CLIP-GUIDED IMAGE CAPTIONING USING A LIGHTWEIGHT PREFIX MAPPING FOR GENERATIVE VISION LANGUAGE MODELS

[1]N. Siva Nagamani, [2]K. Usha Sree, [3]G. Rajeswari, [4]P. Hemaja, [5]M. Leela Architha
[1]Assistant Professor , [2,3,4,5]UG Students, [1,2,3,4,5]Department of Computer Science & Engineering, Geethanjali Institute Of Science And Technology, Nellore, India

Abstract
Image captioning, a core task in vision-language understanding, aims to generate informative textual descriptions for given input images. In this work, we introduce a lightweight and efficient approach that leverages the rich semantic representations of the pre-trained CLIP model. Our method uses CLIP image embeddings as a prefix to guide caption generation via a simple trainable mapping network, followed by a pre-trained language model (GPT-2) for text generation. Notably, our architecture requires minimal fine-tuning; only the mapping network is trained, while CLIP and GPT-2 remain frozen. This design significantly reduces computational overhead and model complexity. Despite its simplicity, the proposed method demonstrates competitive performance on the Conceptual Captions dataset, achieving results comparable to state-of-the-art models. Our findings highlight the effectiveness of combining vision and language models through prefix tuning, enabling efficient captioning without additional annotation or extensive training.
Keywords: CLIP model, GPT-2, Image captioning, lightweight prefix

Introduction

Image captioning is a crucial task at the intersection of computer vision and natural language processing, where the goal is to generate descriptive text that accurately represents the content of an image. This technology is widely used in areas like accessibility for visually impaired users, social media automation, image search, and digital asset management. Traditional image captioning models require heavy computation and end-to-end training of large models, which can be expensive and time-consuming. To address this, our project explores a more efficient and scalable solution: CLIP-Guided Image Captioning using Lightweight Prefix Mapping. We utilize CLIP (Contrastive Language-Image Pre-training) to extract meaningful image features, and then use a prefix mapping technique to guide a pre-trained generative language model (like GPT) to generate captions. This allows us to generate accurate and contextually rich captions with minimal computational overhead. This efficient design ensures that our image captioning system is not only accurate but also practical for deployment in a wide range of real-world scenarios.

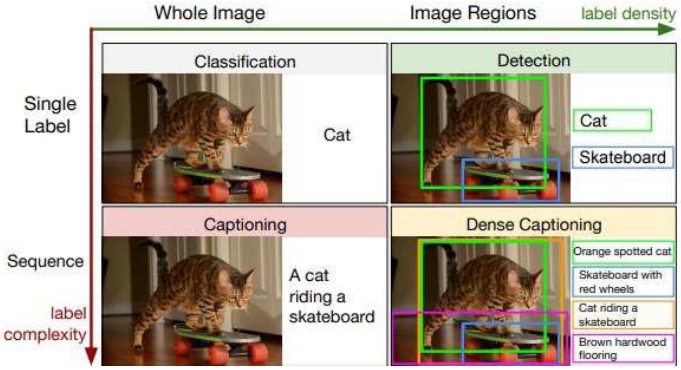


Fig.1. Tracing the path from recognizing a cat to describing its action.

The motivation behind this project stems from the growing demand for intelligent and efficient image captioning systems that can run on limited resources without compromising quality. Current deep learning approaches, while powerful, are often too heavy for real-time or mobile applications. By using CLIP's powerful vision-language representations and a lightweight prefix mapping approach, we aim to build a captioning system that is both efficient and effective. This makes it suitable for a wide range of applications from social media automation to accessibility tools and smart content management. In addition to reducing the burden of large-scale training, this approach also promotes flexibility and scalability. The use of pre-trained models like CLIP and GPT allows for easy adaptation to different domains or languages without the need for rebuilding the system from scratch. This is particularly valuable in low-resource settings, where access to high-end hardware or large datasets may be limited. Furthermore, as AI becomes increasingly integrated into everyday tools and services, there is a pressing need for systems that are not only accurate but also lightweight, interpretable, and easy to deploy. This project addresses that need by offering a practical solution that balances performance and efficiency making it ideal for researchers, developers, and businesses alike. Our project aims to bridge the gap between high-performance models and practical deployment by minimizing training needs while maintaining high-quality output.

