SzegedAI at SemEval-2023 Task 1: Applying Quasi-Symbolic Representations in Visual Word Sense Disambiguation

Gábor Berend[♠◇]
[♠]ELKH-SZTE Research Group on Artificial Intelligence
[◇]University of Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract

In this paper, we introduce our submission in the task of visual word sense disambiguation (V-WSD). Our proposed solution operates by deriving quasi-symbolic semantic categories from the hidden representations of multi-modal text-image encoders. Our results are mixed, as we manage to achieve a substantial boost in performance when evaluating on a validation set, however, we experienced detrimental effects during evaluation on the actual test set. Our positive results on the validation set confirms the validity of the quasi-symbolic features, whereas our results on the test set revealed that the proposed technique was not able to cope with the sufficiently different distribution of the test data.

1 Introduction

Multi-modal text-image encoders, such as CLIP (Radford et al., 2021) have a demonstrated capability of performing zero-shot image classification tasks over diverse benchmark datasets. The contrastive learning objective applied during the pre-training of such models make them a natural choice for the V-WSD shared task (Raganato et al., 2023), where the goal is to choose from a set of candidate images the one which suits the best some query expression including a potentially ambiguous term. That is, the shared task problem can be framed as an image retrieval task, for where the transfer learning capabilities of CLIP can be arguably exploited.

The kind of continuous hidden representations produced by neural models, including CLIP, are limited by their lack of translucency from a human perspective. Our earlier work demonstrated that by using an unsupervised technique relying on sparse coding, it becomes possible to connect static word embedding representations with human interpretable concepts and properties (Balogh et al., 2020). In (Berend et al., 2018), we successfully employed these representations for hypernym discovery (Camacho-Collados et al., 2018), surpassing the performance obtained by relying on traditional and less interpretable dense word embeddings. In (Berend, 2020), we demonstrated that the technique can be extended to contextualized word representations, substantially improving the performance in classic unimodal, text-only word sense disambiguation.

Our approach for inducing implicit semantic information to words within their context built on the observation that performing sparse coding over the contextual representations of pre-trained language models result in representations that align well with semantic categories. With our shared task participation, our goal was to investigate if similar benefits that we saw for the classic uni-modal, text-only WSD task can be observed for multi-modal CLIP-based representations in V-WSD.

Our experiments delivered mixed results, as the sparse multi-modal representations provided substantial performance gain during the development phase, when evaluation was carried out on a held-out part of the training set, but the same quasi-symbolic features proved to be detrimental for the unseen test set with a distribution shift relative to the training set.

2 Our approach

We first introduce our notation (§2.1), then describe our methodology for determining multi-modal quasi-symbolic representations of texts and images (§2.2). Finally, we reveal how we incorporated those during ranking (§2.3).

2.1 Problem setting

We are given a short query text snippet t containing an ambiguous word and a collection of 10 candidate images for that text snippet $\mathcal{I}_t = \{i_1, \ldots, i_{10}\}$, and the task is to find the image $i_p \in \mathcal{I}_t$, such that the semantic overlap between the input expression t and image i_p ($p \in \{1 \ldots 10\}$) is maximal.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1965–1971
July 13-14, 2023 ©2023 Association for Computational Linguistics

2.2 Obtaining quasi-symbolic representations

CLIP-style multi-modal encoders determine a vector representation of images and texts in a shared latent space, where images and corresponding texts describing them are supposed to be encoded by similar vectors. We shall refer to the vector representation of some input object with $\mathbf{e}(\cdot)$, i.e., $\mathbf{e}(t)$ is the embedded representation of the query text t from the text encoder part of CLIP, whereas $\mathbf{e}(i)$ refers to that from the image encoder for image i.

The way we introduced our quasi-symbolic representations is similar to (Berend, 2020). That is, we turn the dense continuous latent representations produced by some neural encoder into a sparse vector by performing sparse coding over them. This results in alternative representations, that have a high fraction of coefficients being precisely 0.

The sparse vectors that we determine can be naturally interpreted in a quasi-symbolic way, i.e., one can assign the original dense input symbol belonging to those discrete categories – each represented by a latent dimension in the transformed space – for which a particular sparse representation contains a non-zero coefficient. Moreover, as demonstrated in (Balogh et al., 2020; Berend, 2020), these new dimensions are often interpretable by humans. In Section 3.3, we shall illustrate this property of our quasi-symbolic properties for the image domain.

