University of Hildesheim at SemEval-2023 Task 1: Combining Pre-trained Multimodal and Generative Models for Image Disambiguation

Sebastian Diem, Chan Jong Im, Thomas Mandl
University of Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract

Multimodal ambiguity is a challenge for understanding text and images. Large pre-trained models have reached a high level of quality already. This paper presents an implementation for solving an image disambiguation task relying solely on the knowledge captured in multimodal and language models. Within task 1 of SemEval 2023 (Visual Word Sense Disambiguation), this approach managed to achieve an MRR of 0.738 using CLIP-Large and the OPT model for generating text. Applying a generative model to create more text given a phrase with an ambiguous word leads to an improvement in our results. The performance gain from a bigger language model is larger than the performance gain from using the larger CLIP model.

1 Introduction

Ambiguity is one of the big challenges for Natural Language Processing (NLP) and can often be solved by context. In a multimodal setting, the ambiguity can be solved by identifying the correct visual information. This is the goal of the system implemented in this work.

The goal of *Task 1: Visual Word Sense Disambiguation (V-WSD)* is to find the most similar image among ten candidate images based on textual input (see Task Overview (Raganato et al., 2023)). The scientific aim is the development of technology for disambiguation in a multimodal context.

The textual input consisted of A: the ambiguous target word, and B: an additional phrase that clarified the intended meaning of the target word.

Some of the candidate images share visual similarities, which increases the importance of a precise representation of the model. The task was split into three different languages: English, Farsi, and Italian. The training dataset consists of nearly 13,000 training examples, where each includes an ambiguous word, an additional phrase, and ten candidate images. The correct image was located in a separate file. The test dataset has a similar structure, including 8,100 images for 463 examples (Raganato et al., 2023).

For this work, we focused only on the English dataset. Our approach requires no training and uses only pre-trained models. The progress in large scale language as well as multimodal models in recent years has been rapid. These models capture the semantics of language and images and allow many applications (Gan et al., 2022).

The main focus of our approach aims at the reduction of the ambiguity of the textual input by using "large" generative language models to create more text when given the input phrases. We used four different text inputs, including the phrase, the ambiguous word, the phrase without the ambiguous word, and the input created by the generative model. By using the multimodal model CLIP (Radford et al., 2021), we mapped these texts into the joint embedding space. Then we used the closest image from the given set to identify the correct candidate image.

2 Background

Word sense disambiguation is an important task in Natural language processing. It typically involves the identification of a word in a given context and requires the selection of the correct sense given several options (Navigli, 2009).

Textual ambiguity has been approached from several different perspectives and with various goals. In Information Retrieval, the ambiguity of a query can lead to unsatisfactory results for users (Cronen-Townsend and Croft, 2002). Ambiguity is also very frequent in technical domains (Frainay et al., 2021). Even in patents, it is widespread, as a large-scale analysis showed (Bertram and Mandl, 2017). Ambiguity is often used for expressing humor but also to express hate by exploiting multimodal semantics (Kalkenings and Mandl, 2022).

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 130–135
July 13-14, 2023 ©2023 Association for Computational Linguistics

Visual word sense disambiguation can be seen from different angles. An image could indicate the correct sense for a word. In the Visual Word Sense Disambiguation, several senses of a word are represented by images. The correct image is indicated by an additional context word for the ambiguous word (Raganato et al., 2023).

Progress in large scale multimodal models facilitates visual word sense disambiguation. Modern multimodal models are able to combine different input modalities and perform a variety of tasks. By mapping input from different modalities into a joint embedding space, such models create a bridge between a visual and textual representation (Radford et al., 2021). These multimodal models can be used for a variety of applications. These include visual question answering, image generation from text (e.g. Dall-e) and image to caption generation (Barraco et al., 2022).

Besides CLIP, there are other models like Google Align (Jia et al., 2021) and Microsoft UniCL (Yang et al., 2022). They differ in their training data or procedures as well as in their architecture.

One big obstacle often mentioned for supervised multimodal models is the availability and quality of natural language training data. Datasets had to be heavily curated (Radford et al., 2021) or must be extremely large when relying on noisy text image relations. For example, for Google ALIGN, over 1.8 billion examples were used (Jia et al., 2021) to generate useful joint embeddings in which visual and textual representations can be aligned. OpenAI's CLIP is the oldest but most renown model in this field, released in February 2021. Shortly after, Google's ALIGN model was released in May 2021. Both models build pairs of text inputs and images to gain joint representations. Microsoft's UniCL from April 2022 combines a third layer by also using labeled images, which are used e.g. for image classification, to gain a more robust image representation (Yang et al., 2022). We decided to use CLIP as the basis for our approach because various model sizes are made available with the capability to modify inputs and outputs rather easily.