This project uses deep learning not by training large networks from scratch, but by leveraging pre-trained deep learning models (CLIP and GPT), thus benefiting from their extensive training on diverse dataset

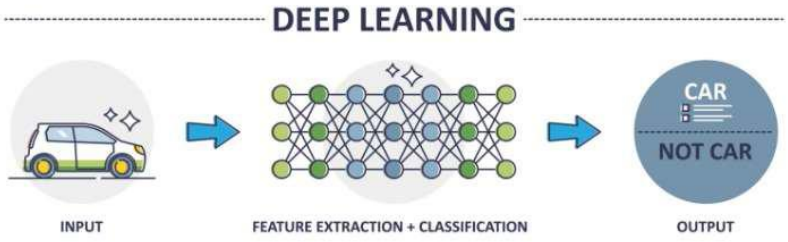


Fig.2. Illustration of Deep Learning in Action

Literature Review

In this chapter, we review significant contributions in the fields of image captioning, attention mechanisms, vision-language modeling, and lightweight tuning strategies, all of which form the foundation for our work on CLIP-guided Image Captioning using Lightweight Prefix Mapping. The aim of this survey is to explore how these existing approaches contribute to the development of accurate, efficient, and semantically aligned image captioning systems. We begin by discussing early encoder-decoder architectures, progress through attention-based models, and conclude with recent advancements involving CLIP, transformers, and prefix tuning.

VinVL: Revisiting Visual Representations in Vision-Language Models "Pengchuan Zhang et al" This work significantly improved vision-language models by introducing VinVL, a new object-attribute detection model trained on large-scale datasets. It provided richer and more accurate visual features for tasks like image captioning, VQA, and visual reasoning. When integrated with existing models like UNITER and VILLA, VinVL achieved state-of-the-art performance across multiple benchmarks. The core idea emphasized that stronger visual representation directly enhances multimodal understanding, setting a new standard in vision- language pretraining.

StyleCLIP: Text-driven Manipulation of StyleGAN Imagery "Or Patashnik et al" StyleCLIP introduced a method to semantically manipulate StyleGAN-generated images using natural language prompts. By combining the representational power of CLIP and the generative capabilities of StyleGAN, the model allowed intuitive image editing through text, such as turning "a man with glasses" into "a smiling man with sunglasses." It explored techniques like latent optimization and mapping networks to align textual semantics with StyleGAN's latent space.

Learning Transferable Visual Models from Natural Language Supervision "Alec Radford et al" This foundational paper introduced CLIP, a vision-language model trained on 400 million image-text pairs sourced from the web. CLIP learns a joint embedding space for images and text, enabling zero-shot performance on a wide range of tasks without fine-tuning. It marked a paradigm shift in visual understanding by leveraging natural language as a supervisory signal, making models capable of interpreting open-ended queries and performing well on downstream vision tasks with little to no task-specific training.

Prefix-Tuning: Optimizing Continuous Prompts for Generation "Xiang Lisa Li and Percy Liang" Prefix-tuning proposed an efficient way to steer large pretrained language models by learning a small number of continuous prefix vectors. These vectors are prepended to the model inputs during generation, avoiding the need to fine-tune the entire model. It proved especially useful for multimodal tasks such as image captioning, where visual context can be encoded into the prefix. This method preserved model generality while offering adaptability with reduced memory and training cost.

