ECNU_MIV at SemEval-2023 Task 1: CTIM - Contrastive Text-Image Model for Multilingual Visual Word Sense Disambiguation

Zhenghui Li [1], Qi Zhang [1], Xueyin Xia [2], YinXiang Ye [2], Qi Zhang[2], and Cong Huang[3]

[1]School of Computer Science and Technology East China Normal University, China
[2]School of Communication and Electronic Engineering East China Normal University, China
[3]School Of Software JiangXi Agricultural University, China
{ zhli, xyxia, qzhang, yxye, qzhang}@stu.ecnu.edu.cn
HCong@gmail.com

Abstract

This paper describes the performance of MIV team in SemEval-2023 Task-1-Visual Word Sense Disambiguation (Visual-WSD). This task is to give a potentially ambiguous word and some limited textual context and select among a set of candidate images the one which corresponds to the intended meaning of the target word, which plays a critical role in human language understanding. Our team focuses on the multimodal domain of images and texts, we propose a model that can learn the matching relationship between text-image pairs by contrastive learning. More specifically, We train the model from the labeled data provided by the official organizer, after pre-training, texts are used to reference learned visual concepts enabling visual word sense disambiguation tasks. In addition, the top results our teams got have been released showing the effectiveness of our solution. We release our code at https://github.com/Insaner1004/VWSD.

1 Introduction

Word sense disambiguation (WSD) consists of associating words in context with their most suitable entry in a predefined sense inventory, which is a historical and hot point in natural language processing and artificial intelligence. Supervised (Luo et al., 2018; Barba et al., 2021) and knowledge-based(Moro et al., 2014; Chaplot and Salakhutdinov, 2018) are two common research methods that are widely used. However, visual word sense disambiguation aims to select among a set of candidate images the one which corresponds to the intended meaning of the target ambiguous word, which is completely different from the previous work(Raganato et al., 2023). In the real world, information often exists in different modalities, therefore, the fusion of multimodal information plays an important role in the research of deep learning.

This task is also useful for vision-and-language tasks (Lu et al., 2019) and text-to-image generator(Wang et al., 2022). In this paper, we focus on the computing of similarity between text and image track. As shown in Figure 1, given the ambiguous word "andromeda" and some limited textual context "andromeda tree", select one of the best corresponding images from candidate pictures to match the semantics according to the similarity. Apart from that, it is worth noting that this task includes datasets in three languages: English, Farsi, and Italian.

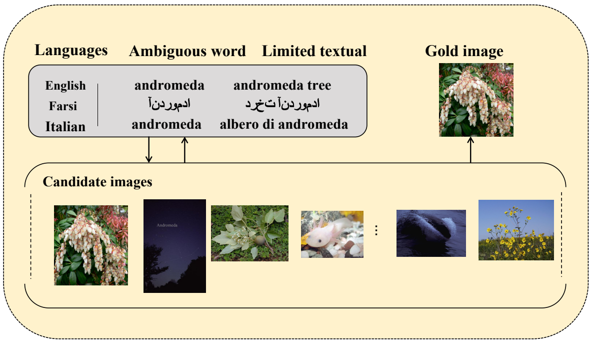

Figure 1: An example of the visual word sense disambiguation task. Given a potentially ambiguous word and some limited textual context, it is required to select a gold image form candidate images.

In recent years, pre-training methods have revolutionized NLP and achieved state-of-the-art performance in many fields(Sarzynska-Wawer et al., 2021; Devlin et al., 2018; Raffel et al., 2020). To solve the above problems effectively, we use a self-supervised pre-training method, the text is used as a supervisory signal to train a better visual model(Li et al., 2017; Radford et al., 2021). For the first, we use an image encoder and text encoder to extract feature representations of each modality separately. The second is joint multimodal embedding and calculating cosine similarities. The experimental results show that our approach has achieved competitive results.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 101–107
July 13-14, 2023 ©2023 Association for Computational Linguistics

2 Related Work

2.1 Word Sense Disambiguation

Word sense disambiguation (WSD) is the process of identifying the meaning of a word in a sentence or other contextual fragment. Previous researches mainly focus on English only, which can be divided into two groups: supervised systems and knowledge-based systems.

