Multimodal Machine Translation for Low-Resource Indic Languages: A Chain-of-Thought Approach Using Large Language Models

Pawan Kumar Rajpoot* pawan.rajpoot2411@gmail.com

Nagraj N Bhat* nagbhat25@gmail.com

Ashish Shrivastava* ashish3586@gmail.com

Abstract

This paper presents the approach and results of team v036 in the English-to-Low-Resource Multi-Modal Translation Task at the Ninth Conference on Machine Translation (WMT24). Our team tackled the challenge of translating English source text to low-resource Indic languages, specifically Hindi, Malayalam, and Bengali, while leveraging visual context provided alongside the text data. We used InternVL2 for extracting the image context along with Knowledge Distillation from bigger LLMs to train Small Language Model on the tranlsation task. During current shared task phase, we submitted best models (for this task), and overall we got rank 3 on Hindi, Bengali, and Malyalam datasets. We also open source our models on huggingface.[1]

1 Introduction

With the recent advances in text generative AI Achiam et al. (2023); Dubey et al. (2024); Yang et al. (2024) and Diffusion based Dhariwal and Nichol (2021); Nichol and Dhariwal (2021); Saharia et al. (2022); Ramesh et al. (2022) models, multimodal approaches have gained significant traction. The concept of a model to understand both text and visual contexts provides a unique advantage for these models to understand the real world. On the other end, Machine Translation has been one of the most important task in NLP world. Since its origin, the MT task has undergone large shifts from rule based Nirenburg (1989); Chen et al. (2007) to complex Neural network based approaches and recently Transformer Vaswani (2017); Yin and Read (2020); Xu et al. (2024) based approaches. In the recent days with the advancements in the field of NLP, Multimodal Machine Translation (MMT) has evolved as an important research field, wherein the Model utilizes both vision and text information to achieve the translation task. This would better equip the model with additional context information and thus reducing the issues due to polysemy or missing text context. MMT finds its application in various fields like Media, Retail, Automobile etc. In this work we explore the problem of English to Lowres Multimodal Translation for Hindi, Bengali and Malyalam languages. The task requires translating a short English caption of the rectangular region to one of these languages, given the image context. There are multiple approaches possible which can be largely classified into :

- Text-only translation (Source image not used)
- Image captioning (English source text not used)
- Multi-modal translation (uses both the image and the text)

We strongly feel that Multi-modal translation approach would best solve the problem due to more context information. In this paper, we propose a novel unconstrained approach to solve the Lowres MMT task for Hindi, Bengali and Malyalam languages. Our solution tries to merge the best of both text and language contexts. In particular, our key contributions are:

- Fusing Multimodal image context with improved language understanding : We provide a concise yet effective approach to combine context information from vision to text description
- Advanced Chain of Thought reasoning for language translation: Our approach to step by step reasoning ustilizing the COT, gives a whole new perspective to enhance the ability of the model to comprehend better.

[1]https://huggingface.co/team-v036

*Authors contributed equally to this work

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 833–838
November 15–16, 2024.

- Custom finetuning : Our approach of custom finetuning on target languages on training samples, equips the model to better perform on the MMT task.

2 Data

Visual Genome introduced by Krishna et al. (2017) is a rich dataset to enable the modelling of complex cognitive interactions and relations between objects in an image. Based on this dataset, Parida et al. (2019) introduce the Hindi Visual Genome dataset, which is a multi-modal dataset consisting of text and images suitable for English-Hindi multimodal machine translation task. They select short English segments (captions) from Visual Genome along with the associated images and automatically translate them to Hindi with a careful manual post-editing(Parida et al. (2019)) The dataset takes into account ambiguous English words based on the embedding. similarity and manual selection of certain cases where image helps to resolve the ambiguity(Parida et al. (2019)). Hence this is a perfect dataset suited for the task. Similarly Sen et al. (2022) propose the Bengali Visual Genome Dataset which is manually labelled on HVG samples and Parida et al. (2019) curated the malyalam Visual Genome Dataset.

All three (Hindi, Bengali and Malyalam) dataset consists of 29k training samples, 1k dev set, 1.6k evaluation set and 1.4k challenge set.

The evaluation of the models were performed with BLEU metrics (Papineni et al. (2002)) on challenge and evaluation set independently. Along with these a manual labeller evaluation is also performed, subject to availability.

