teamPN at SemEval-2023 Task 1: Visual Word Sense Disambiguation Using Zero-Shot MultiModal Approach

Nikita Katyal * Siam Makro
Pawan Rajpoot * MUST Research
Subhanandh Tamilarasu * Rakuten India Pvt. Ltd.
Joy Mustafi MUST Research

Abstract

Visual Word Sense Disambiguation shared task at SemEval-2023 aims to identify an image corresponding to the intended meaning of a given ambiguous word (with related context) from a set of candidate images. The lack of textual description for the candidate image and the corresponding word's ambiguity makes it a challenging problem. This paper describes teamPN's multi-modal and modular approach to solving this in English track of the task. We efficiently used recent multi-modal pre-trained models backed by real-time multi-modal knowledge graphs to augment textual knowledge for the images and select the best matching image accordingly. We outperformed the baseline model by 5 points and proposed a unique approach that can further work as a framework for other modular and knowledge-backed solutions.

1 Introduction

As per SemEval-2023 task 1 definition for Visual-WSD Raganato et al. (2023). "Given a word and some limited textual context, the task is to select among a set of candidate images the one which corresponds to the intended meaning of the target word." The final evaluation data in this task has 3 language tracks English, Italian, and Farsi.

An initial trial data set was released with 16 samples, whereas the final training dataset was released with 12869 samples, where each pair of text and contextual text were mapped to a target image that has to be retrieved from a set of 10 images. The images attached to each of the training samples overlap with other samples. And thus, a total of 12999 unique images were found in the training data pool. These images vary from a wide variety of genres. A sample input data example can be seen in fig. 1. Here, 'white' is the term, and 'white yolk' is the context. The model has to retrieve the right image from the set of input images.

*Equal Contribution

The text consists of words and context covering

Figure 1: Example visual word sense disambiguation, the ambiguous word in this example is "white" and the context is "white yolk". Here the positive image is highlighted by green and negative images are highlighted in pink.

various domains; even the context's intentions vary with different samples. A word cloud of the text involved in the train data can be seen in 2. However, a significant proportion of the dataset looks to retrieve images based on their scientific terms, especially plants, birds and rocks. This can be clearly interpreted from the word cloud containing the terms genus, family, herb, shrub, and minerals.

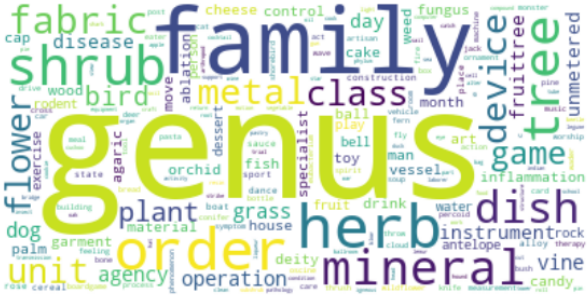


Figure 2: Word cloud of dataset

2 Related Work

Starting from Radford et al. (2021) there has been a recent increase in multi-modal vision and language research (Li et al., 2022; Jia et al., 2021; Li et al., 2021). Also, Transfer Learning methods have aided in the advancements of these models (Sung et al., 2022). Complex models trained on multi-modal

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 457–461

data on web scale can be fine-tuned to domain-specific and task-specific problems. Recent works like (Li et al., 2022) has pushed the boundaries of general and open-domain understanding of these models and they outperform traditional domain-specific trained models when used in zero-shot settings. For this specific task, the baseline model uses CLIP to compute the text and image embeddings, and rank the candidate images based on the cosine similarity scores between the text and image pairs. Although these models' understanding is good in describing the image, there is a lack of image-related background knowledge which can provide more context to the image description.

3 Approach

3.1 Module 1: BLIP model

With the development of a plethora of advanced models, we chose BLIP (Li et al., 2022) a recent vision-language pre-trained model for unified image-grounded text understanding and generation tasks. We preferred BLIP over CLIP (Radford et al., 2021) (both were conveniently available at the time of this work) because of its ability to reduce noisy captions by using filters and also strong generalization ability, by which it can provide great results in a zero-short setup.

