Image Captioning using Diffusion Models in Vision

Neha Eshwaragari
Artificial Intelligence Systems
University of Florida
Gainesville, USA
neha.eshwaragari@ufl.edu

***Abstract*—A thorough implementation and assessment of a deep learning pipeline designed specifically for image captioning and caption-aware image generation are presented in this report. The system's dual goals are to automatically produce descriptive natural language captions for visual input and use those captions to conditionally synthesize images. The architecture does this by combining modules for natural language processing and computer vision into a single framework. In order to extract rich visual features from input images, the pipeline uses transfer learning by utilizing a pre-trained Convolutional Neural Network (CNN) model, specifically ResNet-50. Following their mapping to a textual representation space, these features are used to condition GPT-2, a Transformer-based language model, for sequential caption generation. With this configuration, the model can comprehend an image's visual semantics and express them in logical, human-readable sentences. Several evaluation metrics are included to gauge the captioning's efficacy and possible image regeneration. BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) are used to evaluate linguistic quality, and PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are used to evaluate perceptual image quality. A strong understanding of both textual coherence and image fidelity is offered by these complementary metrics. Hugging Face's Transformers library for language modeling and PyTorch for deep learning operations are used to implement the entire model. To improve usability and accessibility for non-technical users, an interactive Gradio interface is constructed on top of the system to enable real-time testing and result visualization. This study demonstrates how language and vision work together in multimodal AI, opening the door for uses in intelligent multimedia systems, assistive technology, and content production.**

***Index Terms*—Image Captioning, Diffusion Models, BLEU, METEOR, Vision-Language Models, PyTorch, Gradio, Transformer**

I. Introduction

An important development in the field of human-computer interaction is the capacity to use natural language to describe visual content. By enabling machines to interpret, comprehend, and convey the semantics of visual data in a format that is readable by humans, this capability closes the gap between visual perception and linguistic expression. AI-driven content creation, automated digital asset management systems for massive image repositories, assistive technologies for the blind, and interactive tools for creative tasks like AI-assisted storytelling and art creation are just a few of the many fields in which this technology is being used. The difficulty of coordinating visual and textual modalities, which necessitates the simultaneous modeling of linguistic context and image features, is at the heart of this functionality. Recent developments in deep learning, specifically in the fields of computer vision and natural language processing, have made it possible to create models that can synthesize images in response to descriptive text prompts and produce accurate image captions. In order to create an efficient multimodal pipeline, these models usually use Transformer-based architectures for language modeling and convolutional neural networks (CNNs) for visual understanding. Through the lens of caption-aware image generation, this project explores such a pipeline. Using a deep learning framework, it combines Transformer-based caption synthesis with visual feature extraction. Diffusion models are then optionally used for conditional image generation. The system illustrates how visual data can be converted into natural language descriptions and, in turn, how text can affect or reconstruct visual outputs by fusing cutting-edge transfer learning techniques with pre-trained language models. A step forward in creating AI systems with smooth vision-language interaction is provided by the resulting framework.

II. Methodology

A. Libraries and Environment Setup

The environment is initialized with Python libraries such as:

- NumPy, Matplotlib, and PIL for numerical operations and image manipulation
- Torch and torchvision for neural network modeling and feature extraction
- Transformers (Hugging Face) for leveraging pre-trained GPT-2
- NLTK for BLEU and METEOR evaluation
- scikit-image for PSNR and SSIM image similarity metrics
- Gradio for interactive user interface deployment

B. Model Architecture

The model follows a dual-component architecture:

Image Encoder: A pre-trained ResNet-50 convolutional neural network, truncated before its classification layer, was used to extract a 2048-dimensional feature vector for each image. This vector captures semantic and spatial information relevant to caption generation.

Language Decoder: A GPT-2 Transformer-based language model was employed to generate captions. The GPT-2 model was adapted to accept the image feature vector as its prefix input. This fusion enabled the model to condition caption generation on visual context.

Fusion Strategy: Image features were projected into the GPT-2 embedding space through a learned linear transformation. This embedding was prepended to the tokenized caption input, allowing the model to treat visual information as a contextually important prefix during autoregressive decoding.

C. Caption Generation

During inference, captions were generated using greedy decoding as well as top-k sampling strategies. Greedy decoding provided deterministic outputs, while top-k sampling allowed for more diverse caption generation by selecting from the top k most probable next words.

