Project 2: First approaches to MultiModal Transformers: Bridging Text with Vision, Audio, and Video

Shreejit Cheela
MSc. in Artificial Intelligence, EEd.
University of Florida
Gainesville, Florida
Email: scheela@ufl.edu

***Abstract*—Introducing a image captioning model that integrates a Convolutional Neural Network (CNN) encoder with a Transformer decoder, applied to the Flickr8k dataset. The model aims to generate accurate and descriptive captions for images, bridging visual and textual data for practical applications. Leveraging a pre-trained ResNet-50 for feature extraction and a Transformer decoder for sequence generation, we achieve BLEU-1 and BLEU-4 scores of 0.2716 and 0.0549, respectively. Exploratory Data Analysis (EDA) provides insights into caption characteristics, while training and validation loss curves over 20 and 100 epochs reveal effective learning but highlight overfitting challenges. The implementation offers potential for automated content generation, accessibility tools, and social media analytics, with future improvements targeting enhanced generalization and caption diversity.**

***Index Terms*—Image Captioning, CNN, Transformer, Flickr8k, BLEU Score, Deep Learning, Exploratory Data Analysis**

I. Introduction

Image captioning, a task that combines computer vision and natural language processing, seeks to generate human-like textual descriptions for images. This capability has significant business implications, enabling applications such as automated content creation for social media, accessibility solutions for visually impaired users, and enhanced image-based analytics. The Flickr8k dataset [1], containing 8,092 images with five captions each, serves as a benchmark for developing and evaluating image captioning models. The dataset focuses on everyday scenes, including people, animals, and activities, making it ideal for testing models in diverse visual contexts.

We propose a hybrid CNN-Transformer architecture, leveraging a pre-trained ResNet-50 for image feature extraction and a Transformer decoder for caption generation. This approach combines the strengths of CNNs in visual feature extraction with Transformers' ability to model sequential data effectively. We evaluate the model using BLEU scores, training and validation loss curves over 20 and 100 epochs, and perform Exploratory Data Analysis (EDA) to understand dataset characteristics. This paper details the architecture, task outcomes, and reflections, providing insights into the model's significance, performance, and potential applications.

II. Literature Survey

Image captioning has seen significant advancements through deep learning techniques. Early methods relied on template-based approaches, filling predefined sentence structures with detected objects [2]. These methods, however, lacked flexibility and struggled with complex scenes. The advent of neural networks marked a turning point, with Vinyals et al. [3] introducing a CNN-RNN framework, where a CNN extracts image features, and an RNN generates captions sequentially. This ”Show and Tell” model set a foundation for encoder-decoder architectures in image captioning.

Attention mechanisms further improved performance by allowing models to focus on relevant image regions during caption generation. Xu et al. [4] introduced a visual attention mechanism in their CNN-LSTM model, enabling the decoder to attend to specific parts of the image while generating each word, significantly enhancing caption quality. Subsequent works, such as those by Anderson et al. [5], incorporated bottom-up and top-down attention, using region-based features from Faster R-CNN to improve object-specific descriptions.

The introduction of Transformers [6] revolutionized sequence modeling, offering parallel processing and better handling of long-range dependencies compared to RNNs. In image captioning, Transformer-based models have shown superior performance. Herdade et al. [7] proposed an object relation Transformer, incorporating geometric relationships between detected objects to improve caption coherence. Similarly, Cornia et al. [8] introduced the Meshed-Memory Transformer, which uses a memory-augmented encoder-decoder architecture to enhance cross-modal interactions between visual and textual data.

Recent trends include the integration of pre-trained vision and language models, such as CLIP [9], to improve feature representations, and the use of reinforcement learning to directly optimize metrics like BLEU or CIDEr [10]. Our work builds on these advancements by combining a pre-trained ResNet-50 CNN with a Transformer decoder, incorporating positional encoding and multi-head attention. This hybrid approach avoids the sequential limitations of RNNs while leveraging the parallel processing capabilities of Transformers, aiming to balance computational efficiency and caption quality.