BLIP is pre-trained with joint optimization on three objectives. Out of which, two loss functions; Image-Text Contrastive Loss (ITC) and Image-Text Matching Loss (ITM) helped the model understand the text-image similarity. **Image-Text Matching Loss (ITM)** activates an image-grounded text encoder that makes the model predict if an image-text pair is positive or negative given their multimodal feature.

We instantiated **BLIP w/ ViT-L** model to preprocess the input and provide inference using the ITM head. And, for each input sample record, we paired each of the images in the sample to the contextual text. These pairs are passed to the BLIP model. The model measures the similarity of all these images to the input contextual text and provides a probability ranging from 0 to 1.

After running through the inference on train data, the images were ordered on the basis of similarity score to measure top K accuracy. The results are tabulated in 2.

Since the terms and their contextual meanings can be further expanded with meaningful words, we decided to leverage **Term expansions** for each

Top K	accuracy
Top 1	63%
Top 3	85%
Top 5	96%

Table 1: Top K accuracy on train dataset using BLIP w/ ViT-L.

of the input text data during inference. In the past, there had been several approaches tht focused on Query expansions to improve the effectiveness of Information retrieval (Azad and Deepak, 2019). There are multiple open data APIs to support term expansion. We experimented with Wikipedia API to search for synonyms and hypernyms for input text queries (both terms and contextual terms).

Thus tackle this, a semantic filter was introduced as an intermediate stage, it helped in ordering the keywords in a semantic sense and thereby provided a way to filter out less relevant terms from the expansion set. This **semantic filter** was powered by a sentence transformer. Here we used all-MiniLM-L6-v2 which is a version of miniLM trained on retrieval data (Wang et al., 2020; Joshi et al., 2017; Bowman et al., 2015; Fan et al., 2019) to perform text similarities within the semantic filter component.

Here, we list a few examples of contextual texts along with their term expansions along with semantic similarity scores:

- gym dip: Climbing gym (0.54), Outdoor gym (0.50), Street workout (0.46), Exercise equipment (0.44), Sport climbing (0.38), Indoor climbing (0.33)

- fledged Cygnus: Cygnus (genus) (0.71), Birds of Patagonia (0.31), Black-necked swan (0.24), Bald eagle (0.22), Greylag goose (0.22)

Only Contextual terms were used for all the above approaches due to their better retrieval performance. In order to incorporate terms that may add more generalized meaning to the retrieval mechanism, we came up with a **Fusion switch** mechanism that considered the prediction results of the text-image pairs instead of contextual word-image pairs when the top 1 accuracy value of these pairs is higher than the latter for any input record.

All the above-mentioned test results are aggregated in 3.

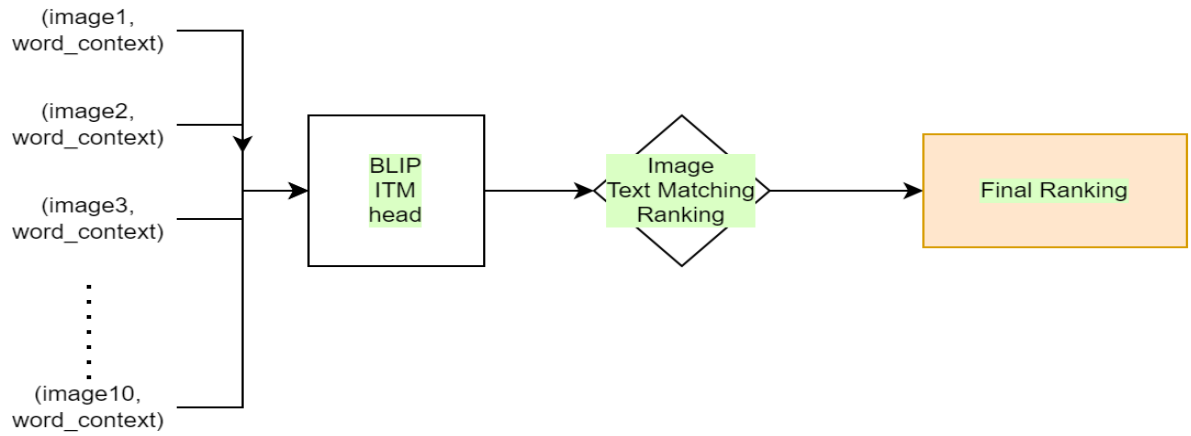


Figure 3: Module1: BLIP ITM Approach

Top K	accuracy
Top 1	49.02%
Top 3	65%
Top 5	81%

Table 2: Top K accuracy on train dataset using MMKG Grounding.

