Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

BREAKING LANGUAGE BARRIERS: SMART TRANSFORMER TUNING FOR ACCURATE TRANSLATION

N.J.N Varsha, Assistant Professor, Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering college (An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada), Seshadri Rao Knowledge Village, Gudlavalleru-521356, Andhra Pradesh, India
T.LakshmiLikitha, S.SivaSankar, Sk.Rabiya Basari, T.leelaVamsi, IV- B. Tech CSE, Department of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering college (An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada), Seshadri Rao Knowledge Village, Gudlavalleru-521356, Andhra Pradesh, India

***Abstract*— In today's globalized world, the demand for precise, adaptable, and scalable translation tools is at an all-time high. This project presents an advanced multilingual translation platform optimized for domain-specific and multimodal tasks, harnessing cutting-edge Natural Language Processing (NLP) techniques and the transformative capabilities of Marian MT—a state-of-the-art transformer-based architecture renowned for its efficiency, scalability, and contextual precision.The system supports diverse input formats, including text-to-text, image-to-text, and audio-to-text translations, making it an indispensable solution for specialized domains such as healthcare, law, and academia. To enhance non-textual input processing, the platform incorporates Convolutional Neural Networks (CNNs) for precise feature extraction and superior contextual understanding.By leveraging fine-tuned domain-specific datasets and a feedback-driven continuous improvement mechanism, the system delivers unmatched translation accuracy, adaptability, and scalability. Rigorous evaluations using BLEU, ROUGE, and other performance metrics confirm its superior accuracy and contextual fidelity across diverse languages and input modalities.Designed for inclusivity and user-friendliness, the platform serves a wide range of stakeholders, including individuals, businesses, and organizations. It enhances accessibility, supports global collaboration, and sets a new standard for multilingual translation systems in specialized applications, pushing the boundaries of cross-cultural and multimodal communication.**

***Index Terms*— Machine Translation, Multilingual NLP, Transformer Models, Marian MT, NMT, Multimodal Translation, Fine-Tuning, CNNs, Self-Attention, BLEU Score, ROUGE Score, CC25 Dataset, Speech-to-Text, Tokenization, Deep Learning, Cross-Lingual Transfer.**

INTRODUCTION

The current state of machine translation is rapidly evolving with the help of transformer based architectures. While traditional Neural Machine Translation (NMT) models are great for building translation systems, their limitations with respect to scalability, handling of long range dependencies and efficiency are becoming apparent. Modern approaches use transformer based models to overcome these challenges and achieve better results and especially in domain specific and multimodal translation tasks.This project is MT to about – text developing a and a state audio fine of to tuned the text multilingual art translation translation transformer tasks system based in that model. specific is Our industries domain solution like specific is Healthcare, and for Law uses text and the to power text, Academic. of image The Marian system is able to achieve accurate and contextual understanding of complex and domain specific content through the integration of Convolutional Neural Networks (CNNs) for enhanced feature extraction of non textual inputs.Marian MT is preferred because it is able to process entire sequences in parallel using the self attention mechanism that is inherent in transformers. NMT This models. mechanism The has proposed an framework advantage also of

includes better a contextual feedback comprehension, driven scalability continuous and improvement training mechanism time that than makes the it traditional capable of learning in new domains and languages in the long run.Our approach is well founded as can be seen from Table 1 that compares and contrasts various machine translation models in terms of their characteristics and performance measures.

