Prompt Engineering Analysis on Image Captioning

Vaishnav Anugrah [1]
[1]02132107
University of Massachusetts Lowell

Abstract

This project explores a comprehensive pipeline for generating and evaluating image captions, focusing on prompt engineering. The central hypothesis tests whether prompts that include chain-of-thought reasoning (Caption B) lead to captions that are more preferred by various evaluators—large language models (LLM), CLIP Small, CLIP Large, and humans—compared to prompts without chain-of-thought reasoning (Caption A).

The pipeline begins by generating image captions using prompts with and without chain-of-thought reasoning. These captions are then evaluated through a multi-stage process. CLIP Small and CLIP Large assess semantic alignment and coherence, while a different image-text multimodal model (LLM) provides additional automated evaluation. The final stage involves human surveys, where evaluators choose their preferred captions for a given set of images.

The results reveal variations in preference among the evaluators. CLIP Small and CLIP Large tend to align more closely with human evaluators, while the LLM's preference patterns differ. The agreement rates between the evaluators reflect the varying levels of preference for captions with and without chain-of-thought reasoning, contributing to a deeper understanding of prompt engineering in image captioning.

This project's multi-faceted approach to generating and evaluating image captions, combined with an exploration of prompt engineering, provides insights into the effectiveness of chain-of-thought reasoning in prompts. The findings offer valuable guidance for researchers and developers seeking to improve the accuracy and relevance of language model output.

Motivation

Image captioning is the process of generating descriptive text for visual content, with applications ranging from accessibility tools to automated content indexing. This task requires a careful balance between semantic accuracy and contextual relevance, as captions must effectively capture the essence of an image while remaining meaningful to a human audience. Recent advancements in deep learning (Radford and others 2021)(Karpathy and Fei-Fei 2015)(Vinyals et al. 2015), particularly with image-text multimodal models, have transformed the field of image captioning, allowing for more accurate and contextually appropriate descriptions.

Central to the development of effective image captioning systems is the concept of prompt engineering. Prompt engineering involves designing and refining the textual prompts used to generate responses from language models like GPT-3 and similar models. The quality and relevance of the generated text depend heavily on the structure and content of the prompts, making prompt engineering a crucial aspect of model development.

This project explores the impact of prompt engineering on image captioning, with a specific focus on chain-of-thought reasoning in prompts. The hypothesis is that prompts incorporating chain-of-thought reasoning result in captions that are more preferred by various evaluators—LLM, CLIP Small, CLIP Large, and human evaluators—compared to prompts without this approach.

The pipeline begins with generating image captions using prompts designed with and without chain-of-thought reasoning. These captions undergo a multi-stage evaluation process. In the initial stages, automated evaluations are conducted using CLIP Small and CLIP Large, assessing semantic alignment and coherence. Additionally, a different image-text multimodal model (LLM) is employed to provide further automated evaluation. The final stage involves human surveys, where evaluators select their preferred captions from a set of images.

Understanding the alignment between automated model evaluations and human preferences is crucial for improving image captioning systems. By analyzing the agreement rates between different evaluators, this project seeks to validate the hypothesis and determine the effectiveness of chain-of-thought reasoning in prompt engineering.

Related Work

The field of image captioning has experienced significant growth, driven by advances in deep learning and multimodal models that integrate visual and textual data. This section discusses key works that have influenced the development of image captioning, focusing on the evolution of techniques, models, and evaluation methods.

Interactive approaches to image description offer diverse multimodal controls (Wang et al. 2023), allowing users to

guide the captioning process based on their input and preferences. This flexibility adds a new dimension to image captioning, creating more personalized and contextually relevant descriptions. These interactive techniques enable a closer alignment between the generated captions and the specific requirements of users.

Prefix-based methods (Mokady, Hertz, and Bermano 2021) for image captioning leverage the power of multimodal models like CLIP, known for their contrastive learning capabilities. By using a prefix to guide the captioning process, these approaches achieve a streamlined and efficient generation of high-quality captions. This technique demonstrates the potential of utilizing model-specific strengths to optimize the image captioning process, resulting in more coherent and accurate descriptions.

