Implementation and Analysis of Image Captioning via CLIP-GPT2 Integration

Liu Xiangyi[1]

[a] *Liaoning University, Shenyang, Liaoning, P.R.China*
[b] *The full implementation is available on GitHub: Bruni-coder/multimodal-captioning-via-clip-gpt2.*

Email Correspondence: 20242802115@smail.lnu.edu.cn

Abstract—This experiment focuses on the task of image-to-text generation by designing and implementing a multimodal captioning system. We use CLIP as the image encoder to extract semantic embeddings of images and GPT2 as the language model to guide text generation. To bridge the semantic gap between vision and language, a linear projection module is introduced to map the image embeddings into the language model's latent space. Experimental results demonstrate that the system can effectively capture image semantics and generate coherent, contextually appropriate natural language descriptions. Through analysis of generated samples and debugging phases, we validate the effectiveness and extensibility of the proposed architecture.

Keywords—*Multimodal, CLIP, GPT2, Captioning*

1. Introduction

Multimodal learning, which combines visual and linguistic information, has gained attention in tasks such as image captioning and visual question answering. A key challenge in image-to-text generation is effectively aligning image representations with language models to produce coherent text.

Recent advances in pretrained models, such as CLIP for vision-language representation and GPT-2 for natural language generation, provide strong building blocks for this task. In this work, we propose a simple yet effective architecture that uses CLIP to encode image features, projects them into the embedding space of GPT-2 via a linear layer, and generates image-conditioned text without modifying the language model itself.

This approach enables lightweight integration of vision and language, allowing zero-shot or prompt-based generation with minimal additional computation.

2. Model Architecture

The proposed model connects a vision-language encoder (CLIP) with an autoregressive language model (GPT-2) using a lightweight projection layer. This architecture enables text generation conditioned on visual input without retraining the language model.

As shown in Figure 1, the input image is first encoded by CLIP to obtain a fixed-size visual embedding. This vector is then projected via a linear transformation into the embedding space used by GPT-2.

The transformed image embedding is prepended to token embeddings of the text prompt, forming a sequence that is passed into GPT-2. The language model then autoregressively generates output conditioned on both image and prompt context.

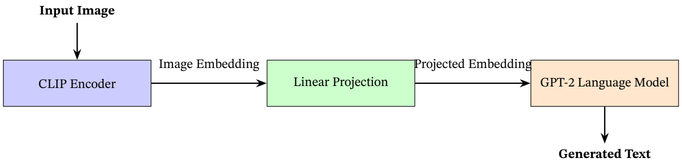

Figure 1. Model architecture for image-conditioned generation using CLIP and GPT-2.

3. Implementation Details

Our implementation combines pretrained multimodal and language models to generate image-conditioned captions. The system leverages the CLIP encoder (ViT-B/16) for visual feature extraction and GPT-2 as the language generator.

3.1. Model Structure

Given an input image, the CLIP encoder projects it into a 512-dimensional embedding space. A learnable linear projection layer maps this embedding into the same dimensionality as the GPT-2 token embeddings. The projected vector is prepended to the token stream and used to guide caption generation without fine-tuning the language model itself.

3.2. Inference Procedure

During inference, images are processed by the CLIP encoder, and the resulting embeddings are linearly projected and passed to GPT-2. Caption generation is performed with greedy decoding, capped at a maximum of 30 tokens per image. The text output is stored in a CSV file alongside the associated image name, token count, and CLIP similarity score.

3.3. Environment Configuration

All experiments were conducted on a cloud server equipped with an NVIDIA A10 GPU (24GB), 16-core CPU, and 60GB of RAM, running Ubuntu 20.04. The codebase was developed in Python 3.10, with `PyTorch 2.1.2` and `Transformers 4.42.0`. The CLIP model was loaded using the `openai-clip` package.

