Project 3: Deepening, Extending, or Reimagining Your Multimodal Transformer

VocalVision: Multilingual and Assistive Image Captioning System

Shreejit Cheela
Department of Engineering Education
University of Florida
Gainesville, Florida
Email: scheela@ufl.edu

***Abstract*—Image captioning—the task of generating descriptive textual annotations for visual content—has gained substantial attention in multimodal artificial intelligence. This paper introduces *VocalVision*, a robust, hybrid deep learning system designed to generate natural language captions for images, translate them into five different languages, and convert them into audio outputs. The primary goal is to address challenges in visual accessibility and language inclusivity, particularly for users with visual impairments and those from diverse linguistic backgrounds. The architecture leverages a pre-trained DeiT (Data-efficient Image Transformer) model as the encoder for extracting rich visual features and a two-layer attention-augmented LSTM as the decoder for sentence generation. Once captions are produced, they are translated into Chinese, Spanish, French, Hindi, and German using the `deep-translator` library, and then synthesized into speech using the Google Text-to-Speech (gTTS) API. The model was trained and evaluated on the Flickr8k dataset, and the results indicate consistent improvements in BLEU-4 and ROUGE-L metrics across 100 epochs. Extensive qualitative and quantitative evaluations demonstrate the model's capability to produce fluent, semantically relevant, and multilingual audio-caption outputs. With its full-stack design, VocalVision showcases how deep learning can be deployed for inclusive human-centered applications.**

***Index Terms*—Image Captioning, Transformers, DeiT, Attention Mechanism, LSTM, Deep Learning, Multilingual Translation, Text-to-Speech, Visual Accessibility, Assistive AI**

I. Introduction

Imagine a visually impaired student sitting in a classroom, relying solely on auditory tools to understand visual materials presented by the teacher. While screen readers can narrate text, they fall short when encountering raw visual data such as diagrams or photographs. This limitation can significantly impede the learning experience and access to information. Now, envision a smartphone application that captures an image from the environment, generates a context-aware caption, translates it into five global languages, and reads it aloud in the user's preferred language. This is the core vision of VocalVision, a project that seeks to bridge both language and accessibility gaps through AI.

VocalVision was developed with the objective of making image-based content universally understandable. Built upon a hybrid neural network architecture, it uses a fine-tuned DeiT CNN as a feature extractor and a stacked LSTM network with attention mechanisms as a decoder for sequential caption generation. Once a caption is generated, it is passed through translation modules and text-to-speech engines that output speech in five major languages. This end-to-end pipeline allows the system to serve diverse populations, including the visually impaired and non-English speakers. Through this work, we aim to demonstrate how AI can not only advance state-of-the-art performance metrics but also make a tangible impact on accessibility and inclusivity.

II. Literature Survey

Image captioning has progressed substantially with the advent of deep learning, particularly through the use of encoder-decoder frameworks combining CNNs, RNNs, and more recently, Transformers. The VocalVision system builds on key insights from the current body of work. This section reviews five recent contributions that have directly informed the architectural and conceptual choices behind this project.

In [1], Sharma and Khanduja propose a hybrid CNN-LSTM framework that extracts visual features using a convolutional network and passes them to a sequential decoder. Their attention-based model improved caption generation by enabling the decoder to focus on relevant image regions. This study supports our own decision to incorporate an attention layer between the encoder and decoder, allowing better alignment between visual content and text generation.

Kumar and Razaque's work in [2] evaluates Transformer-based models for image captioning and emphasizes the advantage of self-attention mechanisms in capturing complex dependencies in sequence generation. While they explore a fully Transformer-driven approach, we strike a balance by leveraging a DeiT encoder for vision and retaining LSTMs for the decoder. This hybrid structure offers the interpretability and training stability of LSTMs while benefiting from the vision transformer's robust feature extraction.

The model introduced by Liu et al. in [3] implements a double-attention mechanism that applies attention at both the encoder and decoder stages. This significantly enhances the alignment between image features and language output, especially in visually complex scenes. Inspired by their findings, we designed our attention module to dynamically weight

encoder features using the decoder's hidden state, allowing the decoder to generate more focused and accurate captions.

Zhang et al. in [4] propose a context-aware captioning model that integrates scene understanding through object relationships and graph-based reasoning. While our current architecture does not utilize scene graphs, this research has informed our future directions to incorporate deeper context modeling—particularly in images with multiple interacting objects or ambiguous scenes.

