ML Mob at SemEval-2023 Task 1: Probing CLIP on Visual Word-Sense Disambiguation

Clifton A. Poth* **Martin B. Hentschel***
Tobias Werner **Hannah Sterz** **Leonard Bongard**
Technical University of Darmstadt
{cliftonalexander.poth, martin.hentschel}@stud.tu-darmstadt.de

Abstract

Successful word sense disambiguation (WSD) is a fundamental element of natural language understanding. As part of SemEval-2023 Task 1, we investigate WSD in a multimodal setting, where ambiguous words are to be matched with candidate images representing word senses. We compare multiple systems based on pre-trained CLIP models. In our experiments, we find CLIP to have solid zero-shot performance on monolingual and multilingual data. By employing different fine-tuning techniques, we are able to further enhance performance. However, transferring knowledge between data distributions proves to be more challenging.

1 Introduction

Determining the correct meaning of an ambiguous word in a given textual context is a fundamental element of successful understanding of natural language. Formulated in the task of Word Sense Disambiguation (WSD), it remains a key challenge of Natural Language Processing (NLP). Recent successful approaches tackling this task often leverage large pre-trained language models (Bevilacqua et al., 2021). With the rise of large pre-trained vision-and-language (V&L) models such as CLIP (Radford et al., 2021), the question arises how well these models can tackle WSD in a multimodal setting.

Towards answering this question, Task 1 of SemEval-2023 proposes a visual WSD task (Raganato et al., 2023), combining ambiguous contextualized words with images encoding different word senses. Instead of selecting the correct word sense from a sense inventory, a system now has to retrieve the best matching image from a set of candidate images, thereby bridging the gap between vision and language.

In this paper, we present our contributions to the SemEval-2023 Visual-WSD task. Our work focuses on empirically comparing the WSD performance of multiple CLIP models in different transfer learning settings. This includes the investigation of zero-shot transfer and *adapters* (Rebuffi et al., 2017; Houlsby et al., 2019) as a parameter-efficient means of transfer learning. We especially aim to highlight to which degree simple techniques are sufficient for successful task performance. Our best submitted system ranks 9th out of 54 on the official task leaderboard (counting only one submission per person).

2 Background

2.1 Word Sense Disambiguation (WSD)

WSD is the task of determining the exact meaning of an ambiguous word in a given context. Typically, this is achieved by selecting the most suitable word sense from a pre-defined, static word sense inventory such as WordNet (Miller et al., 1990). Existing solutions for this task mainly follow two approaches: *Knowledge-based approaches* exploit the knowledge encoded into lexical resources, such as the graph structure of WordNet. *Supervised approaches* aim to learn a parameterized model, such as a neural network, that directly predicts the correct sense s given a word w and a context c in a classification setting (Hadiwinoto et al., 2019). Training these models in a supervised setup requires large sets of word-context pairs annotated with correct sense labels. One such dataset is SemCor (Miller et al., 1993), consisting of 200,000 sense annotations based on the WordNet sense inventory. Recent state-of-the-art WSD systems can often be described as a combination of knowledge-based and supervised approaches. E.g., EWISER (Bevilacqua and Navigli, 2020) combines a pre-trained BERT encoder (Devlin et al., 2019) with a synset graph extracted from WordNet. ESCHER (Barba et al., 2021) on the other hand augment the word-context pair with textual definitions for all

* equal contribution

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1463–1469
July 13-14, 2023 ©2023 Association for Computational Linguistics

possible senses from a sense inventory.

2.2 SemEval-2023 Visual-WSD Task

Existing work on WSD almost unanimously focuses only on the text domain. The SemEval-2023 Visual-WSD task (Raganato et al., 2023) extends the scope of WSD by formulating it as a multimodal V&L problem. Similar to the classical WSD formulation, a pair (w, c) consisting of an ambiguous word w and a short textual context c is given as input. The goal now is to select an image from a set of candidates. More concretely, from a set of ten candidate images $\{I_0, \ldots, I_9\}$, a system has to select the image I_y that best represents the intended meaning of w given c.

The Visual-WSD task provides a labeled English-language training dataset consisting of 12,869 (w, c) pairs along with 12,999 unique images representing word senses. The test split contains samples in three languages, including English (463 samples), Italian (305 samples) and Farsi (200 samples). Samples of all three languages share a set of 8100 test images.