Upon the determination of the quasi-symbolic multi-modal representations, we first perform dictionary learning (Mairal et al., 2009), during which we solve

$$\min_{\boldsymbol{D},\boldsymbol{\alpha_i}\in\mathbb{R}^k_{\geq 0}} \sum_{i=1}^{N} \frac{1}{2}\|\boldsymbol{h_i} - \boldsymbol{D}\boldsymbol{\alpha_i}\|_2^2 + \lambda\|\boldsymbol{\alpha_i}\|_1, \quad (1)$$

with $\boldsymbol{D} \in \mathbb{R}^{d\times k}$ being a dictionary matrix, the column norms of which do not exceed 1, $\boldsymbol{\alpha_i} \in \mathbb{R}^k$ is a sparse vector of coefficients that indicate the extent to which the components of $\boldsymbol{D}$ are used for the reconstruction of $\boldsymbol{h_i} \in \mathbb{R}^d$, which is the unit-normalized continuous hidden representation of some image i. λ is a regularization coefficient for controlling the sparsity of $\boldsymbol{\alpha_i}$.

For determining $\boldsymbol{D}$, we relied only on the images of the training set of the shared task. That is, we did not use the latent representations corresponding to the textual inputs for obtaining $\boldsymbol{D}$. The reason for not treating the $\mathbf{e}(t)$ hidden representations as potential placeholders for $\boldsymbol{h_i}$ upon solving (1) is that when we did otherwise, the quality of the dictionary matrix $\boldsymbol{D}$ dropped. The utility of $\boldsymbol{D}$ was likely inferior in that case as different column vectors from $\boldsymbol{D}$ specialized for either taking part in the reconstruction of text or image input representations. With different latent dimensions specializing towards different modalities, the idea of a joint latent quasi-symbolic space would have been corrupted.

Determining a sparse, quasi-symbolic representation of some CLIP-encoded (image or text) input $\boldsymbol{h_j}$ – once the dictionary matrix $\boldsymbol{D}$ from the first phase has been completed – corresponds to solving a LASSO optimization on the hidden representation of the input object in the joint latent multi-modal space $\boldsymbol{h_j}$ as

$$\min_{\boldsymbol{\alpha_j}\in\mathbb{R}^k_{\geq 0}} \frac{1}{2}\|\boldsymbol{h_j} - \boldsymbol{D}\boldsymbol{\alpha_j}\|_2^2 + \lambda\|\boldsymbol{\alpha_j}\|_1. \quad (2)$$

In contrast to (1), where a hidden vector $\boldsymbol{h_i}$ refer to some image encoding produced by a CLIP model, in (2) $\boldsymbol{h_j}$ could refer to the encoding of some text or image modality produced by the same CLIP model $\boldsymbol{D}$ was determined for. We obtain a sparse k-dimensional vector representation of a text snippet or an image by replacing $\boldsymbol{h_j}$ by either $\mathbf{e}(t)$ or $\mathbf{e}(i)$, respectively, and solving (2). We shall denote the sparse representations produced by (2) for a given text or image input as $s(\mathbf{e}(t))$ and $s(\mathbf{e}(i))$, respectively.

The way we determine the discrete quasi-symbolic representations out of the sparse vectors is by determining the set of coordinates for which they have a non-zero coefficient. That is, we define the set of latent symbols describing an image i with a sparse representation $s(\mathbf{e}(i))$ as $S(i) = \{m | s(\mathbf{e}(i))[m] > 0, \forall m \in \{1, \ldots, k\}\}$, i.e., its coordinates for which the coefficient determined by (2) is greater than zero. $S(t)$, the set of quasi-symbolic representation for a text snippet t, can be obtained in an analogous way to $S(i)$ by using the neural encoding of t originating from a CLIP model.

2.3 Ranking framework

Given that our goal was to base our ranking on features that are quasi-symbolic and more human interpretable, we decided to employ a simple and intrinsically interpretable linear model for performing the ranking of candidate images $\mathcal{I}_t$ for a given query term t. As such, we employed a logistic regression classifier from `scikit-learn` (Pedregosa et al., 2011) the training of which we describe in the followings.

For a given input tuple $(t, \mathcal{I}_t)$, with one of the images $i_p \in \mathcal{I}_t$ being the image expected to be ranked first, we determined a feature vector $\phi(t, i)$ for all $i \in \mathcal{I}_t$ and a corresponding class label y, which was considered to take up the label `Negative` for $\mathcal{I}_t \setminus i_p$, i.e., for all the candidate images apart from i_p, while treating the class label corresponding to the expected image i_p as `Positive`. With this model, we could rank the images in $\mathcal{I}_t$ according to their predicted probability of belonging to the `Positive` class, which gave our ranking for the test cases.