Cross-modal learning (Li et al. 2022) has emerged as an effective approach to vision-language tasks, where skip connections between different modalities are used to enhance learning efficiency. This innovative technique improves the integration of visual and textual information, leading to more accurate and contextually relevant image captions. By effectively combining different modalities, these methods contribute to the development of more robust vision-language models.

The concept of visual attention (Xu et al. 2015), has significantly influenced image captioning models. This approach allows models to focus on specific regions of an image when generating captions. An important addition to this field is the BLIP (Bootstrapping Language-Image Pre-training) (Li and others 2022) model, which extends the multimodal approach by leveraging large-scale pre-training and bootstrapping techniques to improve visual-language alignment. BLIP's unique strategies make it an essential model in contemporary image captioning.

Together, these works form the basis for modern image captioning methodologies and the role of prompt engineering in generating high-quality textual responses from language models. By integrating insights from these key studies, this project aims to contribute to the ongoing advancement of image captioning, with a focus on understanding the impact of prompt engineering and chain-of-thought reasoning in text generation.

Proposed Approach

The primary objective of this project is to evaluate the impact of chain-of-thought reasoning in prompt engineering on image captioning quality. The hypothesis posits that captions generated using prompts with chain-of-thought reasoning will be more preferred by various evaluators compared to those generated with prompts lacking this approach. This section outlines the process pipeline, the models used for generating captions and evaluation, and the methodology for human surveys.

Hypothesis

The core hypothesis posits that prompts with chain-of-thought reasoning will result in captions that are more favorably received by various evaluators, including LLMs, CLIP Small, CLIP Large, and human evaluators, when compared to captions derived from prompts without chain-of-thought reasoning.

Generating Image Captions

LLaVa-Mistral 7B (Liu et al. 2023b)(Liu et al. 2024) from Hugging Face `llava-hf/llava-v1.6-mistral-7b-hf` (Liu et al. 2023a) is used for generating captions for images due to its robust language model capabilities and multimodal integration. This model is designed to leverage large-scale pre-training on diverse text and image datasets, allowing it to understand complex relationships between visual and textual information. This deep learning-based model can generate rich and contextually relevant captions for a wide range of images, providing flexibility and accuracy in text generation.

One of the key reasons for using LLaVa-Mistral 7B(Liu et al. 2024) is its ability to handle various prompt engineering techniques, such as chain-of-thought reasoning. This capability makes it ideal for exploring different approaches to prompt structure and understanding how these affect caption generation quality. Additionally, LLaVa-Mistral 7B's extensive pre-training ensures that the model has a broad understanding of language and visual contexts, making it suitable for generating captions that accurately reflect the content of images. Its versatility and robustness contribute to its effectiveness in generating high-quality captions for this project, offering a reliable foundation for subsequent evaluations and human surveys.

Captions are generated using an image-text multimodal model. Two types of prompts are used:

- **Without Chain-of-Thought**: "Generate a caption for this image in 50 words."
- **With Chain-of-Thought**: "Generate a caption for this image, and the description should include the number of objects in the image without explicitly mentioning it in 50 words."

Evaluation

The evaluation process combines automated model-based assessments with human-based evaluations, providing a comprehensive view of the effectiveness of the generated captions.

Automated evaluations: The automated evaluations have three main components: CLIP Small (ViT-B/32), CLIP Large (ViT-L/14), and LLaVa-Mistral 7B from Hugging Face. These models are used to evaluate the semantic alignment and coherence of the generated captions (see Figure 1), offering a consistent and scalable method to assess their quality.

1. **CLIP Small and CLIP Large**: The CLIP (Contrastive Language–Image Pre-training) models are used for evaluation in this project due to their unique ability to align visual and textual data, making them highly effective in assessing semantic coherence and alignment. CLIP models, including CLIP Small (B/32) and CLIP Large (L/14),

```
('Caption (A): In the heart of a bustling bookstore, two individuals are '
 'immersed in a world of literature. The person on the left, clad in a black '
 'coat, stands behind a table laden with an array of books. Their gaze is.')
('Caption (B): In the heart of a bustling bookstore, two individuals are '
 'immersed in the world of literature. The store is a treasure trove of '
 'knowledge, with books of various sizes and colors neatly arranged on shelves '
 'that stretch from floor to ceiling.')
```

