SLT at SemEval-2023 Task 1: Enhancing Visual Word Sense Disambiguation through Image Text Retrieval using BLIP

Mohammadreza Molavi
Department of Computer Engineering
Amirkabir University of Technology
mmdreza.molavi@aut.ac.ir

Hossein Zeinali
Department of Computer Engineering
Amirkabir University of Technology
hzeinali@aut.ac.ir

Abstract

Based on recent progress in image-text retrieval techniques, this paper presents a fine-tuned model for the Visual Word Sense Disambiguation (VWSD) task. The proposed system fine-tunes a pre-trained model using ITC and ITM losses and employs a candidate selection approach for faster inference. The system was trained on the VWSD task dataset and evaluated on a separate test set using Mean Reciprocal Rank (MRR) metric. Additionally, the system was tested on the provided test set which contained Persian and Italian languages, and the results were evaluated on each language separately. Our proposed system demonstrates the potential of fine-tuning pre-trained models for complex language tasks and provides insights for further research in the field of image text retrieval.

1 Introduction

Visual word sense disambiguation (VWSD) organized by (Raganato et al., 2023)[1] is the task of selecting the correct image from a set of candidate images that corresponds to the intended meaning of a target word, given a limited textual context. SemEval 2023 Task 1 focuses on VWSD in the context of images and textual descriptions. VWSD helps to improve the performance of many natural language processing (NLP) applications that involve both textual and visual information. For example, in image captioning(Hossain et al., 2019), the system needs to identify the correct sense of a word in a given context to generate an accurate and meaningful caption. Similarly, in visual question answering (VQA) (Antol et al., 2015), the system needs to understand the meaning of the question and the visual context to provide a correct answer.

Image retrieval aims to retrieve images that are relevant to a given textual query, such as a caption or a keyword. This involves learning a representation of both images and text in a shared embedding space, where the similarity between images and text is determined by their proximity in the embedding space. The objective is to locate the most relevant images based on the textual query and present them in a ranked order. (Datta et al., 2008)

Vision-language pre-training has emerged as a powerful approach for multimodal understanding in recent years (e.g., CLIP (Radford et al., 2021), ALBEF (Li et al., 2021), SimVLM (Wang et al., 2021)). This approach involves jointly training models on large-scale image and text datasets to learn a shared representation that captures the semantic relationship between visual and textual inputs. Task organizers selected CLIP as a baseline approach. CLIP has emerged as a powerful approach for multimodal understanding and leverages a vast amount of supervision by pre-training on a dataset of 400 million (image, text) pairs collected from the internet. Using CLIP as a baseline, we aim to demonstrate the effectiveness of our proposed approach for the VWSD task. Vision-language pre-training has achieved significant success in various downstream tasks, such as image captioning, visual question answering(Zhou et al., 2020), and image-text retrieval(Liu et al., 2021). By leveraging the joint representation learned during pre-training, these models can effectively reason about the relationships between visual and textual information(Chen et al., 2023), leading to improved performance on these tasks.

Our approach to VWSD in this task is based on Vision-Language pre-training, specifically Bootstrapping Language-Image Pre-training (BLIP) (Li et al., 2022). BLIP trains a model to predict the relationship between an image and its corresponding textual description, and has been shown to be effective for a range of vision-language tasks, including VWSD.

In this paper, we describe our system for Se-

[1] https://raganato.github.io/vwsd/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1921–1925
July 13-14, 2023 ©2023 Association for Computational Linguistics

mEval 2023 Task 1, which uses BLIP pre-training to generate image embeddings that capture the semantic information of the images. Our system then utilizes these embeddings to perform VWSD by computing the similarity between the target word's embedding and the embeddings of the candidate images. That also incorporates an attention mechanism to further refine the similarity scores.

We implemented the Bootstrapping Language-Image Pre-training (BLIP) method by adapting an open-source implementation available on GitHub[2]. We made some modifications to the code to suit our specific requirements, like finetuning and evaluating the method on the VWSD task.

2 Background

2.1 Task Definitions

The VWSD task involves selecting the correct image from a set of candidate images given a phrase that may contain an ambiguous word, which can have multiple meanings based on the context. In this case, the system's task is to disambiguate the meaning of the word and select the correct image based on that meaning. For example, consider the phrase “bank erosion". The word "bank" can refer to a financial institution or the side of a river, and the word "erosion" can refer to the gradual destruction of something over time. In the context of the VWSD task, the system's task would be to disambiguate the meaning of the word "bank" and select the appropriate image based on the intended meaning. If the system identifies "bank" as referring to the side of a river, it would select an image of a riverbank undergoing erosion.