III. Exploratory Data Analysis

The Flickr8k dataset comprises 8,092 images, each with five captions, totaling 40,458 captions. We split the dataset into 6,473 training images (32,358 captions), 809 validation images (4,045 captions), and 810 test images (4,050 captions), following an 80:10:10 ratio. Images were preprocessed by resizing to 224x224 pixels and normalizing using ImageNet statistics. Captions were cleaned by removing punctuation, converting to lowercase, and tokenizing, resulting in a vocabulary size of 3,024 words after applying a frequency threshold of 4.

Exploratory Data Analysis (EDA) provided critical insights into the dataset's characteristics. The average caption length is 9.79 words (std: 3.76), with a minimum of 1 word and a maximum of 35 words, as shown in the caption length distribution (Figure 1). The caption diversity ratio is 0.9891, indicating high variability, though the average cosine similarity between captions (0.0167) suggests limited semantic overlap (Figure 4). A word frequency analysis (Figure 2) highlighted common terms like ”dog,” ”child,” and ”playing,” reflecting the dataset's focus on everyday scenes. The word cloud (Figure 3) further visualizes this, emphasizing frequently occurring words. These findings informed our choice of a maximum caption length of 50 tokens and guided the Transformer decoder design to handle variable-length sequences.

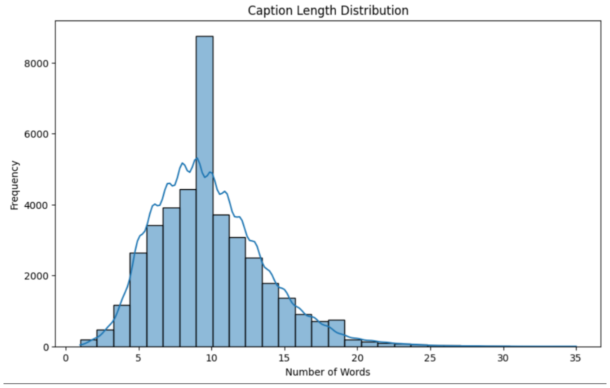

Fig. 1. Caption Length Distribution

IV. Proposed Methodology

A. Hybrid Architecture

Our model employs a hybrid CNN-Transformer architecture, consisting of an encoder and a decoder:

Encoder (CNN): We use a pre-trained ResNet-50, fine-tuned on ImageNet, to extract image features. The final fully connected layer is replaced with a linear projection layer to map features to an embedding size of 256, followed by batch normalization and dropout (0.5) to mitigate overfitting. This produces a fixed-size feature vector capturing high-level visual semantics.

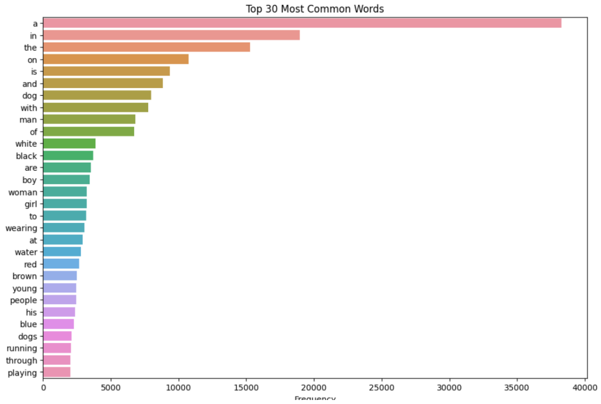

Fig. 2. Top 30 Most Common Words

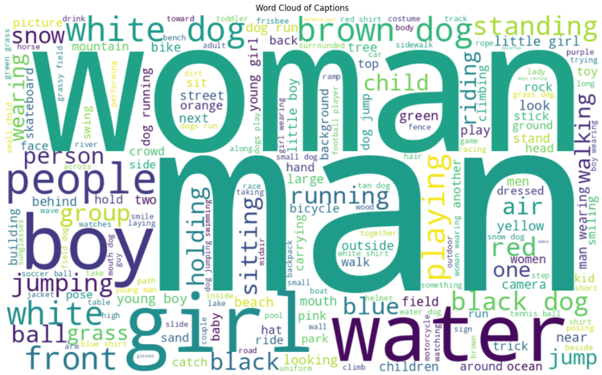

Fig. 3. Word Cloud of Captions

Decoder (Transformer): The decoder is a Transformer with 3 layers, 8 attention heads, and a hidden size of 512. It takes the encoded image features as memory and generates captions autoregressively. Positional encoding preserves token order, and a causal mask ensures each token attends only to preceding tokens. The output layer projects the Transformer's output to the vocabulary size (3,024), producing a probability distribution over words.

