DCU ADAPT at WMT24: English to Low-resource Multi-Modal Translation Task

Sami Ul Haq, Rudali Huidrom, Sheila Castilho
ADAPT Centre, Dublin City University, Dublin, Ireland
{sami.haq, rudali.huidrom, sheila.castilho}@adaptcentre.ie

Abstract

This paper presents the system description of "DCU_NMT's" submission to the WMT-WAT24 English-to-Low-Resource Multimodal Translation Task. We participated in the English-to-Hindi track, developing both text-only and multimodal neural machine translation (NMT) systems. The text-only systems were trained from scratch on constrained data and augmented with back-translated data. For the multimodal approach, we implemented a context-aware transformer model that integrates visual features as additional contextual information. Specifically, image descriptions generated by an image captioning model were encoded using BERT and concatenated with the textual input.

The results indicate that our multimodal system, trained solely on limited data, showed improvements over the text-only baseline in both the challenge and evaluation sets, suggesting the potential benefits of incorporating visual information.

1 Introduction

The increasing prominence of multimodal content in the machine translation (MT) community highlights its potential to improve translation quality by incorporating visual context, which is otherwise inaccessible through textual information alone. This approach has significant implications for commercial applications, including the translation of image captions in online news articles and the translation of product descriptions in e-commerce platforms (Belz et al., 2017; Calixto et al., 2017; Lala et al., 2017; Zhou et al., 2018). By integrating visual information, multimodal MT systems can achieve more accurate and contextually appropriate translations.

Despite MT achieving near-human performance for many high-resource languages, significant challenges remain, particularly for low-resource languages (Popel et al., 2020; Costa-jussà et al., 2022). In recent years, the integration of additional modalities, such as images, into MT systems has gained prominence as a critical area of research (Sulubacak et al., 2020; Parida et al., 2021b,a). This multimodal approach seeks to address the limitations of traditional text-only MT by incorporating supplementary contextual information, thereby improving translation accuracy and expanding the applicability of MT across a broader spectrum of languages and specialised domains.

The WMT-WAT 2024 Shared Task[1] introduces the "English to Lowres Multi-Modal Translation Task," utilizing the Hindi, Bengali, Malayalam, and Hausa Visual Genome datasets. Participants are given an image, a specific rectangular region within it, and a short English caption describing the region. The task is to translate the caption into one of the target languages: Hindi, Bengali, Malayalam, or Hausa.

In this system description paper, we explain our approach for the tasks in which we participated in English (EN) to Hindi (HI) (i) Text only and (ii) Multimodal translation. We released the code and data produced during research through GitHub[2].

2 Dataset

We use the data sets provided by the organizers for the relevant tasks. The Visual Genome datasets for Hindi, Bengali, Malayalam, and Hausa include 29,000 training examples, 1,000 examples for development, and 1,600 examples for evaluation. These datasets are based on a shared set of images, with some variations due to independent sanity checks conducted for each language. For evaluation, the WMT-WAT 2024 Multimodal Shared Task utilises 1,600 examples from the evaluation set and 1,400 examples from the challenge

[1] https://www2.statmt.org/wmt24/multimodallowresmt-task.html

[2] https://github.com/sami-haq99/DCU_NMT_WMT-WAT24

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 810–814

set. In this submission, we denote evaluation set as "EV" and challenge set as "CH" respectively. The statistics of the dataset are shown in Table 1. Due to time constrained, We only trained our systems for English-Hindi language pair.

		Tokens	
Set	Sentences	English	Hindi
Train	28930	143164	145448
D-Test	998	4922	4978
E-Test	1595	7853	7852
C-Test	1400	8186	8639

Table 1: Statistics of our data used in the English→Hindi Multimodal translation task.

3 Experimental Details

In this section, we present our experimental details for the tasks we participated in.

3.1 Text-only translation

For the EN-HI text-only translation task, we have two submissions: one restricted and the other using additional monolingual data.

Back-translation enables the effective use of monolingual data to improve the MT system, especially in a low-resource context (Sennrich et al., 2016; Ul Haq et al., 2020) where model struggles to learn reliable alignments from limited parallel data. For our experiments, we used backtranslated data generated from Flickr8k image captioning data set to enrich text-only data (Parida et al., 2022). For our text-only baseline, we trained the sentence-level transformer model from scratch using all training data until convergence.

3.2 Multimodal translation

For the EN-HI multimodal translation system, we employ a context-aware model, an extension of the Transformer architecture designed to incorporate additional contextual information during translation. Unlike traditional neural machine translation (NMT) models that translate sentences independently, context-aware NMT relaxes this assumption by conditioning the translation not only on the current source sentence but also on auxiliary information from within or outside the document. Given that the HVG data set is limited to the caption translation of specific image regions, we hypothesize that providing the model with additional context, such as a comprehensive description of the entire image, could enhance the accuracy of the generated translations. To take advantage of visual features, we extracted image captions from HVG image dataset and used them as additional context for translation. Additionally, we used pre-trained BERT as additional encoder to encode and aggregate contextual features (Wu et al., 2022).

