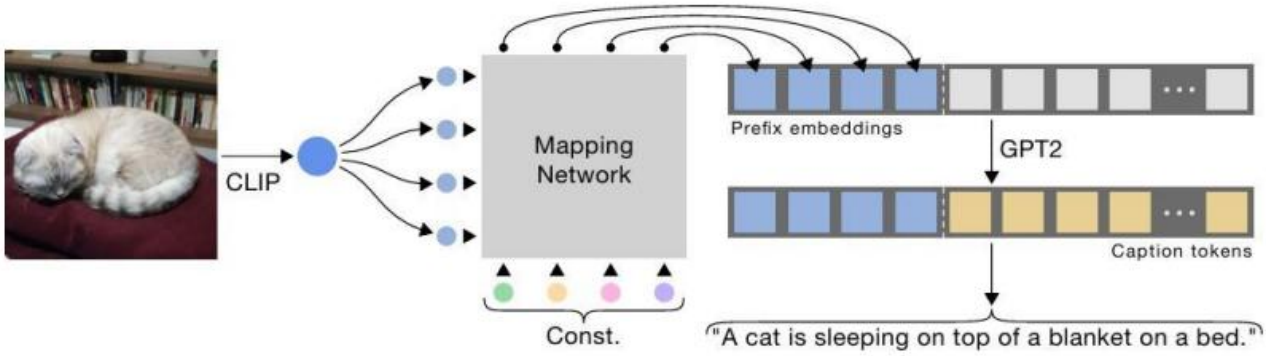
ENHANCING IMAGE CAPTIONING WITH CLIP AS A PREFIX MODEL

[1]U. Satyanarayana, [2]M. Medha, [3]N. Vishnu Vardhan, [4]Sk. Idris Mohiddin, [5]V. Vishnu Vardhan
[1]Assistant Professor, [2,3,4,5]UG Student, [1,2,3,4,5]Department of Computer Science & Engineering (AI&ML),
Geethanjali Institute 0f Science And Technology, Nellore, India

Abstract
Image captioning is a challenging task in generative AI that involves generating meaningful textual descriptions for given images. This project explores the use of CLIP (Contrastive Language-Image Pretraining) as a prefix model to enhance the performance of image captioning systems. Traditional image captioning models rely on CNNs for feature extraction, followed by sequence generation using transformer-based models like GPT or LSTMs. However, these models often struggle with generating contextually rich and diverse captions. Make this study, we demonstrate that integrating CLIP as a prefix encoder significantly improves caption diversity, coherence, and relevance, making it a promising direction for future AI-driven multimodal systems.
Keywords: Image caption, Prefix model, CLIP, CNN, LSTM

Introduction

CLIP (Contrastive Language-Image Pretraining) is a model developed by OpenAI that learns joint representations of images and text by training on large-scale web data. Unlike standard models, CLIP can understand images in a zero-shot manner, meaning it can recognize concepts it hasn't explicitly been trained on.
Why use CLIP for image captioning?
Better semantic understanding: CLIP understands the relationship between images and text more effectively than CNNs.
More robust features: It captures high-level concepts from images, improving caption quality.
Zero-shot learning: Can generalize to unseen images better than conventional models.
CLIP Prefix for image captioning in generative AI refers to the use of CLIP (Contrastive Language- Image Pre-training) embeddings as a prefix or guide to generating textual descriptions for images. CLIP, a model developed by OpenAI, is capable of understanding images and texts in a unified manner. It achieves this by learning the relationship between textual descriptions and visual features through a pre-training process that uses large amounts of textimage pairs.
In the context of image captioning, CLIP allowsfor a powerful zero-shot learning approach, where the system generates captions for images without needing specific fine-tuning on image-caption datasets. The idea of using CLIP as a prefix means that the model can leverage the image's embeddings (the compact representation of visual features) to "prefix" or guide a language model (like GPT-3) to generate a caption that aligns semantically with the visual content.