Integrate the context and glosses of the target word in order to make full use of both labeled data and lexical knowledge(Luo et al., 2018), a flexible toolkit (Papandrea et al., 2017) be designed for feature extraction. Besides, lexical knowledge(Blevins and Zettlemoyer, 2020) or graph structure(Bevilacqua and Navigli, 2020) has been applied to achieve impressive performance in many supervised systems (Loureiro and Jorge, 2019; Scarlini et al., 2020). Knowledge-based systems typically leverage WordNet(Miller, 1998) or BabelNet(Navigli and Ponzetto, 2012) as semantic networks.

In recent years, multilingual word sense disambiguation has become a hot research topic. OneSeC(Scarlini et al., 2019) proposed a method, which extract hundreds of thousands of sentences in Wikipedia to generate multilingual datasets automatically, Hauer(Hauer et al., 2021)apply machine translation to transfer existing sense annotations to other languages. With the unified sense representations, (Su et al., 2022) transferring annotations from rich sourced languages to poorer ones to address the annotation scarcity problem.

2.2 Contrastive Learning

In the early days of deep learning, superfluous labeling work and fixed prediction object category may restrict generality to models that transfer to other tasks. Recently, self-supervised learning has attracted much attention due to its success in computer vision(Doersch et al., 2015). Furthermore, contrastive learning is a kind of self-supervised learning method, it is also worth conducting further research.

In recent work, Zhirong Wu(Wu et al., 2018) learns a good feature representation that captures the apparent similarity among instances and saves numbers of negative samples in a discrete memory bank. According to the features of the same instance from different data augmentations should be invariant, Mang Ye(Ye et al., 2019) introduce a novel instance feature-based softmax embedding method. CPC(Oord et al., 2018) demonstrates that their approach can learn useful representations to achieve a strong performance on different modalities. MoCo(He et al., 2020)build a dynamic dictionary with a queue and a moving-averaged encoder to facilitate contrastive unsupervised learning and have a profound impact on the subsequent research work. SimCLR(Chen et al., 2020) (A simple framework for contrastive learning of visual representations) without requiring specialized architectures or a memory bank.

With previous contrastive learning research methods, different BYOL(Grill et al., 2020)achieves a new state of the art without negative samples. DINO(Caron et al., 2021) design a simple self-supervised approach that can be interpreted as a form of knowledge distillation with no labels to address the difficulties encountered by Vision Transformers (ViT)(Dosovitskiy et al., 2020).

3 System overview

3.1 Task Definition

In this subsection, we redefine the visual word sense disambiguation task to an image-text similarity task. Given a potentially ambiguous word w and some limited textual context T. The goal of this task is to select one image among a set of candidate images $M = \{\mathrm{M}_1, ..., \mathrm{M}_{|I|}\}$, which corresponds to the intended meaning of the target word. Calculate the Cosine similarity between the text embedding and the image embedding then rank the candidate images based on the calculated result. We illustrate the framework of our model in Figure 2, which consists of three components: **Multilingual Machine Translation**, **Text Encoder**, and **Image Encoder**.

3.2 Multilingual Machine Translation

We input the Italian and Farsi limited textual context in the test data into the text encoder directly and achieved poor results. The main reason is that the text encoder pre-trained focuses heavily on English and the training data in this task is English. Therefore, it is difficult to achieve the desired result when using other languages. So, it is necessary to preprocess the given context in different languages. We used a ready-made solution (multilingual machine translation model) to overcome this difficulty. M4(Arivazhagan et al., 2019) model is a massively multilingual neural machine translation