3 System Overview

The key components used in our approach are CLIP and the text generation frameworks GPT-2 and OPT. The generated text is additionally used in CLIP to disambiguate the given terms since only a few words were provided by the task. We assume that text generation models are powerful at providing additional text. By generating an additional short paragraph of text based on the ambiguous target word and the provided phrase, we exploit the knowledge included in the language generation models. The objective of this approach is that the generative model solves the ambiguity and adds more words for the correct semantics.

We used the CLIP model, which contains extensive knowledge due to the pre-training on the relationship between text and visual information. We applied the Visual Transformer, which was pretrained on ImageNet. Again, we assume that the pretrained model contains the relevant knowledge.

The processing pipeline for our approach contains the following stages.

- *Data processing*: The target word, phrase, and names of candidate images are extracted.
- *Input organization*: All images are loaded and resized to ViT size (224,224). The additional phrase is used for text generation.
- *Joint embedding construction*: CLIP is used for processing the text and image inputs. The model produces joint embeddings that contain information from both input types.
- *Ranking formulation*: A ranking of the images for each term is derived from the produced embedding vector.

The pipeline is configured in various settings that particularly use multiple textual inputs. Overall, four text inputs are made available within our approach. The following list shows them with an example in italics.

- Given input phrase: *goal, football*
- Ambiguous word: *goal*
- Phrase without ambiguous word: *football*
- Generated text (from the input phrase): *Football goals are the best I'm a goalie and I agree. I'm a goalie and I agree with you. I'm a goalie and I agree with you.*

4 Experimental Setup

We implemented several configurations to examine the effect of using various textual inputs. Our different experiment setups are displayed in table

ID	Models	Inputs	Sequ. leng.	Penalty
CB-1	CLIP-B	Target	-	-
CB-2	CLIP-B	Target, phrase, second word	-	-
CB-3	CLIP-B, GPT2	Target, phrase, second word, generated text	50	-
CB-4	CLIP-B, OPT	Target, phrase, second word, generated text	40	1.0
CB-5	CLIP-B, OPT	Generated Text	40	1.0
CL-1	CLIP-L	Target	-	-
CL-2	CLIP-L	Target, phrase, second word	-	-
CL-3	CLIP-L, OPT	Target, phrase, second word, generated text	40	1.0
CL-4	CLIP-L, OPT	Generated text	40	1.0
CL-5	CLIP-L, OPT	Target, phrase, second word, generated text	60	1.5
CL-6	CLIP-L, OPT	Generated text	60	1.5
CL-7	CLIP-L, OPT	Target, phrase, second word, generated text	70	2.0
CL-8	CLIP-L, OPT	Generated text	70	2.0

Table 1: Experiment overview

1 which includes all relevant configurations. We provide "IDs" for our models for better comparison with the results in Section 5.

We mainly divided our results between the smaller CLIP-B model and the larger CLIP-L model. The increase is due to the scaling of the attention heads and hidden states. For a better comparison to later experiments with generated text inputs, we measured the performance of CLIP based on only the ambiguous word, the input phrase, and the phrase without the ambiguous word (CB-1, CB-2, CL-1, CL-2).

Since the CLIP model is trained on sequences of words as inputs and not only keywords, we conducted separate runs with only the generated text as input and the generated text, target word, additional phrase, and the non-ambiguous word of the additional phrase. The generative models all used the same seed set for better comparison when possible.

The text generation models we initially used were GPT-2 and later OPT. We used the GPT-2 model with 355 million parameters and the OPT model with 2.7 billion parameters. We additionally included two model-specific configurations to observe: the maximum number of tokens to generate and the penalty for repetition. The maximal sequence length was thus set to 40, 60, and finally 70 tokens.

To fit all of our inputs into CLIP, we had to limit the text generation to a maximum of 70 tokens, since the default limit for CLIP is 77 tokens. OPT has the possibility of penalizing repetition within a generation. By increasing the penalty, the model is forced to write more diverse sentences.

When inferring with multiple textual inputs, every input creates a softmax distribution for the ten candidate images. By averaging the results over all four inputs, we created the final ranking for the candidates.

During the processing, some issues were observed. A few images in the training dataset are corrupted in some way, mainly truncated. We imported truncated images as they are, which might affect the model's results. Text has to be imported with the right encoding since the data includes letters from different languages like Chinese, Arabic, Latin, and emojis.

The metrics proposed by the task organizers are used to evaluate the performance of the experiments. The first metric is the hit rate. It measures the correctness of the top-1 prediction. The second metric is the mean reciprocal rank (MRR). It is a measurement that reciprocally considers the correct ranking position across all results.

