Ebhaam at SemEval-2023 Task 1: A CLIP-Based Approach for Comparing Cross-modality and Unimodality in Visual Word Sense Disambiguation

Zeinab Taghavi [1], Parsa Haghighi Naeini[*1], Mohammad Ali Sadraei[*1], Soroush Gooran[1],

Ehsaneddin Asgari[2], Hamid Reza Rabiee[1] and Hossein Sameti[1]

[1]Department of Computer Engineering, Sharif University of Technology
[2]AI Innovation Center, Data:Lab, Volkswagen AG, Munich, Germany

Abstract

This paper presents an approach to tackle the task of Visual Word Sense Disambiguation (Visual-WSD), which involves determining the most appropriate image to represent a given polysemous word in one of its particular senses. The proposed approach leverages the CLIP model, prompt engineering, and text-to-image models such as GLIDE and DALL-E 2 for both image retrieval and generation. To evaluate our approach, we participated in the SemEval 2023 shared task on "Visual Word Sense Disambiguation (Visual-WSD)" using a zero-shot learning setting, where we compared the accuracy of different combinations of tools, including "Simple prompt-based" methods and "Generated prompt-based" methods for prompt engineering using completion models, and text-to-image models for changing input modality from text to image. Moreover, we explored the benefits of cross-modality evaluation between text and candidate images using CLIP. Our experimental results demonstrate that the proposed approach reaches better results than cross-modality approaches, highlighting the potential of prompt engineering and text-to-image models to improve accuracy in Visual-WSD tasks. We assessed our approach in a zero-shot learning scenario and attained an accuracy of 68.75% in our best attempt.

1 Introduction

Visual Word Sense Disambiguation (Visual-WSD) is a challenging task in natural language processing that aims to disambiguate words or phrases with multiple possible meanings and retrieve the image that best matches the semantic content of the input text. The SemEval 2023 shared task on "Visual Word Sense Disambiguation (Visual-WSD)" provides a valuable benchmark for evaluating the performance of Visual-WSD systems (Raganato et al., 2023). In this paper, we propose an approach to Visual-WSD that utilizes the CLIP (Radford et al., 2021) model for image retrieval and prompt engineering with OpenAI's completion model for generating images from textual descriptions. We describe our methodology and experimental results, including our use of zero-shot learning and comparison of two text-to-image models, GLIDE (Nichol et al., 2021) and DALL-E 2 (Ramesh et al., 2022), for image synthesis. Furthermore, we compare the performance of our approach with image retrieval using CLIP encoders for both input text and candidate images, finding that our best result using generated images reaches better results than using text and images for evaluation. Our findings suggest that prompt engineering and text-to-image models have the potential to significantly improve the accuracy of image retrieval tasks. Finally, we discuss the implications of our results for future research in this area. We should note that we only test our model with English texts, and not the optional Persian and Italian languages as part of the competition.

The outline of the paper consists of four sections: Background, System Overview, Experimental Setup, and Conclusion. In the Background section, the authors provide an overview of the problem of Visual Word Sense Disambiguation and its importance in natural language processing. The System Overview section outlines the proposed approach, which involves using CLIP, prompt engineering, and text-to-image models for image retrieval and synthesis. The Experimental Setup section describes the evaluation methodology and the results obtained using different combinations of tools. Finally, the Conclusion section summarizes the findings and highlights future research opportunities in Visual-WSD.

2 Background

Visual Word Sense Disambiguation (Visual-WSD) is a challenging problem in natural language processing that aims to disambiguate word meanings using visual cues. Traditional WSD methods rely

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1960–1964

on textual data, but they may not always be effective in multiple meanings or ambiguous contexts. Visual-WSD approaches use visual information, such as images or videos, to provide additional context for word disambiguation (Barnard et al., 2003).

The CLIP model is a state-of-the-art model for zero-shot image classification and retrieval that has the potential for Visual-WSD tasks. CLIP stands for Contrastive Language-Image Pre-training, and it was developed by OpenAI. CLIP is a neural network that is pre-trained on a large corpus of text and images and is able to encode both modalities into a shared embedding space. This allows for efficient retrieval of images that are semantically similar to input text, even in zero-shot settings (Radford et al., 2021).

Prompt engineering is a technique used to generate more effective prompts for models such as CLIP (**?**). In the context of Visual-WSD, prompt engineering involves generating textual prompts that include limited textual context and the target word or phrase. These prompts are used to synthesize images that are semantically related to the input text (Liu et al., 2023).
OpenAI's Completion models are one example of a prompt engineering technique that can generate high-quality prompts (Liu and Chilton, 2021).

Text-to-image models like GLIDE and DALL-E 2 can be used for image retrieval and synthesis in Visual-WSD tasks. GLIDE is a text-to-image model developed by OpenAI that is able to generate high-quality images from textual descriptions (Nichol et al., 2021). DALL-E 2 is another text-to-image model developed by OpenAI that can generate complex images from textual prompts (Ramesh et al., 2022). These text-to-image models can be used to generate images that are semantically related to the input text, providing additional context for word disambiguation.