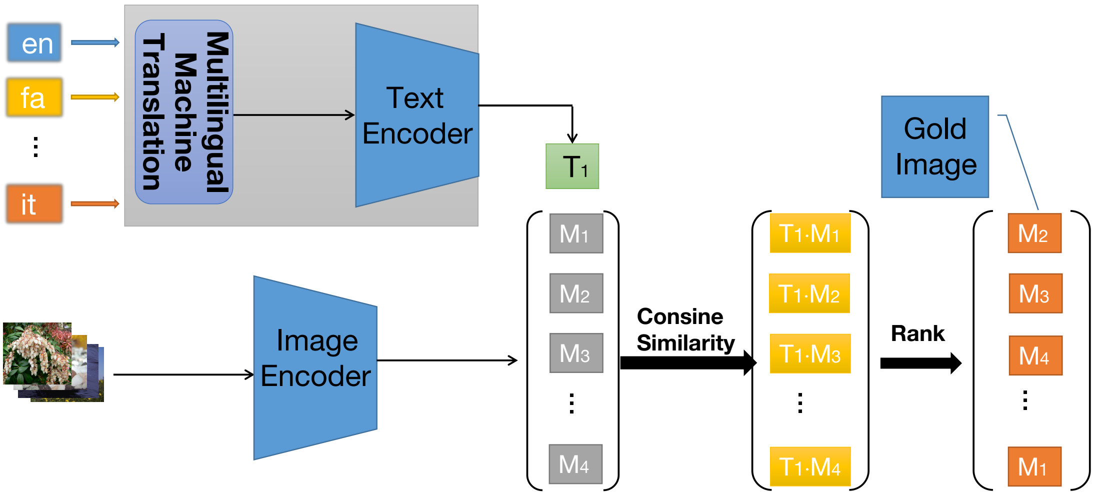

Figure 2: Overview of our approach. The text encoder embeds limited text in multiple languages into a vector space. The image encoder extracts each feature representation from candidate images and correlates to the text embeddings and then calculate the cosine similarity and rank the candidate images based on the results.

system, released by Google Research that can translate between 103 languages. The encoder converts sentences in different languages into a semantic representation vector, and the decoder gradually translates them into the corresponding target language. Specific steps are as follows: Preprocessing stage: Segment, tokenize, and encode the input text so that the model can process it.

Encoder: Encode the input text to generate a semantic representation.

$$h_t = f_{enc}(x_t, h_{t-1}) \tag{1}$$

Decoder: Use semantic representation to generate text in the target language.

$$s_{t+1} = f_{dec}(y_t, s_t, c) \tag{2}$$

Loss function:

$$L = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, x) \tag{3}$$

3.3 Text Encoder

We use a BERT(Devlin et al., 2018) network as our core architecture to extract text features. The input text is segmented, embedded, and position encoded, and then the context representation of each word is calculated by a multi-layer Transformer encoder. Finally, the representations of these words are pooled on average to obtain a vector representation of the entire sentence, which is the output of the text encoder. This vector is represented as a fixed-size text embedding vector, which can be used to calculate the similarity between texts and can also be used to calculate the similarity between text and images with the image embedding vector obtained by the image encoder.

Word embedding layer: Map each word x_i in the input text sequence $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ to a d-dimensional vector e_i, where d is the dimension of the word embedding. The calculation formula of the word embedding layer is:

$$\mathbf{E} = [e_1, e_2, \ldots, e_n] = \text{Embedding}(\mathbf{x}) \tag{4}$$

Multi-head self-attention layer: Perform multi-head self-attention calculation on the word embedding vector to obtain an output vector $\mathbf{A}$ of $n \times d$, where n is the length of the text sequence, d is the hidden layer dimension. The calculation formula of the multi-head self-attention layer is:

$$\mathbf{MHA}(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O \tag{5}$$

$$head_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$

$$\mathbf{A} = \mathbf{MHA}(\mathbf{E}, \mathbf{E}, \mathbf{E}) \tag{7}$$

Among them, $Q = K = V = \mathbf{E}$ is the input of the self-attention mechanism, and W_i^Q, W_i^K, W_i^V are the weight matrices acting on Q, K, V respectively, h is the number of heads, Attention is the self-attention mechanism formula, Concat is the output of all heads spliced together, W^O is the weight matrix of the output vector.

Feedforward neural network layer: perform feedforward neural network calculation on the output vector $\mathbf{A}$ of the multi-head self-attention layer, and obtain an output vector $\mathbf{T}$ of $n \times d$. The calculation formula of the feedforward neural network layer is:

$$\mathbf{T} = \max(0, \mathbf{A}W_1 + \mathbf{b}_1)W_2 + \mathbf{b}_2 \tag{8}$$

Among them, $W_1, \mathbf{b}_1, W_2, \mathbf{b}_2$ are learnable weight matrices and bias vectors, and $\max$ is the ReLU activation function. Ultimately, the output vector of the text encoder is $\mathbf{T}$ for $n \times d$.