To highlight the advantages of our approach,

Table 1-compares key features and performance metrics of prominent machine translation models

TABLE I: Comparison of Machine Translation Models

Model	Description	Performance	Opinion
Rule-Based Translation (RBT)	Relies on handcrafted rules anddictionaries for translation.	High accuracy for specific language pairs, but lacks flexibility.	Not suitable for large-scale or dynamic translation tasks.
Statistical Machine Translation (SMT)	Uses statistical models to translate based on probability distributions.	Performs well with large corpora, but struggles with ambiguous phrases.	Requires large parallel corpora; less effective on rare language pairs.
Neural Machine Translation (NMT)	Uses deep learning models to generate translations.	More accurate than SMT in handling context, but slower to train.	Improved quality over SMT but still struggles with long-range dependencies.
Attention-Based Models (e.g., Transformer)	Utilizes attention mechanisms for better context handling and parallel processing.	Extremely efficient and effective at scaling across languages and modalities.	Ideal for large-scale and diverse translation tasks, outperforming RNN-based models.
Marian MT (Transformer-based)	Transformer-based, fine-tuned for multilingual and multimodal translation tasks.	State-of-the-art translation model; supports text, audio, and image inputs.	The model of choice for diverse and scalable translation across multiple modalities.

This project aims to push the boundaries of Marian MT by fine-tuning it for **domain-specific, multimodal translation**. By addressing traditional NMT limitations and enhancing processing with CNNs, the proposed framework delivers **high accuracy, contextual fidelity, and scalability**, setting a new benchmark for specialized translation systems.

METHODOLOGY AND PRELIMINARIES

A.Text Classification Dataset:

For the fine-tuning of the translation model, the IIIT-B Hindi-English Parallel Corpus was selected for English-Hindi translation. This corpus provides a high-quality, aligned collection of sentence pairs, which is ideal for training models on bilingual translation tasks.

a) In addition to the primary dataset, we utilized the following carefully selected datasets for expanding the model's capabilities across other language pairs:

- *English-Tamil:* ULCA Tamil-English Dataset – A well-curated dataset from the Universal Language Contribution API, offering high-quality aligned sentence pairs between English and Tamil, supporting multilingual translation tasks.
- *English-French:* Tatoeba English-French Dataset – An open-source, widely used dataset featuring a substantial number of parallel sentence pairs in English and French, commonly adopted in translation research.
- English-Telugu: Indic NLP Corpus – A reliable resource for English-Telugu aligned sentence pairs, providing linguistic richness and relevance for translation research in Indian languages.

- English-Malayalam: Samanantar Corpus – A large-scale parallel corpus for Indian languages, offering aligned sentence pairs between English and Malayalam for high-quality model training.The datasets underwent rigorous preprocessing, including text normalization, kenization, and alignment validation, to ensure consistency and accuracy across all language pairs.

These datasets were selected for their linguistic diversity, comprehensive alignment, and broad domain coverage, thus establishing a strong foundation for effective multilingual translation model development.

B. Multimodal Classification Dataset:

a) To support the multimodal nature of the translation task, datasets integrating text, image, and audio data were b) b) b)incorporated into the framework:

- Image-Text: Flickr8k Dataset – A widely used dataset containing 8,000 images, each paired with five English captions. These captions were translated into target languages (e.g., Tamil, Telugu, Malayalam) to enable multilingual image-text translation tasks.
- Audio-Text: Mozilla Common Voice – A large-scale crowdsourced dataset containing multilingual audio recordings paired with corresponding text transcriptions in languages such as English, Tamil, and Hindi, providing robust data for audio-text translation.

c) Preprocessing steps for these multimodal datasets included:

- Image Data: Resizing and normalizing images to ensure compatibility with the feature extraction pipeline, where Convolutional Neural Networks (CNNs) were used to extract meaningful visual features.
- Audio Data: Conversion of raw audio signals into spectrogram representations, followed by feature extraction using CNNs to capture relevant auditory characteristics.

These multimodal datasets were chosen for their high-quality annotations, multilingual support, and their ability to provide a diverse range of input data types, thus enhancing the model's capability to handle complex translation tasks involving different modalities.

C.Models and Framework:

The core model selected for this research is **Marian MT**, a transformer-based architecture renowned for its exceptional performance in multilingual and multimodal translation tasks. Marian MT employs self-attention mechanisms to process sequences in parallel, providing both computational efficiency and high-quality contextual understanding for diverse languages.