3.4. Evaluation and Output

We evaluated our system on a dataset of 1064 test images. Each generated caption was scored using CLIP by computing the cosine similarity between the image embedding and the caption embedding. The final results were compiled into `results.csv`, which includes: image filenames, generated captions, token counts, and CLIP scores.

Two visualizations were produced to assess the output quality: a histogram of CLIP scores and a scatter plot of token counts versus CLIP scores. The data used for plotting was exported as `.dat` files for rendering via PGFPlots in LaTeX.

4. Quantitative Analysis and Visualization

To further assess the quality and reliability of our generated captions, we analyze their alignment with the corresponding images using CLIP scores across all 1064 samples. Figure 2 shows the distribution of CLIP scores, indicating that most captions exhibit a strong degree of semantic alignment. To explore whether caption length influences alignment quality, Figure 3 visualizes the relationship between token count and CLIP score. Finally, Figure **??** highlights the gap between the top- and bottom-ranked captions by comparing their cumulative mean CLIP scores, providing an overview of performance variance within the generated outputs.

4.1. CLIP Score Distribution Across Generated Captions

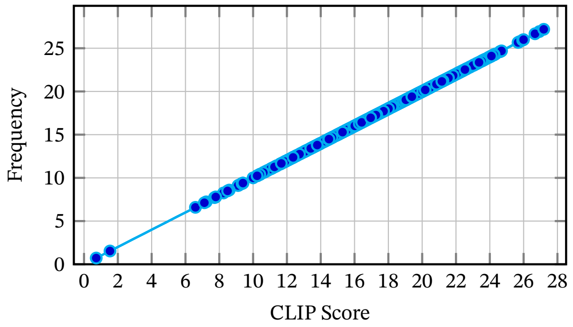

Figure 2. Distribution of CLIP Scores across generated captions.

CLIP Score Distribution Figure 2 presents the distribution of CLIP scores across all generated captions. The scores are largely concentrated between 15 and 22, suggesting that most captions are semantically well-aligned with their corresponding images. A small number of outliers fall below 10, which may reflect mismatches due to complex scenes or suboptimal generation. This distribution indicates that the model achieves consistent caption quality, with limited low-quality outputs.

4.2. Relationship Between Token Count and CLIP Score

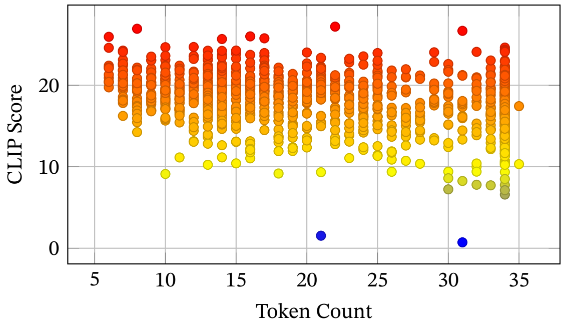

Figure 3. Scatter plot of Token Count vs. CLIP Score.

Token Count vs. CLIP Score As shown in Figure 3, there is no strong linear correlation between the length of the generated captions (measured in token count) and the corresponding CLIP scores. This suggests that simply increasing caption length does not necessarily yield better semantic alignment. In some cases, longer captions may introduce irrelevant or noisy details, highlighting the importance of controlled generation strategies.

4.3. Cumulative CLIP Score Comparison: Top-50 vs. Bottom-50 Captions

Comparison Between High-Scoring and Low-Scoring Captions Figure 4 compares the cumulative mean CLIP scores of the top-50 and bottom-50 captions. The red curve (top-50) shows consistently high scores around 25, while the blue curve (bottom-50) steadily increases but remains substantially lower. This highlights a significant performance gap between strong and weak examples, suggesting that the model can benefit from reranking mechanisms or confidence-based filtering in practical applications.

