

A Multimodal Approach for English to Telugu Translation with FNet and Transformer Models

Joshita Malla¹[0000–1111–2222–3333], Ujwala Kanumuri²[1111–2222–3333–4444],
and Swathi Mytreya Chaganti³[2222–3333–4444–5555] Swaminathan
J⁴[3333–4444–5555–6666]

^{1,2,3,4}Department of Computer Science and Engineering, Amrita School of
Computing,
Amrita Vishwa Vidyapeetham, Amritapuri, India

Abstract. This research paper develops a multimodal translation system that is versatile in nature specifically designed for converting English to Telugu language, which addresses accessibility issues across various applications. Our method involves using ResNet50 to ensure that we have robust image feature extraction while using word tokenization to generate English and Telugu tokens. The translation is done through F-Net encoder-decoder architectures with the multimodal encoder acting as the de facto combiner of textual and visual information and the decoder generating the telugu translation to ensure the results are both accurate and contextually appropriate. Based on extensive experimental observations, considerable improvements were observed when compared with other unimodal approaches especially in terms of translation quality. Our framework goes beyond assistive devices for the blind and has the possibility of being used as a navigation aid, communication tool, or educational resource thereby improving access and inclusion among individuals who speak different languages.

Keywords: Multimodal translation, ResNet50, Transformer-based encoder-decoder, accessibility, inclusivity.

1 Introduction

Due to the increasing globalization in today's world, there is an increasing demand for easily accessible technology especially in areas with several languages. Countries like India, with a large number of languages and dialects, present unique challenges in making technology accessible to all as the translation of these languages is not an extensively researched area compared to more dominant languages. The syntactical and semantical differences between dominant languages such as English and regional languages like Telugu result in a translation gap, which is one of the primary challenges preventing the proper usage of devices. Overcoming this gap is crucial for encouraging inclusion and equitable access to technology among different types of end users.

Existing translation models have limitations[16], such as a lack of high-quality training data and syntactic differences between English and Telugu. This process

is further complicated by structural differences in word arrangement and sentence structure between these two languages. This demonstrates that effective translation algorithms capable of handling multilingual settings are required because they involve a wide range of assistive devices, educational tools, navigation systems, and communication devices.

Traditional approaches frequently miss out on visual cues because they rely on text-only based input. Different machine translation techniques like statistical and neural machine translation [1] have been explored to deal with these challenges. Attention mechanisms have also gained significant attention because of their ability to improve translation quality [2,17]. However, incorporation of multimodal elements such as images has started being viewed as a promising approach towards improving the quality of translations generated. While textual inputs just represent the linguistic content, images provide context, cultural references, and visual cues that can help understand the meaning and enhance translation accuracy. Nonetheless, the exploration of multi-modal translation techniques in less-resourced languages like Telugu is still relatively limited. In fact, most research on this topic has been done using English, Spanish, and Chinese as target languages thus not taking into account language diversity and richness of languages like Telugu.

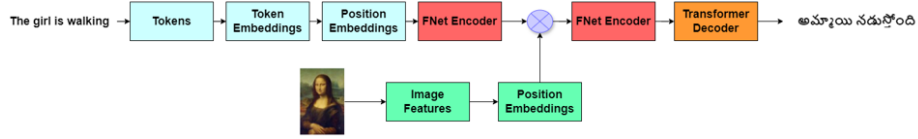


Fig. 1. Proposed model flow diagram

So in this research paper, we aim to address these problems by proposing a new translation strategy based on a multimodal approach as seen in Fig. 1. based on the studies [13,14,15] We first created a dataset consisting of different types of content and language complexities represented by texts in Telugu and English with their corresponding images. This dataset is used to train multimodal translation models which are then evaluated with respect to their performance at generating the Telugu translations. Multimodal machine translation models can effectively use visual input even in cases with limited textual context [3]. So, we developed this algorithm by combining textual features with visual elements in order to improve translation accuracy and meet the linguistic needs of a wide range of users. Although there have been tremendous improvements in machine translation systems, the diversity of language nuances and intricate visual contexts present new challenges that require better approaches. The effectiveness of the algorithm in generating precise translations and improving technology accessibility will be thoroughly evaluated through an analysis of its performance.