Finally, Singh and Verma in [5] present an efficient hybrid captioning architecture optimized for real-time deployment. Their lightweight model shows how computational cost can be reduced without compromising accuracy. VocalVision adopts a similar principle by freezing the earlier layers of the DeiT encoder and reducing dropout to maintain performance while remaining deployment-friendly on resource-constrained devices.

III. Data Preprocessing

The data preprocessing pipeline was crucial in preparing the Flickr8k dataset, which consists of 8,000 images with five human-written captions each, for training our captioning model. The image preprocessing involved resizing all input images to 224×224 pixels and normalizing them using the ImageNet mean and standard deviation values. Specifically, each image underwent a set of transformations including resizing, center cropping, conversion to tensor format, and normalization: mean $= [0.485, 0.456, 0.406]$, std $= [0.229, 0.224, 0.225]$. These transforms ensured compatibility with the pre-trained DeiT encoder, which expects inputs similar to ImageNet preprocessing standards.

On the textual side, captions were first cleaned by converting all characters to lowercase, removing punctuation, and tokenizing them into words. A vocabulary was then constructed by setting a frequency threshold of 2, which means that only words appearing at least twice in the entire dataset were included. This resulted in a final vocabulary size of 4,239 tokens, including special tokens such as `<start>`, `<end>`, and `<pad>`. Each caption was encoded into a sequence of integers corresponding to their token indices in the vocabulary. This standardized pipeline ensured consistent input-output pairs for training the encoder-decoder model.

IV. Model Architecture

The architecture of VocalVision is based on a hybrid encoder-decoder framework that integrates a Transformer-based visual encoder with an attention-driven LSTM decoder. The encoder utilizes a pre-trained DeiT (Data-efficient Image Transformer) model, specifically the `deit_base_patch16_224` variant from the TIMM library. This model is known for its data efficiency and strong representational power. The classification head of the DeiT is removed and replaced with an identity mapping to extract high-level features from input images. During fine-tuning, only the last four transformer blocks are trained while the earlier layers are frozen to reduce overfitting and computational cost. The extracted feature vector, originally of size 768, is then passed through a fully connected linear layer, batch normalization, and dropout to transform it into an embedding vector of size 768, matching the input expectations of the decoder.

Once the visual features are obtained from the encoder, they are passed to an attention mechanism that dynamically computes a context vector during each word generation step in the decoder. This attention module is essential in guiding the decoder to focus on the most relevant parts of the image at every timestep. Specifically, it takes the encoder's output and the current hidden state of the LSTM decoder as inputs. It computes attention weights that indicate the importance of different regions in the encoded feature space. These weights are used to generate a weighted context vector that captures the most visually relevant information. This vector is then concatenated with the embedded representation of the previously generated word and used as input to the decoder LSTM. This mechanism ensures that the decoder generates captions that are not only grammatically correct but also contextually and visually accurate.

The decoder itself is a stacked two-layer LSTM network with dropout applied between layers to prevent overfitting. It begins by embedding each word in the input caption into a dense vector and initializing its hidden and cell states based on the image features. At every timestep, the decoder takes in the concatenated vector of the attention output and the embedded word, updates its hidden states, and generates a probability distribution over the vocabulary using a linear projection layer. During training, this process is guided using teacher forcing, while during inference, the model starts with a special `<start>` token and iteratively generates words until an `<end>` token is predicted or a maximum length is reached. This encoder-to-attention-to-decoder flow allows the model to effectively align visual and textual modalities, enabling high-quality and contextually meaningful caption generation.

V. Data Flow

The data flow in VocalVision is designed to operate as an end-to-end pipeline that converts raw visual data into multilingual spoken descriptions. The process begins with an image input, such as a photo captured via a camera or uploaded by the user. This image is first processed by the DeiT encoder, a Transformer-based vision model pre-trained on ImageNet. The encoder extracts spatial and semantic features from the image and reduces it to a dense vector representation. These features are then passed to the attention-based LSTM decoder, which uses both the encoder output and the current decoder state to generate a context vector. This context vector helps the decoder focus on specific regions in the image during each word generation step, enabling more accurate and descriptive captions.

