SRCB at SemEval-2023 Task 1: Prompt Based and Cross-Modal Retrieval Enhanced Visual Word Sense Disambiguation

Xudong Zhang[✉], **Tiange Zhen**[*], **Jing Zhang, Yujin Wang, Song Liu**
Samsung Research China - Beijing (SRC-B)
{xudong.z1, tiange.zhen, jing97.zhang, yujin.wang, s0101.liu}@samsung.com

Abstract

The Visual Word Sense Disambiguation (VWSD) shared task aims at selecting the image among candidates that best interprets the semantics of a target word with a short-length phrase for English, Italian, and Farsi. The limited phrase context, which only contains 2-3 words, challenges the model's understanding ability, and the visual label requires image-text matching performance across different modalities. In this paper, we propose a prompt based and multimodal retrieval enhanced VWSD system, which uses the rich potential knowledge of large-scale pretrained models by prompting and additional text-image information from knowledge bases and open datasets. Under the English situation and given an input phrase, (1) the context retrieval module predicts the correct definition from sense inventory by matching phrase and context through a biencoder architecture. (2) The image retrieval module retrieves the relevant images from an image dataset. (3) The matching module decides that either text or image is used to pair with image labels by a rule-based strategy, then ranks the candidate images according to the similarity score. Our system ranks first in the English track and second in the average of all languages (English, Italian, and Farsi).

1 Introduction

Word Sense Disambiguation (WSD) is a challenging task of Natural Language Processing, which aims at identifying the polysemic word with correct interpretation (Bevilacqua et al., 2021). It has numerous applications, such as machine translation, search engines, and text understanding. The traditional WSD task provides the predicting word with its context, which is supposed to be a sentence containing the word, and chooses the proper sense from the predefined inventories (Miller, 1995; Navigli and Ponzetto, 2012). The model is asked to extract the semantic feature of the target word from the particular text and match it with the most practical sense. The WSD task can be processed as a classification problem among the senses as labels in the given inventory by taking advantage of the neural language model's powerful semantic understanding capability. The appearance of Visual WSD connects the different modalities and gives the word sense a new presentation form. The sense inventory now is a set of images instead of text definitions, and the target word has to be paired with the corresponding image. The Vision-Language model, CLIP (Radford et al., 2021), has shown its outstanding performance on unifying image-test modalities, and it provides potential knowledge for semantic understanding required by disambiguation.

The SemEval-2023 Task-1 Visual Word Sense Disambiguation (VWSD) (Raganato et al., 2023) is a multilingual WSD task, for English, Italian, and Farsi, with the phrase as textual context and images as visual senses. Most of the system setup in this paper is under the English track. The other two languages are translated to English and only apply a partial sub-system without image sense involvement. Under the English situation, the phrase contains the target word and has a length of 2-3, which differs from the traditional WSD task with the sentence as context. This challenges the encoder to extract sufficient semantic information from the target word from the short text. Image sense labels require a highly unified image-text connection to produce the final output. For Farsi and Italian, although the specific language version of CLIP and multilingual CLIP are available (Bianchi et al., 2021; Sajjad Ayoubi, 2022; Carlsson et al., 2022), the actual performance in the current task is limited by the scale of pre-training data.

Our system makes the best use of the potential knowledge of the pre-trained model by prompting rather than continuous training and fine-tuning. We

[*]Contribution during Internship in Samsung Research China-Beijing.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 439–446