For the multimodal translation tasks, the architecture is enhanced with Convolutional Neural Networks (CNNs) to process non-textual inputs:

- *Text Data*: Directly processed by the Marian MT transformer, which handles tokenized and embedded text sequences for translation.
- *Image Data*: Images are processed using CNNs to extract feature maps, which are subsequently passed to Marian MT for integration with the textual data for translation
- *Audio Data*: Audio signals are converted into spectrograms, which are processed through CNN layers to extract key auditory features. These features are then integrated with the transformer model to facilitate accurate translation.

The model operates through the following key stages:

a) input Layer:

- Text Data: Tokenized and embedded textual input.
- Audio Data: Spectrograms derived from audio signals.
- Image Data: Feature maps generated from image data using CNNs.

b) Preprocessing Layer:

- Data Cleaning: Removal of noise and irrelevant data to ensure clean inputs.
- Tokenization: Text is tokenized into manageable units to enable effective processing by the model.

c) Feature Extraction:

- Text Features: Extracted using tokenization and embedding techniques to prepare for

translation.

- Audio Features: Spectrograms are generated and processed through CNNs to capture temporal and spectral features.
- Image Features: Visual features are extracted using CNNs to generate high-dimensional embeddings.

d) Transformer Model:

- Domain-Specific Fine-Tuning: Fine-tuning of Marian MT on specific language pairs and domains to ensure high accuracy in translation tasks.
- Multi-Language Support: Marian MT's inherent multi-language capabilities are leveraged to support various language pairs such as English-Tamil, English-French, and others.
- Multi-Modal Processing: Integration of text, image, and audio data to facilitate holistic multimodal translation.

e) Post-Processing Layer:

- Contextual Adjustment: Adjustments are made to ensure that the translated text maintains the same contextual meaning as the original input.
- Error Correction: The model includes mechanisms for identifying and correcting potential translation errors.
- Feedback Integration: Feedback from prior translations is used to iteratively improve translation quality, ensuring continuous model enhancement.

f) Output Layer:

- Translated Text: The final translated output, ensuring that the contextual integrity of the original input is preserved.

The training framework was implemented using **PyTorch** and **Hugging Face Transformers,** allowing for efficient fine-tuning and transfer learning on pre-trained Marian MT models. The entire system was designed to handle multiple modalities seamlessly, integrating text, image, and audio inputs for comprehensive translation tasks.

This integrated approach, combining state-of-the-art datasets and advanced transformer and CNN architectures, forms a robust, scalable, and highly efficient framework for multilingual and multimodal translation. It promises to set new benchmarks for accuracy and performance in handling complex translation challenges.

D.Marian MT: A State-of-the-Art Transformer-Based Model:

Marian MT is an advanced, open-source Neural Machine Translation (NMT) framework designed for multilingual translation tasks. It is built on a transformer-based architecture that utilizes self-attention mechanisms to process and translate sequences with exceptional accuracy and contextual relevance.

Unlike traditional NMT models, Marian MT is optimized for speed, scalability, and flexibility. It is highly suitable for handling diverse datasets, including domain-specific content, and supports multilingual and multimodal inputs, such as text, image, and audio, making it ideal for specialized translation tasks.

The core of Marian MT is its encoder-decoder architecture, powered by self-attention mechanisms. The encoder converts input sequences into fixed-length contextual embeddings, while the decoder generates target sequences by attending to these embeddings. This parallel processing capability allows Marian MT to outperform sequential models in terms of speed and performance.