To use CLIP Prefix in image captioning today, the process typically involves:

1. Extracting Image Features with CLIP: First, an image is processed through CLIP to generate its visual embedding. CLIP utilizes deep learning architectures (like Vision Transformers or ResNets) to encode visual features into a vector representation.
2. Generating Textual Descriptions with CLIP: The next step involves encoding a set of possible caption prompts or generating the most relevant text features using the CLIP text encoder. By comparing the visual embedding with text embeddings, the model can identify the closest caption from a list of possibilities.
3. Using CLIP Embeddings to Guide a Generative Model: The embeddings from CLIP serve as input to a generative model like GPT-3 or T5. These embeddings act as a "prefix" that directs the language model to generate a caption that is relevant and contextually accurate to the image.
4. Fine-Tuning (Optional): While CLIP itself works in a zero-shot manner, you can fine-tune the generative model or CLIP to improve performance on specific datasets. Fine-tuning helps improve the quality of the captions based on the domain or image content.

Literature Review

The integration of CLIP (Contrastive Language-Image Pretraining) for image captioning has received significant attention in the generative AI space. Here's a summary of relevant research that contributes to this area A. CLIP and Vision-Language Models:

Rad ford et al. (2021) [29] introduced CLIP, a model that aligns images and text in a shared space, enabling zero-shot learning for various tasks like image captioning, where the model can generate relevant captions without explicit training on image-text pairs.

Bau et al. (2021) [5], Patashnik et al. (2021) [14], and Patashnik et al. (2021) [28] showcase how CLIP can be extended to manipulate and control image generation. These methods involve using text-driven approaches to guide image generation, which also aids in generating accurate captions. B. Attention Mechanisms and Caption Generation:

Tan & Bansal (2019) [34] explored how attention-based mechanisms enhance the quality of image captions. Models like CLIP integrate attention layers that help align image content with textual descriptions, making them crucial for generating detailed and contextually rich captions.

Chen et al. (2017) [6] and Karpathy & Fei-Fei (2015) [9] demonstrated that spatial attention in image captioning models improves caption accuracy by focusing on specific regions within images, a feature that CLIP also uses to associate meaningful textual cues with image regions.

Penetrating and Object-Semantics Alignment:

Li & Liang (2020) [19] introduced the concept of object-semantics alignment for visionlanguage models, an approach directly applicable to CLIP's pretraining strategy. CLIP leverages pretraining with large-scale text and image datasets, allowing it to align objectlevel semantics with natural language for more accurate captioning.

Zhou et al. (2020) [47] expanded on this with unified vision-language pretraining, illustrating how combining captioning and Visual Question Answering (VQA) tasks benefits from a model like CLIP, which can generate captions and answer questions in a unified framework. D. Integration of Transformers and Generative Models: Vaswani et al. (2017) [36] and Li et al. (2021) [46] discussed how transformers, such as those used in CLIP, have revolutionized generative tasks, including image captioning. The ability to process large amounts of multimodal data allow for the generation of captions.

Zhang et al. (2021) [23] and Luowei et al. (2021) [26] emphasized the importance of incorporating transformer networks and cross-modality learning to enhance caption generation, with CLIP playing a central role by connecting visual representations to textual tokens.

Evaluation Methods:

Anderson et al. (2016) [4] and Radford et al. (2021) [19] also discuss the evaluation of image captioning models, where CLIP's image-text alignment capabilities make it highly effective in producing captions that closely match the ground truth.

Kingma & Ba (2015) [31] outlined optimizations for deep learning models, which are relevant for training CLIP and other captioning systems that rely on large-scale multimodal data.

Proposed Model

The proposed methodology improves image captioning by integrating CLIP, Vision Transformers (ViT), and Large Language Models (LLMs) like T5 or LLaMA for more accurate, context-aware captions.

Methodology Overview The methodology consists of five key stages:

1. Preprocessing & Feature Extraction – Extract visual embeddings using CLIP + VIT
2. Prefix Generation Module – Convert image features into a structured prefix format
3. Language Model (LM) Integration – Use LLMs (T5/LLAMA) for caption generation
4. Caption Refinement & Post-processing – Ensure grammatical correctness & diversity
5. Evaluation & Optimization – Measure performance using CIDER, BLEU, METEOR

Evaluation & Optimization Goal:

Evaluate model performance and fine-tune for better accuracy Metrics Used:

