BITS-P at WAT 2023: Improving Indic Language Multimodal Translation by Image Augmentation using Diffusion Models

Amulya Ratna Dash[†*] p20200105@pilani.bits-pilani.ac.in
Hrithik Raj Gupta[†] f20190995@pilani.bits-pilani.ac.in
Yashvardhan Sharma[†] yash@pilani.bits-pilani.ac.in
†BITS Pilani, Pilani, Rajasthan, India
*IQVIA, Bangalore, Karnataka, India

Abstract

This paper describes the proposed system for mutlimodal machine translation. We have participated in multimodal translation tasks for English into three Indic languages: Hindi, Bengali, and Malayalam. We leverage the inherent richness of multimodal data to bridge the gap of ambiguity in translation. We fine-tuned the 'No Language Left Behind' (NLLB) machine translation model for multimodal translation, further enhancing the model accuracy by image data augmentation using latent diffusion. Our submission achieves the best BLEU score for English-Hindi, English-Bengali, and English-Malayalam language pairs for both Evaluation and Challenge test sets.

1 Introduction

Machine Translation (MT) is the NLP task of translation between language pairs. Multimodal Machine Translation (MMT) is the translation process that utilizes information from multiple modalities, not just text. The most popular approach is to use extra visual context in addition to the source text input. The visual context is presented in the form of images that are relevant to the text to be translated and may help in cases of ambiguity.

In 10th Workshop on Asian Translation (WAT 2023), we investigate Multimodal Machine Translation for English to Hindi, Bengali, and Malayalam languages by fine-tuning the 'No Language Left Behind' (NLLB) pre-trained machine translation model by Costa-jussà et al. (2022) and using image data augmentation techniques. Figure 1 shows an example where the visual context of an image helps generate accurate machine-translated text.

The rest of the paper is organized as follows: Section 2 presents the review of related works. The data and system are briefly described in Section 3. Section 4 reports the results, followed by the future scope and conclusion in Section 5.

2 Related Work

In the literature survey, there are some multimodal machine translation works that take both text and image as input, and learn joint multimodal representations from images and text (Specia et al., 2016). Huang et al. (2016) incorporated an object detection system, extracting local and global image features as additional inputs to the encoder and decoder. Lin et al. (2020) utilized

Proceedings of the 10th Workshop on Asian Translation, pages 41–45
September 4, 2023, Macau SAR, China

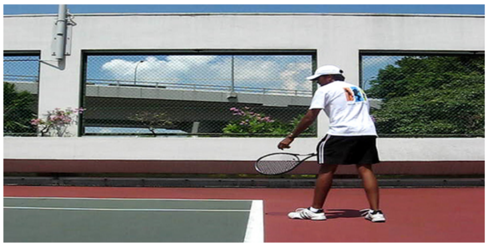

Image caption: A large pipe extending from the wall of the court.
Hindi translation: कोर्ट की दीवार से निक्ली हुई एक बड़ी पाइप
Bengali translation: কোর্টের দেয়াল থেকে প্রসারিত একটি বৃহত পাইপ।
Malayalam translation: കോർട്ടിന്റെ മതിലിൽ നിന്ന് നീളുന്ന ഒരു വലിയ പൈപ്പ്.

Figure 1: Example of Multimodal translation

Dynamic Context-guided Capsule Network (DCCN) for iterative extraction of related visual features. Most of them investigated multimodal MT for high-resource European languages. There are only a few works for Indic languages.

Dutta Chowdhury et al. (2018) who employed synthetic data for training and used multimodal, attention-based MT incorporating visual features into the encoder and decoder (Calixto and Liu, 2017). Su et al. (2019) further demonstrated the advantage of jointly learning text-image interaction rather than modeling them separately using attentional networks.

Parida et al. (2019) proposed a subset of the Visual Genome dataset (Krishna et al., 2017) for multimodal translation between English and Hindi, a less explored language pair in this context. Parida et al. (2021) used the Bengali Visual Genome (Sen et al., 2022) and adopted the ViTA (Gupta et al., 2021) approach, with mBART (Liu et al., 2020) for encoding English sentences with object tags and decoding Bengali translations.

Our current work builds upon these foundations, introducing a novel approach using NLLB and Stable Diffusion (Rombach et al., 2022) to multimodal translation, with a specific focus on the English to Hindi, Bengali, and Malayalam language pairs.

3 System Overview

In this section, we describe the dataset, data augmentation technique, and model approach for the proposed system we use for the multimodal translation task.

3.1 Dataset Description

The primary datasets utilized for training are the Hindi Visual Genome (HVG), the Bengali Visual Genome (BVG), and the Malayalam Visual Genome (MVG). The HVG dataset comprises of around 29K parallel English-Hindi sentence pairs, each associated with an image. Each data point in the multimodal dataset contains an image alongside a textual description of a certain rectangular portion of the image, delineated by provided coordinates. The task is to translate these descriptions using contextual support from the images.