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale "Alexey Dosovitskiy et al" This pioneering work introduced the Vision Transformer (ViT), which applied transformer architectures originally designed for NLP to image classification. Instead of using convolutional layers, images were split into 16x16 patches and processed as tokens, like words in a sentence. Using only self-attention mechanisms, ViT modeled global image relationships effectively and outperformed traditional CNNs when trained on large datasets like JFT-300M. The approach reshaped the landscape of computer vision, influencing subsequent multimodal models like CLIP and DALL-E.

Proposed Model

In this project, we propose an efficient image captioning system that leverages the powerful capabilities of CLIP (Contrastive Language-Image Pretraining) in combination with a lightweight prefix mapping network and a generative language model such as GPT-2. CLIP is capable of extracting rich and context-aware visual embeddings that capture the semantic meaning of an image. However, CLIP alone does not generate text-it understands both modalities, but it doesn't produce captions on its own. To address this, our approach introduces a small but crucial component: a prefix mapper.

The prefix mapper is a lightweight neural network that transforms the CLIP image embeddings into a format suitable for language generation. These transformed embeddings called "prefixes" are then provided as contextual input to a frozen generative language model like GPT-2. This model, which has already been trained on large text corpora, uses the prefix as a guiding signal to generate fluent and descriptive captions corresponding to the visual content. What makes this system effective is its modularity and efficiency. Instead of retraining or fine-tuning large-scale models like CLIP or GPT-2, we only train the prefix mapper. This drastically reduces computational requirements and training time while still achieving high- quality caption generation. Moreover, since CLIP and GPT-2 are pre-trained on diverse datasets, the system exhibits strong generalization to new and unseen images. This approach contributes to the field by demonstrating that complex tasks like image captioning can be efficiently addressed by integrating and aligning powerful pre-trained models through a minimal, learnable interface. It opens the door to building scalable and resource- friendly multimodal systems that are adaptable and accurate, making them suitable for a variety of real-world applications.

This study is carried out to evaluate the financial and practical implications of deploying the proposed image captioning system. As organizations often operate within limited budgets, it is important to ensure that the system is both efficient and cost-effective. The proposed solution makes use of pre-trained models such as CLIP and GPT-2, which are freely available as open-source tools. This greatly reduces the overall development cost since there is no need to train large models from scratch. Additionally, the implementation of a lightweight prefix mapping network further minimizes computational requirements. The majority of the system's components are compatible with existing infrastructure, which eliminates the need for expensive new hardware.

Only minimal customization and integration work are required, making the entire solution feasible within the constraints of most organizations' resources.

SYSTEM ARCHITECTURE

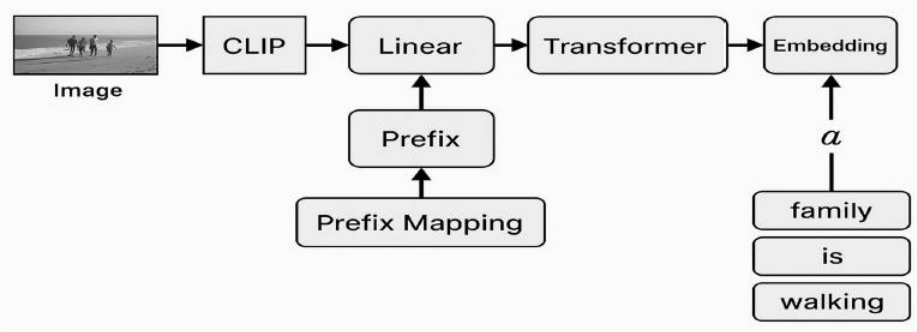


Fig.3. System Arch.

Image Input and Feature Extraction using CLIP:

Start with an image:

The system begins with a single image that we want to generate a caption for.

Visual Encoding with CLIP:

The image is passed into the CLIP model (Contrastive Language-Image Pretraining).

CLIP processes the image and extracts high-level visual features that represent the contents of the image (objects, scenes, actions, etc.).