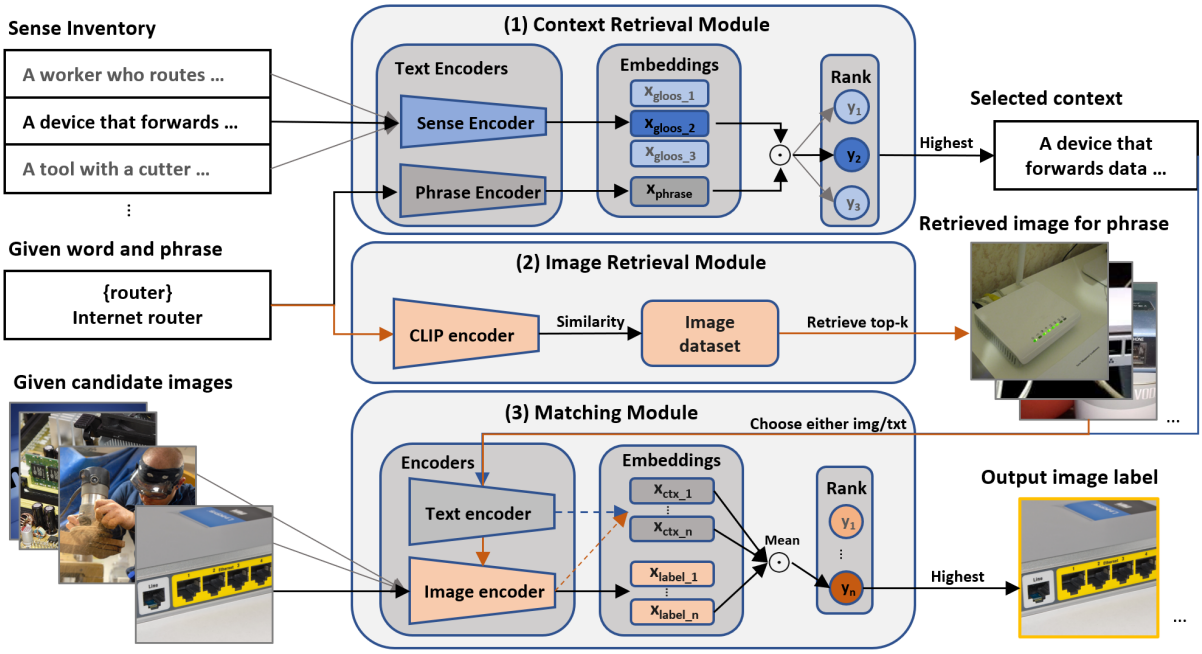

Figure 1: The overall architecture of our system. Three main modules: (1) Context Retrieval module performs disambiguation and retrieves best definitions, (2) Image Retrieval module collects relative images from LAION, and (3) Matching module computes the similarity with candidates image labels.

gather information on the target word from semantic networks and online dictionaries. The most pertinent information is selected by matching the augmented phrase through a biencoder architecture and regarded as the prompt. This biencoder architecture was proposed in BEM (Blevins and Zettlemoyer, 2020),which jointly trains context and gloss in the same vector space. SimCSE (Gao et al., 2022) involves contrastive loss to maximize the similarity of sentences with identical meanings, which makes it a powerful tool for matching phrases and definitions. Therefore, we use SimCSE as the backbone of our biencoder architecture. The later work (Wu et al., 2022) improves the performance of matching text with different lengths. There is a recent multimodal WSD dataset (Calabrese et al., 2020) that uses images to represent the sense of the concept. Beyond the text information, we use additional images from the open image dataset. Different types of words require different modalities to express. Therefore, we design a rule-based process that decides whether text prompt or images are used for final sense representation. Finally, we use a huge version of CLIP to extract features and rank the candidate images according to similarity scores.

2 System Overview

The system description in this section is under the English situation. Italian and Farsi are translated to English and only apply a partial sub-system without image sense involved. For English, we divide the VWSD task into three main modules, including (1) context retrieval module, (2) image retrieval module[1], and (3) matching module, as shown in Figure 1. The two retrieval modules perform disambiguation among senses and are regarded as a WSD-specific modules.

2.1 Context Retrieval Module

In this module, the most suitable text prompt interpreting the meaning of the target word will be selected. We use the definitions and synonyms as the context and match them with the given phrase, which is a similar setup to the traditional WSD task. However, this approach still needs more information on the short-length phrase when matching with context. We use translation to gain extra knowledge by back-translating the phrase from another language to English with multiple online translators (Luan et al., 2020) and concatenate the results. Then, we put the augmented phrase and the customized context into the biencoders, which

[1]This module is not applied on Italian and Farsi.

separately encode phrase and context as shown in the Text Encoder part of Figure 1. To further improve the WSD performance, we apply weighted similarity to minimize the distance to the phrase. We disambiguate the confused senses of the target word and maximize the similarity with the phrase with equation 1,

$$\begin{aligned} Sim_i &= \mathbf{X}_i \cdot \mathbf{X}_{context}^{\top} \\ Sim_{final} &= Sim_{phrase} - \alpha * Sim_{word} \end{aligned} \quad (1)$$

and α determines the disambiguation weight. As a result, compute the similarity between phrase and candidate contexts, which gives the nearest embeddings to represent the correct sense.

