**ABSTRACT**

**Image Captioning With Pre-Trained Models:**

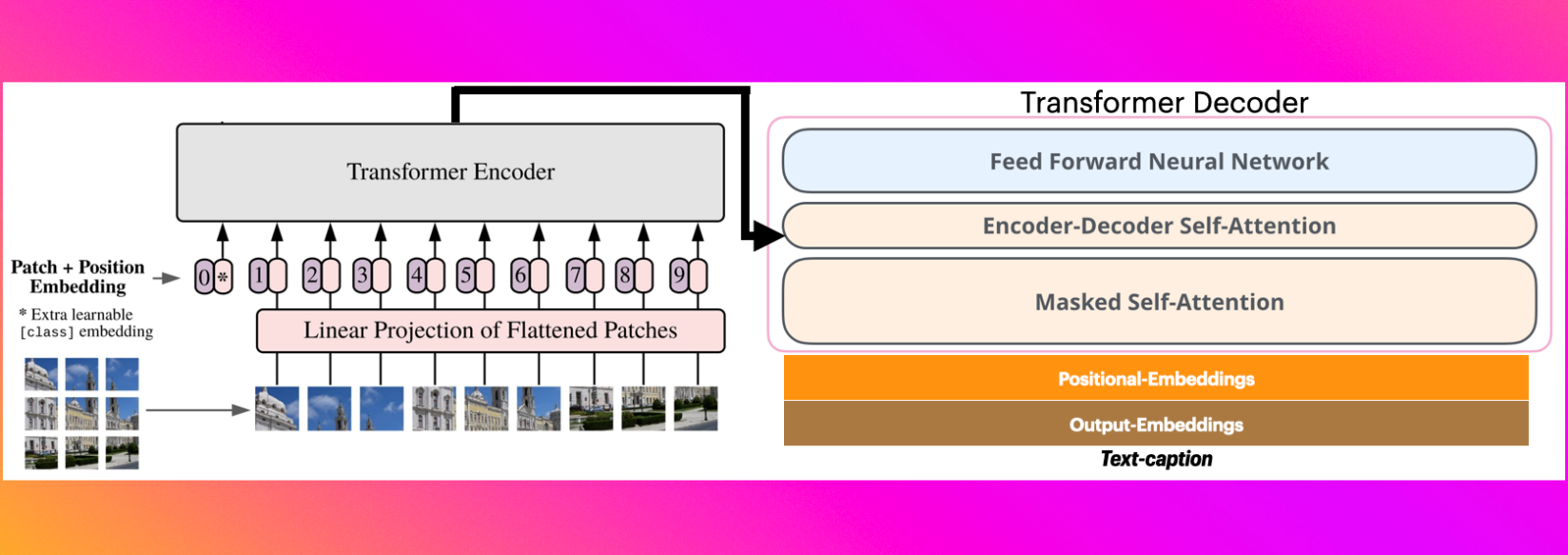


Image captioning, the task of automatically generating textual descriptions for images, presents a significant challenge at the intersection of computer vision and natural language processing. This project explores the use of pre-trained deep learning models to enhance the accuracy and quality of image captions, addressing the primary problem of generating contextually relevant and precise descriptions for a wide variety of images. Accurate image captioning is crucial for applications such as assistive technologies, image indexing, and social media automation.

This project focuses on developing an image captioning system utilizing the pre-trained model "nlpconnect/vit-gpt2-image-captioning." This model integrates the Vision Transformer (ViT) for image encoding and GPT-2 for caption generation through causal language modeling. The ViT component effectively captures the visual features of input images, while the GPT-2 decoder translates these features into coherent and contextually relevant captions. By leveraging the strengths of both components, this approach aims to achieve accurate and meaningful descriptions of visual content.

The ViT model is designed to handle the complexity of image data by transforming input images into a sequence of embedded patches. This transformation enables the model to process images in a manner akin to how transformers process sequences of text, allowing it to capture intricate visual details and contextual information. The embedded patches are then fed into the GPT-2 model, which is fine-tuned for the task of generating natural language captions.

The integration of ViT and GPT-2 allows the model to benefit from the advanced capabilities of both vision and language processing. ViT's robust image representation ensures that the visual features are captured with high fidelity, while GPT-2's language generation capabilities ensure that the captions are grammatically correct, contextually relevant, and coherent. This combination addresses the limitations of previous image captioning models, which often struggled with either visual understanding or language generation.

In this project, we conduct extensive experiments to evaluate the performance of the "nlpconnect/vit-gpt2-image-captioning" model. The evaluation includes quantitative metrics such as BLEU, METEOR, and CIDEr scores, as well as qualitative assessments to ensure the captions generated are both accurate and meaningful. The results demonstrate the model's ability to generate high-quality captions across a diverse range of images, highlighting its potential for real-world applications.

Overall, this project showcases the potential of combining advanced vision and language models in the task of image captioning. By leveraging the pre-trained "nlpconnect/vit-gpt2-image-captioning" model, we demonstrate significant improvements in the quality and accuracy of generated captions. This approach paves the way for future advancements in assistive technologies, image indexing, and social media automation, where precise and contextually relevant image descriptions are paramount.