This research has major implications for the development of translation technology with special emphasis on reducing the gap between languages that differ by linguistic resources. The approach we have taken is more adaptable than mere translation. Including visual cues in the process of translation expands opportunities for various applications such as aiding the visually impaired as seen in the study [4] education, navigation, and communication tools thus improving user experiences across domains.

2 Related Works

The study [5] by Lekshmy O H and Swaminathan J employs a multi-modal machine translation framework guided by images and text to enhance translation correctness into Malayalam. The research employs the image-guided Multi-Modal Machine Translation methodology with LSTM models as a means to enhance translation accuracy and scene distributions in Malayalam. However, it should be noted that using sequential models such as Long Short Term Memory (LSTM) has a drawback because of the vanishing gradient problem leading to reduced efficiency in translation.

The authors further reflect on major data resources, evaluation campaigns as well as state-of-the-art approaches for both end-to-end and pipeline research designs. Performance measurement issues are also highlighted within this paper, along with future research needs including larger datasets and more input-output multimodality. However, ethical concerns especially on privacy and confidentiality in education have received little focus in this study

The study [6] by Paul Pu Liang et al. focuses on Multimodal machine learning. In this area, several other areas, such as video understanding and multisensor fusion in healthcare and robotics, have emerged, bringing up computational challenges due to data heterogeneity. It provides insights into advancements and open research directions through historical and recent perspectives by identifying key principles and technical challenges.

The study [7] by Suxia Lei et al. involves the development and application of a machine English translation evaluation system using a BP neural network algorithm to improve translation efficiency and quality. This model will help to get rid of some limitations inherent in conventional MT systems, including poor data utilization as well as huge model parameters through optimizing translation performance via applying neural network algorithms that may be hampered by such matters as inadequate information processing and immense model sizes.

The study [8] by Loitongbam Sanayai Meetei et al. proposes a method of multimodal English-to-Hindi machine translation that uses news images and captions. These are different from normal NLP images, which show significant events to give context. The text and image inputs are processed in parallel by separate encoders. Two tests were done to assess the influence of the image contents on translation systems. BPE for text representation and VGG-19 for visual feature extraction.

The study [9] by Sahinur Rahman Laskar et al focuses on participation in the WAT2022 workshop. This work examines neural machine translation (NMT) for low-resource language pairs and multimodal approaches that improve translation accuracy. The transliteration-based phrase pairs in English to Hindi multi-modal translations were improved through an approach developed by Team CNLP-NITS-PP during the WAT2022 workshop which led to a significant performance increase. However, this research only uses transliteration, thus it ignores much of what translated text is all about.

In comparison with existing studies in multimodal machine translation, our research represents a significant advancement in English-to-Telugu translation that includes both images and texts. Many of the previous inquiries in this area had to overcome several challenges. The most usual stumbling blocks were small datasets, vagueness in language, and lexical & semantic differences. Such issues have always prevented accurate translations of both text and visual information.

To tackle these problems, we propose a multimodal approach for translation tasks from English to Telugu architecture. Conventional multimodal translation models have been based on traditional encoders like convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for text. However, we implemented an FNet encoder-transformer decoder architecture for our multimodal machine translation model. Instead of complicated sequential models making use of sequential encoders and decoders, the FNet offers a simpler yet effective alternative for encoding images using its Fourier-based transformation method. The model then utilizes a transformer decoder that has proved useful in many natural language processing tasks for capturing long-range dependencies effectively through self-attention mechanisms to merge information from images and texts seamlessly. The outcome of the new combination not only advances translation accuracy but also assists in deepening the comprehension of how visuals and text interact during translations.

3 Dataset

There are no existing datasets for multimodal machine translation from English to Telugu. Hence, the dataset used in this task is a Telugu Visual Genome which we translated using the existing Hindi Visual Genome. In this dataset, the train data was created using 28,928 parallel English-Telugu sentences and 28,928 images. Each dataset item comprises an image, a rectangular region in the image, a source text in English describing the image region, and its translation in Telugu. After removing duplicate sentences having ID numbers 2391240, 2385507, 2328549 from parallel data and one image having ID number 2326837 (since the corresponding text is not present in parallel data), the parallel and image train data was reduced to 28,927. The selected rectangular portion of the image is represented using image coordinates (X, Y. Width, Height). The Testing and Validation sets consist of 1595 and 998 images respectively. Another text dataset containing English to Telugu text translation with 155798 sentences was

also employed and added to the token dictionary to generate more tokens and include all possible variations of Telugu sentences and words.