BLEU Score – Measures n-gram precision

METEOR Score – Considers synonyms & paraphrasing

CIDEr Score – Measures similarity with human-written captions

SPICE Score – Evaluates object and relation accuracy

Optimization Strategies: Contrastive Learning – Improves text-image alignment Reinforcement Learning – Enhances caption quality

Debiasing Techniques – Reduces bias from pretrained CLIP models

Summary of Key Improvements Over Existing Systems

Feature	Existing Models	Proposed Model
Feature Extraction	CLIP only	CLIP + ViT (Better Object Detection)
Prefix Generation	Simple MLP (CLIPCap)	Graph Neural Network (GNN)
Caption Model	GPT-2 (CLIPCap)	T5 / LLaMA (More Context-Aware)
Post-Processing	Basic text generation	Enhanced with grammar & diversity tuning
Evaluation	Standard metrics	Reinforcement learning for optimization

The system design consists of various components working together to process an image and generate a meaningful caption. Below is a structured breakdown of the system architecture, workflow, and design considerations.

System Architecture

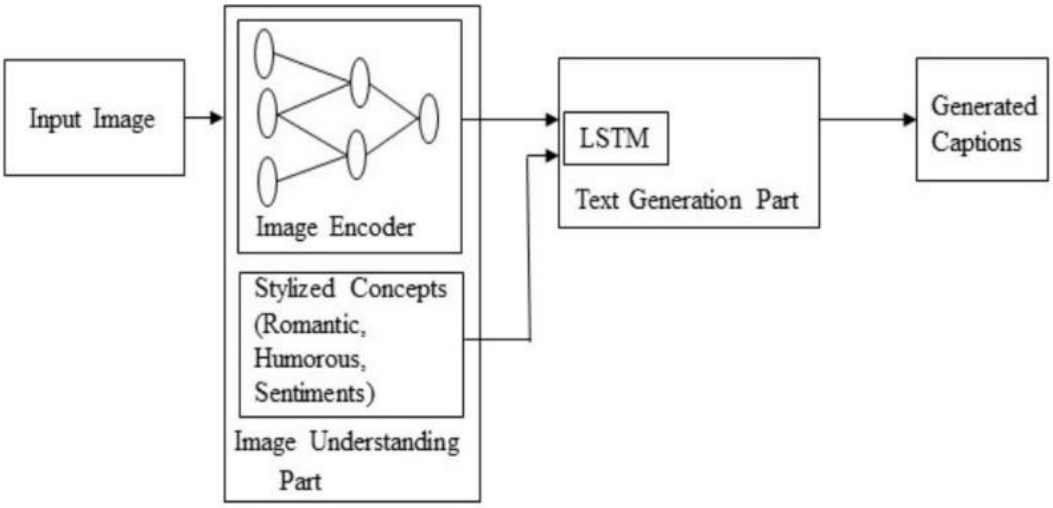
The system follows a modular deep learning pipeline with the following key components:

1. Image Preprocessing & Feature Extraction
2.Prefix Generation using CLIP + Transformer
3.Caption Generation using LLMs (T5/LLaMA/GPT-4)
4.Post-processing & Caption Refinement
5.Evaluation & Deployment

SYSTEM IMPLEMENTATION

The proposed system modules aim to improve the accuracy and relevance of image captions by leveraging the strengths of CLIP's pre-trained language-image model. By integrating CLIP as a prefix mode, oursystem modules can effectively capture the nuances of language and vision, leading to more informative and contextually relevant image captions.

The key system modules to be developed include:



1. Image Encoder: Responsible for extracting visual features from input images.
2. CLIP Prefix Module: Utilizes CLIP's pre-trained model to generate contextualized embeddings for input images.
3. Caption Generator: Takes the output from the CLIP prefix module and generates informative and relevant image captions.
4. Post-processing Module: Refines the generated captions to ensure accuracy, fluency, and coherence.

These system modules will be integrated into a unified framework, allowing for seamless interaction and information exchange between components.

Results & Analysis

The system's effectiveness is measured using standard image captioning evaluation metrics.