Figure 1: Captions generated for an image. Caption (A) - Without COT, Caption (B) - With COT

are pre-trained on a large dataset that combines images and text, allowing them to understand the relationship between the two modalities. CLIP's pre-training allows it to evaluate how well a given caption corresponds to its associated image, offering a robust measure of semantic accuracy. By using CLIP models for evaluation, the project can assess the generated captions' quality and coherence in an automated manner, ensuring that they accurately represent the content and context of the images. This automated evaluation is critical in providing a consistent and scalable method to validate the hypothesis and determine the effectiveness of different prompt structures in image captioning.

These models use contrastive learning to evaluate the semantic alignment between images and their corresponding captions. The CLIP models are pre-trained on large datasets, allowing them to understand a wide range of visual and textual information. The evaluations focus on how well the generated captions represent the content and context of the images.

2. **LLaVa-Mistral 7B**: This model serves as an additional automated evaluation to further assess the quality of the generated captions. LLaVa-Mistral is known for its robust language model capabilities, making it suitable for evaluating the contextual relevance of the captions.

The automated evaluations are used to compare the generated captions' quality, coherence, and semantic alignment. The agreement rates between these models and human evaluations provide insights into the effectiveness of the captions generated with different prompts.

Human-Based evaluations: Human-based evaluations involve conducting surveys to gather qualitative feedback from human evaluators. This stage is crucial in understanding which captions resonate best with actual users, providing a more nuanced assessment of caption quality. In this stage, human evaluators are asked to select their preferred captions from a set of generated captions.

Metrics for evaluation:

The evaluations are assessed using various metrics to determine the effectiveness of the generated captions. Some of the key metrics include:

- Semantic Alignment: This metric measures how well the generated captions align with the semantic content of the images. CLIP Small and CLIP Large are primarily used for this evaluation.
- Contextual Relevance: This metric assesses the contextual appropriateness of the captions. LLaVa-Mistral 7B is used to evaluate the relevance of the captions to the visual content.
- Human Preference: This metric is derived from human surveys, indicating which captions are more preferred by human evaluators.

Experimental Results

Dataset Description

The data used for this study is sourced from the Visual Genome dataset, a comprehensive collection of images designed to support various computer vision and natural language processing tasks. For this study, a subset of the Visual Genome dataset is used to evaluate the quality of image captions.

Results

Human evaluations: Participants choose their preferred caption from a set of images. Figure 2 shows that Caption B (with chain-of-thought reasoning) was selected more frequently (49 times) than Caption A (without chain-of-thought reasoning) (34 times), indicating a preference for the more detailed and contextually rich approach.

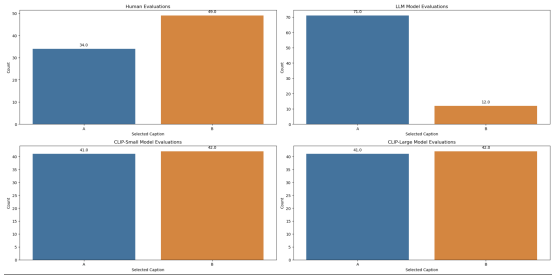

Figure 2: Caption preference for different evaluators

Automated evaluations: Automated evaluations used CLIP Small, CLIP Large, and an LLM to assess semantic alignment and coherence. The following results highlight the outcomes from these evaluations:

- CLIP Small and CLIP Large: These models showed a slight preference for Caption B, indicating a similar alignment with human evaluations. CLIP Small selected Caption B 42 times and Caption A 41 times, while CLIP Large selected Caption B 42 times and Caption A 41 times.

- LLM: In contrast to CLIP models and human evaluations, the LLM showed a strong preference for Caption A, selecting it 71 times compared to 12 times for Caption B.

Agreement rates: The agreement rates between different evaluators provide insights into how well the automated evaluations align with human assessments:

- LLM and Human Evaluations: The agreement rate was 40.96%, indicating a relatively low alignment between the LLM and human evaluations.
- CLIP-Small and Human Evaluations: The agreement rate was 57.83%, suggesting a closer alignment with human preferences.
- CLIP-Large and Human Evaluations: The agreement rate was 50.60%, indicating moderate alignment.