2.2 Image Retrieval Module

Instead of constraining in text space, we are inspired to obtain cross-modal information by BabelPic (Calabrese et al., 2020), a multimodal dataset for non-concrete concepts. Although the domain of BabelPic does not match with the task dataset, it is observed that there are a considerable amount of specific entities in both training and testing datasets that benefit from image information. Therefore, we collected the extra image data using clip retrieval (Beaumont, 2022), which retrieves images according to the similarities between phrase and LAION image embeddings.

2.3 Matching Module

With the text context and image sense as the inputs, the final module will match them with candidate image labels and produce the image ranking according to similarities. The decision function chooses either text context or image sense to be used for final matching. We investigated the BabelNet information and found that it provides valuable properties of target words. We designed a series of rules to decide using text or images based on the word properties. The selection strategy is designed according to the sense properties of BabelNet shown in Figure 2, and an overall decision process is based on the error analysis of training data. For instance, named entities of geographic places and biological creatures benefit from images and have a higher similarity than text context. However, concepts have various visual representations, and images might involve more mistakes than text prompts. Then the final context is processed through the corresponding encoder with the image labels. The encoders are from a large pre-trained VLM, and we believe that the pre-training data contains potential knowledge that benefits WSD purpose and the alignment capability, which is significant to obtain correct text-image pairs. Therefore, we use a huge version of CLIP trained on LAION-5B as our module backbone for feature extraction.

# of Senses	Train	Test$_{EN}$
1	**73.5%**	10.8%
2	11.9%	16.9%
≥ 3	14.5%	**72.2%**

Table 1: The estimated number of senses in training and testing data on English

2.4 Sense Inventory and Data Augmentation

The WSD task resource is an inventory containing all possible senses of the target word, and the widely-used sources are semantic networks such as WordNet and BabelNet. As the given phrase is too short for the model to extract information, we collected extra data from Wikipedia, WordNet, BabelNet, and other online dictionary using the target word (phrase if available). The details of data collection are introduced in Section 3.1. The online dictionaries involve extra text knowledge over the traditional knowledge base, which is proved to be useful (Bevilacqua et al., 2021). The quality of sense inventory significantly affects the WSD results, and a good inventory should contain sufficient and understandable information for the model to match with phrases and image labels. Our sense inventory contains abundant and high-quality information on the predicting words after a post-processing strategy.

3 Experimental Setup

This section will discuss the detailed data usage and experiment setup. As there is a distribution gap between training and testing data, most of the setup is based on the testing and customized validation sets.

3.1 Data Usage

Data splits There is a huge sense distribution gap between training and testing data, as shown in Table 1. The number of senses is referred to the online dictionaries rather than BabelNet, which prefers to contain as many senses as possible. Most of the training set is monosemy, which rates over 70% and most are under biology, rather than pol-

