Rutgers Multimedia Image Processing Lab at SemEval-2023 Task-1: Text-Augmentation-based Approach for Visual Word Sense Disambiguation

Keyi Li[1], Sen Yang[2], Chenyang Gao[1], Ivan Marsic[1]
[1]Rutgers University
[2]Waymo LLC
{kl734,sy358,cg694,marsic}@rutgers.edu

Abstract

This paper describes our system used in SemEval-2023 Task-1: Visual Word Sense Disambiguation (VWSD). The VWSD task is to identify the correct image that corresponds to an ambiguous target word given limited textual context. To reduce word ambiguity and enhance image selection, we proposed several text augmentation techniques, such as prompting, WordNet synonyms, and text generation. We experimented with different vision-language pre-trained models to capture the joint features of the augmented text and image. Our approach achieved the best performance using a combination of GPT-3 text generation and the CLIP model. On the multilingual test sets, our system achieved an average hit rate (at top-1) of 51.11 and a mean reciprocal rank of 65.69.

1 Introduction

Polysemous words are common in human language. These words are ambiguous and can be interpreted in variant ways under different contexts. Although it is easy for humans to distinguish different word senses, machines need to transform the word senses into data structure and analyze the differences. Word sense disambiguation (WSD) is a task for machines to identify the meaning of words given limited text contexts (Navigli, 2009; Bevilacqua et al., 2021). WSD is widely used in different NLP tasks, including information retrieval, machine translation, information extraction, content analysis, and lexicography. With the rapid growth of the multimodal datasets, WSD task has expanded from language to the visual field, which aims to improve the tasks such as image description, visual question answering, object detection, and image retrieval (Chen et al., 2015; Gella et al., 2016; de Guevara et al., 2020; Calabrese et al., 2021). For example, when we search for an image of "Andromeda tree" with an ambiguous word "Andromeda", the search engine needs to identify the meaning of "Andromeda" under the context "tree", i.e., a species of plant instead of the Greek goddess or the galaxy. The SemEval-2023 Task-1 describes this task as the visual word sense disambiguation task (VWSD): given an ambiguous target word with limited textual context, select among a set of candidate images the one which corresponds to the intended meaning of the target word (Raganato et al., 2023).

Word embedding has been widely used to represent words in numerical vectors for machines to interpret the word sense (Mikolov et al., 2013). A word is transformed into high-dimensional vector space where each dimension represents a unique feature of the word. Word embeddings can capture the semantic and syntactic relations between words, e.g., words with similar meanings will be closer in the space, and antonyms words will be orthogonal to each other. However, word embeddings have a limitation that polysemous words with multiple meanings cannot be captured by a single vector. Embedding the words with context is important for a model to identify the specific sense of the word (Kumar, 2021; Reisinger and Mooney, 2010).

Extending to the vision-language domain, the vision-language pre-trained (VLP) models aim to represent the joint features for both text and image. The textual and visual information is transformed into a shared feature space where the text and image with similar meanings are close to each other. Existing VLP models were usually pre-trained on large-scale corpus of image-text pairs where the text data is usually a sentence that describes the image (Radford et al., 2021). To find the best matching image-text pairs for the VWSD task, we propose an approach that extends the context of the target word using different text-augmentation methods, such as prompting (Gao et al., 2020), the WordNet synonyms set (Miller, 1995), and text generation (Brown et al., 2020). For image selection, we experimented with different VLP models to extract the features for the augmented text and the candidate images (Radford et al., 2021; Li et al.,