3 Related Work

MMT has gained increasing attention in recent years as a way to leverage visual information to improve translation quality. Several shared tasks and datasets have been introduced to advance research in this area, with a particular focus on low-resource languages. The Workshop on Asian Translation (WAT) has played a key role in promoting MMT research for Asian languages. Parida et al. (2019) introduced the first Hindi Visual Genome task at WAT 2019, using the Hindi Visual Genome 1.0 dataset (Parida et al. (2019)). This dataset contains English image captions paired with Hindi translations and associated image regions. The task evaluates systems on their ability to translate from English to Hindi while incorporating visual context. Subsequent iterations of WAT expanded the Hindi Visual Genome used an updated Hindi Visual Genome 1.1 dataset and introduced new evaluation tracks, including Hindi image captioning. The latest WAT 2021 (Nakazawa et al. (2021)) further refined the Hindi task and introduced a new English-Malayalam MMT task using the Malayalam Visual Genome dataset (Parida and Bojar, 2021). This represented the first multimodal translation dataset for Malayalam. For the Hindi task, recent approaches have focused on leveraging object tags extracted from images (Gupta et al. (2021)) and region-specific captioning (Parida et al. (2021)) to enhance translation quality. The introduction of the Malayalam task provides an opportunity to evaluate MMT techniques on a new low-resource language. While Hindi and Malayalam have been addressed in shared tasks, Bengali has seen less attention for MMT despite being widely spoken. The creation of a Bengali Visual Genome dataset, following the model of Hindi and Malayalam, would fill an important gap and enable MMT research for another major South Asian language. Overall, the development of these language-specific visual genome datasets has been crucial for advancing MMT for low-resource Indian languages. They provide much-needed benchmarks and drive innovation in incorporating visual context for translation. Expanding to additional languages like Bengali represents an important direction for broadening the scope of MMT research in the Indian context.

4 Approach

Our overall approach follow a three step process as seen in Figure 1.

4.1 Stage 1: Fusing Multimodal image context with improved language understanding

In this stage, we first extract context from cropped visual data using a powerful open-source Multimodal Large Language Model (MLLM)- InternVl2-8B (Chen et al. (2023, 2024)). This model demonstrates powerful capabilities in handling complex multimodal data and achieves state of art numbers on many open VQA tasks. Figure 2 shows a sample image and its description. We feed the output of segement description as an input into a Rapid Automatic Keyword Extraction (RAKE) algorithm (Rose et al. (2012)) which is an efficient keyword extraction algorithm. The top extracted key phrases

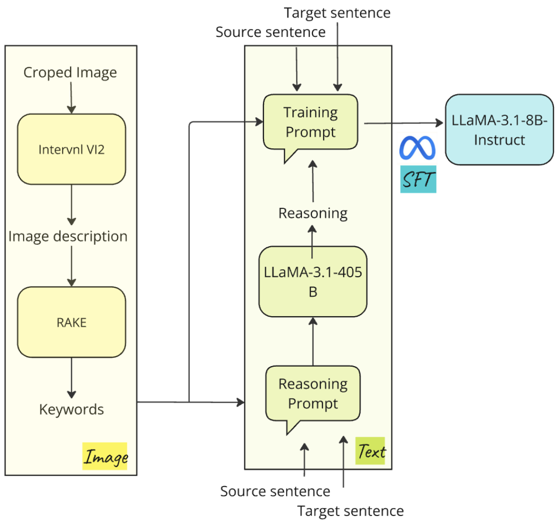

Figure 1: Overall Approach.

are selected and used as hastags to provide context to the source English text. This way we condense the full description into short and concise information for the next stage. This ensures that the further step do not completely rely on the image context also but rather use the original text but still use the relevant information from the image descriptors. This step is common for both train and evaluation process.

4.2 Stage 2: Advanced Chain of Thought reasoning for language translation

Chain-of-thought (CoT) (Wei et al. (2022)) prompting enables complex reasoning capabilities through intermediate reasoning steps. It is shown that the models ability is substantially improved by making them produce step by step reasoning. We employ this ability of Large Language model to solve the task in a more understandable and reasonable approach by decomposing the problem into multiple sub-problems. We use State of the Art LLaMa 405B (Dubey et al. (2024)) model to generate the CoTs for the training data. The model is provided with the English caption text that needs to be translated along, the hashtags generated in previous step, and the target language caption. The model is then asked to generate the step by step reasoning for converting the source caption text to target language text along with the condensed context information provided. Following prompt (Table 1) template is used to get the CoT from the bigger model.