We used BLIP[3], an image caption model, to generate a description of the HVG image dataset. As HVG contains short descriptions of specified regions of images in English and Hindi, we generate captions of entire image to be fed as additional information to multimodal context-aware model. Since our context-aware model expects context during the training and evaluation stage, we generated captions for the entire HVG dataset, including the evaluation (EV) and Challenge (CH) test sets. The overview of our multimodal translation system is depicted in Figure 1.

Two step training strategy is followed, we first train a strong sentence-level transformer model using all the training data until convergence, then the context-aware model is initialized from best checkpoint and fine-tuned on context-aware data. We select the best model on the validation data. Contextual features are encoded using pre-trained *bert-base* model released under transformers package [4].The model incorporates two special tokens: [CLS], which is added at the beginning of a sentence, and [SEP], which is employed to separate different sequences. The context is concatenated with sentences as follows:

> x_{ctx} = *[CLS] surfer on a surfboard riding a wave in the ocean [SEP] man surfing in ocean [SEP]*

Several techniques exist for context integration (Castilho et al., 2020; Wu et al., 2022; Haq et al., 2022), we used $1-fixed-sequence$ on the source and target side as context. In this approach, a single previous sentence or external sequence is considered context for current sentence being translated. After that Bert encoded features are extracted as defined in equation 1. Although the context-aware multi-encoder models are exposed to additional contextual information, the translation is still performed at sentence level.

$$C = BERT(x_{ctx}) \quad (1)$$

[3] https://huggingface.co/Salesforce/blip-image-captioning-large

[4] https://github.com/huggingface/transformers

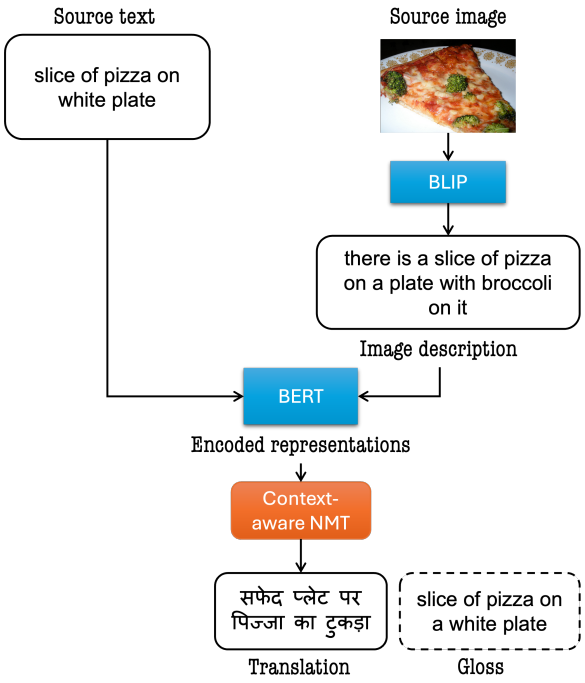

Figure 1: Overview of multimodal translation system.

Our translation models are based on transformer architecture with 6 encoder/decoder blocks, 512 embedding input, and 1024 FFN layer dimension size. Dropout rate is 0.3 for all tasks. We use the Adam optimizer and 5×10^{-4} learning rate schedule with 4000 warmup steps. Model training was conducted on two GPUs, with a batch size of 6000 tokens per GPU. Our Transformer implementation is based on the Fairseq (Ott et al., 2019) toolkit.

4 Results

Our results for EN-HI text-only and multimodal translation are presented in Table 2.

Modality	System	*BLEU* EV	*BLEU* CH
text-only	Transformer	40.20	29.20
text-only	Transformer$_{bt}$	**42.70**	**35.90**
multimodal	Context-aware$_{src_tgt}$	40.60	28.60
multimodal	Context-aware$_{src}$	40.60	**30.30**

Table 2: WMT_WAT2024 Automatic evaluation results for EN→HI on Evaluation (EV) and Challenge (CH) test sets. "Transformer$_{bt}$" denotes NMT model trained with back-translated data. For multimodal task, "src_tgt" represents context-aware model with visual contextual features used on both encoder and decoder side while src indicates context used only on the encoder side.

For text-only translation, the baseline system (Transformer) obtains BLEU scores of 40.20 on the evaluation set (EV) and 29.20 on the challenge test (CH). In contrast, the Transformer$_{bt}$ (Transformer with back-translated data) system demonstrates improved results, with BLEU scores of 42.70 for EV and 35.90 for CH. This improvement suggests that back-translation enhances translation quality by incorporating additional synthetic data, which is particularly advantageous for the challenge set (CH).