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1483–1490

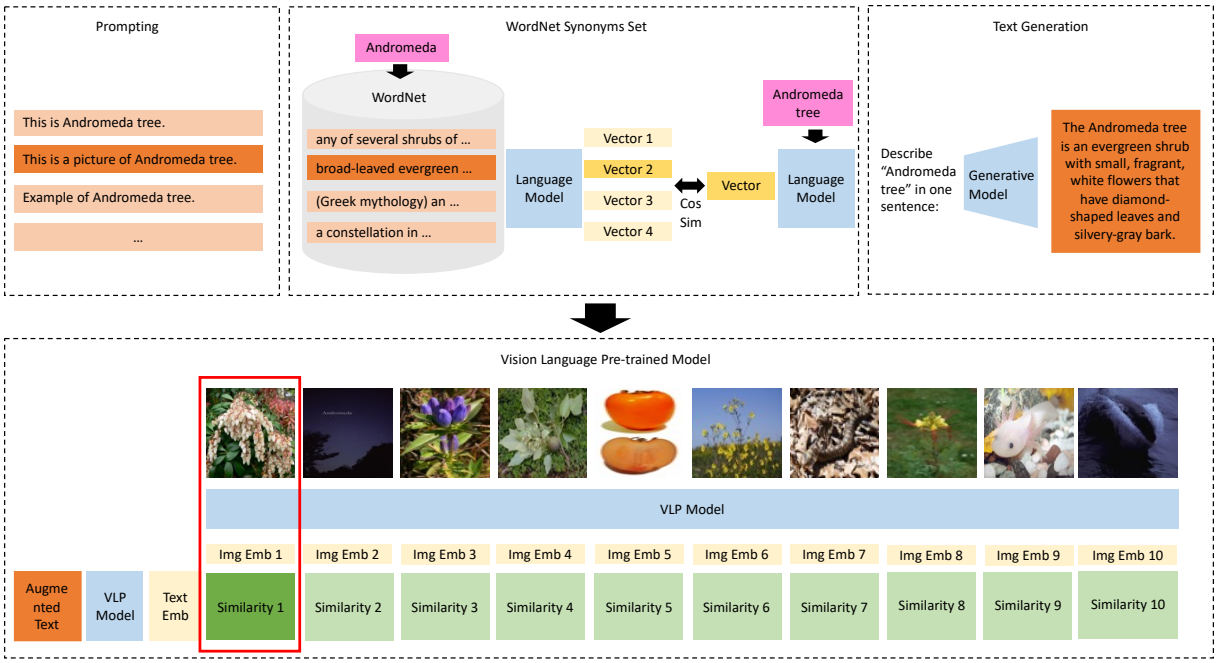

Figure 1: The architecture of our text-augmentation-based VWSD approach. Three text augmentation methods were used for extending the context of the target word, including prompting, WordNet Synonyms Set, and text generation. The augmented text and the candidate images are then input to the VLP model to obtain the multi-modal embeddings. The image with the highest similarity to the augmented text is selected as the result.

2021, 2022c, 2023). The image corresponding to the target word is selected based on the similarity between the embeddings of the augmented text and the images. Our contributions are:

- We introduced three text augmentation approaches to reduce the ambiguity of polysemous words for VWSD task.
- We experimented with various vision-language pre-trained (VLP) models for the VWSD task and focused on their zero-shot learning abilities. Our approach of combining GPT-3 text generation and CLIP model achieved the best performance with a hit rate of 51.11 and a mean reciprocal rank of 65.69.

2 System Overview

Our proposed approach consists of two parts (Figure 1): (1) text augmentation and (2) image selection. Here we describe our text augmentation approaches and the VLP models we used in our experiments.

2.1 Text Augmentation

Prompting. The VLP models are usually pre-trained on large-scale corpus of image-text pairs. For example, the pre-training dataset of the CLIP model contains about 400 million image-text pairs obtained from the Internet. The text in these pairs are usually sentences that describe the images. Thus, the authors of the CLIP model proposed that prompting the single word or phrase can improve the efficiency and enable the zero-shot ability of the VLP models (Radford et al., 2021). Considering the clarity of the prompted text and the variable image categories (e.g., photographs, illustrations, charts and diagrams, etc.) in our VWSD dataset, we designed some prompts manually such as "This is ...", "This is a picture of ...", "Example of ..." and selected the one with the best performance. In our experiment, we used "This is a picture of ..." as our prompt.