2.2 Dataset and Metrics

The dataset for the VWSD task was provided by the task organizers and consists of 12,869 ambiguous textual phrases and 12,999 accompanying images. The dataset covers a diverse range of concepts and domains and has a substantial size of 17GB. The evaluation metrics used in the task are mean reciprocal rank (MRR) (Craswell, 2009) and hit rate. MRR stands for Mean Reciprocal Rank and is a commonly used evaluation metric in information retrieval and natural language processing tasks. It is often used in tasks where the goal is to retrieve a ranked list of items based on a given query. In the case of VWSD, MRR is used to evaluate how well the system can retrieve the correct image corresponding to the given ambiguous phrase. Specifically, for each ambiguous phrase, the system ranks the candidate images in order of their relevance to the phrase. MRR is then calculated as the average of the reciprocal ranks of the correctly identified images across all the test instances. The MRR metric has the advantage of considering the rank of the correct image in the retrieved list, rather than just whether the system retrieved the correct image or not. This makes it a more nuanced measure of performance than binary evaluation metrics like accuracy.

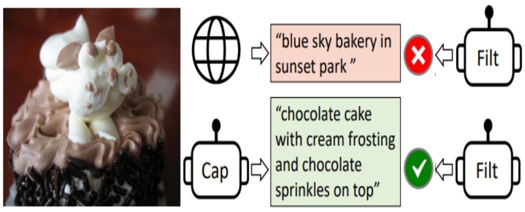
"blue sky bakery in sunset park"
Filt
Cap
"chocolate cake with cream frosting and chocolate sprinkles on top"
Filt

Figure 1: A synthetic caption generator, referred to as Captioner (Cap), is utilized to generate captions for web images, while a noise filter, referred to as Filter (Filt), is applied to eliminate any noisy captions. Image reference: https://github.com/salesforce/BLIP.

3 System Overview

BLIP is a Vision-Language Pre-training (VLP) framework that transfers flexibly to both vision-language understanding and generation tasks. It is designed to effectively utilize noisy web data by bootstrapping the captions. The framework consists of two stages: bootstrapping and refinement. During the bootstrapping stage, synthetic captions are generated for noisy image-text pairs from the web. The captioner network is trained to generate captions for a given image, and a filter network is used to remove noisy captions. This process is shown in Figure 1. This results in a dataset of high-quality image-text pairs for further pre-training. In the refinement stage, the pre-trained model is fine-tuned on downstream tasks such as image captioning, visual question answering, and image text retrieval. This stage involves training the model on a specific task dataset to fine-tune the pre-trained model parameters to improve task performance.

3.1 Model Architecture

A multi-task model called Multi-modal Mixture of Encoder-Decoder (MED) was proposed in the BLIP's paper to pre-train a model that can both

[2] https://github.com/salesforce/BLIP

comprehend and generate. The model can operate in three functionalities: (1) uni-modal encoder, (2) image-grounded text encoder, and (3) image-grounded text decoder. The MED's text encoder is similar to BERT (Devlin et al., 2018), where the input text is prepended with a [CLS] token to summarize the sentence. Meanwhile, in the image-grounded text encoder, an additional cross-attention layer is added between the self-attention layer and the feed-forward network for each transformer block of the text encoder to inject visual information. Finally, the MED model's image-grounded text decoder differs from the image-grounded text encoder in that it utilizes causal self-attention layers in place of bidirectional self-attention layers. Additionally, the decoder uses a "decode" token to indicate the start of a sequence and an "end-of-sequence" token to signal its conclusion.

3.2 ITC and ITM Losses

BLIP's paper describes a pre-training approach that optimizes three objectives simultaneously with two understanding-based objectives and one generation-based objective. The paper employs Image-Text Contrastive Loss (ITC) to align feature spaces of the visual and text transformers. This loss function promotes similar representations for positive image-text pairs and dissimilar representations for negative pairs. This approach has proven to be effective in enhancing vision and language comprehension. The proposed method follows the ITC loss introduced by Li (Li et al., 2021). In this method, a momentum encoder generates features, and soft labels are produced from the momentum encoder to serve as training targets, accounting for potential positives in negative pairs. The Image-Text Matching Loss (ITM) activates the image-grounded text encoder to develop an image-text multi-modal representation that captures fine-grained alignment between vision and language. It involves a binary classification task where the ITM head of the model predicts whether an image-text pair is positive or negative based on their multi-modal feature. To select more informative negatives, the paper applies the hard negative mining strategy proposed by Li et al. (2021). This strategy chooses negative pairs with higher contrastive similarity in a batch to compute the loss.

Table 1: Different parts of the dataset

Set	# Phrases	Percentage
Train	12295	80 %
Validation	1286	10 %
Test	1286	10 %

3.3 Image-text Retrieval

The system fine-tunes the pre-trained model using ITC and ITM losses. In order to speed up inference, a candidate selection approach is employed. First, k candidates are selected based on the similarity of their image-text features. Then, the selected candidates are reranked based on their pairwise ITM scores. This approach allows for faster and more efficient processing of image-text pairs, enabling the system to generate accurate and relevant captions for a wide range of web images. Figure 2 shows this process. We utilized the text-image retrieval module provided by the BLIP framework to tackle the visual word sense disambiguation (VWSD) task. To adapt the pre-trained BLIP model to the VWSD task, we fine-tuned it on the task-specific dataset. Specifically, we used the training set to optimize the model's parameters using the ITC and ITM losses described earlier. During fine-tuning, the model learned to better capture the semantic relationships between images and text in the VWSD domain. This demonstrates the effectiveness of our approach in learning task-specific image-text representations.