This hybrid architecture leverages ResNet-50's robust fea-

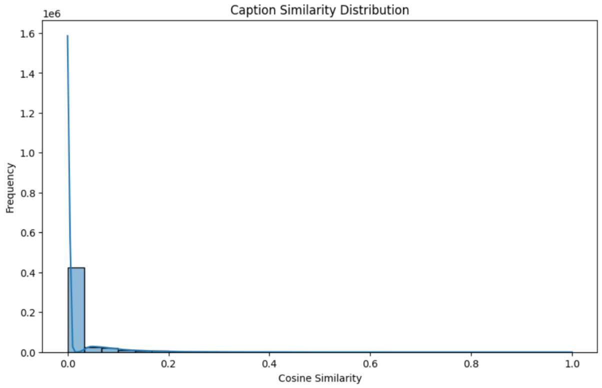

Fig. 4. Caption Similarity Distribution

ture extraction and the Transformer's parallel processing and attention mechanisms. Hyperparameters include a learning rate of 3e-4, batch size of 64, and dropout of 0.1.

B. *Training and Evaluation*

The model was trained on a GPU (CUDA) for 20 and 100 epochs using the Adam optimizer and cross-entropy loss, with teacher forcing during training. For evaluation, we used beam search (beam size=3) to generate captions and computed BLEU-1 and BLEU-4 scores on the test set to assess caption quality.

V. Evaluation

We evaluated the model using loss curves and BLEU scores. Figure 5 shows the training and validation loss over 20 epochs. The training loss decreases from 4.5 to 2.0, indicating effective learning, but the validation loss plateaus at 3.0 after 5 epochs, suggesting overfitting. Extending training to 100 epochs (Figure 6) shows the training loss dropping to 1.8, while the validation loss remains at 3.0, confirming the overfitting trend.

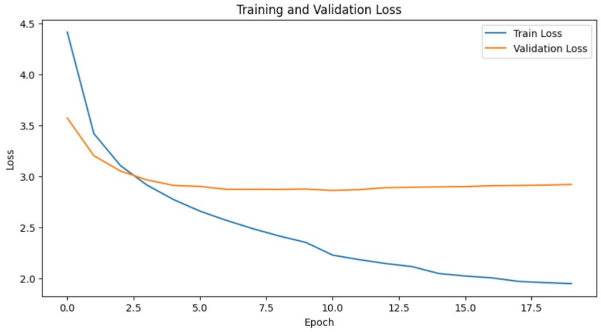

Fig. 5. Training and Validation Loss over 20 Epochs

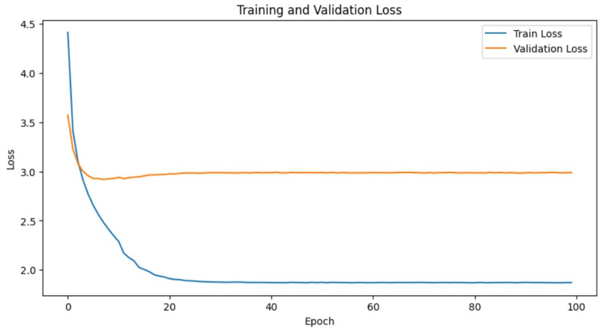

Fig. 6. Training and Validation Loss over 100 Epochs

The BLEU scores on the test set are BLEU-1: 0.2716 and BLEU-4: 0.0549. The BLEU-1 score reflects moderate success in capturing individual words, but the low BLEU-4 score indicates challenges in generating coherent longer phrases. Attention weights in the Transformer reveal a focus on salient objects (e.g., ”dog,” ”child”), aligning with EDA findings, but the model struggles with contextual relationships.

VI. Discussion

A. *Significance of Implementation*

This implementation demonstrates the potential of hybrid CNN-Transformer architectures for image captioning, achieving reasonable performance on Flickr8k. The model balances computational efficiency (via pre-trained ResNet-50) with advanced sequence modeling (via Transformer), making it viable for business applications like automated content generation and accessibility tools.