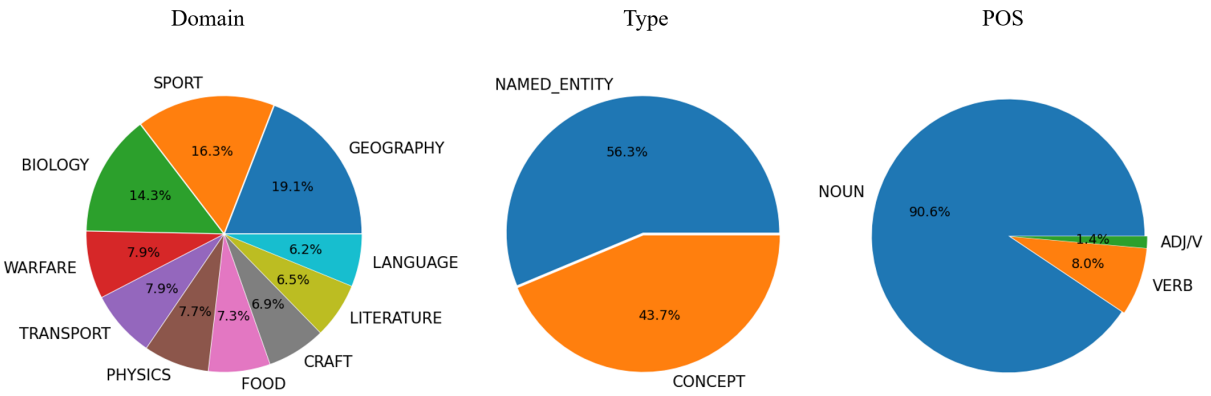

Figure 2: The properties of target senses in testing set

ysemy, which is expected to be processed by the WSD task. The reversed situation is applied to the testing set. To bridge this gap, we randomly select target words with certain number of senses and construct a validation set with the same distribution as the testing set.

Extra data The external gloss inventory is formed by context from multiple sources. For Wikipedia and online dictionaries, we only retrieve the definitions of the target word. For semantic networks, including WordNet and BabelNet, we gather synonyms, categories, domains, and types besides definitions.

3.2 Experimental Setup

The prompt template decides the exploring depth of the pre-trained model, and it requires the right information with the right position. Although the learnable prompt is proven to benefit downstream tasks, it has limited generalization on zero-shot problems (Zhou et al., 2022b,a). This task data is proved as zero-shot and long-tail, and we applied a fixed prompt template. The essential elements of the prompts are synonyms and glosses, which contributes a lot to matching phrase and images. By adding target phrase and connecting words, the final template is

[Phrase], as known as [Synonyms], is [Gloss].

e.g. Reflecting glass, also known as looking glass, is a mirror; usually a ladies' dressing mirror

and the order of the elements has slight influences on the results.

We use Microsoft, Google translates, and DeepL to perform back-translation on phrases and concatenate the result to pair with prompt context through a SimCSE-based biencoder. For image collection, clip retrieval (Beaumont, 2022) is a tool that allows users to retrieve images with given text, which encodes text and images to clip embeddings and calculates the distances. Finally, the huge version of open-clip trained on LAION-5B (Cherti et al., 2022) is applied to match text-image or image-image pairs and rank the candidates by similarity scores.

The evaluation measures below are used as marking metrics:

- **Hit rate** is the accuracy of correctly ranking the gold image in the first place.
- **Mean reciprocal rank** (MRR) is a measurement evaluating a ranked list as shown in the following equation 2.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (2)$$

4 Results and Discussion

In this session, we mainly analyze the English results of the approaches in our system and discuss how it contributes to the results. The training and testing data are further compared and analyzed in the behavior of different training strategies.

4.1 Final results

Our system achieves a hit rate of 84 on the English testing set, which ranks 1st out of 56 teams on the English leaderboard. The baseline approach uses the Most Frequent Sense from sense inventory as the input context and matches it with image labels. There is a huge gap between the results of the

