai Meetei et al., 2019) used cropped rectangular portion of images to assist the system to translate. (Laskar et al., 2020) utilizes pre-train word embeddings of the monolingual corpus and additional parallel data (IITB corpus). In (Laskar et al., 2021), they made attempts to utilize phrase pairs (Sen et al., 2021) to enhance the translation performance. (Gain et al., 2021a) used a combination of systems, one with pre-training and one without pre-training for text along with cropped rectangular portions of images. We perform our experiments on $NMT_{src+img}$ method as described in (Calixto et al., 2017). For sake of better comparison with other methods, we do not pre-train the systems using additional data except for learning byte-pair-encoding described in Section 3

4.3 Text Only Transformer

We experiment on the standard transformer (Vaswani et al., 2017) model using its fairseq (Ott et al., 2019) implementation. We byte-pair-encode texts by HindEnCorp and Visual Genome 1.1 training set using fastBPE. We do not use any image features on this method.

5 Evaluation Results

For every method, we train the model with 100 epochs and pick the model with best validation set result to generate translations. We keep the same hyper-parameters as reported in the respected paper. We translate all text from English to Hindi by using beam size of 5. For BLEU and RIBES score, we tokenize all texts with Indic-tokenizer and calculate BLEU with Moses *multi-bleu.perl* and RIBES v1.02.4 [2]. We calculate BLEU scores using sacreBLEU (Post, 2018) using two different tokenizers [3]. We report our results in Table 1. First five rows contains official results from WAT 2021. Row 6-10 contain results by other methods experimented by us. For the test set, the object tag based system (Gupta et al., 2021) secured the best result with 44.21 BLEU score and 0.8186 RIBES score. It is to be noted that, results obtained using other methods are close. Even text-only system generated good results and only two systems generated better results than text-only system. Considering text-only system do not use any pre-training and other systems use sophisticated methods for pre-training, we suggest that image features have minimal impact on the test set by most of the systems.

The challenge set was created by searching for (particularly) ambiguous English words based on the embedding similarity and manually selecting those where the image and surrounding text help to resolve the ambiguity. Object tag based system performs excellently compared to the other methods and achieves 52.02 BLEU points. Most of the systems perform better than the text-only MNMT. Thus, image features have significant contribution in the performance of the challenge set. It is to be noted that, while image features add extra information to clarify ambiguous sentences, it also drops translation results in non-ambiguous sentences, acting as noise. In future, it will be important to build generalized system that can handle both of these properties.

Furthermore, we notice that cropped images helped to generate better results than that of full images.

[2] http://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/automatic_evaluation_systems/automaticEvaluationHI.html

[3] sacreBLEU signature is: *BLEU+case.mixed+numrefs.1 +smooth.exp+tok.(13a,intl)+version.1.4.13*

6 Research Directions

Multimodal Machine Translation approach proposed by (Yao and Wan, 2020) (Multimodal Transformer) used cross-attention mechanism in place of standard self attention to model the interdependence between textual and visual modalities. Cross attention mechanism is used as a generalization of self-attention mechanism inside transformer models where the input Query (Q) is different compared to same Key (K) and Value (V) vectors. The standard mechanism used in Multimodal transformer was to input concatenated representation of textual and visual tokens as Query vector whereas visual tokens are input as both Key and Value vectors. We propose the following directions for future experiments.

- Implement cross-attention mechanism to model dependence between textual and visual representation by using different input modalities as Key and Query vector, respectively. This should be different as modeled by Multimodal transformer.
- Model text-image and image-text relationship with the help of cross-attention, by different choice of Q and K. For text-image relationship, we use Q as visual tokens and K as text tokens and vice versa for image-text relationship.
- Concatenating text-image and image-text representations as more compact measure of interdependence between the two.

We aim to do these experiments in the next iterations of the paper.

7 Conclusion

Neural machine translation is a very challenging task as one sentence can be translated into multiple ways. Multimodal translation is used to introduce information from different modalities to assist the system to translate. There are several systems that translate multimodal information from one language to the other. These methods are proven to be helpful to generate better translation on high resource languages including English, German, French, etc. However, it was unknown how useful these methods are, specifically on low resource languages like Hindi. We show comparisons of different state-of-the-art MNMT systems. We observe that multimodal information is useful to translate ambiguous sentences on Hindi. Furthermore, we found that multimodal information can act as noise and may not be worth using in case of non-ambiguous sentences. We hope those will serve as reference for future MNMT systems on low resource settings. In future, we would like to propose different MNMT methods for low resource languages maximizing multimodal information to improve translation while minimizing the same acting as noise. Furthermore, we would like to extend our work on more existing methods and adapt them to more low resource languages.

References

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

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, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021a. IITP at WAT 2021: System description for English-Hindi Multimodal Translation Task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 161–165, Online. Association for Computational Linguistics.

Baban Gain, Rejwanul Haque, and Asif Ekbal. 2021b. Not all contexts are important: The impact of effec-

tive context in conversational neural machine translation. In *2021 International Joint Conference on Neural Networks (IJCNN)*.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.

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.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. Improved English to Hindi multimodal neural machine translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 155–160, Online. 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.

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, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Online. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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.

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering*, 27(3):271–292.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Shaowei Yao, Tianming Wang, and Xiaojun Wan. 2020. Heterogeneous graph transformer for graph-to-sequence learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7145–7154, Online. Association for Computational Linguistics.