5 Results

Overall, our approach reached the 77th rank out of 98 submissions within the task. It was ranked 55th in the rankings for only English, which we focused on. Our pipeline achieved robust performance in zero-shot settings, that is, without training.

When compared to the leader board in English only, our approach is positioned one rank below the baseline (BL in table 2). The challenge baseline achieved 60.47% hit rate and 0.7387 MRR. The top result from Samsung Research China - Beijing (SRC in table 2) recorded a 84.017% hit rate and

0.8955 MRR. In total, there were 6 submissions that achieved over 80% hit rate and 13 submissions above 70%. The majority of submissions scored between 60% and 70% hit rate, with 33 submissions scoring greater. 21 submissions were above 50% with our results leading this part of the leader board. The remaining 25 results were below the 50% hit rate. Table 2 shows all experiment results performed on train and test sets.

Model	Train		Test	
ID	Hit %	MRR	Hit %	MRR
CB-1	64.7	0.76	32.0	0.51
CB-2	70.8	0.81	54.8	0.70
CB-3	68.9	0.80	55.4	0.70
CB-4	71.1	0.82	56.3	0.71
CB-5	64.8	0.76	48.3	0.66
CL-1	73.2	**0.82**	33.3	0.53
CL-2	81.9	0.88	52.1	0.68
CL-3	82.5	**0.89**	**57.3**	**0.72**
CL-4	74.6	83.1	53.0	69.1
CL-5	**83.2**	**0.89**	56.7	**0.72**
CL-6	74.3	0.83	52.3	0.68
CL-7	80.6	0.87	56.7	0.72
CL-8	10.5	0.3	52.8	0.69
BL	-	-	60.5	0.74
SRC	-	-	84	0.90

Table 2: Comparison of results

In general, the results reveal significant differences between the performance of our approach on the two sets, although we performed no specific training. This tendency is observed regardless of the inclusion of generated text. In accordance with the reported CLIP results in (Radford et al., 2021), robust zero-shot performance can be observed from the training set. The best performance is seen from CL-5 with a hit rate of 83.2%. Comparing these results to the available training phase results on the competition page of the task, this would have been in the top three results out of 26 submissions [1]. However, a much lower performance can be noticed on the final test set. The hit rate of 57.3% is seen from CL-3. Other submissions had some very good results on the test data, like Samsung Research China with 84% hit rate. Specific optimization for the task can increase performance.

The results also reveal that additional textual inputs do improve the disambiguation performance of the model. In particular, the use of averaged embeddings of multiple text inputs tends to show performance improvements. However, the text that is generated with a greater repetition penalty and number of tokens seems to make the model lose focus on the target term's semantics. Further improvements are noticed when the averaged embeddings are used in a larger CLIP model. Interestingly, the improvements are clearly identified only from the training results.

Finally, the use of larger language generative models improved the disambiguation ability by a small margin. The effect of using different language models for generation is noticed by comparing the results of CB-3 and CB-4. A slight performance improvements are identified from both train and test sets. The OPT model, which consists of a much larger number of parameters (2.7 B) than GPT-2 (355 M), is considered suitable for target term disambiguation based on its extensive knowledge of word semantics.

The third entry of table A1 shows a more abstract example. The given input phrase is "lift, raising" and the target image displays a person lifting another person. Since this describes an activity, it is very hard for our approach to put this input in the correct context, and defaults to the most common example. An image captioning model like BLIP might be able to describe the activity in the scene and give more valuable insights.

The best results for the test data were produced with a shorter generated sequence of 40 tokens and no repetition penalty (1.0). Even though the low repetition penalty model scored the best on the test dataset, the generated text can be misleading for the model's prediction. In the globe example of table A1 and table A2 the model predicted the wrong candidate image. This is likely due to the additional attribute "wood" in the generator text. The higher penalized model instead includes the part "[...] a spherical object that represents earth in space." Such an explanation of a globe with a focus on visual aspects can direct the model to the correct answer. As the prediction accuracy drops with a higher penalty, as earlier described, this could indicate that the more complexly generated texts are misleading in some cases. It can be concluded that the text generation models provide more knowledge for ambiguous words than CLIP itself. They are able to put the ambiguous word into the right context and, as a result, improve overall prediction

[1] https://codalab.lisn.upsaclay.fr/competitions/8190#results

accuracy.

6 Conclusion

This paper showed a robust approach for image disambiguation relying only on pre-trained models and without any training. In future work, we intend to linearly weigh the four text inputs differently and find optimal settings through a grid search. We expect that such a straightforward approach, which was applied similarly for other cases with several text inputs (Madhu et al., 2023) can improve the performance.