Our goal in this study is to contribute to the growing body of research on Visual-WSD and to inspire further development in this important area of natural language processing.

3 System Overview

Our Visual-WSD system consists of three main components: prompt engineering, text-to-image systems, and evaluation using CLIP. Figure 2 provides a high-level overview of our system architecture.

3.1 Prompt Engineering

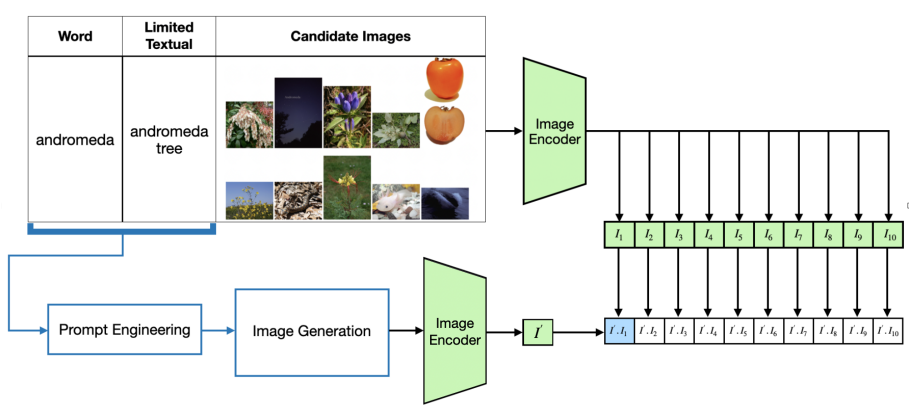
Prompt engineering is a process in natural language processing that involves creating well-formed and informative prompts to guide a language model toward producing relevant and accurate outputs(Liu et al., 2023). In the context of Visual Word Sense Disambiguation (Visual-WSD), this process involves generating a set of prompts that can be used to retrieve relevant images using pre-trained image generation models such as DALL-E 2 or GLIDE. Two prompt types were used: Simple prompts and Generated prompts. Simple prompts involved concatenating the word and a limited textual context together as **'I want to describe ' + *word* + ' as it means ' + *limited textual context*** while "Generated prompts" used OpenAI's completion models as ***Simple prompt* + ', can you describe it as picture?'** then the output of the model will be the input of the text-to-image model as you can see in Figure 1a (Logan Iv et al., 2022). Figure 2a shows an example of simple and generated prompt. The experiments show that prompt engineering improved the accuracy of image retrieval tasks and reduced the need for large amounts of labeled data.

3.2 Text-to-Image Systems

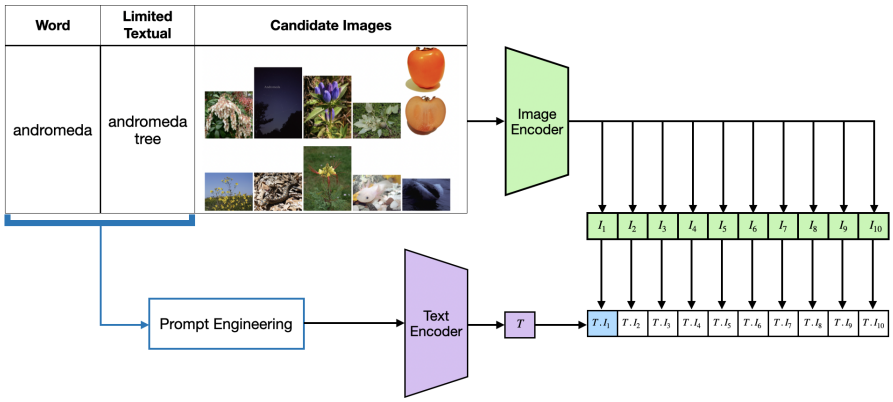
The text-to-image systems subsection is a crucial component of our approach to Visual Word Sense Disambiguation. In this subsection, we aim to convert textual descriptions of a word or phrase into images that are in the same data modality as the candidate images and can be used for comparison with them. This conversion process involves using text prompts to generate images that match the semantic content of the input text. We evaluate the performance of two different text-to-image models, DALL-E 2 and GLIDE, to determine which one can best generate images that accurately reflect the intended meaning of the input text prompts. By comparing the performance of these models, we can determine which one is best suited for our Visual-WSD task and can help improve the accuracy of image retrieval tasks.

3.3 Evaluation using CLIP

Finally, we evaluate the accuracy of our Visual-WSD system using CLIP, a state-of-the-art model for zero-shot image classification and retrieval. We explore the benefits of cross-modality evaluation between text and candidate images using CLIP encoders for both input text and images. Our experi-



(a) Image-modality comparison method. Using CLIP encoders for both input text and candidate images. The CLIP model is used to retrieve the candidate images that best match the input text prompts, and the retrieved images are evaluated against the ground truth images.



(b) Cross-modality comparison method. Text prompts are used to generate images using text-to-image models. The CLIP image encoder is used to encode both the generated images and candidate images, and the closest match is selected.