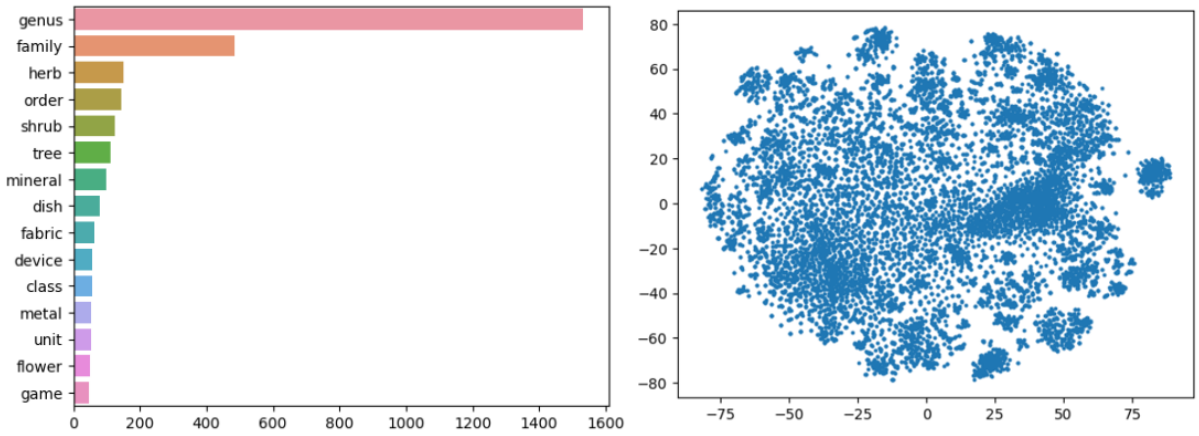

Figure 3: Left: the word frequency of training phrases containing prefix and suffix words, indicating the long-tail domain distribution. Right: the distribution of CLIP text features, showing no major cluster information

Index	Approach	English		
		train	val	test
(1)	MFS gloss	85.5	75.3	57.3
(2)	Gloss + Biencoder	87.7	77.2	65.6
(3)	(2) + Synonym	88.6	81.3	74.0
(4)	(3) + Augmentation	89.1	83.3	75.5
(5)	(4) + Weighted SIM	89.0	84.8	78.2
(6)	Image matching	-	75.4	69.4
Final	(5) + (6)	-	**89.3**	**84.0**

Table 2: Performance of different approaches on English measured in hit rate, where MFS refers to Most Frequent Sense.

training and the testing set, and the constructed validation set has limited ability to eliminate the gap, as shown in Table 2. By applying a SimCSE-based biencoder to select the correct sense, the results significantly increase the testing set, which is related to the more multi-sense words in the testing set than the training set listed in Table 1. Another major improvement is adding synonyms to the prompt, which has contributed a lot to WSD ability. After analyzing the CLIP results, we find out that the synonyms can match the pre-training data of CLIP and transfer the potential knowledge of pre-trained models. To further improve the WSD performance, we back-translate the phrase for augmentation and add weighted similarity to minimize the distance to the phrase, which leads to a hit rate of 78 with text only.

The involvement of image representation provides cross-modal information for disambiguation. The basement results of the image are not as expected, but the overlap of wrong instances is small, which means the combination of image and text covers a broader range of correct instances. Those words represented well by visual information, such as named entities, are replaced by images and, as a result, we achieve the final result of 84 hit rate on the testing set.

4.2 Distribution Shift

The distribution of senses determines the contribution ratio of the WSD module and matching module. The training set has more single-sense words, with a proportion of 70%, than multi-sense words, and the testing set has the reversed situation as indicated in Table 1. The training data is more likely to be a text-image matching task rather than a WSD task with minor WSD contribution, and the modification of the matching module has more influence on the results. However, the testing data involves more ambiguity than the training set, and most words have multiple meanings. Therefore, after the testing data was released, we changed our strategy and put more effort into WSD performance, as shown in Table 2 index 1-5.

4.3 Fine-tuning and Pre-training

To investigate the value of training data, we tried both fine-tuning and continuous pre-training on CLIP with the training set. As the accuracy of the selected gloss cannot be guaranteed, the phrase context is the only available text data to be learned. Moreover, the testing set covers two more languages over English, making fitting more challenging.

Fine-tuning the CLIP classification layers with training data give 1% improvement on the validation set, which has the probability of dropping

down the zero-shot capability on unseen words. The training set shows its properties, including open domain, long tail, and zero-shot, as shown in Figure 3. The classification layers have limited capability of dealing with exceptionally sparsely distributed data.

Continuous Pre-training the CLIP with a contrastive loss between given phrase and images gives better results than fine-tuning classification layers, which has over 2% improvement. Rather than adapting specific data distribution, continuous training could be regarded as injecting more knowledge into the model, which might benefit generalization.