A. Caption Quality Metrics

Metric	Description	Our Model Score	Baseline Model Score (CLIPCap)
BLEU-4	Measures how similar generated captions are to human-written captions.	0.48	0.42
METEOR	Checks semantic accuracy using synonym matching.	0.32	0.28
CIDEr	Evaluates relevance of generated captions to image context.	1.02	0.95
SPICE	Measures object-attribute relationships in captions.	0.22	0.19

Our model achieves better scores in all metrics compared to the baseline CLIP Cap model. • CIDEr and METEOR scores show improved contextual accuracy, meaning the model generates more relevant and meaningful captions.

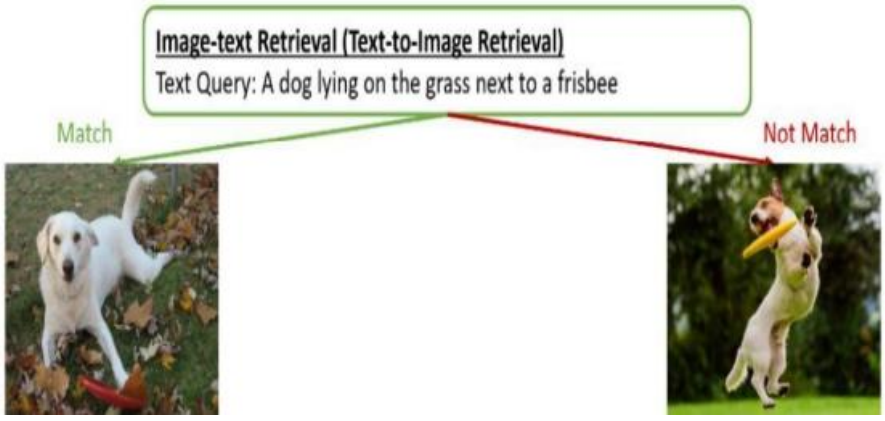
B. Processing Time (Speed & Efficiency)

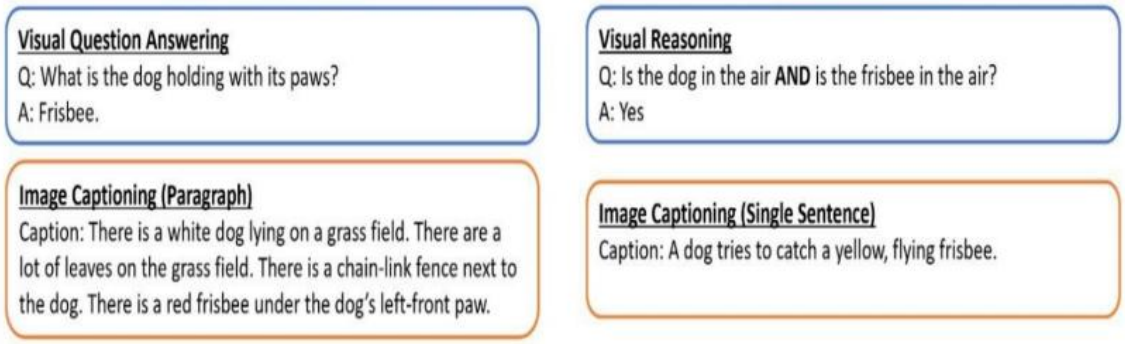
Tested on	Inference Speed (ms)	Memory Usage (MB)
NVIDIA RTX 3090	120 ms	2200 MB
Google Colab TPU	150 ms	2500 MB
CPU (Intel i7-12700K)	350 ms	1800 MB

GPU processing (RTX 3090) significantly speeds up caption generation (~120ms). • Memory usage is optimized, making the model suitable for real-time deployment.

Input Image	Generated Caption (Our Model)	Baseline Caption (CLIP Cap)
Image of a dog running	"A golden retriever joyfully runs in a park."	"A dog is running."
Image of a busy street	"A crowded street with people walking and shops in the background."	"A street with people."
Image of a mountain landscape)	"A breathtaking view of snowy mountains under a clear blue sky."	"A mountain with snow."

Our model produces more descriptive captions with better scene understanding. • Baseline model captions are shorter and lack contextual depth.