4 Proposed Methodology

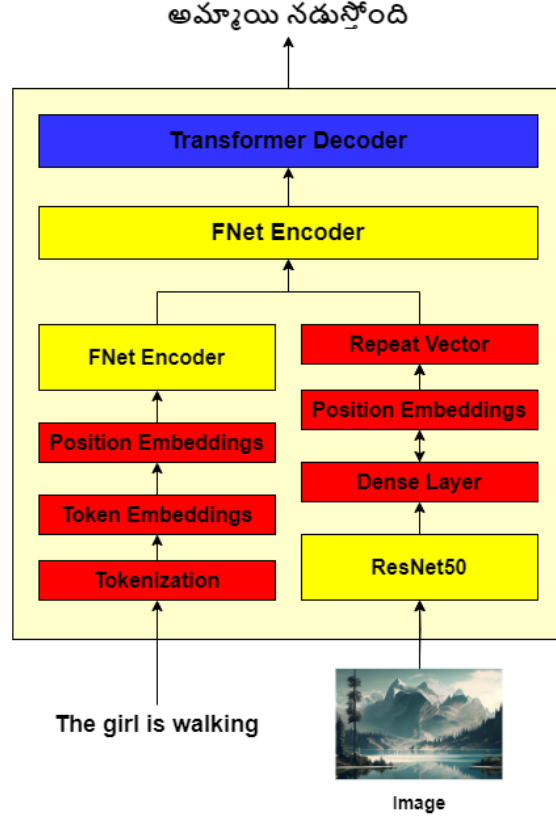


Fig. 2. Model Architecture

In order to improve translation accuracy and contextual relevance, we propose a multimodal translation approach in which we leverage both visual and textual information for translation.

As shown in Fig. 2., we start by tokenizing the text and integrating position embeddings, which contain the information on the position of each token in the sentence, to capture sequential patterns. Simultaneously, the pre-trained ResNet-50 model is used to extract rich visual features from the images. The textual and visual features are then fed into our multimodal encoder comprising

two FNet layers, where they undergo a fusion. The fused representations are then passed through a Transformer decoder translating these representations, which includes attention mechanisms that enable the model to focus on relevant parts of the input sequence and feed-forward networks that provide additional non-linearity and computational depth. This decoder can thus handle complex linguistic structures while producing contextually appropriate Telugu translations.

4.1 Text Processing

Tokenization is the process of breaking down text into smaller units called tokens. The types of tokenization are word tokenization which breaks down the text into individual words based on spaces, character tokenization which breaks text into individual letters and subword tokenization which creates units that are larger than characters but smaller than the entire word.

As Character-level tokenization may result in loss of semantic information and context subword tokenization was employed initially. However, current algorithms couldn't accurately perform subword tokenization for Telugu due to its complex structure. So, as a practical solution word tokenization has been implemented.

Word Tokenization Word tokenization effectively handled the Telugu vocabulary within the translation framework. It made sure that the model captures the semantic meaning of words. The Tokenizer class from the TensorFlow Keras library was used to implement word tokenization. This involves initializing the Tokenizer and fitting it to the corpus. This will create a dictionary where each word is mapped to a unique index. The preprocessed text is then split into words based on spaces and each word is represented by its corresponding index converting the text into tokenized sequences. This resulted in a comprehensive coverage of vocabulary of around 50000 tokens for Telugu and 17000 tokens for English.