3.2 Module 2: Multi-Modal Knowledge Graph Grounding

For augmentation we ground the images to VisualSem Knowledge Graph (Alberts et al., 2020), we used pre-trained CLIP (Radford et al., 2021) based embedding to extract the central entity from the image. Entities are extracted by measuring similarity between MMKG node definition's and input Images. As discussed in the original VisualSem paper When there are more than one English definition associated to a node, the best ranking across all glosses as the ranking of the retrieved node is used. For Image context, we used the entity definition and related entity names from MMKG. We then match the context of images with the complex words (with context). For this, we use the "msmarco-distilbert-base-tas-b" semantic search model trained on MS-MARCO retrieval dataset(Bonifacio et al., 2021) from huggingface. We treat all image contexts as documents and use complex words plus their context as keywords to find the most relevant document.

3.3 Aggregation

We found that because of more relevant context module 2 wherever predicted with greater than 0.9 confidence, and the majority of results were true, also some of the results where module 2 was very confident, module 1 struggled with less context and hence gave less confidence score. Therefore we aggregated the results and after optimizing we finalized on a weighted average where more weight was given to the confidence score of module 2(grid search weight tuning on trial data). Our final weights were given using

$$Confidence = 0.6 * module2 + 0.4 * module1 \quad (1)$$

The equation 1 states aggregation for confidence score calculation.

4 Results

The results of model performance on various approaches are tabulated in Table 3

4.1 Model Performance

Running BLIP w/ ViT-L inference on test data, we were able to get results with top 1 accuracy of 61.1% and Mean Reciprocal Rank (MRR) of 75.4%. Adding term expansion to the existing contextual terms, we improved top 1 accuracy to 62.4% and MRR of 75.2%. Term expansions improved the precision of retrieving the top 1st relevant image, but failed to improve MRR. We found BLIP ITM similarity scores for most of the top 5 and above text-image pairs were almost zero, highly regularizing the results.

All these above-mentioned metrics were reported only for contextual text-image pairs as they performed well against text-image pairs. Fusing the results of text-image and contextual text-image led to improving the top 1 accuracy to 64.7% and MRR to 77.6%.

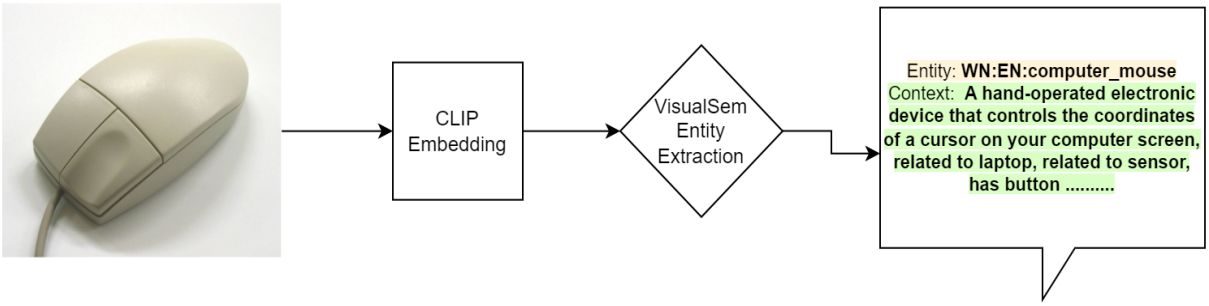


Figure 4: Entity Grounding and Context creation using VisualSem MMKG grounding

Approach	Top 1 accuracy	MRR
Baseline	60.47%	73.87%
BLIP ITM	61.1%	75.4%
BLIP ITM + TE	62.4%	75.2%
BLIP ITM + FTE	64.7%	77.6%
MMKG	49.0%	65.5%
MMKG + BLIP ITM + FTE	66.9%	78.6%

Table 3: Model performance on test dataset (FTE = Fusion Term Expansion, TE = Term Expansion)

Multi-modal KGG was able to get the Top 1 Accuracy of 49.0% with a Mean Retrieval Rank of 65.5%. However, aggregation of Multi-modal KGG, BLIP ITM, and Fusion Term expansion yielded in Top 1 Accuracy of 66.9% with an MRR of 78.6%

4.2 Ablation study

As per our analysis, most of the cases where the BLIP ITM model failed to retrieve correct images, belong to **scientific terms**. This constitutes more than 35% of the data in the training dataset.