Figure 1: System overview of our Visual-WSD approach

mental results show that the proposed approach outperforms cross-modality approaches and highlights the potential of prompt engineering and text-to-image models to improve accuracy in Visual-WSD tasks.

The evaluation using the CLIP aims to compare the performance of different methods in Visual Word Sense Disambiguation (Visual-WSD) using the CLIP model for image retrieval. We use the CLIP encoders to represent both the generated image or textual prompts and candidate images and compute the cosine similarity between them to find the most relevant image. We evaluate the effectiveness of this approach in comparison to other methods, including prompt engineering and text-to-image models such as DALL-E 2 and GLIDE. On the overall performance of the system, you can see in Figure 2. Our experiments show that using the same modality of data, the CLIP model for image retrieval can achieve competitive performance in Visual-WSD tasks in comparison to cross-modality image retrieval.

4 Experiment Setup

The study involved an image selection task using a dataset that requires selecting an image that corresponds to a given word and limited textual context. Two types of prompts were used: "Simple prompts", which involved concatenating the word and limited textual context to create a text prompt, and "Generated prompts," which involved using OpenAI's "text-davinci-003" Completion model to "Generated prompts" by adding the phrase ", can you describe it as a picture?" to the end of the "Simple prompts".

Two text-to-image models, GLIDE and DALL-E 2, were used to generate images from the text prompts. The generated images and candidate images were then processed using the "ViT-B/32" CLIP model to obtain embeddings. Cosine similarity was used to compare the embeddings and determine the closest image to the input text.

Simple prompt: 'I want to describe andromeda as it means andromeda tree'

Generated prompt (Input of Completion Model: 'I want to describe andromeda as it means andromeda tree , describe it as picture') :
'The Andromeda tree is an evergreen shrub or small tree native to North America. It is also known as the mountain laurel, spoonwood, or calico-bush. The tree grows to a height of 10-15 feet and has a spreading, open form. Its leaves are leathery, glossy, and dark green in color, and its bark is reddish-brown and scaly. The Andromeda tree produces clusters of white, bell-shaped flowers in the spring, followed by dark blue, waxy berries in the summer. The berries are edible and can be used to make jellies, jams, and wines. The tree is also popular as an ornamental plant due to its attractive foliage and fragrant flowers.'

(a) Simple Prompt and Generated Prompt example.

(b) The left image is the main paired image with the word "Andromeda" in the given dataset. The center image was generated by DALL-E 2, and the right image was generated by GLIDE using simple prompts for the same word.

(c) The image generated by DALL-E 2 (left), and the image generated by GLIDE (right) using Simple prompts for the word "Andromeda"

Figure 2: Example of prompts and generated Images. We can see the difference in results due to the difference in prompt engineering and text-to-image models.

5 Results and Conclusions

We conducted experiments to evaluate the performance of our Visual Word Sense Disambiguation (Visual-WSD) system using two different text-to-image models: GLIDE and DALL-E 2. The system was evaluated with two different prompts: a "Simple prompt" that combines the target word and limited textual context, and a "Generated prompt" that is generated using a completion model and also conduct an experiment to evaluate a system that uses "Simple prompt" and candidate images for Visual-WSD task.

Our results show that the "Generated prompt" with DALL-E 2 leads to the highest accuracy of 68.75%, while the lowest accuracy of 18.75% was achieved using the "Simple prompt" with GLIDE. In our research, it was observed that DALL-E 2 outperformed GLIDE with statistically significant results when evaluated using "Simple prompts" and "Generated prompts", exhibiting a respective improvement. While using cross-modality embeddings achieve 56.25% accuracy.

Moreover, we observed that changing the prompts from "Simple prompts" to "Generated prompts" can substantially improve the accuracy of GLIDE from 18.75% to 43.75%, and DALL-E 2 from 62.5% to 68.75%.

The results are summarized in Table 1, which shows the accuracy of the system using different methods. Overall, the findings suggest that prompt engineering and text-to-image systems are effective in improving the accuracy of Visual-WSD systems and that DALL-E 2 reaches better results than GLIDE in this task. Also, we found that in the same dataset, converting the text modality into the image modality, and finding the most similar image when input and candidate images are in the same modality instead of cross-modality evaluation, we can improve the accuracy. This issue can

Table 1: Results of Visual Word Sense Disambiguation using different methods

Data Modality	Prompt Type	Text-to-Image model	Accuracy
Image Modality	Simple prompt	GLIDE	18.75%
Image Modality	Simple prompt	DALL-E 2	62.5%
Image Modality	Generated prompt	GLIDE	43.75%
Image Modality	Generated prompt	DALL-E 2	68.75%
Cross Modality	Generated prompt	-	56.25%

be an inspiration that sometimes we can get different results without changing the input content and only by converting the modality. We hope that this idea can be followed in the future to improve our disambiguation systems.

References

Kobus Barnard, Matthew Johnson, and David Forsyth. 2003. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data -*, volume 6, pages 1–5, Not Known. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35.

Vivian Liu and Lydia B. Chilton. 2021. Design Guidelines for Prompt Engineering Text-to-Image Generative Models.

Robert Logan Iv, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.