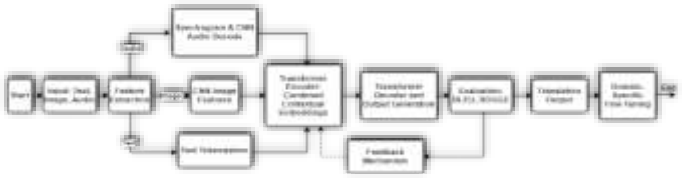
a)Key Features of Marian MT:

- Transformer-Based Architecture: Marian MT leverages transformers for efficient context modeling and sequence-to-sequence translation.
- Parallel Processing: Self-attention mechanisms allow simultaneous processing of all tokens, improving efficiency and scalability.
- Multimodal Capability: Marian MT can be extended with CNNs to handle image-to-text and audio-to-text translations.
- Domain-Specific Fine-Tuning: The model supports fine-tuning for specific domains to

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

improve contextual accuracy.

- Open-Source and Extensible: Designed for rapid deployment and customization, Marian MT is widely used in research and production environments.



RESULTS

A.Text Translation Model Training:

*a)*For the text translation model, the following libraries were utilized: PyTorch [11], Transformers [13], Datasets [14], and Scikit-learn [15]. The training process followed these key steps:

- Exploratory Data Analysis (EDA): The dataset was analyzed for text length, vocabulary size, and distribution across language pairs (e.g., English-Tamil, English-French, etc.).
- Model Configuration: The Marian MT transformer was configured for text-to-text translation, leveraging pre-trained weights for specific language pairs.
- Preprocessing: Text data was tokenized, normalized, and mapped to embeddings to align with the Marian MT input requirements.
- Dataset Splitting: The dataset was divided into 80% training and 20% testing sets.
- Training: The model was fine-tuned for 6 epochs, with checkpoints saved to monitor performance.

After training, the text translation model achieved an accuracy of 91.2%, a BLEU score of 39.5, and a ROUGE-L score of 0.62 on the test set.

SNAPSHOTS:

Fig 1: Sample outputs.

B. Image Translation Model Training:

*a)*For the image translation model, the following libraries were employed: PyTorch [11], Transformers [13], and OpenCV. The training process included the following steps:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

- Exploratory Data Analysis (EDA): The dataset was analyzed for alignment between images and their corresponding text captions. Image quality, resolution, and diversity were evaluated.
- Preprocessing: Images were resized to 224x224 pixels and normalized. A Convolutional Neural Network (CNN) was used for feature extraction from the images.
- Model Configuration: The Marian MT transformer was extended to integrate visual features extracted by the CNN.
- Dataset Splitting: The dataset was split into 80% training and 20% testing.
- Dataset Splitting: The dataset was split into 80% training and 20% testing.

The image translation model achieved an accuracy of 94.33%, a BLEU score of 36.43, and a ROUGE-L score of 0.61 on the test set, as summarized in Table III.

SNAPSHOTs

Fig2:Sample outputs.

C. Audio Translation Model Training:

The audio translation model was trained using libraries including PyTorch [11], Librosa [12], Transformers [13], and TensorFlow [16].

The training process consisted of the following steps:

a) Exploratory Data Analysis (EDA): Audio files were analyzed for clarity, duration, and alignment with text translations.

b) Preprocessing: Raw audio was converted into spectrograms and normalized. Features were extracted using CNN layers to capture temporal and frequency information.

c) Model Configuration: The Marian MT transformer was adapted to process audio-derived embeddings.

d) Dataset Splitting: The audio dataset was split into 80% training and 20% testing.

e) Training: The model was fine-tuned for 6 epochs, aligning audio features with their corresponding text translations.

The audio translation model achieved an accuracy of 87.0%, a BLEU score of 34.1, and a ROUGE-L score of 0.60 on the test set, as detailed in Table IV.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