In multimodal translation, the context-aware$_{src_tgt}$ approach achieves BLEU scores of 40.60 for EV and 28.60 for CH. Compared with a text-only restricted baseline, the EV score slightly exceeds that of the Transformer (40.20), the CH score is lower, indicating that while the multimodal context benefits the evaluation set, it does not consistently improve performance on the challenge set. Conversely, the context-aware$_{src}$ (source only context) method achieves BLEU scores of 40.60 for EV and 30.30 for CH, showing a modest improvement for the challenge set compared to the src_tgt and text-only methods (except Transformer$_{bt}$).

5 Discussion

The baseline Transformer model achieves BLEU scores of 40.20 for the evaluation set (EV) and 29.20 for the challenge test (CH). The Transformer$_{bt}$ model shows marked improvement, with BLEU scores of 42.70 for EV and 35.90 for CH, highlighting back-translation's effectiveness in enhancing performance, particularly in challenging scenarios.

In multimodal translation, the context-aware$_{src_tgt}$ method, which utilises visual context on both the encoder and decoder sides, scores 40.60 for EV and 28.60 for CH. It slightly outperforms the baseline Transformer on EV but underperforms on CH, suggesting that while visual context can help in simpler cases, it may complicate results in more difficult scenarios.

The context-aware src only approach, using visual context only with the source text, achieves BLEU scores of 40.60 for EV and 30.30 for CH. It shows modest improvement over src_tgt for CH but does not surpass the Transformer + Back-translation in overall performance. This is obvious because multimodal translation systems are trained on constrained resources while Transfomer$_{bt}$ use 8k additional synthetic parallel sentences for training.

These findings underscore the value of back-translation in improving text-only translation, especially for more challenging tasks. For multimodal translation, while visual context can be beneficial, its effectiveness varies with context integration choice. The results suggest that different methods may be better suited to different types of translation challenge, indicating a need for further research to optimize the use of visual context.

6 Conclusion

Our results have showed that the Transformer with back-translated data consistently outperforms the text-only and multimodal systems in both evaluation tasks, demonstrating a significant benefit of back-translation, particularly for challenging scenarios. Our multimodal systems, despite not utilizing back-translated data, still outperformed the text-only baseline, highlighting the potential of visual context in improving translation accuracy. However, multimodal systems employing visual context on both the encoder and decoder sides do not exhibit a clear advantage over the text-only model or other multimodal approaches. Notably, the multimodal method shows diminished effectiveness for the challenge test (CH), suggesting that while additional visual context may enhance performance in certain cases, it can also introduce complexities that potentially undermine translation accuracy. These findings highlight the need for further investigation into optimizing the integration of visual context to improve translation outcomes across varying task difficulties.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The Authors also benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Anja Belz, Erkut Erdem, Katerina Pastra, and Krystian Mikolajczyk. 2017. Proceedings of the sixth workshop on vision and language. In *Proceedings of the Sixth Workshop on Vision and Language*.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017. Using Images to Improve Machine-Translating E-Commerce Product Listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 637–643, Valencia, Spain. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Sami Ul Haq, Sadaf Abdul Rauf, Arslan Shaukat, and Muhammad Hassan Arif. 2022. Context-aware neural machine translation using selected context. In *2022 19th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pages 349–352. IEEE.

Chiraag Lala, Pranava Madhyastha, JK Wang, and Lucia Specia. 2017. Unraveling the contribution of image captioning and neural machine translation for multimodal machine translation. In *The Prague Bulletin of Mathematical Linguistics*, volume 108-1, pages 197–208. De Gruyter Open.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Satya Prakash Biswal, Ketan Kotwal, Arghyadeep Sen, Satya Ranjan Dash, and Petr Motlicek. 2021a. 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.

Shantipriya Parida, Subhadarshi Panda, Stig-Arne Grönroos, Mark Granroth-Wilding, and Mika Koistinen. 2022. Silo NLP's participation at WAT2022. In *Proceedings of the 9th Workshop on Asian Translation*, pages 99–105, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Ketan Kotwal, Amulya Ratna Dash, Satya Ranjan Dash, Yashvardhan Sharma, Petr Motlicek, and Ondřej Bojar. 2021b. Nlphut's participation at wat2021. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 146–154.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147.

Sami Ul Haq, Sadaf Abdul Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document level NMT of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 442–446, Online. Association for Computational Linguistics.

Xueqing Wu, Yingce Xia, Jinhua Zhu, Lijun Wu, Shufang Xie, and Tao Qin. 2022. A study of bert for context-aware neural machine translation. *Machine Learning*, pages 1–19.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. *arXiv preprint arXiv:1808.08266*.