WordNet Synonyms Set. Polysemous words need context to determine which sense is used. The senses of one polysemous word can be very different, resulting in long distances of their embeddings in the feature space. Given that one word only has one embedding, this embedding usually corresponds to the most popular meaning of the word. To obtain an embedding close to the specific meaning of the word, more contextual information is needed to select the embedding.

WordNet is a lexical database for the English

language, and it groups English words into synonyms sets (synsets) (Miller, 1995). Given a target word, the synset gives a set of synonyms of the target word, as well as the corresponding descriptions of the synonyms (e.g., the synset of the word "book" includes different senses of "book", and also includes other synonyms like "script", "reserve", etc.). Our aim is to identify the specific sense with limited context, i.e., a context word in our VWSD task.

Our approach first calculates the sentence embeddings of all the synonyms' descriptions and the target phrase (Figure 1). We experimented with different pre-trained language models to extract the embeddings, including:

- **Bert**: The original Bert model pre-trained on large corpus of English dataset (Devlin et al., 2018). We extract the last layer feature as the embedding of the sentence.
- **Transformer-XL**: The Transformer-XL model pre-trained on the WikiText-103 to handle the long-range dependencies of the words in the description sentences as well as the different lengths between sentences and the context word (Dai et al., 2019).
- **MiniLMv2**: The MiniLMv2 is a distilled pre-trained model which has small model size and fast inference speed (Wang et al., 2020). Based on the multi-head self-attention relation distillation, the MiniLMv2 model has good performance in the downstream language understanding tasks.
- **MP-Net**: The MP-Net unifies the masked language model and permuted language model to capture the relations between words and the information of the whole sentence (Song et al., 2020).

After we obtained the embeddings, we calculated the cosine similarities between the target phrase and the descriptions. The most similar description to the context word was selected as our augmented text.

Text Generation. Although the WordNet has the ability to give an accurate description of the target word, the performance of the WordNet approach is limited by the representation ability of the pre-trained language models. If the description is not accurately selected, the subsequent image selection task may not perform well. To mitigate this gap, our aim was to get the most accurate description of the given phrases.

We designed a text augmentation approach based on the large text generation models to ensure the description was expressing the target phrase (Brown et al., 2020; Ouyang et al., 2022). We used the newest version of the GPT-3 model family, i.e., the text-davinci-003 model, to generate the descriptions. To obtain a short and clear description, we designed a prompt for the GPT-3 model: "Describe '(target phrase)' in one sentence: ...", and used the generated sentence as the augmented text.

2.2 Vision-language Pre-trained Models

Vision language pre-training aims to learn the semantic correspondence between vision and language modalities through self-supervised learning (Li et al., 2022b; Ericsson et al., 2022). The pre-trained vision language representation achieved promising results in various VL downstream tasks with fine-tuning, and had strong performance in zero-shot learning. With the augmented text we obtained from our text augmentation approaches, we formulated the VWSD task as an image-text retrieval task. We explored the zero-shot learning capability over several open-source VLP models on the VWSD task.

CLIP. The CLIP model was pre-trained on a dataset of 400 million image-text pairs collected from the Internet (Radford et al., 2021). The model consists of two encoders: a vision transformer and a transformer-based language model, to capture the features of images and texts, respectively. For each batch that contains n samples, an image-text pair is considered as a positive sample, and the n-1 samples that are not pairs are considered as negative samples. The embeddings of the image-text pairs become gradually closer through the image-text contrastive (ITC) learning. CLIP model achieved high performance in many vision-language tasks, especially in the image-text retrieval task.

ALBEF. The ALBEF model consists of a vision transformer encoder, a BERT-based text encoder, and an additional multi-modal encoder (Li et al., 2021). Considering that the CLIP model did not have the multi-modal fusion layer, and the 400 million internet image-text pairs might contain noisy samples (e.g., the text and image might not match, or some matching image and text are not paired), it might not perform well on some fine-grained vision-language tasks. The ALBEF model added the image-text matching (ITM) task and masked language modeling (MLM) task for

Table 1: The performance of different combinations of augmented texts and different VLP models (HIT/MRR reported respectively).