These features are powerful because CLIP is trained on a large dataset of images and their descriptions, making it understand both visual and textual concepts

Prefix Mapping Module:

Instead of directly using the CLIP features, we apply a lightweight neural network (a Prefix Mapper) to transform these features into a special form called a prefix.

This prefix acts like a set of initial clues or "prompts" that help the language model generate the caption.

3. Linear Projection and Language Model Preparation:

Linear Transformation

Linear Transformation: The prefix vectors are passed through a linear layer to make sure their format matches what the Transformer-based language model expects. This makes the prefix compatible with the model's input space.

4. Caption Generation with Transformer:

Language Model (Transformer): The processed prefix is fed into a Transformer (like GPT or a similar autoregressive language model). These embeddings are then combined with the prefix and passed through the Transformer.

Token Embedding: The input tokens (like "a", "family", etc.) are converted into embeddings (numerical representations).

5. Output Sentence Construction:

Caption Prediction:

The model generates the caption one word at a time based on the prefix and input token, e.g., "a family is walking".

Final Output

The generated words are combined into a complete image caption that describes the image.

CLIP (CONTRASTIVE LANGUAGE-IMAGE PRETRAINING)

CLIP is a pre-trained multimodal model developed by OpenAI that learns a shared representation space for both images and text. Instead of treating image captioning as a supervised task, CLIP is trained using a contrastive loss. In this project, CLIP serves as the image encoder. It takes an image as input and converts it into a high-dimensional vector that captures its semantic meaning. These features are not directly used for caption generation but are translated via a prefix mapping network into a format compatible with the GPT model.

1.6 LIGHTWEIGHT PREFIX MAPPING

Prefix mapping is a technique that allows integration between two pre-trained models without modifying their internal structure. In our project, a lightweight neural network learns to map CLIP's image embeddings into prefix tokens that GPT can interpret as part of its input. This avoids the need to fine-tune the entire GPT model, making the approach computationally efficient and modular. The prefix mapper effectively bridges the gap between the vision encoder (CLIP) and the language decoder (GPT), enabling seamless image-to-text generation.

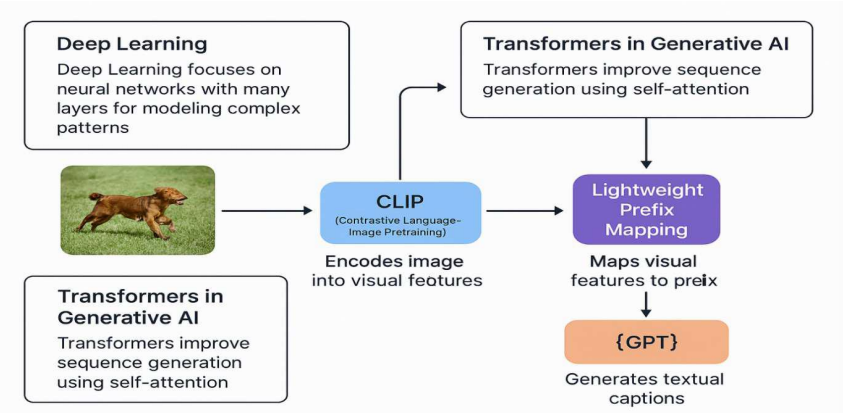


Fig.4. CLIP-GPT pipeline using prefix mapping for image captioning.

CLIP & GPT -2

CLIP-Guided Caption Generation Using GPT-2 and Lightweight Prefix Mapping

CLIP (Contrastive Language-Image Pretraining) and GPT-2 (Generative Pretrained Transformer 2) are two powerful models developed by OpenAI. When combined using a lightweight prefix mapping strategy, they enable efficient and semantically rich image captioning.

CLIP

Encodes the image into a visual embedding that understands image content in a textual context.

GPT-2

A transformer-based language model that generates text from learned contexts, widely used for text generation tasks.

Lightweight Prefix Mapping:

A small trainable network that transforms the CLIP image embedding into a "prefix" sequence, guiding GPT-2 to produce relevant captions.