We next define the kind of features that we determined for a (text, image) input pair (t, i) based on their original CLIP-based representations $\mathbf{e}(t)$ and $\mathbf{e}(i)$, and their corresponding sparse representations, the determination of which we discussed earlier in §2.2.

The most trivial feature to be induced by $\phi(t, i)$ is the cosine similarity of their original CLIP-derived embeddings $\mathbf{e}(t)$ and $\mathbf{e}(i)$. We considered the model that used this only feature as our baseline. The behavior of the ranking obtained by a model that was trained using this unique feature is essentially the same as if we performed ranking based on a simple nearest-neighbor search among the candidate images for some input term t. Notice, that this was the kind of baseline the shared task organizers also proposed.

There was an extended baseline solution we gave, which used as features not only the cosine similarity between the unit-normalized embeddings $\mathbf{e}(t)$ and $\mathbf{e}(i)$, but also their dimensionwise product of coordinates. This means that for $\mathbf{e}(t), \mathbf{e}(i) \in \mathbb{R}^d$, we defined d distinct features to be induced by $\phi(t, i)$, each corresponding to the elementwise product of the two vectors. Notice that these features can be regarded as a generalization of the cosine similarity feature, because when all the elementwise products are taken into consideration with weight one, the sum of these features is able to reconstruct the cosine similarity feature. By treating the individual dimensions separately, it becomes possible to weight the various latent dimensions differently.

The features introduced so far did not make use of the quasi-symbolic representations that we introduced in §2.2, so we introduce those next. Two additional features corresponded to (i) the sum of elementwise products of the sparse embeddings $s(\mathbf{e}(t))$ and $s(\mathbf{e}(i))$ and the Jaccard similarity between the sets of quasi-symbolic representations $S(\mathbf{e}(t))$ and $S(\mathbf{e}(i))$. Up until $d + 1$ features have been introduced based on the original dense CLIP-based representations, and 2 additional features that were based on the representations introduced in § 2.2.

Finally, we introduce k further features introduced by $\phi(t, i)$ as the sparse representations $s(\mathbf{e}(t)), s(\mathbf{e}(i)) \in \mathbb{R}^k$. The k individual features reflected by

- a value of $+1$ if a particular dimension had a nonzero coefficient for both $s(\mathbf{e}(t))$ and $s(\mathbf{e}(i))$,
- a value of -1 if a particular dimension had a nonzero coefficient for precisely one of $s(\mathbf{e}(t))$ and $s(\mathbf{e}(i))$,
- and 0 otherwise.

From the perspective of the set representation of the text and image inputs, these feature values indicated if a particular quasi-semantic property was in the intersection of both (value $+1$), or in their symmetric difference (value -1), or neither (value 0). The idea for learning a separate weight for each such latent semantic properties was that arguably the different latent aspects of words and images play a different role in the determination of the fitness of an image to a query expression, and this could be learned by the logistic regression classifier.

3 Experiments

In the following subsections, we first introduce the datasets (§3.1) and report our quantitative and qualitative results (§3.2 and 3.3).

3.1 Dataset

The training dataset included $12,869$ instances of text query and candidate image set pairs, $(t, \mathcal{I}_t)$ where the text inputs were exclusively written in English. As there was no dedicated development set released, we split these $12,869$ instances into two disjoint parts, i.e., a reduced training set of 10,000 instances and a development set of the remaining $2,869$ instances. For better comparability between the development and test set results, we report all our figures for the case when only the reduced training set consisting of 10,000 instances were used for training the logistic regression model as described in §2.3.

	Baseline		Extended baseline	
	Hit rate	MRR	Hit rate	MRR
Dev. set	0.829	0.890	0.840	0.898
Test set	0.700	0.809	0.702	0.812

Table 1: Evaluation performance of the baseline approaches not utilizing the quasi-symbolic features over the development set and the official test set.