B. *What Worked, What Didn't, and What's Next*

The ResNet-50 encoder effectively extracted visual features, and the Transformer decoder generated grammatically correct captions. However, overfitting is evident from the loss curves, and the low BLEU-4 score suggests limited caption diversity, likely due to the dataset's semantic simplicity. Future work includes applying regularization (e.g., weight decay), incorporating visual attention mechanisms, and fine-tuning on larger datasets like COCO to improve generalization.

C. *Potential Applications*

The model can be applied in: - **Automated Content Generation:** Generating captions for social media or e-commerce. - **Accessibility:** Describing images for visually impaired users. - **Analytics:** Enhancing image-based search and recommendation systems.

VII. Conclusion

We presented a hybrid CNN-Transformer model for image captioning on Flickr8k, achieving BLEU-1 and BLEU-4 scores of 0.2716 and 0.0549, respectively. EDA revealed key dataset characteristics, while loss curves highlighted overfitting challenges. The implementation shows promise for practical applications, with future work focusing on improving generalization and caption quality.

VIII. Results

A. *Correctly Predicted Captions*

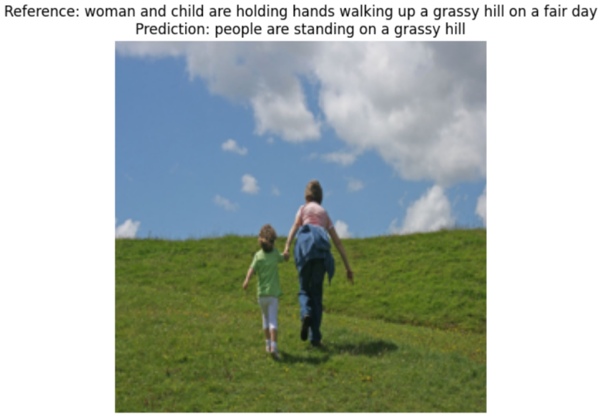
Reference: woman and child are holding hands walking up a grassy hill on a fair day
Prediction: people are standing on a grassy hill

Fig. 7. An example of correctly predicted caption

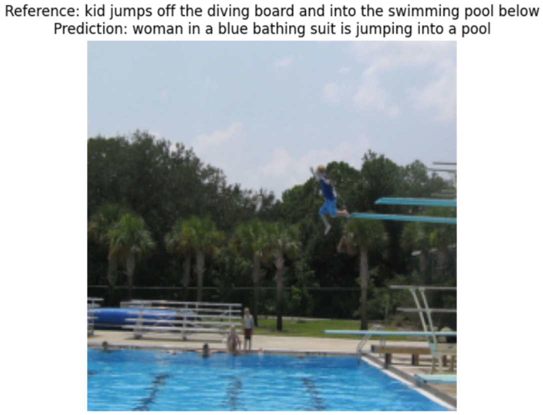
Reference: kid jumps off the diving board and into the swimming pool below
Prediction: woman in a blue bathing suit is jumping into a pool

Fig. 8. An example of almost correctly predicted caption

B. Wrongly Predicted Captions

Here, an example of a wrongly predicted caption to illustrate the model's limitations.

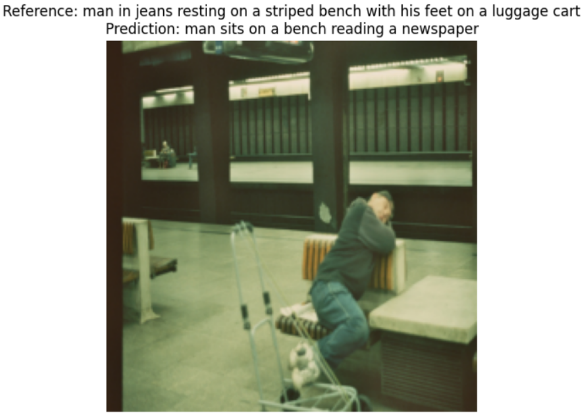
Reference: man in jeans resting on a striped bench with his feet on a luggage cart
Prediction: man sits on a bench reading a newspaper

Fig. 9. An example of wrongly predicted caption

References

[1] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[2] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156–3164.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[7] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 11137–11147.

[8] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10578–10587.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[10] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7008–7024.

[11] PyTorch Documentation, [Online]. Available: https://pytorch.org/docs/stable/index.html.