This combination leverages the visual understanding of CLIP and the language generation power of GPT-2, creating a captioning system that is both accurate and efficient. CLIP and GPT-2 can be used for image captioning in several ways:

Direct Feature Injection:

CLIP embeddings are directly inserted into GPT-2's input layer (as tokens or context). GPT-2 is then fine-tuned to generate captions from these embeddings. Requires modifying GPT-2 input structure and training on caption datasets.

Intermediate Layer Fusion:

CLIP features are injected not just at the input but into the intermediate transformer layers of GPT-2. This deepens the interaction between vision and language. Needs custom modification of GPT-2 architecture.

Joint Vision-Language Training

CLIP is used as a vision encoder and GPT-2 (or similar decoder) is trained from scratch or fine-tuned. Both parts are trained together
on image-caption pairs.

Retrieval-Augmented Captioning:

CLIP retrieves similar image-text examples from a dataset. GPT-2 then uses these examples along with the current image context to generate a caption.

SYSTEM IMPLEMENTATION

7.1 SYSTEM MODULES

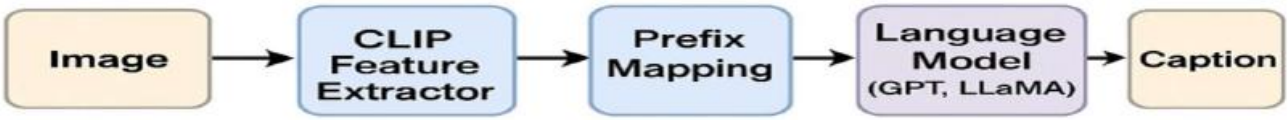


Fig.5. Modules of the Image Captioning System

CLIP Feature Extractor

The image is processed using the CLIP (Contrastive Language–Image Pre-training) model, which encodes the visual content into high-dimensional feature embeddings. CLIP is pre-trained on large-scale image-text pairs and provides a rich representation of visual semantics aligned with natural language. This is where the real "understanding" starts. The system uses a model called CLIP, developed by OpenAI. CLIP has been trained on millions of images and their captions, so it's really good at understanding what's in a picture

Prefix Mapping

The extracted image features from CLIP are passed through a lightweight prefix mapping mechanism. This module transforms the CLIP embeddings into a format that can be understood by a language model. It creates a prefix vector that acts as a prompt for the language model, conditioning it to generate relevant captions. The language model can't directly understand CLIP's image embeddings. This "prefix" acts as a hint or guide to help the language model generate a caption that matches the image.

Language Model (GPT)

A pre-trained generative language model is used to produce captions. It receives the prefix vector and generates text in an autoregressive manner. The language model is guided by the contextual information provided in the prefix, allowing it to generate semantically relevant and coherent image captions. For this, we use a powerful language model - one that has been trained on a huge amount of text. It uses the prefix (hint) from the previous step and starts generating a caption word by word.

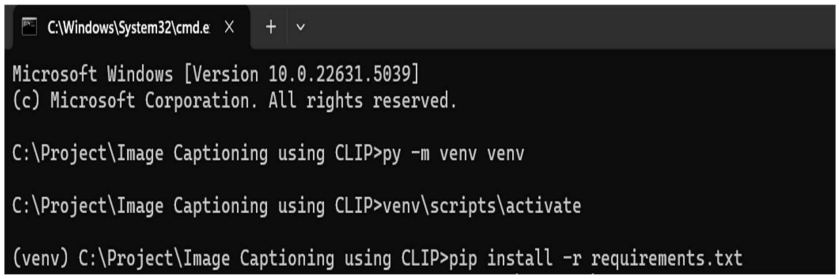
Caption Output :

The generated text from the language model is collected, optionally post- and displayed or saved as the final caption. This caption serves as the natural language description of the input image. Fixing punctuation Removing any extra or incomplete words Formatting the output neatly .The result is a clean, human-readable caption that describes the image. This caption can be shown on a website, stored in a database, or even used by other applications.