3.4 Image Encoder

We leverage Vision Transformer (ViT)(Dosovitskiy et al., 2020) model as our core architecture. Vision Transformer model has received extensive attention in the field of computer vision and achieved excellent performance in multiple tasks, which may indicate that the golden age of convolutional neural networks (CNN) in the field of computer vision last for several years will be replaced by the Vision Transformer model.

Due to the wide application of the VIT model and its great success in the field of vision, we used the VIT model in our model to extract feature representations for every candidate image. Assume $H \times W$ is the input image size and C refers to the channel. The output is a dense embedding $I \in \mathbb{R}^{H \times W \times C}$.

3.5 Cosine Similarity

After extracting feature representations including text and image, scaling cosine similarities of each modality. Cosine similarity uses the value of the angle between two vectors in the multi-modal embedding space, which compared to other measures, cosine similarity pays more attention to the difference in direction of two vectors, rather than distance or length.

$$\text{Similarity }(\mathrm{T}, \mathrm{M}_{|I|}) = \frac{\mathrm{T} \cdot \mathrm{M}_{|I|}}{\|\mathrm{T}\| \times \|\mathrm{M}_{|I|}\|} \tag{9}$$

In all cases, our model can be trained on multilingual textual context and supports flexible mounts of candidate images through the image encoder.

4 Experiments

4.1 Experimental Setup

Datasets The Visual Word Sense Disambiguation (VisualWSD) task organizers share datasets in two different phases. In the trial phase, datasets include trial data in English only, which contains 12869 train samples and 16 trial samples, including gold keys. The evaluation phase differs greatly from to trial phase, this phase is the main phase of the competition where input test sets in three different languages are available (English/Farsi/Italian). Each dataset is divided into three sections: ambiguous word, limited textual context, and candidate images. A depiction of the dataset shared by the organizer is shown in Table 1.

Table 1: The dataset provided by the organizer.

Languages	En	Fa	It
Trial phase	12869	null	null
Evaluation phase	463	200	305

Concretely, in terms of data, we are allowed to use the training/trial data set provided by the organizers, pretrained vision and language models, as well as other sources of training data (we sholud be explain in detail the data sources).

Implementation Details For the text encoder, we utilize HuggingFace's implementation of the BERT model, which has been pre-trained on English sentences for the task of language understanding. As a base size, we use a 63M-parameter 12-layer 512-wide model with 8 attention heads. The text sequence will be extracted and treated as the feature representation of the context, which is layer normalized and then linearly projected into the multimodal embedding space.

For the vision encoder, we have re-implemented the ViT model using Google's repository as a reference. We select a ViT-B/32(Dosovitskiy et al., 2020) as our core architecture due to less computation. During our experiments, we observed poor performance when directly inputting unprocessed Italian and Farsi limited textual context in the test data into the text encoder. Therefore, we introduced a multilingual machine translation module which translated the other two languages into the language that the model was pre-trained and trained on, resulting in significantly better results.

We train our model for 32 epochs and keep the model that performs best regarding HR on the trial set. We use Adam optimizer with the learning rates of 1e-5. The dimensions of word embedding are 128. The max sequence length is 512. The dropout is 0.1. The reported test results are based on the parameters that obtain the best performance on the development.

4.2 Evaluation Metrics

Visual word sense disambiguation is a ranking problem, where each instance contains a query of the target phrase in the text and multiple candidates of images, and you must predict the candidate described by the query. The organizers evaluate the quality of our model by computing the mean reciprocal rank (MRR) and hit rate (HR).

Mean Reciprocal Ranking (MRR) is a measure to evaluate systems that return a ranked list of answers to queries. For a single query, the reciprocal rank is $\frac{1}{\text{rank}}$ where rank is the position of the highest-ranked answer. If no correct answer was returned in the query, then the reciprocal rank is 0.

$$\text{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\text{rank}_i} \tag{10}$$

Hit Rate at 1 (HR) is defined as each sample having only one gold key, and the gold key should be classified correctly.

$$\text{HR} = \frac{\sum_{i=0}^{N-1}(f(x_i) == y_i)}{N} \tag{11}$$

where $\mathbf{x}_\mathrm{i}$ is the sample and the sample's gold key is $\mathrm{y_i}$,the prediction function is f. Meanwhile, the sample in the test sets and the size-N present the total number of the test sets list.