4.2 Image feature extraction

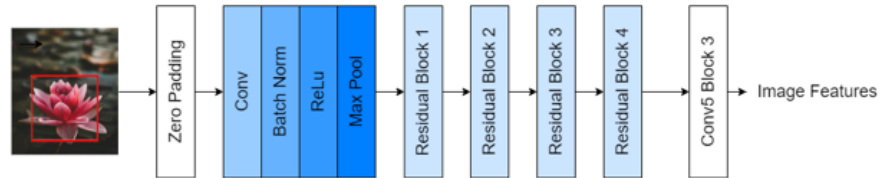


Fig. 3. Image Feature Extraction using ResNet50

In the current approach, based on the studies [10][11][12] we have used the pre-trained ResNet-50 model to extract image features. The pre-trained ResNet-50 is one of the most powerful models that has been trained thoroughly on the ImageNet dataset. We have trained this pre-trained CNN with an image dataset containing 28,928 images as well as their ROI coordinates for optimum results by removing noise. Usually, Pre-trained models like CNN are primarily used for image classification, and they are composed of mainly two parts - feature extraction and classification. In the case of a ResNet-50, the architecture has a global average pooling layer which is followed by a single fully connected layer without activation, used for classification tasks. As displayed in Fig 3., we have adopted this architecture for feature extraction by using layers up to the fully connected stage. This model accepts the input of shape (224, 224, 3) and returns the feature vectors of shape (, 2560). These feature vectors are further reshaped by passing through fully connected layers. Finally, a repeat vector layer is introduced to add one more dimension to the model, which is equal to the length of the maximum sentence in the English language which is 34, therefore our output vector will be of shape (, 34, 512).

4.3 FNet Encoder

Taking the tokens and image features as input, our encoder architecture comprises two FNet encoder layers, which effectively extract features from the sentences, providing higher-quality translations. In the FNet architecture seen in Fig. 4., the self-attention mechanism of the traditional Transformer models is replaced with a leaner mechanism: a parameterized Fourier transformation-based linear mixer for input tokens.

This approach is more effective and scalable, while also ensuring that the loss of information during the transformation is minimal. Before passing the textual data into the FNet layer, the text data is first passed through an embedding and position encoding layer where all the input tokens are mapped to their corresponding dense vector representations along with their positional information. These textual embeddings are then passed as input to the first FNet encoder layer. An early multimodal fusion of the resulting embeddings and image features is then performed and this is fed as input into the subsequent FNet layer. So the encoder extracts all the meaningful features and encodes them into a comprehensive representation for subsequent translation.

4.4 Transformer Decoder

Our intended pipeline contains a Decoder for generating the translations from the embedded sequences obtained from the encoder output. The Decoder part works with attention mechanisms and feed-forward networks utilizing a transformer-based architecture which can be visualized in Fig. 4. The decoder's ability to attend to all parts of the input sequence simultaneously enhances its capability to handle complex linguistic structures and idiomatic expressions, resulting in

more accurate translations, The decoder also uses attention mechanism to identify important parts of the source text, which includes long-range dependencies and context more efficiently. Due to this the decoder model does not suffer from the vanishing gradient problem that sequential models face. Furthermore, the integration of feed-forward networks into this model enables it to learn intricate relationships between input and output sequences that in turn promote the production of coherent as well as fluent translations. In this way, our Decoder goes beyond legacy sequential decoders thereby offering improved translation performance and guaranteeing much more accurate results given their context.



Fig. 4. Model Architecture

4.5 Sampling

In order to generate the translations from the resulting probability distribution sampling is used. At first, in order to achieve proper translation within our

pipeline, a greedy sampling technique was used. Though it is simple and efficient, this method often produced more than one sentence that meant the same thing, thus reducing the quality of translations as a whole. In response to this, we tried beam sampling as our next step toward improvement. By considering several competing translations at once and identifying the best ones using a predefined beam width, repetitions were effectively eliminated from beam sampling and the translated texts became more fluent with high coherence level.

5 Experiments and Results

The efficiency of various encoder-decoder architectures was explored in our experiments on multi-modal English to Telugu translation. The main objective was to improve translation quality from classic LSTM, GRU, and Transformer encoders. We found that by using a FNet encoder and a transformer decoder we got the best results as shown in Table 1.

Metric	LSTM model	GRU model	TED	FETD
Bleu score	28.9677	35.6099	46.2206	54.7341

Table 1. Result of LSTM, GRU, Transformer Encoder Decoder and FNet Encoder-Transformer Decoder model using BLEU score

TED = Transformer Encoder-Decoder model

FETD = FNet Encoder-Transformer Decoder model

The model was trained using a training dataset which contains 28928 rows of data and monitored using a validation dataset containing 998 rows of data respectively.