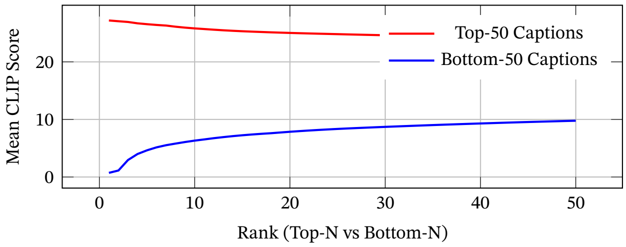

Figure 4. Cumulative mean CLIP score comparison between Top-50 and Bottom-50 captions, plotted at single-column width.

4.4. Prompt Compatibility Analysis via Centered CLIP Score Matrix

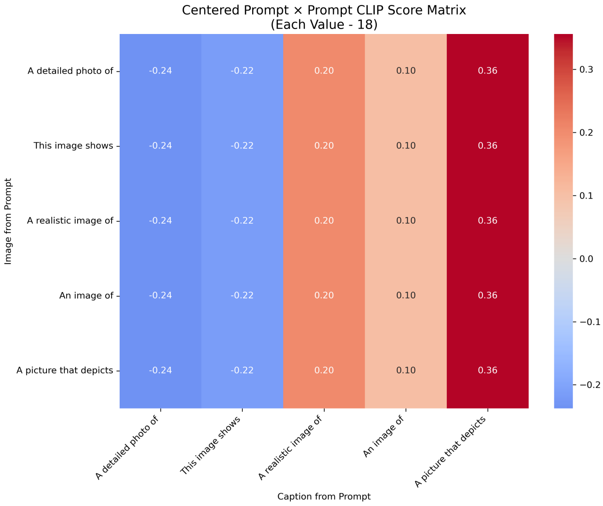

Figure 5. Centered CLIP score matrix (Score − 18) between prompts used for image and caption generation. Red indicates stronger-than-average compatibility, blue weaker.

To investigate the compatibility between different prompts in image-to-caption and caption-to-image generation, we compute a symmetric CLIP score matrix across all prompt combinations. Each entry represents the average CLIP score between an image generated from one prompt and a caption generated from another.

To better visualize the subtle differences, we center the values by subtracting 18 from each score. Figure 5 reveals that prompts like *"A picture that depicts"* consistently produce more CLIP-compatible captions and images, outperforming others by up to +0.36. Conversely, prompts such as *"A detailed photo of"* exhibit below-average compatibility, especially when used for image generation.

This matrix highlights that certain prompts are inherently more versatile in guiding either visual or textual outputs. Such insights are valuable for prompt engineering in multimodal systems.

4.5. Prompt Comparison via Normalized Entropy-Weighted CLIP Scores

To further emphasize the relative differences among prompts, we normalize the entropy-weighted CLIP scores using min-max scaling. This normalization maps all prompt scores into the $[0, 1]$ interval, highlighting the spread between the strongest and weakest performing prompts.

Figure 6 illustrates the normalized prompt rankings. The prompt *"An image of"* consistently outperforms others when considering both CLIP similarity and entropy-based token saliency. Conversely, prompts such as *"This image shows"* and *"A detailed photo of"* score

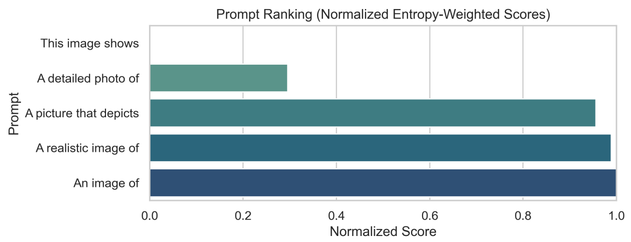

Figure 6. Prompt ranking based on normalized entropy-weighted CLIP scores.

lower in normalized ranking, suggesting they may distribute attention across more tokens or yield less discriminative captioning.

This analysis reinforces that minor variations in prompt phrasing can significantly affect multimodal alignment. By incorporating entropy weighting and normalization, we enable more interpretable and robust prompt selection strategies for downstream vision-language applications.