4.3 Development Results

Our main development experimental results are presented in Table 2, which trains and tests on English datasets. The top 5 HR scores are between 75.00% and 93.75%. Also, most of the scores were concentrated at 60% and 70%. For MRR scores, it ranges from 79.16% to 95.83%, which spans a wide range. Our model obtains good performance on the two evaluation metrics. Particularly, our model obtains the third place in HR.

Table 2: Top 5 Results for the trial phase.

User	HR	MRR
thanet.markchom	93.7500 (1)	94.3750 (2)
zywiolak	93.7500 (1)	95.8333 (1)
deniskokosss	81.2500 (2)	87.5000 (3)
WeiJinroad	81.2500 (2)	87.5000 (3)
leon_0712	**75.0000 (3)**	**83.3333 (8)**

4.4 Evaluation Results

The best models verified on the development sets were used on the test sets which are the official competition sets. Table 3 shows the results on the test sets for our module and compares them with the baseline, which shows that our approach significantly better than the baseline model including three different languages. Moreover, Our model has a huge improvement over the baseline model in Farsi and Italian. Table 4 displays the results of the evaluation phase.

Table 3: Results of the evaluation phase and compared with the baseline results.

Languages	Baseline	Our Model
En	60.475 / 73.876	**62.203 / 75.531**
Fa	28.500 / 46.697	**65,799 / 54.098**
It	22.623 / 42.606	**54.098 / 68.965**
Avg	37.199 / 54.393	**56.434 / 70.098**

Table 4: Results for the evaluation phase.

User	HR	MRR
ardriuno	71.8254 (1)	80.7175 (1)
zywiolak	70.4927 (2)	79.8041 (2)
Rahul	69.5666 (3)	76.5793 (4)
Chicky	68.5137 (4)	78.7965 (3)
yangqihao	64.2802 (5)	74.5810 (5)
tara101	62.3607 (6)	74.1990 (6)
Pinal-Patel	60.8968 (7)	71.7063 (7)
mabehen	58.7976 (8)	71.4432 (8)
calpt	58.0495 (9)	71.2657 (9)
arshandalili	57.4591 (10)	71.1322 (10)
leon_0712	**56.4338 (11)**	**70.0980 (11)**

5 Conclusions

In this paper, we propose a unified and general model to deal with the visual word sense disambiguation task, which was constructed so that different modalities could be colluded including the image and text. Moreover, effective interaction and fusion between multimodal information play a key role in the creation and perception of information in computer vision and deep learning research.

Different from the previous word sense disambiguation studies, we transform this task into an image-text similarity task, and the model shows competitive results in most languages. As part of future work, we plan to further improve our model better handle the problems of the multilingual task.

Acknowledgements

We express our gratitude to the organizers for providing such an inspiring competition and resolving our questions patiently. We would like to express our sincere gratitude to those who have provided detailed suggestions for our paper. Their diligent efforts are greatly appreciated.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

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.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Devendra Singh Chaplot and Ruslan Salakhutdinov. 2018. Knowledge-based word sense disambiguation using topic models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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*.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Bradley Hauer, Grzegorz Kondrak, Yixing Luan, Arnob Mallik, and Lili Mou. 2021. Semi-supervised and unsupervised sense annotation via translations. *arXiv preprint arXiv:2106.06462*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2017. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. *arXiv preprint arXiv:1906.10007*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. *arXiv preprint arXiv:1805.08028*.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Simone Papandrea, Alessandro Raganato, and Claudio Delli Bovi. 2017. Supwsd: A flexible toolkit for supervised word sense disambiguation. In *Proceedings of the 2017 conference on empirical methods in natural language processing: system demonstrations*, pages 103–108.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2019. Just "onesec" for producing multilingual sense-annotated data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 699–709.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8758–8765.

Ying Su, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. Multilingual word sense disambiguation with unified sense representation. *arXiv preprint arXiv:2210.07447*.

Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. 2022. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*.

Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219.