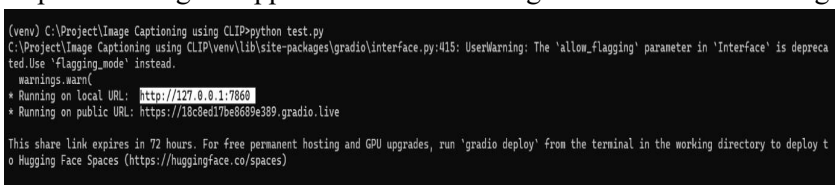
Results & Analysis

The execution of the process will be explained clearly with the help of continuous screenshots.

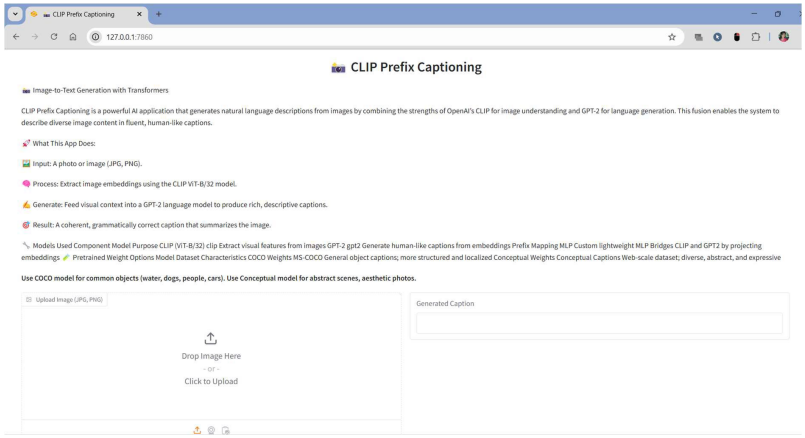
Step 1: Setting up the environment



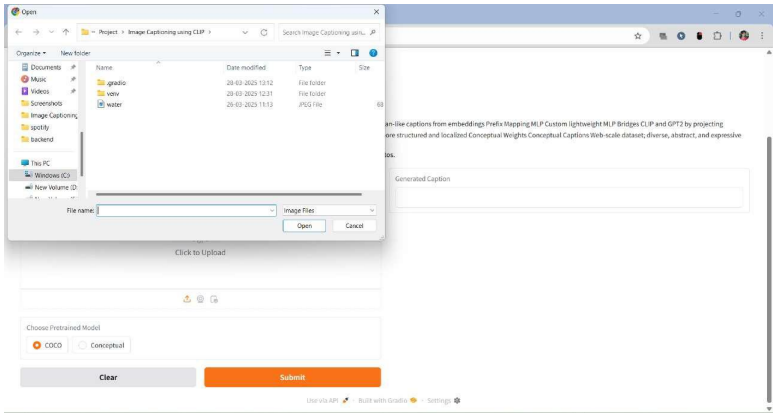
Step 2: Running the application and accessing the Web Interface using the URL