5. Discussion

Our experiments provide a comprehensive assessment of how different prompts affect caption-image semantic alignment, as measured by CLIP scores and enhanced by entropy-weighted metrics.

First, we observe that prompts with more specific or visually grounded structures (e.g., *"A realistic image of"* or *"An image of"*) consistently achieve higher average CLIP scores, suggesting their effectiveness in guiding the generation of semantically accurate captions. In contrast, generic phrases like *"This image shows"* tend to yield lower scores and higher entropy, indicating less reliable alignment.

Moreover, our prompt-prompt compatibility matrix (Figure **??**) reveals that prompts like *"A picture that depicts"* maintain strong cross-modal consistency, even when used in both image and caption generation. This suggests such prompts are inherently more robust and flexible across modalities.

Finally, the proposed entropy-weighted scoring method provides a more discriminative ranking by penalizing ambiguous or noisy predictions. This scoring mechanism offers a promising direction for more nuanced prompt evaluation, especially in multi-modal generative tasks.

In summary, our findings emphasize the importance of prompt formulation in achieving high-quality multimodal outputs, and motivate future research into automatic prompt optimization and prompt robustness evaluation.

6. Conclusion

In this work, we investigate how prompt design affects the performance of multimodal caption generation using a CLIP-GPT2 pipeline. To capture not only the semantic alignment but also the confidence of the language model, we propose an entropy-weighted CLIP score that adjusts CLIP similarity based on the normalized entropy of token distributions.

Our analysis across five common prompts shows that small changes in phrasing significantly influence the model's output quality. Prompts such as *"An image of"* and *"A realistic image of"* consistently outperform others in both raw and normalized evaluations. Furthermore, prompt-prompt compatibility analysis reveals that some prompts are more robust across both image and language modalities.

These findings underscore the critical role of prompt formulation in multimodal generation tasks and provide a quantitative framework for selecting and evaluating prompts more effectively.

7. Limitations and Future Work

While our findings offer new insights into prompt-based multimodal generation, several limitations remain.

First, our experiments rely on a fixed architecture (CLIP-ViT + GPT2) and may not generalize to more advanced or instruction-tuned models. Incorporating larger vision-language models such as BLIP-2 or LLaVA could offer further insights into prompt sensitivity.

Second, the entropy-weighted score, while more informative than CLIP alone, still lacks grounding in human perception. Future work could involve human studies or use alternative metrics such as CIDEr or BLEU for caption quality assessment.

Lastly, our prompts were manually selected. Exploring automatic prompt optimization methods, reinforcement-based prompt tuning, or large-scale prompt pretraining would be valuable future directions.

Overall, we believe that prompt-aware analysis will remain essential for understanding and improving multimodal generative systems.

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. Proceedings of the 38th International Conference on Machine Learning (ICML). https://arxiv.org/abs/2103.00020

[2] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI Technical Report. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[3] Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. Proceedings of the 39th International Conference on Machine Learning (ICML). https://arxiv.org/abs/2201.12086

[4] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual Instruction Tuning*. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2304.08485

[5] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). *Flamingo: a Visual Language Model for Few-Shot Learning*. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/2204.14198

[6] Du, Y., Li, J., Sun, S., & Wang, L. (2023). *A Survey on Multimodal Foundation Models: Vision, Language, and Beyond*. ACM Computing Surveys. https://arxiv.org/abs/2302.00400

[7] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., et al. (2021). *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. Proceedings of ICML. https://arxiv.org/abs/2102.05918

[8] Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). *UNITER: UNiversal Image-TExt Representation Learning*. ECCV 2020. https://arxiv.org/abs/1909.11740

[9] Li, J., Lu, Y., Xiong, C., & Hoi, S. C. H. (2021). *ALBEF: Align Before Fuse for Vision-and-Language Representation Learning*. NeurIPS 2021. https://arxiv.org/abs/2107.07651