Table 1: Prompt for CoT reasoning generation from bigger model (LLAMA 3.1 405B)

TASK:
ASSUMING YOU ARE A ENGLISH - HINDI translation expert, For given context of an image related to a original sentence, English sentence and translation of the sentence in hindi. Give reason on why this translation is the correct translation....ASsume that you secretly know the answer......DO NOT TRY TO FIX Translation...reason for whatever is given only.....reason SHOULD be proper Chain of Thought format in properly divided steps for the answer......give maximum 5 steps which are most important ones ....
Context: {RAKE HASTAGS}
English Sentence: {SOURCE TEXT}
Hindi Sentence: {TARGET TEXT}

Table 2: Prompt for SFT LLAMA 3.1 8B

TASK:
ASSUME YOU ARE AN ENGLISH-HINDI translation expert. Given an image description in English, image context and reasoning/CoT in English, translate the image description in Hindi. Use the image context to solve ambiguity if required. Note: DO NOT USE the image context in translations, just use them for disambiguation.
IMAGE DESCRIPTION:
{SOURCE TEXT}
IMAGE CONTEXT:
{RAKE HASHTAGS}
REASONING:
{GENERATED CoT}
RESPONSE:
{TARGET TEXT}

4.3 Stage 3: Custom fine-tuning

In stage 3 we train a smaller model to perform the task of translation, we finetune a LLaMA 3.1-8B-Instruct model on training samples using data from previous stages. The model is trained by providing the English caption along with hashtag contexts, CoT reasoning and the final answer. A sample prompt is shown in Table 2. We use LORA finetuning with rank=64 and alpha=128. We use following template for the training data so that the CoT step is more aligned to as what humans think, that is first source, then CoT and finally the target text.

During **inference**, we provide the finetuned model with source English caption and context and ask it to come up with the Reasoning and the answer. We then use a post processor script to filter out the final answer from the model output.

These 3 fundamental steps are performed for all 3 languages and we curate one PEFT model for each language.

Figure 2: Sample Data with reference the image segment, its corresponding source and target text along with the key phrase extraction from Internvl2 descriptions

5 Experimental setup

In our experimental setup, we fine-tuned the LLaMA 3.1-8B-Instruct model for the translation task using Quantized Low-Rank Adaptation (QLoRA) (Hu et al. (2021); Dettmers et al. (2024)). The LoRA configuration was carefully selected to balance performance and computational efficiency. We set the rank (r) to 64 and the alpha parameter to 128, with a lora_alpha value of 0.05. Notably, we applied LoRA to all target modules in the model architecture, ensuring a comprehensive adaptation across the entire network.

For the optimization process, we employed a learning rate (lr) of 0.003, coupled with a cosine learning rate scheduler. This scheduling strategy allowed for dynamic adjustments to lr, potentially aiding in convergence and generalization. The model was trained for two epochs, striking a balance between sufficient learning and computational constraints.

The chosen LoRA hyperparameters strike a balance between model capacity and computational efficiency, with the rank of 64 providing sufficient expressiveness for the adaptation.

By leveraging Quantized LoRA and carefully selected training parameters, we aim to achieve high-quality translation performance while minimizing computational resources and training time. We used A100 40GB VRAM and 84GB RAM single node machine to fine tune our models.

6 Results

The results show that our Multimodal approach of using multistage image description extraction clubbed with CoT is an effective approach to solve this task leveraging the knowledge of Large language models. Table 3 shows the results for all 3 Indic languages on Evaluation and Challenge set. Our numbers are very close to SOTA numbers. The SOTA (baseline) approach is based on a fine-tuning of NLLB model on captions of Object tags of original along with synthetic images using DETR model. However, we do not use any additional image set in our process

Table 3: Results (BLEU Scores) on languages comparing to SOTA .