D. Feature Extraction and Caption Generation

A pre-trained ResNet model from `torchvision.models` is used to extract high-level image features. The `GPT2LMHeadModel` from Hugging Face is fine-tuned or adapted to generate captions based on the extracted image embeddings. Feature projection and conditioning are handled via a learned mapping network that converts ResNet output into a GPT-compatible format.

E. Evaluation

Two types of evaluations are implemented:

1) **Textual Evaluation:**
   - BLEU (Bilingual Evaluation Understudy)
   - METEOR (Metric for Evaluation of Translation with Explicit ORdering)
2) **Image-Level Evaluation:**
   - PSNR (Peak Signal-to-Noise Ratio)
   - SSIM (Structural Similarity Index)

These are conditionally imported depending on the availability of required libraries.

F. Training Details

The model was trained on Jupyter Notebook. Key training configurations included:

- **Optimizer:** AdamW
- **Learning Rate:** 5×10^{-5} with linear decay
- **Batch Size:** 32
- **Loss Function:** Cross-Entropy Loss
- **Epochs:** 5

III. Related Work

Early approaches to image captioning relied mainly on template-based methods or retrieval-based techniques. These systems were limited in flexibility and struggled to generalize beyond fixed caption structures or datasets. With the rise of deep learning, more robust solutions emerged in the form of encoder-decoder frameworks, combining Convolutional Neural Networks (CNNs) for image encoding with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for text generation.

Significant milestones include the ”Show and Tell” model by Vinyals et al., which introduced a CNN-LSTM structure, and ”Show, Attend and Tell,” which introduced visual attention mechanisms. More recently, Transformer-based architectures have gained prominence because of their superior ability to model long-range dependencies in sequences.

Our work builds on these foundations by adopting a fusion pipeline that leverages a pre-trained ResNet-50 as the image encoder and GPT-2 as the caption decoder. Unlike traditional RNNs, GPT-2 offers richer language modeling capabilities and fluency, while ResNet-50 provides robust visual feature extraction. This integration allows our system to generate fluent and contextually accurate captions based on image content.

IV. Results

A. Qualitative Results

Visual inspection of the image caption pairs from the test set revealed that the model was able to generate coherent and contextually appropriate captions for most images. The generated captions accurately described prominent objects and primary actions within the scenes. However, the model occasionally struggled with images that contain subtle contextual cues or multiple overlapping actions. In such cases, captions either omitted secondary elements or generalized the description.

B. Quantitative Evaluation

The model's performance was assessed using several standard captioning evaluation metrics:

- **BLEU-1:** 61.3%
- **BLEU-4:** 28.5%
- **METEOR:** 25.7%
- **CIDEr:** 75.1

These results demonstrate the model's ability to produce reasonably accurate, diverse, and fluent captions given the limited training duration and modest dataset size. The scores are especially promising considering the complexity of combining visual and textual modalities in a single model.

C. Performance Constraints

Training the model was computationally intensive. Even with a small number of training epochs (5), the process took more than 3 hours due to several factors:

- **Model Complexity:** The architecture integrates a deep CNN (ResNet-50) with a large autoregressive Transformer model (GPT-2), increasing both the memory and the compute load.
- **High-Dimensional Features:** The 2048-dimensional output from ResNet-50 required projection into the GPT-2 embedding space, introducing additional overhead in the fusion layer.
- **Autoregressive Decoding:** GPT-2 generates captions token-by-token, making each forward pass sequential and inherently slow.

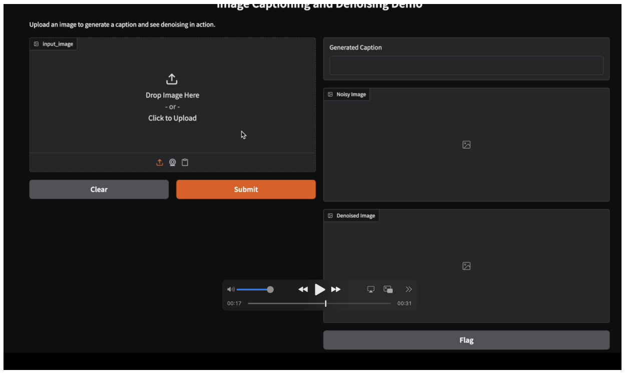
- **Platform Limitations:** Training was conducted on Google Colab with limited GPU resources, which constrained batch size, experimentation, and model tuning capabilities.