Performance of WSD systems on the Visual-WSD task is measured using two metrics. *Accuracy* (or *hit rate at 1*) is used to measure in how many cases a system correctly ranks I_y as the most likely candidate. *Mean reciprocal rank (MRR)* is used to evaluate the overall ranking quality of a system by taking into account at which position it ranks I_y.

3 Methodology

3.1 CLIP

CLIP (Radford et al., 2021) has been proposed as an approach to learn multimodal V&L representations by pre-training on a large-scale dataset of image-text pairs. The CLIP architecture follows a "late interaction" design, where image and text are encoded independently by a vision encoder V and a text encoder T. Both encoders are jointly optimized during pre-training to maximize the cosine similarity between the embeddings of both modalities for matching image-text pairs.

While Radford et al. (2021) show that CLIP has strong zero-shot capabilities on downstream tasks, subsequent work such as CLIP-ViL (Shen et al., 2021) incorporates the frozen CLIP vision encoder into existing architectures for various vision-language tasks. We directly leverage both pre-trained CLIP encoders for transfer to the downstream task. As the formulation of the Visual WSD task is similar to the pre-training task of CLIP, our fine-tuning objective closely follows CLIP's pre-training objective. Instead of passing N image-text pairs as done during pre-training, our input consists of one (w, c) pair alongside 10 images. The symmetric cross entropy loss employed during pre-training therefore collapses to a regular cross entropy loss (CE) over the candidate images. In simplified form, the full fine-tuning objective therefore can be formulated as:

$$\mathcal{L} = \mathrm{CE}(V([I_0, \ldots, I_9]) \cdot T([w, c]), y)$$

where $\cdot$ denotes the dot product. In our work, we select pre-trained checkpoints with Transfomer-based models (Vaswani et al., 2017) for V and T.

3.2 Adapter Methods

Adapters have been introduced to computer vision (Rebuffi et al., 2017) and NLP (Houlsby et al., 2019) as a parameter-efficient alternative to full fine-tuning for transfer learning. Initially focused on simple feed-forward bottleneck modules, the scope of parameter-efficient transfer learning methods has recently broadened to include a wide range of methods (Li and Liang, 2021; Hu et al., 2022; He et al., 2022; Liu et al., 2022) and extensions to new domains such as V&L tasks (Sung et al., 2021; Zhang et al., 2022; Lu et al., 2023). We refer to all of these approaches as *adapter methods*.

All adapter methods have in common that they introduce a small number of new parameters to specific locations in a pre-trained neural network. During training on the downstream task, only the newly introduced parameters are updated while all pre-trained model weights are kept frozen. In this work, we consider the following methods:

Bottleneck adapters introduce bottleneck feed-forward layer modules in each Transformer layer. They consist of a down-projection into a lower dimension, a non-linearity, an up-projection that projects back into the original hidden layer dimension and a residual connection. We follow the approach of Pfeiffer et al. (2021) in adding adapter modules sequentially after the feed-forward component of each Transformer layer.

Low-Rank Adaptation (LoRA) (Hu et al., 2022) injects trainable low-rank decomposition matrices into Transformer layers. For a weight matrix W_0, LoRA performs a re-parametrizations using low-rank matrices $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$ of rank r such that $W = W_0 + BA$. Following the original implementation, we add LoRA to the query and value matrices of each Transformer self-attention layer.

3.3 Intermediate Task Transfer

Due to their modular nature, the parameters of multiple adapters can be composed to leverage the knowledge from multiple downstream tasks. In our work, we study whether enriching our visual WSD system with knowledge learned from a text-only WSD intermediate task s benefits final performance on the target task t. To achieve this, we employ *sequential fine-tuning* of adapters (Poth et al., 2021; Zmarsly, 2022), which first optimizes adapter parameters on s before subsequently fine-tuning the same parameters on t.

For s, we choose the SemCor dataset (Miller et al., 1993), which provides (w, c) pairs annotated with the sense key k best describing w in its context. We train a bottleneck adapter on s by letting a multiple choice model choose k from a list of 10 possible sense keys and transfer its parameters to our models for t.

4 Experimental Setup

4.1 Model Checkpoints

We evaluate the following pre-trained models:

OpenAI CLIP denotes the original model checkpoints provided by Radford et al. (2021). These checkpoints are pre-trained on 400M image-text pairs collected from the Internet by the CLIP authors. For our experiments, we select the model variants using Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the vision encoder. Specifically, we select the variants with ViT-B/32 and ViT-L/14-336 for zero-shot evaluation.