Approach	CLIP	BLIP	BLIP-2	ALBEF
Target word	19.56/38.92	18.49/37.52	19.60/39.58	12.80/31.39
Target phrase	30.31/48.51	26.54/44.93	28.98/47.81	13.68/31.12
Prompt + Target phrase	30.90/48.83	28.34/45.95	29.16/47.77	13.00/32.90
WordNet-C	35.44/53.69	29.56/49.09	31.20/50.22	13.96/33.64
WordNet-T	42.55/58.64	38.97/55.51	40.41/56.90	14.95/34.83
Prompt + WordNet-T	41.91/58.31	37.68/54.74	40.69/57.22	14.31/34.46
GPT3	**49.81/64.76**	47.17/62.81	47.24/62.91	18.96/38.40
Prompt + GPT3	47.18/63.04	44.87/60.81	46.02/61.81	17.60/37.30

Table 2: The performance of different combinations of augmented texts and the variants of CLIP model (HIT/MRR reported respectively).

Approach	CLIP-ViT-B-32	CLIP-ViT-B-16	CLIP-ViT-L-14	CLIP-ViT-L-14-336
Target word	18.17/37.01	18.70/37.67	20.97/39.89	19.56/38.92
Target phrase	28.12/45.38	29.95/47.40	32.12/49.95	30.31/48.51
Prompt + Target phrase	29.53/46.21	30.25/47.34	31.12/49.48	30.90/48.83
WordNet-C	34.39/52.31	33.84/52.73	36.34/54.34	35.44/53.69
WordNet-T	39.23/55.91	41.88/58.18	43.46/59.46	42.55/58.64
Prompt + WordNet-T	41.07/57.05	41.57/57.89	42.40/59.06	41.91/58.31
GPT3	51.06/64.94	**51.11/65.69**	50.48/65.26	49.81/64.76
Prompt + GPT3	47.30/62.15	47.11/62.50	47.99/63.37	47.18/63.04

the multi-modal encoder. ALBEF introduced a momentum model to generate pseudo-targets for ITM and MLM tasks to learn more visual concepts that were not described in the ground-truth text. Pre-trained on only 14 million image-text pairs, the ALBEF model outperformed CLIP model in zero-shot image-text retrieval task on the Flickr30k dataset.

BLIP. The BLIP model bootstrapped the dataset by introducing an image-grounded text encoder (Filter) and an image-grounded text decoder (Captioner) (Li et al., 2022c). The Captioner generated synthetic descriptions for the images in the image-text pairs and the Filter learned to remove the unrelated descriptions. The dataset was then bootstrapped with more related image-text pairs and can be used to pre-train the image and text encoders. The BLIP model unified vision-language understanding and generation. It also achieved better performance in the image-text retrieval task on the Flickr30k dataset than the CLIP model based on 129 million image-text pairs.

BLIP-2. The BLIP-2 model introduced a querying transformer (Q-Former) to bootstrap the dataset and improve the vision language learning model (Li et al., 2023). The Q-Former forced the model to learn the knowledge that the queries asked, and boosted the representation ability of the image and text encoders. As a result, BLIP-2 model achieved the state-of-the-art performance of the image-text retrieval task.

2.3 Image Selection

To select the corresponding image, we calculated the cosine similarity between the image embedding ($\mathbf{I_e}$) and the text embedding ($\mathbf{T_e}$):

$$similarity(I, T) = \mathbf{I_e} \cdot \mathbf{T_e^\intercal} \tag{1}$$

The image with the highest similarity was selected as the corresponding image.