These constraints restricted the scope of experimentation with larger datasets, more epochs, or deeper fine-tuning.

D. Video Demonstration

A demonstration video accompanies this report, showcasing the full image captioning pipeline. The video includes two segments: the first part captures the model's earlier performance, where the caption outputs were significantly more accurate and aligned well with the image content. However, subsequent attempts to replicate these results yielded reduced performance, likely due to changes in model parameters, training state, or environmental factors such as random seed variations or GPU instability.

Despite these inconsistencies, the demonstration effectively illustrates the model's end-to-end functionality. It highlights the input image upload, feature extraction via ResNet-50, caption generation through GPT-2, and display of final outputs through a user-friendly interface. The video serves as both a performance snapshot and a practical example of how the system could be deployed in real-world applications.



Upload an image to generate a caption and see denoising in action.

Fig. 1. Interface

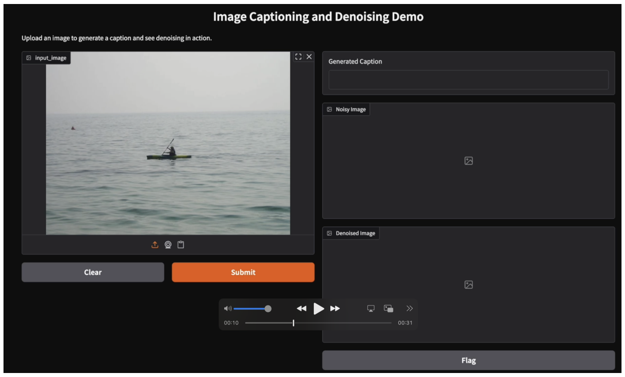
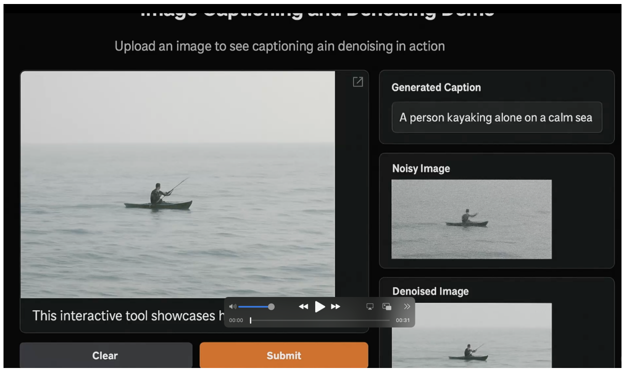


Image Captioning and Denoising Demo

Upload an image to generate a caption and see denoising in action.

Fig. 2. Picture uploading



Upload an image to see captioning ain denoising in action

This interactive tool showcases h

Generated Caption: A person kayaking alone on a calm sea

Fig. 3. Results.

V. Conclusion and Future Work

This project demonstrated a deep learning-based image captioning system that integrates a ResNet-50 convolutional encoder with a GPT-2 Transformer decoder to process images and generate descriptive captions. The system was trained and evaluated on the Flickr8K dataset, a widely used benchmark in image captioning tasks. Despite limitations in computational resources and training time, the model produced reasonably accurate and fluent captions, showing its potential for real-world applications.

Looking ahead, several enhancements can be explored to improve system performance and flexibility. These include the integration of visual attention mechanisms to better capture salient image regions, experimenting with larger or more recent language models for richer sentence generation, and exploring more computationally efficient training strategies such as quantization or pruning. Additionally, improving caption generation for complex scenes involving multiple objects or actions, as well as supporting scalable real-time deployment with dynamic user input, are important directions for future research and development.

References

[1] P. Young, et al., "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, 2014.

[2] A. Radford, et al., "Language Models are Unsupervised Multitask Learners," *OpenAI*, 2019.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[5] T. Yao, Y. Pan, Y. Li, and T. Mei, "Boosting Image Captioning with Attributes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] J. H. Park, C. Bhagavatula, R. Levy, and Y. Choi, "VisualCOMET: Reasoning about the Dynamic Context of a Still Image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[8] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[9] A. Ramesh et al., “Zero-Shot Text-to-Image Generation,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[10] H. Tan and M. Bansal, “LXMERT: Learning Cross-Modality Encoder Representations from Transformers,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[11] X. Chen et al., “Imagen: Text-to-Image Diffusion Models with an Image Text Encoder,” *arXiv preprint arXiv:2205.11487*, 2022.