Term expansions increased the retrieval by adding a performance improvement of 1.3% to the Top 1 Accuracy. MRR remained almost the same with a little decrease of 0.2%. However, these expanded terms did range from highly similar to generic terms and sometimes even irrelevant terms. Semantic filters removed noisy and irrelevant expansions from the expansion term set.

Fusion of term expansions further enhanced the retrieval by adding a performance improvement of 2.3% to top 1 Accuracy and added 2.4% increase to MRR.

Moreover, Term expansions helped in better retrieval performance of scientific terms and general facts with the source of expansion coming from Wikipedia APIs. Fusion helped in better generalization of retrieval by reducing balancing the regularization effect of the contextual text input query.

For Multimodal Knowledge Graph, we used Entity Gloss along with related Entity names for context, this combination gave us the best results. We also tried with appending extra knowledge for extracted Entities using the "Common Sense Knowledge Graph" (Reimers and Gurevych, 2019), but that decreased the overall performance.

Conclusion

We present a multi-modular solution to the visual-WSD problem with limited resources and a zero-shot approach. In the future, we intend to include more modules for both knowledge augmentation and better sense disambiguation. We also hope that our work motivates future use of large-scale trained visual language models in zero-shot settings in more innovative ways. We thank the organizing committee of SemEval-2023 along with the task-setting team of Task-1 for giving us this opportunity to work on this problem.

References

Houda Alberts, Teresa Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2020. Visualsem: a high-quality knowledge graph for vision and language. *arXiv preprint arXiv:2008.09150*.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. A new approach for query expansion using wikipedia and wordnet. *Information sciences*, 492:147–163.

Luiz Henrique Bonifacio, Israel Campiotti, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: long form question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, page arXiv:1705.03551.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

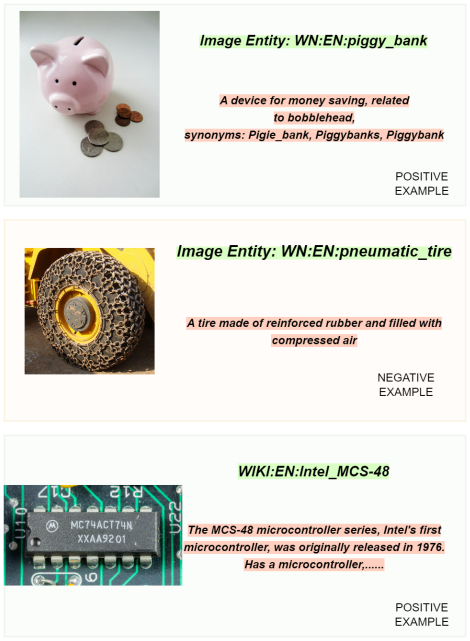
Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 Task 1: Visual Word Sense Disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

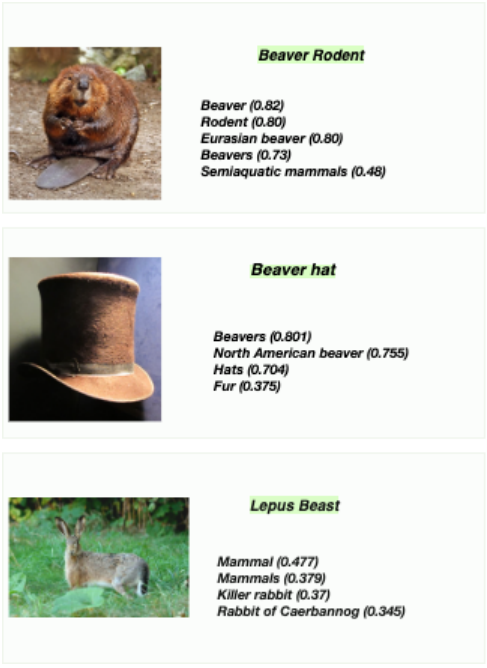
Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

A Appendix: Module Examples



(a) MMKG (Module 2 examples)



(b) Term expansion examples under Module 1