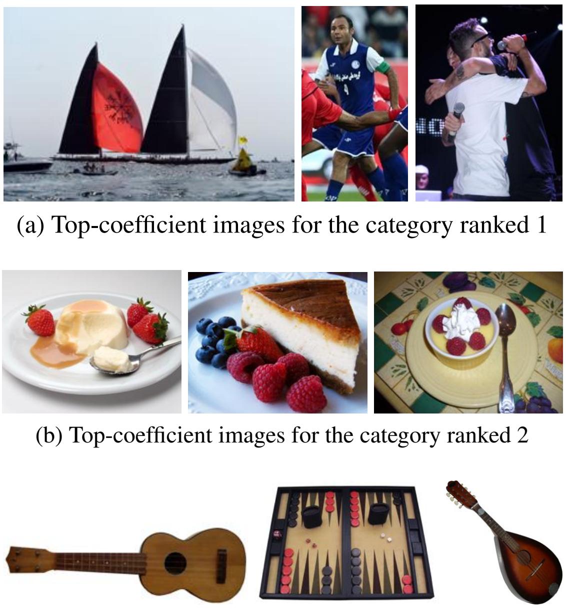
The test set contained three languages, i.e., English (en), Farsi (fa) and Italian (it), each having a corresponding number of 463, 200 and 305 instances. Here, we primarily focus on the English test, as this is the language we can compare the results with our development set results. The two metrics used during evaluation were the hit rate (proportion of test cases, where the expected gold image was ranked first), and mean reciprocal rank (MRR).

3.2 Quantitative results

We relied on the CLIP model (Radford et al., 2021) that uses XLM RoBERTa large (Conneau et al., 2020) as the text encoder, using the OpenClip (Ilharco et al., 2021) implementation[1]. The model was pre-trained over the LAION 5B dataset (Schuhmann et al., 2022), made accessible via the HuggingFace transformers library (Wolf et al., 2020).

In Table 1, we report those performance metrics that we obtained when either using the single feature of the cosine similarity of the CLIP-based encodings of texts and candidate images (Baseline), or additionally their coordinatewise products as well (Extended baseline). Comparing the two baselines we can see, that deriving features from the individual coordinates in of the dense CLIP-representations gave approximately 1 point improvement over the development set, and a marginal one in case if the test instances.

We report next in Table 2 those performance metrics that we obtained when extending our feature space with the quasi-symbolic features that we described in §2.3. As the dictionary learning component for deriving the quasi-symbolic representations involves two hyperparameters (k for the number of symbols introduced and λ controlling the strength of the regularization), we also investigated the effects of choosing them differently in the range of $\{1000, 2000, 3000\} \times \{0.05, 0.1, 0.2\}$.



(a) Top-coefficient images for the category ranked 1

(b) Top-coefficient images for the category ranked 2

(c) Top-coefficient images for the category ranked 3

Figure 1: Top-coefficient training set images of quasi-symbolic categories ranked by the weights of the logistic regression model trained over the first 10K training data.

By comparing the figures in Table 1 and Table 2, we can see that over the development set, we can see a large improvement in both metrics (irrespective of the hyperparameters used for deriving the quasi-symbolic representations), whereas for the test set, our added features caused the performance measures to drop substantially compared to our baseline approaches. The most likely explanation for this drop in performance is that the training set (and the development set that we created from it) had a different distribution of images.

Table 3 provides supportive evidence towards this explanation, in which table we report the correlation coefficients between such logistic regression models that we trained over the reduced training set (not overlapping with the development set), the development set and the test set. We can see that the models learned for the test set behave rather differently, suggesting that different quasi-symbolic features are important for giving the right answer for those instances in the test set.

3.3 Illustrating the quasi-symbolic properties

We next provide an illustration of the human interpretable clusters that our ranking models considered to be the most relevant according to the learned weights for those quasi-symbolic categories. We ranked the categories according to a model that was

[1] https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k

	λ		
k	0.05	0.1	0.2
1000	0.884	0.887	0.887
2000	0.885	0.881	0.886
3000	0.884	0.876	0.878

(a) Hit rate on English development set

	λ		
k	0.05	0.1	0.2
1000	0.930	0.932	0.932
2000	0.931	0.929	0.931
3000	0.930	0.925	0.927

(b) MRR on English development set

	λ		
k	0.05	0.1	0.2
1000	0.672	0.644	0.644
2000	0.616	0.637	0.646
3000	0.588	0.607	0.592