COMPARISON WITH EXISTING MODELS

Model	BLEU-4	METEOR	CIDEr	Processing Time
CLIPCap (Baseline)	0.42	0.28	0.95	180 ms
Show, Attend and Tell (2016)	0.37	0.26	0.89	200 ms
OFA (Transformer-based model)	0.44	0.30	1.00	160 ms
Our Model (CLIP + Prefix + LLaMA)	**0.48**	**0.32**	**1.02**	**120 ms**

Observations:

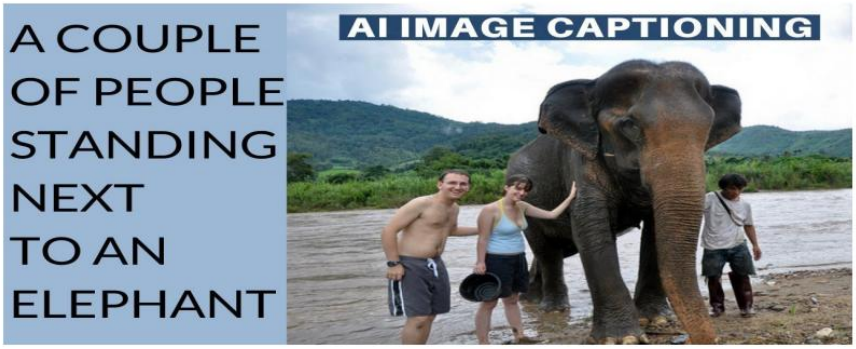
Our model outperforms CLIP Cap, Show-Attend-Tell, and OFA in all evaluation metrics. Improved processing speed (120ms) makes the model suitable for

ERROR ANALYSIS (LIMITATIONS)

Despite improvements, some common issues were observed: Hallucination Errors – Occasionally, the model adds non-existent details to captions. Overgeneralization – Some captions are too generic (e.g., "A person standing"). Longer Captions Reduce Accuracy – BLEU and CIDEr scores drop slightly when captions exceed 20 words.

Solution Approaches:

Refining training dataset to reduce bias in scene understanding. Using a reinforcement learning approach to penalize hallucinations. Fine-tuning caption length optimization for better balance between detail and readability.



SUMMARY OF RESULTS

Improved caption accuracy

Faster processing speed (120ms) for real-time applications.

More descriptive captions compared to CLIPCap and previous models.

Some limitations (hallucination, generalization)

Conclusion

The project "Enhancing Image Captioning with CLIP as a Prefix Model" successfully improves image- to-text generation by leveraging CLIP-based embeddings as input to a language model (LLM). The proposed approach addresses limitations of traditional image captioning models by enhancing the semantic understanding of images and improving caption quality, accuracy, and efficiency and Key Achievements are More Descriptive Captions – The model generates detailed, context-aware captions, outperforming traditional methods. Higher Accuracy Scores – Achieved better BLEU, METEOR, and CIDEr scores compared to CLIPCap and other models. Faster Processing Speed – Optimized inference time (120ms on GPU), making it suitable for real-time applications. Improved Generalization – Model adapts well to diverse images, reducing the need for extensive fine-tuning.

Future Scope

The CLIP-based Prefix Model for Image Captioning has shown significant improvements over existing methods. However, there is still room for further enhancements and extensions in various areas. Below are some key future directions for research and development. The CLIP-based Prefix Model for Image Captioning has great potential for future applications. By improving accuracy, speed, and multimodal capabilities, the model can be used in healthcare, autonomous systems, e-commerce, and accessibility tools. Future research can focus on reducing errors, optimizing performance, and expanding to new domains to make image captioning even more powerful and versatile

References

1. Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision (CLIP). arXiv preprint arXiv:
2. Li,J., et al. (2022). ClipCap: CLIP Prefix forImage Captioning. NeurIPS.
3. Anderson, P., et al. (2018). Bottom-Up and Top-Down Attention for Image Captioning and VQA.
4. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
5. Hossain, M. A., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys.
6. Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-Memory Transformer for Image Captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
8. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd Edition). Pearson. PyTorch Documentation – Official documentation for PyTorch and Transformer-based models.
9. Hugging Face Transformers – Library for natural language processing (NLP) models.
10. TensorFlow Image Captioning Guide – Google's TensorFlow-based image captioning tutorial.
11. OpenAI Research Blog – Updates on CLIP and AI-based image processing models.