The datasets also contain three test sets apart from the training set: a development test set (D-Test), an evaluation test set (E-Test), and a challenge test set (C-Test). The BVG and MVG datasets are structured identically to HVG, with the same images and image captions, but the captions are translated into Bengali and Malayalam, respectively.

We augmented the training data by using Stable Diffusion (v. 1.5)[1] to generate synthetic

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

images based on the English image captions. This technique doubled the image data available for each sentence in the training dataset, providing us with two images for each sentence. We prepared two set of training data, one with 26K data points (Visual Genome) and the other which includes additional 26K data points based on synthetic images. Our final training dataset consists of around 58K unique data points.

3.2 Model Description

DETR (DEtection TRansformer) (Carion et al., 2020) is a transformer-based object detection model that performs end-to-end object detection by directly outputting object detections as sets, eliminating the need for traditional region proposal methods. To extract the image features, we used the DETR object detection system with a ResNet-50 backbone[2]. This allowed us to identify the objects contained within each image. We appended the sentences with a comma-separated list of detected objects preceded by '##' to ensure a clear demarcation between the sentence and the object list.

We fine-tuned the NLLB-200 model[3] for each language with randomly shuffled training dataset, using the D-Test set of our datasets as the validation dataset with 998 data points. We fine-tuned the model for Hindi and Bengali for 100 epochs, while the Malayalam model was fine-tuned for 70 epochs.

The fine-tuning was accomplished utilizing an NVIDIA A100 GPU, and the following training hyperparameters were used: learning rate: 2e-05, batch size: 32, Optimizer: Adam, configured with β_1 = 0.9, β_2 = 0.999, and ϵ = 1e-08.4, and learning rate scheduling: linear.

We trained two models for Hindi language, one with original training dataset and the other with data augmentation for comparative evaluation. Both the models used same hyperparameters and trained for 100 epochs.

4 Experimental Results

Our model achieves promising results during the automated evaluation[4]. To evaluate our model, we used two test sets: i) Evaluation set (E-Test) with 1595 sentences, and ii) Challenge set (C-Test) with 1400 sentences. We test our model on both of these test sets and present the results in Table 1, Table 2, and Table 3 for English-Bengali, English-Hindi, and English-Malayalam tasks, respectively. We use BLEU and RIBES as evaluation metrics.

The English-Hindi multimodal machine translation model trained with additional synthetic image data achieves 52.10 and 45 BLEU score on C-Test and E-Test, while the baseline model trained only on Hindi Visual Genome data achieves 51.20 and 44.90 BLEU score on C-Test and E-Test respectively. The data augmentation technique improved the BLEU score by +0.9 on Challenge set (C-Test).

[2] https://huggingface.co/facebook/detr-resnet-50

[3] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[4] https://lotus.kuee.kyoto-u.ac.jp/WAT/evaluation/index.html

Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7123	**50.60**	**0.814207**
Best-Comp	6743	43.90	0.780669
Challenge Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7122	**48.70**	**0.831946**
Best-Comp	7108	30.50	0.690706

Table 1: Results for EN-BN multimodal translation task

Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7125	**45.00**	**0.829320**
Best-Comp	6428	44.64	0.823319
Challenge Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7124	**52.10**	0.853388
Best-Comp	6430	51.60	**0.859645**

Table 2: Results for EN-HI multimodal translation task

Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7127	**51.90**	**0.799683**
Best-Comp	6936	41.00	0.705349
Challenge Test			
Team	**Data ID**	**BLEU**	**RIBES**
BITS-P	7126	**42.20**	**0.759248**
Best-Comp	6937	20.40	0.533737

Table 3: Results for EN-ML multimodal translation task

5 Conclusion

In this paper, we described our multimodal machine translation system. Our system scored 50.60 and 48.70 BLEU points for Bengali; 45.00 and 52.10 BLEU points for Hindi, and 51.90 and 42.20 BLEU points for Malayalam for Evaluation and Challenge test sets, respectively. The data augmentation strategy of using synthetic images generated by diffusion models improved the system by +0.9 BLEU score. In the future, we would further explore data augmentation approaches for text data using image captioning frameworks.

References

Calixto, I. and Liu, Q. (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.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Dutta Chowdhury, K., Hasanuzzaman, M., and Liu, Q. (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.

Gupta, K., Gautam, D., and Mamidi, R. (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.

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., and Dyer, C. (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.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020). Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Parida, S., Bojar, O., and Dash, S. R. (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.

Parida, S., Panda, S., Biswal, S. P., Kotwal, K., Sen, A., Dash, S. R., and Motlicek, P. (2021). Multimodal neural machine translation system for English to Bengali. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 31–39, Online (Virtual Mode). INCOMA Ltd.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Sen, A., Parida, S., Kotwal, K., Panda, S., Bojar, O., and Dash, S. R. (2022). Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70. Springer.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). 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*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Su, Y., Fan, K., Bach, N., Kuo, C.-C. J., and Huang, F. (2019). Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.