(c) Hit rate on English test set

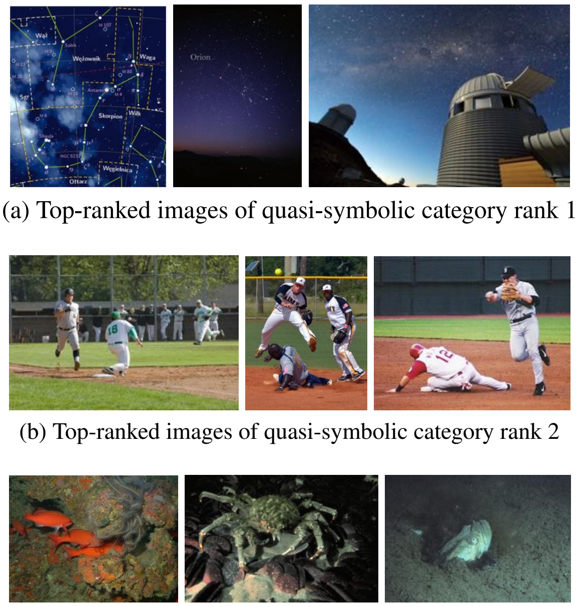
	λ		
k	0.05	0.1	0.2
1000	0.787	0.773	0.768
2000	0.757	0.766	0.769
3000	0.739	0.747	0.740

(d) MRR on English test set

Table 2: Performance of the nearest-neighbor baseline on the development set and the official test set.

	train	dev	test
train	1.00	0.90	0.37
dev	0.23	1.00	0.36
test	0.02	0.06	1.00

Table 3: Correlation coefficient between the model weights of the models trained over different datasets. Values above and under the main diagonal refer to Pearson and Spearman correlation coefficients, respectively.



(a) Top-ranked images of quasi-symbolic category rank 1

(b) Top-ranked images of quasi-symbolic category rank 2

(c) Top-ranked images of quasi-symbolic category rank 3

Figure 2: Top-coefficient training set images of quasi-symbolic categories ranked by the weights of the logistic regression model trained over the English test data.

	en	fa	it
Hit rate	0.678	0.350	0.561
MRR	0.800	0.561	0.711

Table 4: The official results of our final submission.

learned over the 10,000 element reduced training set and the 463 element test set. Figure 1 and Figure 2 provides the top-3 images having the highest coefficient in their $s(\mathbf{e}(i))$ representation along the top-3 categories that received the highest logistic regression weights for the model trained over the reduced training set and the test set, respectively.

Looking at the images we can see that the images with the highest non-zero coefficients along a certain dimension in $s(\mathbf{e}(i))$ are indeed semantically coherent. Besides that, we can also notice that the most important semantic categories for the training set (as depicted in Figure 1) and the test set (as depicted in Figure 2) are rather different, confirming our hypothesis why the introduced features proved to be harmful during testing.

3.4 Final results

We report our official evaluation scores in Table 4. These results were obtained by the predictions of a classifier which was an ensemble over all the 9 hyperparameter combinations of k and λ values that we experimented with. The models were trained differently from the ones analyzed earlier, i.e., we utilized all the official training instances instead of relying on the first 10K instances.

4 Conclusion

In this paper, we have investigated the use of quasi-symbolic representations derived from the latent representations produced by multi-modal, text-image encoders. Our qualitative evaluation suggests that the individual latent representations that we derived from the original representations have translucency from a human point of view. Additionally, the features that we extracted from these quasi-symbolic representations were helpful in boosting V-WSD performance when the distribution of the inputs was closer to that of the data that we used for training our ranking module with. On the negative side, our test scores dropped substantially compared to the ones we observed on the development data. The test set performance could be improved if more diverse training data was used, perhaps, from alternative sources such as MS COCO (Lin et al., 2014) or GCC (Sharma et al., 2018), something we consider as a potential future research direction. Our source code for replicating our experiments can be accessed via `https://github.com/szegedai/vwsd/`.

Acknowledgments

The research was supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. Additionally, we are thankful for the usage of ELKH Cloud (https://science-cloud.hu/) that helped us achieving the results published in this paper.

References

Vanda Balogh, Gábor Berend, Dimitrios I. Diochnos, and György Turán. 2020. Understanding the semantic content of sparse word embeddings using a commonsense knowledge base. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7399–7406. AAAI Press.

Gábor Berend. 2020. Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8498–8508, Online. Association for Computational Linguistics.

Gábor Berend, Márton Makrai, and Péter Földiák. 2018. 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 928–934, New Orleans, Louisiana. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 task 9: Hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 712–724, New Orleans, Louisiana. Association for Computational Linguistics.

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.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA. ACM.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R 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 Datasets and Benchmarks Track*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

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.