4 Experimental Setup and Results

4.1 Experimental Setup

To evaluate the performance of the fine-tuned model, we used the VWSD task dataset, which we further divided into the train, validation, and test sets according to Table 1. We fine-tuned the BLIP's Image text retrieval model for two epochs with a batch size of 32, a learning rate of 1e-5, and a weight decay of 0.05.

4.2 Results

To evaluate the performance of our system, we use the Mean Reciprocal Rank (MRR) metric that is a metric commonly used in information retrieval to evaluate the quality of ranked retrieval results. MRR is calculated as the reciprocal of the rank of the first relevant document retrieved for a query, averaged over all queries. In the context of the

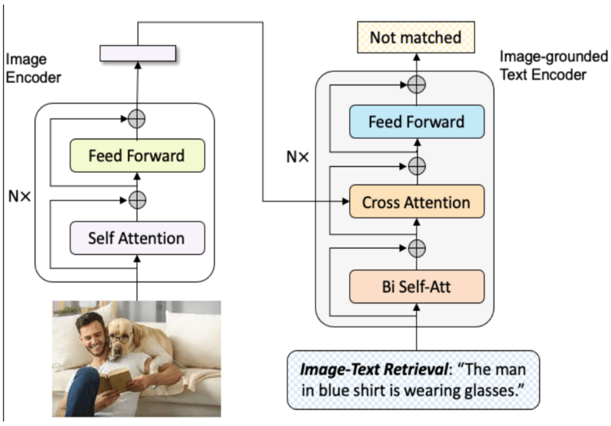
Image Encoder
N×
Feed Forward
Self Attention
Not matched
Image-grounded Text Encoder
N×
Feed Forward
Cross Attention
Bi Self-Att
Image-Text Retrieval: "The man in blue shirt is wearing glasses."

Figure 2: Image-text retrieval system that uses a pre-trained model fine-tuned with ITC and ITM losses, and a candidate selection approach based on image-text feature similarity and pairwise ITM scores, resulting in accurate and relevant captions for web images (Li et al., 2022).

Table 2: Evaluation of different batch sizes for fine-tuning BLIP's Image Text Retrieval on the VWSD task dataset.

Batch size	MRR
8	68.11
12	68.93
16	71.32
32	74.67
38	70.31

VWSD task, MRR is used to evaluate how well the system can retrieve the correct image-text pairs based on the input query. A higher MRR score indicates better performance, as it means that the correct image-text pair is more likely to be ranked higher in the retrieval results. Therefore, MRR is a suitable metric for evaluating the performance of our system in the VWSD task. In order to fine-tune the neural network on the VWSD task dataset, several hyper-parameters were adjusted to optimize the system's performance. Batch size is one of these important hyper-parameters that can have a significant impact on the training process. To find the optimal batch size for our system, four different values were tested in separate tuning files. The performance of the system was then evaluated on a separate test dataset according to Table 1. The results are shown in Table 2, which demonstrates the impact of batch size on the system's Mean Reciprocal Rank (MRR) metric. Based on the results, our final model was tuned with a batch size of 32.

Table 3: Performance Comparison of the Proposed System with CLIP Baseline on the Challenge Test-set in English, Persian, and Italian Languages using MRR (%) as the Evaluation Metric.

Language	CLIP	Proposed system
English	66.8	72.05
Persian (Farsi)	38.9	63.14
Italian	37.5	68.55
Average	47.7	67.91

During the testing phase on the challenge test set, we encountered an issue as the test set contained not only English but also Farsi and Italian languages. To handle this problem, Google Translate was used to translate the non-English text into English. The results of the proposed system for each language as well as the average per for all languages are presented in Table 3. Our experimental results demonstrate that our proposed approach, which involves translating non-English text, such as Farsi and Italian, into English, leads to a considerable improvement based on the MRR metric.

5 Conclusion

Based on the experimental results, we can conclude that our system, which fine-tunes BLIP's Image Text Retrieval model on the VWSD task dataset using the ITC and ITM losses, performs well in estimating similarity between image and text. The candidate selection approach employed in our system also speeds up the inference process, allowing for more efficient processing of image-text pairs. Overall, the proposed system shows promising results in VWSD and has the potential to be applied in various domains such as image retrieval, recommendation systems, and virtual assistants.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Nick Craswell. 2009. *Mean Reciprocal Rank*, pages 1703–1703. Springer US, Boston, MA.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining 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.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134.

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, Jose Camacho-Collados, Asahi Ushio, and Mohammad Taher Pilehvar. 2023. Semeval-2023 task 1: Visual word sense disambiguation (visual-wsd). In *17th International Workshop on Semantic Evaluation (SemEval-2023)*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.