LAION CLIP denotes models pre-trained in an open-source replication of the original CLIP models (Cherti et al., 2022). The best performing models of this family are pre-trained on LAION-2B, the English subset of the LAION-5B dataset (Schuhmann et al., 2022), consisting of 2.3B English language image-text pairs collected from Common Crawl. We select multiple checkpoints using ViT as vision encoder for evaluation, including ViT-B/32, ViT/L-14, ViT-H/14 and ViT-G/14. As of writing this paper, ViT-G/14 has the best zero-shot performance of all publicly available CLIP models[1].

Method	# Parameters	% Parameters
Zero-Shot	0	0.00
Projection Layer Tuning	1,212,417	0.33
LoRA Adapter	1,802,241	0.49
Bottleneck Adapter	3,001,473	0.82
Text Encoder Fine-Tuning	278,665,473	76.11
Full Fine-Tuning	366,121,473	100.00

Table 1: Total number and percentage of trainable parameters for different transfer learning methods of LAION XLM-R Base.

LAION Multilingual CLIP denotes multilingual variants of LAION CLIP pre-trained on the full LAION-5B dataset consisting of 5.85B image-text pairs in over 100 languages. These models use XLM-R (Conneau et al., 2020), a pre-trained multilingual language model, as text encoder. In our experiments, we select two checkpoints, one using ViT-B/32 and XLM-R Base and one using ViT-H/14 and XLM-R Large.

4.2 Transfer Learning Methods

We compare multiple transfer learning methods on the pre-trained CLIP checkpoints. In increasing order of parameters updated during application on the Visual WSD task, these include: 1) *Zero-shot transfer*, i.e. evaluation on the downstream task without any training. 2) *Projection layer tuning*, i.e. fine-tuning of the final projection layers on top of CLIP's vision and text encoders while keeping all other parameters fixed. 3) *Adapter methods* as described in § 3.2. 4) *Partial fine-tuning*, i.e. fine-tuning the full text encoder while keeping the vision encoder frozen. 5) *Full fine-tuning*, i.e. fine-tuning all CLIP parameters on the downstream task.

Their number of trainable parameters is compared in Table 1.

4.3 Training Details

To evaluate our models and compare different approaches during training, we keep a fixed set of 1286 (w, c) pairs (10%) from the training dataset, leaving us 11583 pairs to actually train models on.

In the provided datasets, each context contains the word it is paired with, making it sufficient to only use the tokenized context as input for our text encoder models without further preprocessing. We limit all tokenizers to a maximum sequence length of 20, which is enough to represent all samples from the different datasets.

Preprocessing of the input images consists of rescaling them to have a shorter side of 224 pix-

[1] https://laion.ai/blog/giant-openclip/

Model	Accuracy	MRR
Random Baseline	10.50%	29.91%
Zero-Shot		
OpenAI ViT-B/32	72.94%	82.23%
OpenAI ViT-L/14-336	81.03%	87.75%
LAION ViT-B/32	72.94%	82.23%
LAION ViT-L/14	78.38%	85.98%
LAION ViT-H/14	79.16%	86.43%
LAION ViT-G/14	80.72%	87.31%
LAION XLM-R Base	74.88%	83.51%
LAION XLM-R Large	79.00%	86.89%
LAION XLM-R Base		
Projection Layer Tuning	86.94%	92.15%
LoRA Adapter	85.23%	91.10%
Bottleneck Adapter	86.31%	91.88%
Text Encoder Fine-Tuning	86.78%	92.01%
Full Fine-Tuning	85.23%	91.11%
LAION XLM-R Large		
Projection Layer Tuning	89.11%	93.50%
LoRA Adapter	89.35%	93.78%
Bottleneck Adapter	90.20%	94.18%
Full Fine-Tuning	91.91%	95.37%

Table 2: Evaluation results on the English language validation set

els, cutting out the center to form a 224x224 pixel square, and normalizing them. To save disk space and unnecessary operations during training, we rescale all images beforehand.

Full models are fine-tuned for 5 epochs, using an initial learning rate of 10^{-6}. Adapters are trained for 30 epochs and start with a learning rate of 10^{-4}. In both cases, the learning rate is decreased linearly after every epoch. After training, we keep the checkpoint scoring the highest accuracy on our left-out validation split for inference. We use a batch size of 8 for all training runs.