Language	Evaluation Set	Challenge Set
Hindi	0.446/**0.45**	0.432/**0.534**
Bengali	0.441/**0.506**	0.339/**0.487**
Malyalam	0.427/**0.519**	0.333/**0.422**

The data analysis of the final output revealed a set of cases where the output is technically correct, yet contains variations in tokens compared to the gold set. This suggests that human evaluation could potentially yield higher accuracy, and relying solely on the BLEU score for this task may not fully capture the quality of the output.

Limitations

Given that our approach heavily depends on multiple stages involving large language models, it may not be ideally suited for environments with limited resources. The complexity and computational demands of such models could pose challenges in settings where processing power, memory, or bandwidth are constrained. Additionally, this approach leverages the inherent knowledge embedded within the LLMs being used. The effectiveness of the method is closely tied to the pre-existing information and understanding that these models have acquired during training, which may be influenced by the data used for its training.

Ethics Statement

Our work proposes an innovative approach to addressing the challenge of translating low-resource English to Indic languages - Hindi, Bengali, and Malayalam. In conducting our research, we have carefully considered the ethical implications of data usage. As a result, we have chosen to exclusively rely on the data provided by the Task administrators for our experiments, refraining from incorporating any additional external data sources. This ensures that our approach remains transparent and aligns with the ethical standards expected in this field. However, while using this approach for real world application, data privacy and consent should be given careful considerations.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-engine machine translation with an open-source smt decoder. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 193–196.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at semeval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. *arXiv preprint arXiv:2106.00240*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2021. Overview of the 8th workshop on asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

Sergei Nirenburg. 1989. Knowledge-based machine translation. *Machine Translation*, 4(1):5–24.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Stuart J Rose, Vernon L Crow, Nicholas O Cramer, et al. 2012. Rapid automatic keyword extraction for information retrieval and analysis. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (United States).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu

Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, pages 63–70. Springer.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Kayo Yin and Jesse Read. 2020. Attention is all you sign: sign language translation with transformers. In *Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts*, volume 4.

A Appendix

A sample example of COT generation is shown below:

```
Example 1

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

###TASK:
ASSUME YOU ARE AN ENGLISH-BENGALI translation expert.
Given an image description in English, image context and resoning in English, translate the image description in Bengali. Use the image context to improve the translat
ion, if required.
Note: DO NOT TRY to correct the image description or image context in English.

###IMAGE DESCRIPTION: couples walking in the rain

###IMAGE CONTEXT: #typical rainy day #also dressed casually #white tank top

###REASONING:
Step 1: "Couples" refers to two people in a romantic relationship.
Step 2: In Bengali, "দম্পতি" (dampati) specifically means a married couple.
Step 3: The context shows a casual, relaxed atmosphere, implying a romantic couple.
Step 4: "Walking" is accurately translated to "হাঁটছেন" (hāṭchhen), which is the correct verb form.
Step 5: "In the rain" is correctly translated to "বৃষ্টিতে" (bṛṣṭite), which sets the scene for the action.

###RESPONSE: দম্পতিরা বৃষ্টিতে হাঁটছেন
```

Figure 3: Sample COT training data

A few data analysis samples where we note that the translation is mostly valid but the gold may have different set of words.

source : gold religious cross on top of golden ball
gold: स्वर्ण गेंद के शीर्ष पर स्वर्ण धार्मिक क्रॉस
ours: सोने की गेंद के ऊपर सोने का धातु

source : front springs of motocross motorcycle
gold: मोटोक्रॉस मोटरसाइकिल के सामने स्प्रिंग्स
ours: मोटोक्रॉस मोटरसाइकिल के आगे के स्प्रिंग

source: A display of different phone models
gold: विभिन्न फोन मॉडलों का प्रदर्शन
ours: विभिन्न मोबाइल मॉडलों का एक प्रदर्शन

source: rapid in fast water
gold: तेजी से पानी में
ours: तेज़ पानी में तेज़

source: tail fine view of a red plane
gold: एक लाल विमान की पूंछ का दृश्य
ours: लाल विमान का टेल फाइन व्यू

source : white springs boxes and gears below train engine
gold: सफेद स्प्रिंग्स बक्से और ट्रेन इंजन के नीचे गियर
ours: ट्रेन इंजन के नीचे सफ़ेद स्प्रिंग्स के डिब्बे और गियर हैं।