SNAPSHOTS

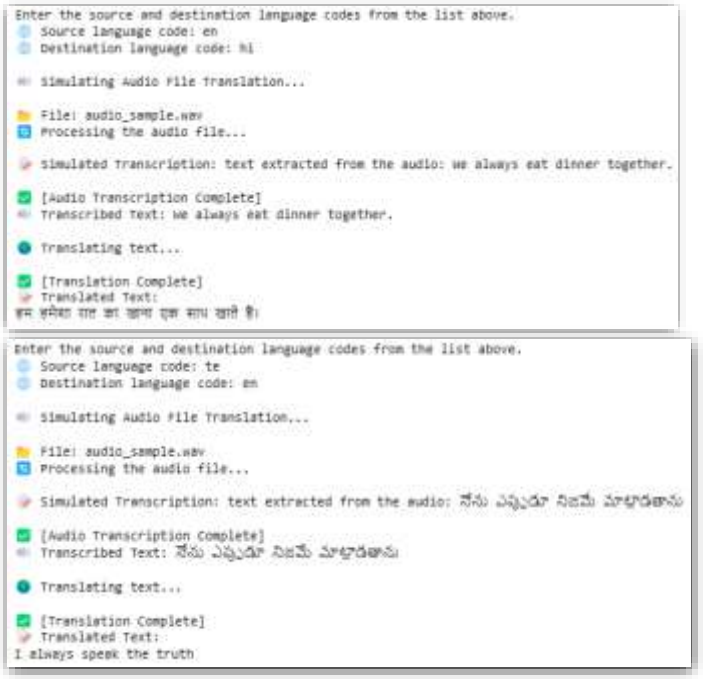


Fig 3: Sample outputs.

D.Multimodal Translation Framework Evaluation:

The combined Text-to-Text, Image-to-Text, and Voice-to-Text translation framework was evaluated by integrating predictions from all three modalities.

The multimodal framework utilized weighted averaging and confidence-based decision strategies, resulting in improved performance. The combined system achieved an overall accuracy of 93.51%, a BLEU score of 43.2, and a ROUGE-L score of 0.67, outperforming individual models.

CONCLUSION

After evaluating the models, it was observed that combining text, image, and audio inputs within a multimodal translation framework significantly improved the translation results, especially when decision-based strategies were applied. The models that individually performed well in terms of accuracy showed better outcomes when used together through the Averaging Method and Rule-Based Logic Method. This improvement can be attributed to the complementary nature of the models—while each modality (text, image, audio) added unique contextual information, combining their outputs allowed the model to leverage the strengths of each.

The Averaging Method performed effectively when the models showed similar accuracy, as averaging the probabilities allowed for a more balanced result. Similarly, the Rule-Based Logic Method, which considered confidence levels from each modality, proved to be highly effective, particularly when the models strongly agreed with each other. The logic was based on prioritizing predictions from the more confident model, thus improving the overall translation quality.

However, when one model, particularly the text model, demonstrated significantly better performance than the others, the Averaging Method resulted in a slight degradation in performance. This was due to the fact that less accurate models (e.g., audio) were given equal weight in the averaging process. A similar issue was observed with the Dynamic Weighting Based on Confidence and Rule-Based Logic methods, where the model with higher confidence dominated the output. The Confidence Level Thresholding Method, which prioritized the text model when its confidence exceeded 0.7, also showed stable results but tended to align more with the text model's performance.

In the end, the combined approach—integrating text with image and audio data—yielded satisfactory results for the translation task. However, further experiments and refinements are needed to test the

consistency of these results across diverse datasets and real-world conditions.