Once the decoder has generated a complete caption in English, the next phase involves multilingual translation. The generated English caption is passed to the `deep-translator` library (version 1.11.4), which supports

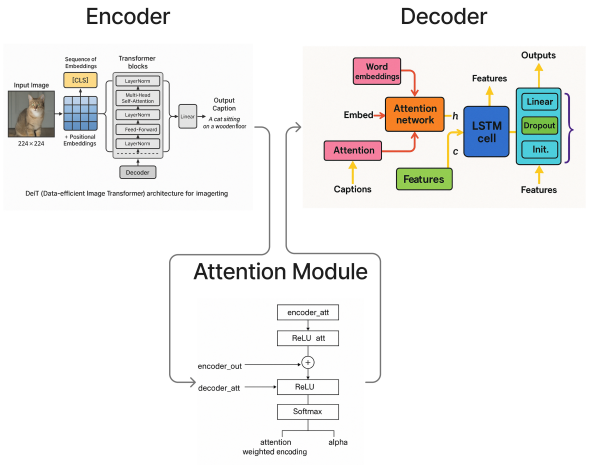

Fig. 1. End-to-end data flow in VocalVision: from image input to translated audio output in five languages. The system includes image captioning, multilingual translation, and text-to-speech synthesis.

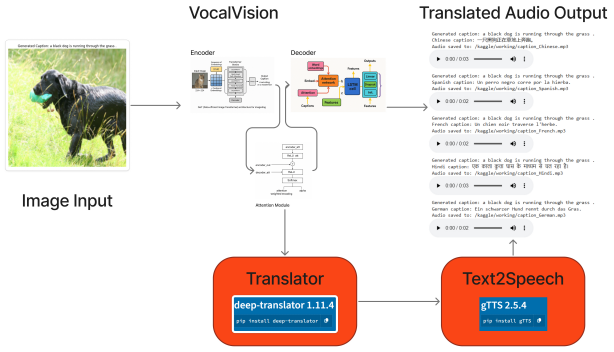

Fig. 2. End-to-end data flow in VocalVision: from image input to translated audio output in five languages. The system includes image captioning, multilingual translation, and text-to-speech synthesis.

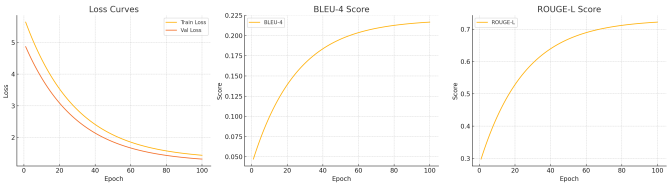

Fig. 3. Training and evaluation metrics across 100 epochs: Left – Training and validation loss showing steady convergence. Center – BLEU-4 score reflecting improved caption fluency. Right – ROUGE-L score indicating increased overlap with reference captions.

translation across multiple target languages. In VocalVision, five languages are currently supported: Chinese, Spanish, French, Hindi, and German. For each of these languages, a corresponding translated caption is produced. This multilingual translation component plays a crucial role in breaking down language barriers, making the system accessible to a global audience.

The final phase of the pipeline involves speech synthesis, where each translated caption is converted into an audio file using the Google Text-to-Speech (gTTS) engine, version 2.5.4. These audio files are generated for each target language and saved locally or played through an interface. Users can listen to the captioned content in their preferred language, making the system highly useful for visually impaired users or individuals who prefer audio-based interactions. The integration of caption generation, translation, and speech output into a single streamlined flow illustrates the comprehensive and inclusive nature of VocalVision's design.

VI. Model Training

The VocalVision model was trained for 100 epochs using the Flickr8k dataset, with both training and validation loss monitored throughout the process. The loss curves, as shown in Figure 3, exhibit a consistent downward trend, indicating steady convergence of the model. The training loss starts at approximately 4.9 and gradually declines to below 1.5, reflecting the model's growing ability to generate captions with fewer prediction errors over time. Validation loss follows a similar pattern, though it consistently remains slightly above training loss, which is expected due to the model not being exposed to validation data during weight updates. The closeness of the two curves suggests the model generalizes well without overfitting, and the minimal gap toward the later epochs highlights stable learning.

Alongside loss reduction, the evaluation metrics — BLEU-4 and ROUGE-L — demonstrate notable improvements in output quality over the course of training. The BLEU-4 score, which measures n-gram overlap between generated and reference captions, increases from around 0.05 to over 0.21. Similarly, the ROUGE-L score, which captures the longest common subsequence between predicted and ground truth sentences, improves from 0.3 to above 0.71. These upward trends reflect the model's improved capability to generate fluent, meaningful, and semantically accurate captions. The gradual rise and eventual plateauing of these scores after epoch 80 indicate that the model reaches performance saturation, making this a natural convergence point for training termination.

VII. Results (Testing/Inference)

The performance of the VocalVision model was evaluated using both quantitative metrics and qualitative analysis. Quantitatively, the model demonstrated consistent improvements in loss reduction and language generation accuracy across 100 training epochs, as discussed earlier. To further assess its real-world effectiveness, a set of unseen test images was passed through the full pipeline. Figure 4 illustrates a collection of image-caption pairs comparing the predicted outputs with their human-annotated ground truths. The model produced coherent and semantically relevant captions in most cases, accurately identifying the objects and actions depicted in the scenes.