2.4 Evaluation

We used the hit rate at top-1 (HIT, the ratio of the "golden" image being the top rank) and the mean reciprocal rank (MRR, the average of the reciprocal ranks of the golden image) to evaluate the performance of our approach:

$$HIT = \frac{\sum_i^N (rank_1 = Golden)}{N} \tag{2}$$

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank_i = Golden} \tag{3}$$

where N is the sample size, and $rank_i = Golden$ means the rank of the golden image, which is i.

Table 3: The performance of WordNet-T on different pre-trained language models in selecting the most similar descriptions (we used the largest CLIP model, i.e., CLIP-ViT-L-14-336. HIT/MRR reported respectively).

Approach	CLIP
Bert	38.98/56.34
Transformer-XL	40.63/57.21
MiniLM	**42.55**/58.64
MP-Net	42.48/**58.84**
Multilingual MP-Net	36.68/54.37

3 Experiment and Results

3.1 Dataset

We used the test dataset provided by SemEval-2023 Task-1 for evaluation. Each sample comprises the target word, the target phrase (i.e., target word with limited context), and ten candidate images. The test dataset consists of 8100 images. Three languages were provided, including English (463 samples), Farsi (200 samples), and Italian (305 samples).

3.2 Baselines

We experimented with different levels of augmented text and with their combinations, including:

- **Target word**: The target ambiguous word.
- **Target phrase**: The target phrase that contains the ambiguous word and limited context.
- **Prompt + Target phrase**: "This is a picture of " + target phrase.
- **WordNet-C**: Target phrase + the first WordNet description of the context word.
- **WordNet-T**: Target phrase + the WordNet description of the target word based on cosine similarity.
- **Prompt + WordNet-T**: "This is a picture of " + target phrase + WordNet-T.
- **GPT-3**: The description generated by GPT-3 (text-davinci-003).
- **Prompt + GPT-3**: "This is a picture of " + target phrase + GPT-3.

In our experiment, considering that the WordNet only contains English words, we experimented with two approaches to deal with the multilingual data: 1. translating both the ambiguous word and target phrase into English using the Google Translator API[1], and 2. translating only the ambiguous word into English, finding the corresponding WordNet descriptions and using multilingual pre-trained language model (we used the multilingual MP-Net in this paper) to calculate the similarities between the English descriptions and the multilingual phrases. For the text generation approach, we tested both multilingual and monolingual texts. For Farsi and Italian text, we used GPT-3 to generate multilingual descriptions by translating prompts "Describe (target phrase) in one sentence: ..." in Farsi and Italian, and monolingual English text by generating the descriptions using the prompt: "Describe (target phrase) in one sentence in English: ...".

We also experimented with the VLP models described in Section 2.2. For the CLIP model, we tested the performances using the different variants, including ViT-B-16, ViT-B-32, ViT-L-14 and ViT-L-14-336, where the ViT model sizes are different in these variants.

3.3 Results

Our results (Table 1) show that augmenting the context of the ambiguous word significantly boosted performance in the VWSD task across all models. We found that using the descriptions generated by GPT-3 in conjunction with image selection by the CLIP model achieves the best performance. Although ALBEF, BLIP, and BLIP-2 models performed better on the Flickr30k dataset for image-text retrieval, our results demonstrate that the CLIP model excels in the VWSD task. We attribute the better performance of the CLIP model to its access to a larger dataset of 400 million image-text pairs, which is more robust for zero-shot learning on new images.

Among the CLIP variants, we observed that larger CLIP models (CLIP-ViT-L models) achieved superior performance when working with shorter texts, such as the target words and target phrases. When working with more comprehensive text descriptions, such as the WordNet descriptions and those generated by GPT-3, we found all CLIP variants performed similarly. We think this is due to the limited size of the test dataset, which did not require the ability of larger CLIP models in the image-text retrieval task. We found the CLIP-ViT-B-16 had the best performance on GPT-3 texts, achieving a hit rate of 51.11 and a mean reciprocal rank of 65.69 (see Table 2).