For all evaluated CLIP models, we rely on PyTorch implementations in the *HuggingFace Transformers* library (Wolf et al., 2020). Adapter method implementations are provided by the *adapter-transformers* library (Pfeiffer et al., 2020).

5 Results and Analysis

5.1 Results on Validation Data

We first analyze the performance of all evaluated methods on our held out validation set (Table 2).

Zero-shot Comparing the zero-shot performance of the selected CLIP checkpoints, it is apparent that all pre-trained models perform reasonably well without fine-tuning on the training set. With accuracy scores between 72% and 81%, they substantially outperform a random baseline. Comparing between the evaluated models, performance increases with model size. The best zero-shot scores are obtained by OpenAI ViT-L/14-336 and LAION ViT-G/14, the largest evaluated representatives of their family. For base size models, we find that the multilingual LAION XLM-R Base slightly outperforms its monolingual counterparts. As we also suspect this model to have the highest potential on the multilingual test data, we focus on evaluating it in the following.

Fine-tuning Switching to the fine-tuned models, we observe that all evaluated methods are able to improve performance on the validation split. In general, accuracy scores on XLM-R Base increase by roughly 12 points compared to zero-shot performance of the same model. Interestingly, we find no substantial performance differences between the different transfer learning methods. Costly full fine-tuning of both the vision and text encoder of CLIP performs similar to various adapter methods and simple fine-tuning of CLIP's vision and text projection layers. The picture for LAION XLM-R Large is similar. Here, full fine-tuning slightly outperforms the other two methods, achieving our overall best scores of 91.91% accuracy and 95.37% MRR. From these results, we can conclude that it is possible to adapt CLIP to visual WSD efficiently without costly fine-tuning of all parameters on task data.

5.2 Results on Test Data

We transfer various approaches for the multilingual models to the test set. Table 3 shows accuracy and MRR scores per language as well as average scores across all languages.

Zero-shot Again, we find that zero-shot performance of the two evaluated multilingual LAION XLM-R models is substantially better than random across all languages. Similarly, they outperform the baseline provided by the task organizers, consisting of a zero-shot OpenAI ViT-L/14-336 for English and a zero-shot variant of OpenAI ViT-B/32 for Italian and Farsi. The latter has been made multilingual using Knowledge Distillation (Reimers and Gurevych, 2019). The best evaluated zero-shot model, LAION XLM-R Large, improves over the

Model / Version	Average		English		Farsi		Italian	
	Accuracy	MRR	Accuracy	MRR	Accuracy	MRR	Accuracy	MRR
Random Baseline	10.10%	29.67%	12.57%	31.67%	11.50%	31.27%	6.23%	26.07%
Zero-Shot								
Organizer Baseline	37.20%	54.39%	60.48%	73.88%	28.50%	46.70%	22.62%	42.61%
LAION XLM-R Base	52.63%	67.04%	68.47%	79.96%	35.00%	52.19%	54.43%	68.97%
LAION XLM-R Large	56.19%	69.73%	69.76%	81.23%	37.50%	54.38%	61.31%	73.59%
LAION XLM-R Base								
Projection Layer Tuning	53.28%	67.56%	69.11%	80.15%	35.00%	52.51%	55.74%	70.02%
LoRA Adapter	55.01%	69.09%	69.33%	80.42%	38.00%	55.49%	57.70%	71.37%
Bottleneck Adapter	52.60%	67.29%	68.90%	80.04%	33.50%	51.83%	55.41%	70.01%
SemCor Adapter	53.04%	67.79%	68.03%	79.81%	36.00%	53.65%	55.08%	69.92%
Full Fine-Tuning	52.64%	66.85%	67.17%	78.85%	35.00%	52.19%	55.74%	69.50%
LAION XLM-R Large								
Projection Layer Tuning	55.59%	69.56%	69.33%	81.24%	34.50%	53.03%	62.95%	74.40%
LoRA Adapter	58.65%	71.44%	72.35%	82.85%	40.00%	55.92%	63.61%	75.56%
Bottleneck Adapter	57.88%	71.34%	70.19%	81.86%	39.50%	56.27%	63.93%	75.90%
SemCor Adapter	58.59%	71.67%	70.84%	82.14%	41.00%	57.29%	63.93%	75.59%
Full Fine-Tuning	58.63%	71.43%	70.63%	82.22%	41.00%	56.23%	64.26%	75.85%

Table 3: Evaluation results on the multilingual test set.

baseline by 9 points on English and Farsi and by 39 points on Italian in terms of accuracy.

Fine-tuning Comparing the fine-tuned LAION XLM-R Base models, we find no meaningful improvements over zero-shot performance on the test set in most cases. Only LoRA improves over zero-shot performance by 2.4 accuracy points and 2 MRR points on average. On LAION XLM-R Large, LoRA and Bottleneck adapters improve slightly over zero-shot performance, reaching MRR scores over 70%.

Intermediate Task Transfer Results corresponding to the method described in § 3.3 are listed as *SemCor Adapter*. They perform similar to our other approaches, with the base variant even being outperformed by LoRA. We can conclude that knowledge transfer between text-only and multimodal WSD is non-trivial and requires further investigation.

Analysis In general, fine-tuning CLIP on the training set does not substantially improve results on the test set in our setup. To investigate potential causes for the limited knowledge transfer between training and test sets, we further analyze the data. Figure 1 compares the distribution of number of WordNet *synsets*[2] per target word in both data splits. We can observe that the majority of target words in the training set are found in zero to two WordNet synsets. In contrast, samples from the test set are found in 6 synsets on average, with some samples occurring in 15 or more synsets. Samples in the training set thus are overwhelmingly specific in their meaning, leaving less room for ambiguity, while test samples show diverse sets of meanings that might prove more difficult to disambiguate. We hypothesize that a WSD system fine-tuned on the Visual-WSD train split using the presented simple techniques therefore could generalize poorly to the test split. We leave further investigation of this issue to future work.

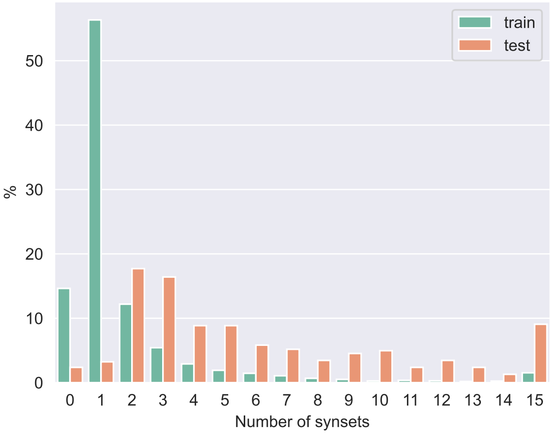

Figure 1: Distribution of number of WordNet synsets per target word in the training set and English test set of Visual-WSD. Values of 15 synsets and above are accumulate.

[2] Synsets are groups of words sharing the same meaning within WordNet.

6 Conclusion

As part of the SemEval-2023 Visual-WSD task, we investigated the multimodal WSD capabilities of various pre-trained CLIP models in a zero-shot and multiple transfer learning settings. We found that large pre-trained CLIP models perform reasonably well in the zero-shot setting, even across languages, yielding the conclusion that these models have substantial inherent visual WSD capabilities. Fine-tuning CLIP checkpoints using different techniques, including parameter-efficient adapter methods, leads to further improvements on in-distribution validation data. However, improvements over zero-shot performance on the multilingual test set are little. We conclude that knowledge transfer to the test set is non-trivial and approaches beyond the simple ones tested might be required for further substantial improvements.

Acknowledgements

We thank Myra Zmarsly and Gianluca Zimmer for their support on this project and reviewing our paper.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*.

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*, abs/2212.07143.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a Unified View of Parameter-Efficient Transfer Learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *ArXiv*, abs/2205.05638.

Haoyu Lu, Mingyu Ding, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Masayoshi Tomizuka, and Wei Zhan. 2023. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *ArXiv*, abs/2302.06605.

George A. Miller, Richard Beckwith, Christiane D. Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross Bunker. 1993. A semantic concordance. In *Human Language Technology - The Baltic Perspectiv*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 *International Conference on Machine Learning*.

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.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *NIPS*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *ArXiv*, abs/2107.06383.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2021. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5217–5227.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhengkun Zhang, Wenya Guo, Xiaojun Meng, Yasheng Wang, Yadao Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. Hyperpelt: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. *ArXiv*, abs/2203.03878.

Myra Zmarsly. 2022. Parameter-efficient language model tuning in the context of empathy and distress prediction. Master's thesis, Technical University of Darmstadt.