LDA Topic modeling: Latent Dirichlet Allocation (LDA)(Blei, Ng, and Jordan 2003) is a technique used to identify and understand the underlying themes or topics in a collection of textual data. It is particularly useful for analyzing large text corpora to uncover patterns and trends in the content. By applying topic modeling to the generated text, we can identify the key concepts and words that frequently appear in captions created with and without chain-of-thought reasoning (see figure 3). This approach allows us to assess the quality and coherence of the generated captions, providing insights into their structure and the recurring elements that may contribute to their preference by evaluators.

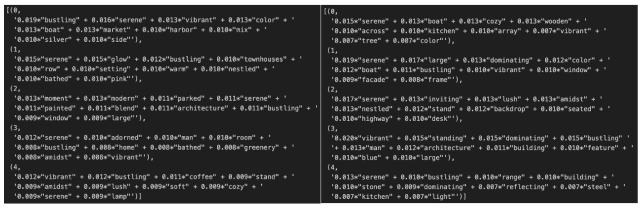

Figure 3: LDA topic modeling - (left) topics for caption A, (right) topics for caption B

The LDA topic model's achieved:

- Coherence Score for Caption A: 0.325
- Coherence Score for Caption B: 0.299
- Perplexity for Caption A: -6.593
- Perplexity for Caption B: -6.433

Discussion

The results from the evaluations provide valuable insights into the impact of prompt engineering, particularly the use of chain-of-thought reasoning, on image captioning quality.

One of the primary findings from this study is the preference for captions generated with chain-of-thought reasoning. Human evaluations indicated a greater preference for Caption B, with 49 selections compared to 34 for Caption A. This suggests that human evaluators tend to favor captions that offer more context and detail stating the number of objects in the image. The CLIP models follow suit to human evaluators.

However, the LLM evaluations diverged from this pattern, showing a strong preference for Caption A, with 71 selections compared to 12 for Caption B. This discrepancy suggests that while the LLM may excel at generating text, it may not always capture the nuances that human evaluators find appealing. This can also seen from the agreement rates between human-LLM evaluations.

Limitation of language models: Despite specifying a 50-word limit in the prompt, the language model (LLaVa-Mistral 7B) often exceeded this limit, generating captions that contained more words than expected. This limitation has several implications:

- Model Constraints: Language models, especially large ones, can struggle with constraints on output length. This could be due to the model's internal structure, where it generates text based on the most probable sequences, leading to longer outputs.
- Prompt Adherence: The inability to consistently limit caption length suggests that prompt engineering may require additional refinement to ensure models adhere to specific constraints. This could affect the usability of the generated captions in applications where brevity is critical.
- Evaluation Impact: If captions exceed the specified word limit, it may influence human evaluations, as longer captions might be perceived as less concise or harder to interpret.

Conclusion

This project investigated the impact of prompt engineering, specifically chain-of-thought reasoning, on image captioning quality. The findings revealed that captions generated with chain-of-thought reasoning were more preferred by human evaluators and automated models such as CLIP Small and CLIP Large, suggesting that adding context and detail enhances caption relevance. However, the LLM showed a significant preference for simpler captions, indicating a divergence in evaluation patterns. A key limitation was the language model's inconsistency in adhering to a 50-word limit, highlighting a need for refined prompt structures and model tuning. The results point to the importance of flexibility in prompt engineering and the value of combining automated and human evaluations to improve the quality of image captioning. Future work should focus on refining prompts, improving model tuning, and exploring additional evaluation methods to enhance the effectiveness of image captioning models.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, 3128–3137.

Li, J., et al. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding. In *arXiv preprint arXiv:2201.12086*.

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; Zhang, J.; Huang, S.; Huang, F.; Zhou, J.; and Si, L. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning.

Radford, A., et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, T.; Zhang, J.; Fei, J.; Zheng, H.; Tang, Y.; Li, Z.; Gao, M.; and Zhao, S. 2023. Caption anything: Interactive image description with diverse multimodal controls.

Xu, K.; Ba, J.; Kiros, R.; et al. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, 2048–2057.