5 Conclusion

We describe our prompt and biencoder-based Visual WSD system and investigate how the potential knowledge of large-scale pre-trained VLM contributes to disambiguation and modal alignment. The well-designed prompt template connects the input phrase and the potential knowledge of the pre-trained model, which also prevents the zero-shot generalization capability. The sense inventory is the fundamental element for WSD tasks, and context quality determines the performance boundary. The involvement of images in the WSD module further extends the coverage of sense interpretation ability. With the combination of text context and image representation, our system achieves a hit rate of 84 and ranks first in the English track.

In future work, more research on the modal fusion strategy will be completed, and the automatic fusion approach that applies to the different datasets will be designed. Rather than selecting a single modality based on rules, extracting useful information from both text and images could benefit the disambiguation performance.

References

Romain Beaumont. 2022. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. `https://github.com/rom1504/clip-retrieval`.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent Trends in Word Sense Disambiguation: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 4330–4338, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. 2021. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*.

Terra Blevins and Luke Zettlemoyer. 2020. Moving Down the Long Tail of Word Sense Disambiguation with Gloss-Informed Biencoders. ArXiv:2005.02590 [cs].

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. *arXiv:1909.11740 [cs]*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. ArXiv:2212.07143 [cs].

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. SimCSE: Simple Contrastive Learning of Sentence Embeddings. ArXiv:2104.08821 [cs].

Yixing Luan, Bradley Hauer, Lili Mou, and Grzegorz Kondrak. 2020. Improving Word Sense Disambiguation with Translations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4055–4065, Online. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

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. ArXiv:2103.00020 [cs].

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.

Amir Ahmadi Sajjad Ayoubi, Navid Kanaani. 2022. Clipfa: Connecting farsi text and images. https://github.com/SajjjadAyobi/CLIPfa.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework.

Xing Wu, Chaochen Gao, Liangjun Zang, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. ESimCSE: Enhanced Sample Building Method for Contrastive Learning of Unsupervised Sentence Embedding. ArXiv:2109.04380 [cs].

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional Prompt Learning for Vision-Language Models. ArXiv:2203.05557 [cs].

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Image Captioning

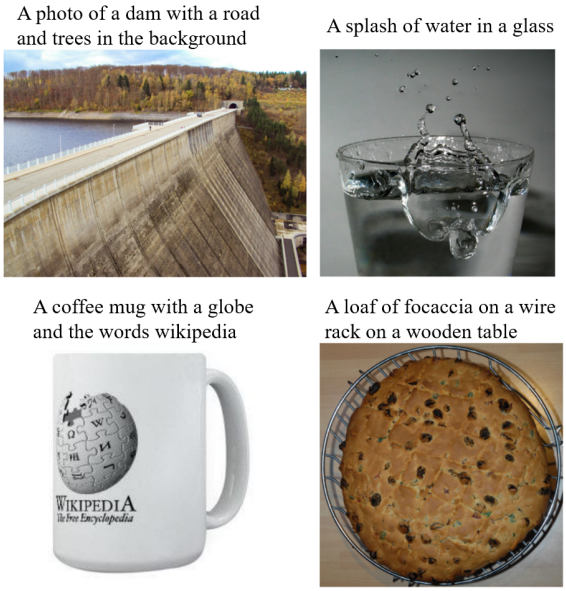

Figure 4: Examples of generated image captions

The captions of image labels are regarded as an augmented way to represent image in textual modality. The generated text of OFA (Wang et al., 2022) has a high quality and can express most features on the image. For exmaple, the generated text of the bottom right image in Figure 4 indicates the specific name of the food and the rack that is mostly hidden. The knowledge reserve and object extraction ability is acceptable. However, the results keep decreasing by fuse a higher weight of captions into the images. Two simple approaches are applied, one is add the text CLIP feature by a ratio to the image feature, in which the result decreases as the fusion ratio increases. Another one is fuse the similarities of image labels and its captions with the given phrase, where same results occurs.

One of the reason might be that the fusion approach is not capable enough to merge information of different modalities. The single-stream model like UNITER (Chen et al., 2020) that fuses modal features through networks instead of dual-stream model like CLIP, where a better feature fusion could enhance the information. Moreover, there is a probability that CLIP has extracted all information contained by the captions and the noise influence the performance. More experiment and research will be done in the future work.