References

Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In *Proceedings of the IEEE/CVF conference on Computer Vision and pattern recognition*, pages 4662–4670.

Jens Bertram and Thomas Mandl. 2017. Ambiguity in patent vocabulary: Experiments with clarity scores for claims and descriptions. In *2017 9th International Conference on Knowledge and Smart Technology (KST)*, pages 365–370. IEEE.

Steve Cronen-Townsend and W. Bruce Croft. 2002. Quantifying query ambiguity. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 104–109, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Clément Frainay, Yoann Pitarch, Sarah Filippi, Marina Evangelou, and Adnan Custovic. 2021. Atopic dermatitis or eczema? consequences of ambiguity in disease name for biomedical literature mining. *Clinical & Experimental Allergy*, 51(9):1185–1194.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Milan Kalkenings and Thomas Mandl. 2022. University of Hildesheim at SemEval-2022 task 5: Combining Deep Text and Image Models for Multimedia Misogyny Detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15,*, pages 718–723. Association for Computational Linguistics.

Hiren Madhu, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. 2023. Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. *Expert Systems with Applications*, 215:119342.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19141–19151. IEEE.

A Appendix

All models are downloaded from huggingface. The following versions of the libraries were used:

- pytorch = 1.12.1
- torchvision = 0.13.1
- transformers = 4.26.1
- pillow = 9.4.0

Table A1 shows some of the predictions of our system for item within the test set.

Table A2 shows generated texts to the previous examples of our system.

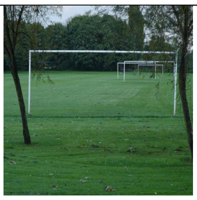
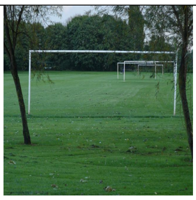
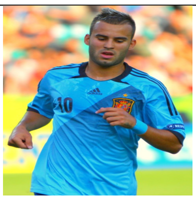
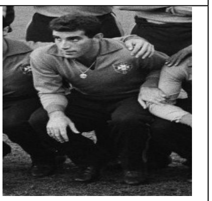
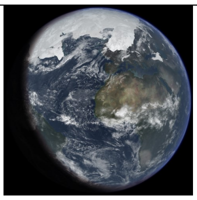
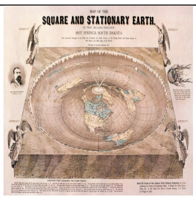
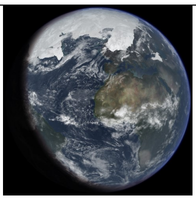
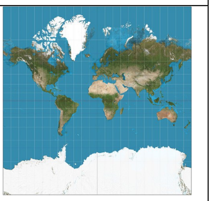
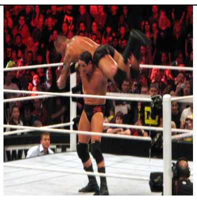
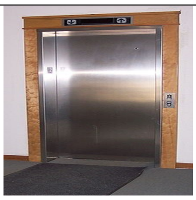
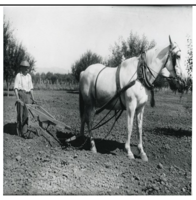
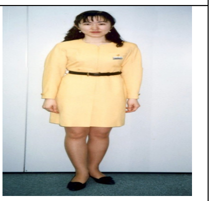
Target word	Phrase	Correct image	First three predicted images		
goal	football, goal				
globe	earth, globe				
lift	lift, raising				

Table A1: Examples for predictions of our system

Input phrase	Low repetition penalty	High repetition penalty
Football, goal	Football goales are the best I'm a goalie and I agree. I'm a goalie and i agree with you. I'm a goalie and I agree with you. I	Football goales are the best I'm a goalie and I agree. It's so much fun to play with your friends, but it's also really hard. I've played soccer for years and I can't even imagine playing goalie. It's not that bad. You just have to be able to read the game well
Globe, earth	The earth globe is a globe that is made of wood. It is a popular item for children to play with. It is made of wood and is made to loo like the	The earth globe is a spherical object that represents the Earth in space. It is used to represent the Earth in astronomy, geology, and other fields of study. The earth globe is made of glass or plastic and is usually mounted on a stand. History The first earth globe was created by John Flamst
Lift, raising	lift raising is a good idea.	Lift raising is a good idea. I would also suggest getting a new belt. The one you have looks like it's been through the wringer. I've had this belt for about 3 years now, and it's still in great shape. It's just that my old belt was too big for me, so I got a

Table A2: Examples for generated texts of our system