5.1 Test Dataset

To evaluate our model we used a comprehensive test dataset that contains 1595 rows of English sentences, their corresponding Telugu translations, and images. The dataset contained unseen data that allowed us to effectively train our model. We used the BLEU score metric to evaluate our model.


<p>Image Id : 1593079</p> 	<p>Source Text (English)</p>	<p>The batter swinging the bat</p>
	<p>Predicted Text (Telugu)</p>	<p>బ్యాట్‌ను బంతి కొట్టే ఆటగాడు బ్యాట్‌ను ఊపుతున్నారు</p>
	<p>Google Translation</p>	<p>ఒక కొట్టు బ్యాట్ ఊపుతున్నాడు</p>

Fig. 5. Results of the Translation

The results made by our model showed significantly better results in comparison to previous sequential models achieving a BLEU score of 54. This can be seen in Fig. 5. which compares the translation displayed by our model on comparison to that of Google Translate. Conversely, other traditional types of sequential models like LSTM and GRU registered much lower BLEU scores at 29 and 35 respectively. Even though Transformer models are well-known for their effectiveness in natural language processing tasks, they were outperformed with a BLEU score of 46.

6 Conclusion and Future Works

This research is an attempt to aid the translation of a low-resource language like Telugu to improve the accuracy as well as preserve the context which may often be lost in unimodal approaches. So our Multimodal English to Telugu Translation approach extracts meaningful features from the text and images and integrates these features to generate more accurate and contextually rich translations as seen in Table 1 when compared to existing sequential models like LSTMs, GRU, or traditional unsequential transformer encoder-decoder models. It finds its vast potential applications in navigation aids, communication tools, or assistive technology. As part of future work, more analysis and research can be done to improve the performance of the machine translation model which can include but is not limited to multimodal fusion strategies such as early fusion, late fusion, or dynamic fusion mechanisms as well as experimenting with other architectures.

References

1. Premjith B, Kumar MA, Soman KP (2019) Neural machine translation system for English to Indian language translation using MTIL parallel corpus. 28::387–398

2. Singh S, Kumar A, Soman KP (2018) Attention based English to Punjabi neural machine translation. 34::1551–1559
3. Caglayan O, Madhyastha P, Specia L, Barrault L (2019) Probing the need for visual context in multimodal machine translation
4. Abhishek S, Sathish H, Kumar A, Anjali T (2022) Aiding the visually impaired using artificial intelligence and speech recognition technology. In: 2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE
5. Lekshmy HO, Jayaraman S (2022) English-malayalam vision aid with multi modal machine learning technologies. In: 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE
6. Liang PP, Zadeh A, Morency LP (2023) Foundations Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions
7. Lei S, Li Y (2023) English Machine translation System Based on Neural Network Algorithm. 228::409–420
8. Meetei LS, Singh SM, Singh A (2023) Hindi to English Multimodal Machine Translation on News Dataset in Low Resource Setting. 218::2102–2109
9. Laskar SR, Singh R, Karim MF (2022) Investigation of English to Hindi Multimodal Neural Machine Translation using Transliteration-based Phrase Pairs Augmentation. In: Proceedings of the 9th Workshop on Asian Translation
10. Li Z, Gu T, Li B (2022) ConvNeXt-based fine-grained image classification and bilinear attention mechanism model. 12::9016
11. Kamal K, Hamid EZ (2023) A comparison between the VGG16, VGG19 and ResNet50 architecture frameworks for classification of normal and CLAHE processed medical images
12. Dhanya R, Paul IR, Akula SS (2019) A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection. In: International Conference on Intelligent Computing and Control Systems (ICCS)
13. Laskar SR, Dadure P, Manna R (2022) English to Bengali Multimodal Neural Machine Translation using Transliteration-based Phrase Pairs Augmentation. In: Proceedings of the 9th Workshop on Asian Translation
14. Parida S, Panda S, Biswal SP (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)
15. Kumar VV, Lalithamani N (2022) English to Tamil Multi-Modal Image Captioning Translation. In: 2022 IEEE World Conference on Applied Intelligence and Computing (AIC). IEEE
16. Kumari D (2013) Translation An Attempt to Deconstruct Historiography in The Brande. 4::1–8
17. Vaswani A, Shazeer N, Parmar N (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA