# LT at SemEval-2023 Task 1: Effective Zero-Shot Visual Word Sense Disambiguation Approaches using External Knowledge Sources

**Florian Schneider** and **Chris Biemann**
Language Technology Group, Department of Informatics, Universität Hamburg, Germany
{florian.schneider-1, chris.biemann}@uni-hamburg.de

## Abstract

The objective of the SemEval-2023 Task 1: Visual Word Sense Disambiguation (VWSD) (Raganato et al., 2023) is to identify the correct image illustrating the indented meaning of a target word and some minimal additional context.The omnipresence of textual and visual data in the task strongly suggests the utilization of the recent advances in multi-modal machine learning, i.e., pretrained visiolinguistic models (VLMs). Often referred to as foundation models due to their strong performance on many vision-language downstream tasks, these models further demonstrate powerful zero-shot capabilities. In this work, we utilize various pertained VLMs in a zero-shot fashion for multiple approaches using external knowledge sources to enrich the contextual information. Further, we evaluate our methods on the final test data and extensively analyze the suitability of different knowledge sources, the influence of training data, model sizes, multi-linguality, and different textual prompting strategies. Although we are not among the best-performing systems (rank 20 of 56), our experiments described in this work prove competitive results. Moreover, we aim to contribute meaningful insights and propel multi-modal machine learning tasks like VWSD.

## 1 Introduction

This paper presents and analyses effective zero-shot approaches for the SemEval-2023 Task 1: Visual Word Sense Disambiguation (VWSD) (Raganato et al., 2023). In traditional word sense disambiguation (WSD), the context or sentence in which ambiguous words, i.e., words with multiple meanings, occur is used for disambiguation by identifying the correct sense in a sense inventory. Frequently used as sense inventories are dictionaries or knowledge bases such as WordNet or DBPedia. As opposed to traditional WSD, in the VWSD shared task, images are used to disambiguate a word given a context. Precisely, given a word and another word serving as context, the task is to identify the image that corresponds to or illustrates the correct meaning in a set of ten images. In the trial phase of the task, 12869 samples in English, including gold labels, were provided. However, besides 463 English samples, the final phase test data also contains 305 Italian and 200 Farsi samples. A random VWSD


Figure 1: An illustration of a random VWSD sample with the target word '*bonxie*', the context '*bonxie skua*', and the correct image highlighted by a a golden border

sample is illustrated in Figure 1.

Due to the multi-modal nature of the task, it requires methods or models to understand textual semantics contained in the target word and context word and visual semantics contained in the images. Therefore, our approach leverages state-of-the-art pretrained visiolinguistic models (VLMs) in a zero-shot fashion, i.e., we do not continue pretraining or finetune. This is motivated by several reasons based on the tasks data: First, the samples are not restricted to a particular topic, e.g., animals or plants, but can belong to any topic (open-domain), which rules out domain adaption strategies. Further, since current VLMs are trained on massive amounts of text-image pairs crawled from the internet, continuing pretraining on additional open-domain data will likely have no effect. Second, the textual context is minimal and contains only a single or, at most, two additional words, which are often rare English words like the Latin names of certain plants. Additionally, it frequently requires expert knowledge to identify the correct image because the set of ten images often contains images very similar to the gold



*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 462–468

image. Due to this, finetuning a VLM on the provided data is ineffective, which we also confirmed in conducted but not reported finetuning experiments. Third, recent pretrained VLMs have proven capable of grasping textual and visual semantics out of the box by demonstrating strong zero-shot performance in many vision-language downstream tasks.

The central strategy of the approaches presented by this work is to utilize given information to acquire additional context from external knowledge sources. A pretrained VLM then computes embeddings for the acquired context and all images to find the image with the maximum similarity. See Section 3 algorithmic details and an illustrative overview.

Our code is publicly available on GitHub[1]

## 2 Background

**Pretrained Visio-Linguistic Models** The combination of recent advances in Natural Language Processing and Computer Vision has greatly increased interest and performance in the emerging field of multi-modal machine learning, especially in visio-linguistic models (VLMs) with strong zero-shot performance on many downstream tasks (Long et al., 2022). In this work, we specifically focus on VLMs referred to as CLIP (Radford et al., 2021), which we utilize to compute semantic representations of text and images in a joint vector space. There exist various versions of CLIP, all having the following in common: A CLIP model implements a dual encoder architecture with two separate encoders: a textual encoder, typically a BERT-based (Devlin et al., 2019) language model, and a visual encoder, typically CNNs like ResNet (He et al., 2016) or ConvNext (Liu et al., 2022) or a transformer (Vaswani et al., 2017) like ViT (Dosovitskiy et al., 2021). The encoders are jointly trained on massive amounts of text-image pairs to maximize the similarity between matching pairs in large batches via contrastive loss in an unsupervised fashion. While there are other VLMs with strong zero-shot capabilities, such as ALIGN (Jia et al., 2021), FLAVA (Singh et al., 2022), or CoCa (Yu et al., 2022), we prefer CLIP because there exist many publicly available pretrained versions, it is conveniently usable via multiple open-source libraries and frameworks like huggingface (Wolf et al., 2020) or OpenClip (Ilharco et al., 2021), and it has an active and large community. Specifically, we evaluate the performance of our VWSD approaches using different sizes of the original model (Radford et al., 2021), models trained on the publicly available datasets LAION (Schuhmann et al., 2022), and multi-lingual versions from SentenceTransformer (Reimers and Gurevych, 2019, 2020) and OpenClip (Cherti et al., 2022). An overview with more details about the CLIP models employed in this work is given in Table 1.

**External Knowledge** Since the provided context in a VWSD sample is minimal, we use different external knowledge sources to acquire additional contextual information. One of the sources is Wikipedia, from which we retrieve article summaries using the target word and the additional context word(s). Another source is a large-scale corpus (Panchenko et al., 2018), containing 252B tokens based on English CommonCrawl data, which we have indexed using ElasticSearch (Gormley and Tong, 2015). The only multi-modal external knowledge source we employ is VisualSem (Alberts et al., 2021), a high-quality knowledge graph containing 90K nodes with 1.3M glosses in 14 languages and 930K images associated with the nodes. Unfortunately, our request to use the large-scale multi-modal knowledge graph BabelNet (Navigli et al., 2021) was rejected. BabelNet arguably would have improved our results significantly since it contains 1.4M senses described by 135M glosses in 500 languages and illustrated by 51M images.

## 3 Approaches

This section provides details for the zero-shot approaches to VWSD presented by this work. Section 4 then analyzes and discusses the evaluation results.

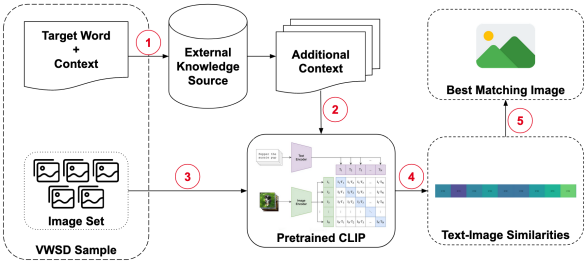


Figure 2: A schematic overview of the general strategy for the VWSD zero-shot approaches presented by this work.

The general strategy, illustrated in Figure 2, comprises five primary steps: In (1), we acquire addi-

[1] https://github.com/uhh-lt/vwsd_semeval_23

| Name | Alias | # P | H | # TS | BS | ML |
|---|---|---|---|---|---|---|
| *sentence-transformers/clip-ViT-B-32-multilingual-v1* | SBCM | 28M | 512 | 400M | 32K | yes |
| *openai/clip-vit-base-patch32* | OAIB | 15M | 512 | 400M | 32K | no |
| *openai/clip-vit-large-patch14* | OAIL | 42M | 768 | 400M | 32K | no |
| *laion/CLIP-ViT-L-14-laion2B-s32B-b82K* | LCL | 42M | 768 | 2B | 82K | no |
| *laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-lai...* | LCH | 119M | 1024 | 5B | 90K | yes |

Table 1: Details on different pretrained CLIP models evaluated in this work. The Alias column describes the alias for the model within this paper; **# P** is the number of parameters; **H** is the embedding dimension; **# TS** and **BS** is the number of text-image pairs, and the batch size used during pretraining, respectively; **ML** indicates whether the model is multi-lingual or not. Note that the names are hyperlinks directing to huggingface for more information.

tional context from an external knowledge source (see Section 2) using the textual information provided in a VWSD sample. In (2) and (3), we leverage a pretrained CLIP model to compute embeddings from the acquired context and all ten images contained in the sample. (4), we compute the cosine similarity of the text embedding and all image embeddings and (5) select the image with the maximum similarity as the best matching image. Depending on the method, we use different external knowledge sources, employ different pretrained CLIP models, or compute the textual or visual embeddings differently.

**Baseline – No External Knowledge** Our baseline method does not use external knowledge but computes the textual embedding only from the target word and or context in a VWSD sample. However, we test different template sentences or prompts to compute the textual embedding (see Table 2).

**Wikipedia Summaries** In this approach, we implemented a multi-stage algorithm to retrieve the summary of the best-matching Wikipedia article for the target word and the provided context. For more details on this algorithm, please refer to the implementation published on our GitHub repository. Although Wikipedia is available in many languages, we translate the Italian and Farsi samples to English using Google Translate[2] for library limitation reasons. If we cannot retrieve a summary for a given VWSD sample, we use a template sentence that contains the target word and the context. We then truncate too-long summaries and use the CLIP text encoder to compute the textual embedding.

**Common Crawl Sentences** In this approach, we query the index Common Crawl corpus (see Section 2) to retrieve all sentences that contain the target word and the context. We use a template sentence containing the target word and context if we cannot find any sentence. Since the corpus only contains English documents, we translate the Farsi and Italian samples into English using Google Translate. Further, because the original authors have cleaned the corpus, i.e., noisy, too long, and too short sentences are removed, we pick the top-5 longest sentences because we assume more contextual information in longer sentences. After truncating the sentences to fit the maximum length of the CLIP text encoder, we compute an embedding for each sentence and average them to obtain the textual embedding.

**VisualSem** Arguably the most sophisticated approaches are based on the multi-modal knowledge graph (KG) VisualSem (see Section 2). There, we first retrieve the best-matching node in the KG for the textual or visual information in a VWSD sample. To do so, we use a pretrained CLIP to compute node embeddings for each node in the KGs and use the FAISS (Johnson et al., 2019) for indexing and efficient similarity search. To compute the node embeddings, we tested four strategies: For the *"single_image"* and *"single_gloss"* strategies, one node in the KG has several embeddings, i.e., we compute an embedding for each associated image up to a maximum of 50 images, and each associated gloss in a particular language. For the *"avg_image"* and *"avg_gloss"* strategies, we compute a single embedding for each node in the KG, which is the average of the respective single embeddings. Then, to retrieve the best matching node(s) for a VWSD sample, we first use the same CLIP model used to compute the KG node embeddings and compute a query embedding from the sample's textual or visual information. When using textual information of a sample, we refer to it as "text_first"; when using visual information, i.e., the images, we refer to it as "image_first". Using the query embedding, we then perform an exhaustive similarity search over

[2] https://translate.google.com/

all nodes to find the best matching node(s). Finally, we find the most similar image, i.e., our prediction for the image with the intended meaning, using the embedding of the retrieved node and the "text_first" or "image_first" embedding.

Since this algorithm has many possible parameters and combinations thereof, it is challenging to describe, hence, please refer to our GitHub repository for implementation details.

## 4 Evaluation and Analysis

In this section, we present and analyze the evaluation results of our approaches described in Section 3. The evaluation is based on the final multilingual evaluation data, including the gold labels released in the Google Group after the competition[3]. Evaluation results for the approaches discussed in this section are depicted in Figure 3.

**Baseline – No External Knowledge** In our first experiments, we tested the performance of different pretrained CLIP models without external knowledge. From the results shown in the first row of Figure 3, it can be observed that all models show strong performance on the English test data. As expected, the largest model, `LCH`, outperforms the smallest model by a significant margin. Noticeable is also the linear decrease in performance with respect to the complexity of the model and the number of text-image pairs in the training data. When inspecting the baseline results for Italian and Farsi languages, a remarkable decrease in performance is noticeable. However, as expected, the multi-lingual CLIP variants significantly outperform English-only versions. Further, a pattern seen across all models and approaches is that the $Hit@3$ score is significantly higher than the $Hits@1$ score. This leads to the conclusion that the samples often contain a few very similar images, which are hard to disambiguate and require expert knowledge.

In another experiment, we measured the performance impact of the employed template string. Therefore we used the 9 different template strings described in Table 2 to compute textual embeddings using the `LCH` model and evaluated the performance of the baseline approach on the English test data. Note that we took inspiration for the template strings from (Radford et al., 2021) From the results depicted in Figure 4, we can see that the most influential parameter of our template strings is whether or not it contains the context information. Template strings containing context information work significantly better than template strings containing only the target word.

| Template String | Alias |
|---|---|
| An image of a "WORD" as in "CONTEXT" . | A |
| A photo of a "WORD" as in "CONTEXT" . | B |
| A picture of a "WORD" as in "CONTEXT" . | C |
| An image of a "WORD" . | D |
| A photo of a "WORD" . | E |
| A picture of a "WORD" . | F |
| An image of a "CONTEXT" . | G |
| A photo of a "CONTEXT" . | H |
| A picture of a "CONTEXT" . | I |

Table 2: Different template strings for English samples. The Alias column defines the alias within this paper.

**Wikipedia Summaries** From the results in the second row of Figure 3, we can notice substantial improvements in performance for the Italian and Farsi data, often on par with English data that improved only slightly or even decreased. From this, we can conclude that the Italian and Farsi translation into English worked reasonably well and that Wikipedia is a promising resource for VWSD. We argue that this approach could be further improved when using Wikipedia in the respective languages directly and when additional information, such as images, is used.

**Common Crawl Sentences** From the third row of Figure 3, we can see that this approach substantially outperforms all other approaches regardless of the employed model or language of the VWSD samples. Especially for Farsi, the improvements compared to our baseline are significant and are in a similar range as the corresponding English samples. This again proves the effectiveness of our simple translation approach. We argue that the translation works so well because only a single or a few words need to be translated, which could be easily done by a dictionary lookup. Another reason this approach works well is arguably the web-scale size of the corpus.

**VisualSem** As shown by the results in the last two rows of Figure 3, the VisualSem approaches did not work. All results are worse than or equal to our baseline results independent of the employed model and language. These are unexpected results since it is the only multi-modal knowledge source we employ and therefore needs further investiga-

[3] See the CodaLab competition page for details.

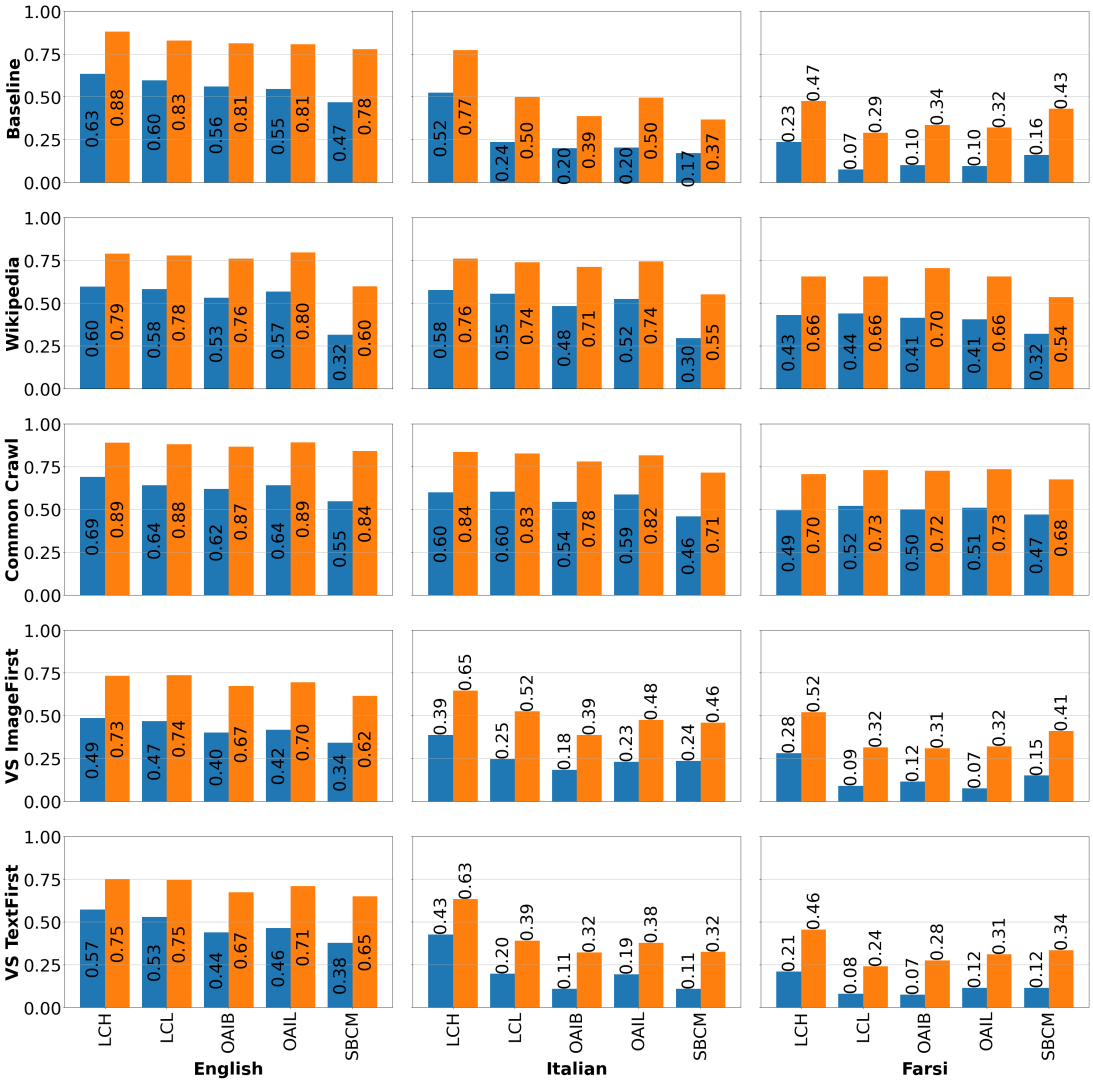


Figure 3: Evaluation results of different zero-shot VWSD approaches presented in this paper. The y-Axis label of each row describes the name of the approach. The x-Axis label of each column indicates the language of the VWSD samples, whereas the x-Axis ticks refer to the alias of the CLIP model used in the experiment (see Table 1).

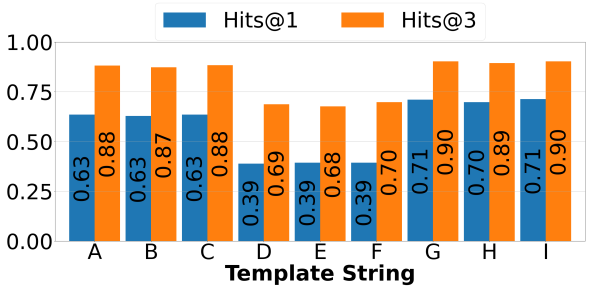


Figure 4: Evaluation results from the baseline approach using different template strings as described in Table 2

tion and error analysis in future work. Probable causes for the poor performance could be algorithmic flaws, the relatively small size of VisualSem, or ignoring meaningful but eventually important information, such as relations between the nodes, in our approaches.

## 5 Conclusion

This work presents various zero-shot Visual Word Sense Disambiguation approaches using different external knowledge sources. Across all approaches, we analyzed different pretrained versions of the CLIP model varying in size, training data, and multi-lingual capabilities. Further, we assessed the suitability of three external knowledge sources: Wikipedia, a large-scale English Common Crawl corpus, and the multi-modal knowledge graph VisualSem. Our best-performing approach involved the Common Crawl corpus which we queried for sentences containing the target word and context, serving as additional context. By translating Farsi and Italian samples into English, we achieved strong competitive results not only for English samples.

## References


Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, and Iacer Calixto. 2021. VisualSem: A High-Quality Knowledge Graph for Vision and Language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL)*, pages 138–152, Punta Cana, Dominican Republic.

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 preprint arXiv:2212.07143*.

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*, pages 4171–4186, Minneapolis, MN, USA.

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)*, Online.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide*, 1st edition. O'Reilly Media, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916, Online.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, New Orleans, LA, USA.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqing Yang. 2022. Vision-and-Language Pretrained Models: A Survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, Vienna, Austria.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten Years of BabelNet: A Survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4559–4567, Online.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2018. Building a Web-Scale Dependency-Parsed Corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

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 (ICML)*, pages 8748–8763, Online.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983, Hong Kong, China.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online.

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. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, New Orleans, LA, USA.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15617–15629, New Orleans, LA, USA.

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 (NIPS)*, volume 30, pages 5998–6008, Long Beach, CA, USA.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45, Online.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research (TMLR)*.