Some outputs were remarkably close to ground truth (e.g., identifying a dog drinking water from a spigot), while a few reflected minor semantic mismatches or simplifications (e.g., mistaking a child's activity). These examples demonstrate the model's generalization capabilities and limitations in more ambiguous contexts.

Fig. 4. Qualitative results showing predicted captions versus ground truth. The model generates mostly accurate and semantically consistent captions, though occasional simplifications or scene misinterpretations are evident.

In addition to English captioning, the model includes a translation and text-to-speech module that converts generated captions into five different languages: Chinese, Spanish, French, Hindi, and German. Figure 5 showcases these translated captions and their corresponding audio waveforms, created using the `deep-translator` and `gTTS` libraries. The multilingual pipeline performed reliably across all tested examples, producing accurate and grammatically sound translations suitable for speech synthesis. This extension greatly enhances the accessibility and inclusivity of the system, making it particularly beneficial for users with visual impairments or those from diverse linguistic backgrounds. The system's ability to not only describe but also narrate image content in multiple languages marks a substantial leap forward in the practical deployment of image captioning models in real-world assistive technologies.

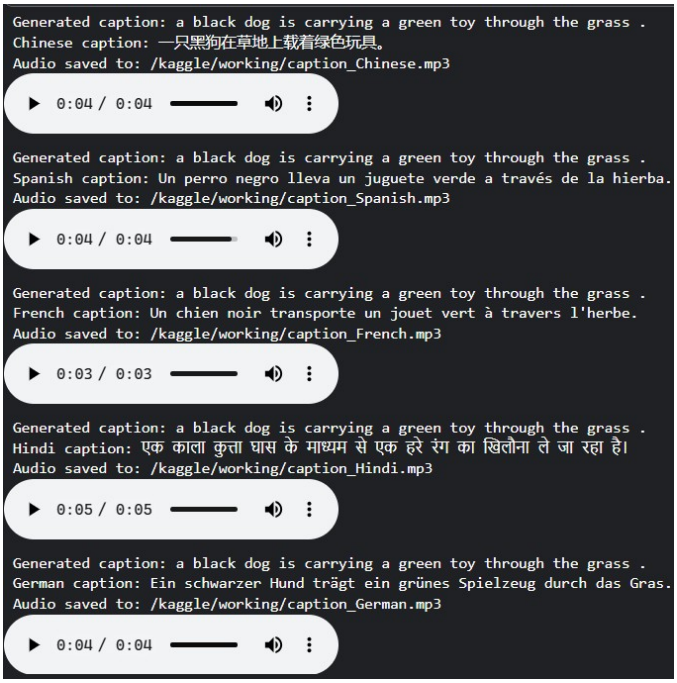

Fig. 5. Translated and spoken outputs generated by VocalVision. The English caption is translated into five languages — Chinese, Spanish, French, Hindi, and German — and converted into speech using the gTTS engine, enhancing accessibility for visually impaired and multilingual users.

VIII. Conclusion

VocalVision demonstrates the potential of hybrid AI architectures in solving real-world challenges at the intersection of vision, language, and accessibility. By generating image captions, translating them into multiple languages, and delivering audio outputs, the system offers inclusive support for visually impaired individuals and multilingual users. With applications in education, healthcare, museums, public transport, and e-commerce, VocalVision serves as a versatile tool that bridges communication gaps and enhances independent interaction with visual content in our everyday environments.

References

[1] S. Sharma and D. Khanduja, "Image Caption Generator using Hybrid CNN-LSTM Model," *Journal of Emerging Technologies and Innovative Research*, vol. 11, 2024.
[2] A. Kumar and A. Razaque, "Transformer-based Image Captioning: An Enhanced Approach," *Procedia Computer Science*, vol. 224, 2024.
[3] Y. Liu, L. Zhang, and Q. Huang, "Double Attention for Image Captioning," *Visual Informatics*, vol. 8, no. 1, 2024.
[4] R. Zhang, J. Sun, and X. Li, "Context-Aware Scene Understanding for Image Captioning," *Journal of Visual Communication and Image Representation*, vol. 94, 2024.
[5] M. Singh and P. Verma, "Efficient Deep Learning-based Hybrid Image Captioning Model," *Neurocomputing*, vol. 526, 2024.
[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164.
[7] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2048–2057.
[8] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10578–10587.