We observed that incorporating prompts to the text did not result in an improvement in VWSD performance (Table 1, Table 2). While adding a

[1]https://pypi.org/project/googletrans/

Table 4: The comparison between the multilingual and monolingual (English) texts using Multilingual-CLIP and original CLIP (both CLIP models are ViT-L-14-336).

Approach	Monolingual	Multilingual
Target word	19.56/38.92	**20.14/39.07**
Target phrase	30.31/48.51	**36.27/53.96**
Prompt + Target phrase	30.90/48.83	**39.84/56.66**
GPT3	**49.81/64.76**	45.50/61.41
Prompt + GPT3	**47.18/63.04**	45.37/61.14

prompt is recommended as means of improving the image-text retrieval performance for existing VLP models (Radford et al., 2021), it did not essentially aid in disambiguation. This indicates that the VLP models are capable of comprehending sentences even without prompts.

Our results indicate that utilizing WordNet augmented text enhanced VWSD performance (Table 1). To calculate the similarity between the target phrase and the descriptions of target word in WordNet, we evaluated different pre-trained language models. We found the MiniLM and MP-Net have comparable performance and outperformed Bert and Transformer-XL model (Table 3). Furthermore, we found that translating Farsi and Italian target phrases into English and using pre-trained models to calculate similarities got better results than the use of multilingual pre-trained models. This suggests that the pre-trained models are more effective in identifying similar English phrases and sentences. Additionally, we found that target word descriptions selected by pre-trained language models outperformed context word descriptions (as shown in Table 1). This implies that providing more specific descriptions can enhance the VLP model's abilities, especially considering that it was pre-trained on internet image-text pairs.

The GPT-3 model's text generation approach exhibited better performance than other approaches in our study. This is likely because the GPT-3 descriptions provided ample context for the VLP model to select images, and they were similar to the text in the pre-training dataset of the VLP models. Using GPT-3 descriptions was more effective than WordNet because GPT-3 eliminated the need for the similarity calculation step, resulting in more precise descriptions.

In addition, we assessed the performance of both multilingual and monolingual CLIP models (Table 4). We observed that the multilingual model outperformed the monolingual model when using word and phrase texts, as well as phrases with prompts. On the other hand, the monolingual model performed better when using GPT-3 texts. This indicates that the GPT-3 model generates more accurate English descriptions compared to other languages.

4 Conclusion

We introduced a text-augmentation based approach for SemEval-2023 Task-1 on visual word sense disambiguation. We explored different text augmentation methods such as prompting, the WordNet synonyms set, and text generation. We experimented with different vision-language pre-trained models using zero-shot learning. Our system achieved the best performance with the combination of GPT-3 text generation and the CLIP model.

Although our team only experimented with the text-augmentation idea, we believe other ideas may also be effective for VWSD task, such as (1) image-augmentation: using the state-of-the-art image generative models to generate images given the target phrase, and calculate the similarity between the generated image and candidate images (Ramesh et al., 2022; Rombach et al., 2022); (2) image captioning: using pre-trained image captioning model to generate descriptions of the candidate images, and calculate the similarity between the descriptions and the target phrase (Li et al., 2022a; Wang et al., 2022); (3) ensemble model: assemble different VLP models to boost the performance from different models.

Acknowledgements

This work is supported by the U.S. National Institutes of Health/National Library of Medicine under grant number R01LM011834.

References

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, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2021. Evilbert: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 481–487.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. 2015. Sense discovery via co-clustering on images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5298–5306.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Manuel Ladron de Guevara, Christopher George, Akshat Gupta, Daragh Byrne, and Ramesh Krishnamurti. 2020. Multimodal word sense disambiguation in creative practice. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 294–301. IEEE.

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*.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint arXiv:1603.09188*.

Abhilasha A Kumar. 2021. Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 28:40–80.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022a. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M Ni, PengChuan Zhang, and Lei Zhang. 2022b. Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022c. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Joseph Reisinger and Raymond Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.