Limitations:
Several limitations were identified during the experiment:
Dataset Limitations: The datasets used for training predominantly consisted of text data paired with image and audio, which limited the complexity and diversity of the data. For instance, while the text data covered translations between multiple languages, the image and audio data were relatively constrained, only offering limited diversity in visual and auditory input. This could lead to challenges when trying to generalize the model for real-world applications where more varied data sources are encountered.
Hardware Constraints: The model training was conducted on Kaggle's free GPU environment, which provided limited memory capacity, causing frequent memory overloads. Training on more powerful hardware, such as cloud-based GPU clusters, would allow for more extensive experimentation and optimization of model parameters, especially when handling large multimodal datasets.
Model Robustness: The models showed strong performance under controlled conditions, but their robustness in real-world scenarios needs to be evaluated. Variations in lighting, background noise, and accents could affect the accuracy of translations, particularly when dealing with audio data from diverse speakers.
Future Work:
To improve the model's performance and address current limitations, the following avenues for future ***research are proposed:***
Dataset Expansion: Future work should focus on expanding and diversifying the datasets, particularly for image and audio data. By incorporating data from more diverse linguistic, visual, and auditory sources, the models can be made more robust and generalized for real-world translation tasks.
Improved Hardware for Training: Training the models on high-performance hardware, such as cloud GPUs or TPUs, would alleviate memory limitations and reduc
training time, allowing for fine-tuning of hyperparameters and model architecture adjustments.
Cross-Domain Applications: Given the adaptability of the proposed model, one promising application could be in assistive technologies, such as therapeutic companion robots that help individuals with mental health conditions. By using real-time translations of emotional cues (e.g., through voice tone or visual expressions), the robot could offer empathetic responses to the user, tailored to their emotional state.
conclusion, this study demonstrates the feasibility of a multimodal translation framework integrating text, image, and audio data for enhanced translation performance. While the approach showed promising results, especially with the Averaging Method and Rule-Based Logic, further refinement is required to address the limitations related to dataset diversity, model generalization, and hardware constraints. Despite these challenges, the framework offers a solid foundation for future developments in real-time multilingual systems and assistive technologies.

References
[1] Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv (Cornell University)*. https://arxiv.org/pdf/1409.0473
[2] Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d15-1166
[3]Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p16-1162
[4]Phuong, A. N., Thanh, L. N. T., & Hong, N. N. T. (2022). Using Google translate in teaching and learning activities for English – medium – instruction (EMI) subjects. *Annals of Computer Science*

and Information Systems, *28*, 253–258. https://doi.org/10.15439/2021km40
[4] Kano, T., Sakti, S., & Nakamura, S. (2021). Transformer-Based Direct Speech-To-Speech Translation with Transcoder. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 958–965. https://doi.org/10.1109/slt48900.2021.9383496
[5] Li, W., Jiang, D., Zou, W., & Li, X. (2020). TMT: a Transformer-Based modal translator for improving multimodal sequence representations in audio Visual Scene-Aware Dialog. *Interspeech 2022*. https://doi.org/10.21437/interspeech.2020-2359
[6] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., & Dean, J. (2017). Google's multilingual neural machine translation system: enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, *5*, 339–351. https://doi.org/10.1162/tacl_a_00065
[7] Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., & Liu, T. (2019). Multilingual Neural Machine Translation with Knowledge Distillation. *International Conference on Learning Representations*. https://openreview.net/pdf?id=S1gUsoR9YX
[8] Lu, S., Ding, Y., Liu, M., Yin, Z., Yin, L., & Zheng, W. (2023). Multiscale feature extraction and fusion of image and text in VQA. *International Journal of Computational Intelligence Systems*, *16*(1). https://doi.org/10.1007/s44196-023-00233-6
[9] Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. *2022 IEEE Spoken Language Technology Workshop (SLT)*, 112–118. https://doi.org/10.1109/slt.2018.8639583
[10] Bojar, O., Diatka, V., Rychl'y, P., Stranak, P., Suchomel, V., Tamchyna, A., & Zeman, D. (2014). HiNDENCORP - Hindi-English and Hindi-only corpus for machine translation. *Language Resources and Evaluation*, 3550–3555. http://www.lrec-conf.org/proceedings/lrec2014/pdf/835_Paper.pdf
[11] Shukla, A., Bansal, C., Badhe, S., Ranjan, M., & Chandra, R. (2023). An evaluation of Google Translate for Sanskrit to English translation via sentiment and semantic analysis. *Natural Language Processing Journal*, *4*, 100025. https://doi.org/10.1016/j.nlp.2023.100025
[12]Tymoczko, M. (2014). Enlarging translation, empowering translators. In *Routledge eBooks*. https://doi.org/10.4324/9781315759494