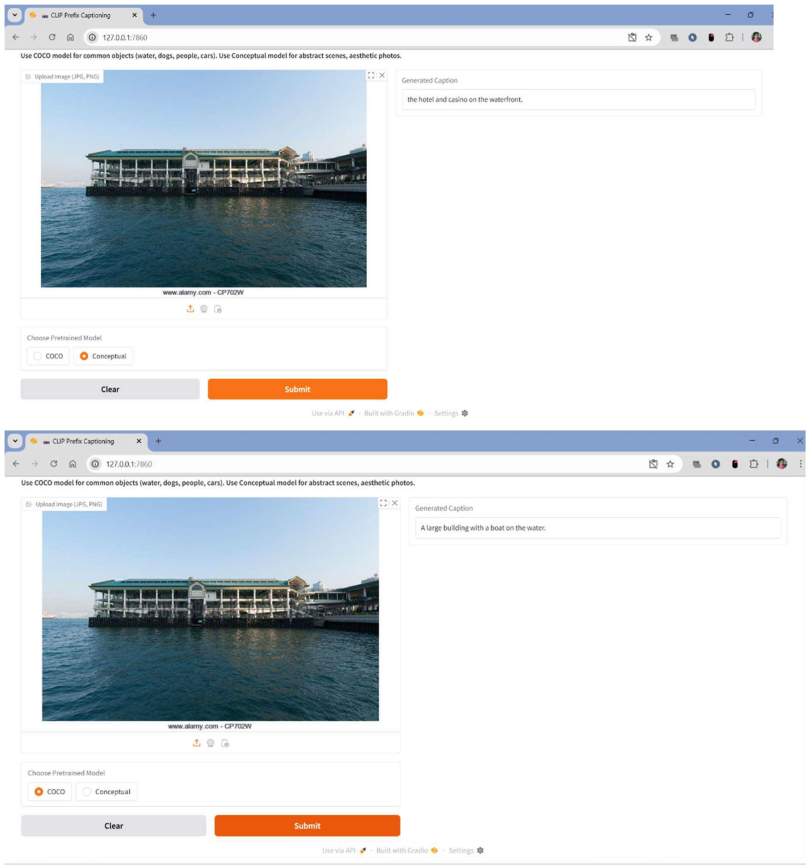
Step 3 : Redirecting to Application's Interface



Step 4 : Uploading Image from system / by allowing access to Webcam



Step 5 : Image captioning using COCO and Conceptual Models



Conclusion

This project successfully demonstrates an efficient and scalable approach to image captioning by integrating CLIP for visual understanding with GPT-2 for language generation, connected through a lightweight prefix mapping network. By leveraging the strengths of powerful pre-trained models and minimizing training overhead, the system achieves high quality, semantically rich captions while remaining resource-friendly. Unlike traditional methods that require extensive training or large model fine-tuning, our method trains only a small mapping component, making it highly efficient without compromising accuracy. The modularity of this approach also allows easy adaptation to various datasets and use cases such as accessibility, content tagging, and visual storytelling. In conclusion, this work offers a practical, low-cost, and effective solution to image captioning, showcasing the potential of aligning vision and language models through minimal yet intelligent integration. Future improvements can include multilingual support, domain specific tuning, and real-time deployment for broader accessibility

Future Scope

The future scope of "CLIP-Guided Image Captioning Using a Lightweight Prefix Mapping" is expansive and transformative. This approach can evolve into a universal multimodal interface, enabling zero-shot applications like visual question answering, scene understanding, and image-grounded dialogue without retraining large models. Its lightweight design makes it ideal for real-time deployment on edge devices such as AR glasses, IoT cameras, and mobile platforms. The system can be extended to support personalized and domain-specific captioning whether for medical imaging, e-commerce, or storytelling by fine- tuning the prefix mapper. By integrating multilingual language models, it can democratize access to AI-powered captioning in low-resource and underrepresented languages. Furthermore, it opens pathways for seamless integration into VR/AR environments and creative tools, generating dynamic scene narratives. This prefix mapping architecture could also serve as a modular bridge for injecting non-textual data into large language models, powering multimodal

agents and assistive AI systems. Ultimately, it lays the groundwork for a new class of efficient, flexible, and scalable vision-language systems.

References

1. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.
2. Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 605–612, 2004.
3. Xinlei Chen and C Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. arXiv preprint arXiv:1411.5654, 2014.
4. Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation, pages 376–380, 2014.
5. Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015.
6. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In European conference on computer vision, pages 382–398. Springer, 2016.
7. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7008–7024, 2017.
8. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, 2018.
9. Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4250–4260, 2019.
10. Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 13041–13049, 2020.
11. Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.
12. Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on image captioning. arXiv preprint arXiv:2107.06912, 2021.