RCLN at SemEval-2023 Task 1: Leveraging Stable Diffusion and Image Captions for Visual WSD

Antonina Mijatovic and **Ekaterina Borisova**
Ecole Polytechnique
Palaiseau, France
antonina.mijatovic@polytechnique.edu
ekaterina.borisova.2023@polytechnique.edu

Davide Buscaldi
LIPN, Université Sorbonne Paris Nord
Villetaneuse, France
buscaldi@lipn.fr

Abstract

This paper describes the participation of the RCLN team at the Visual Word Sense Disambiguation task at SemEval 2023. The participation was focused on the use of CLIP as a base model for the matching between text and images with additional information coming from captions generated from images and the generation of images from the prompt text using Stable Diffusion. The results we obtained are not particularly good, but interestingly enough, we were able to improve over the CLIP baseline in Italian by recurring simply to the generated images.

1 Introduction

Word Sense Disambiguation (WSD) is a Natural Language Processing task consisting in the identification of the correct sense of a word having multiple possible meanings, i.e. an ambiguous word, based on the context in which it appears. SemEval-2023 Task 1 (Raganato et al., 2023) introduces a WSD task that includes a visual representation of the word meaning, named Visual WSD. It consists in determining, out of a set of candidate images, which one is the most pertinent to a given input text. The challenge is that the input text contains an ambiguous word with some of its possible interpretations figured in the set of candidate images. For instance, when presented with the context "andromeda tree" and the focus word "andromeda", a Visual WSD system should choose an image of the Japanese andromeda plant, rather than an image of the Andromeda galaxy, as the former is a better representation of the meaning of "andromeda" in this context.

We approached this task from two directions: the first one was to use image captions to compare to the input text; the second one was to use the input text as a prompt to create an image using a generative model. Image captioning is the task of describing the content of an image in natural language. Various models have been proposed in the past to address this task, from description retrieval to template filling and hand-crafted natural language generation techniques. These techniques have been superseded by modern deep-learning based generative models (Stefanini et al., 2022). Current image captioning models are based on an encoder-decoder architecture. The encoder can be any pre-trained vision transformer, whose purpose is to produce one or multiple one or multiple feature vectors. The decoder is based on a large language model (usually BERT or GPT and derivatives) which uses the embeddings produced in the previous step to produce a sequence of words. In this work, we applied the captioning model that combines the ViT image transformer (Dosovitskiy et al., 2020) with GPT-2 (Radford et al., 2019) as encoder and decoder respectively, proposed in (Kumar, 2022). In Figure 1 we show examples of captions obtained with this model for some of the images in the trial dataset. Note that the model in some cases gives a description that only partially matches the content of the picture (e.g. "fruits and vegetables" instead of "melons"). However, we can take advantage of word and sentence embeddings to obtain a good score even if the match is not perfect. For instance, "andromeda tree" has a relatively higher similarity score with the image depicting flowers, as flowers and tree are close in the embedding space.

Image generation from textual prompts has been the object of increasing attention recently. In particular, models such as DALL-E[1] and Stable Diffusion[2] have captured the attention of researchers and the general public for their ability to produce realistic images with short textual inputs. For this work we used Stable Diffusion given its availability in the Huggingface repository[3]. Stable Diffusion itself is based on an encoder-decoder architecture in

[1] https://openai.com/research/dall-e
[2] https://stability.ai/blog/stable-diffusion-v2-release
[3] https://github.com/huggingface/diffusers

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2174–2178
July 13-14, 2023 ©2023 Association for Computational Linguistics

"a flower arrangement in a flower pot in a garden", 0.173

"a lake with a river and a bench", 0.086

"a variety of fruits and vegetables on a table", 0.135

Figure 1: Examples of automatically generated captions from some of the images in the trial dataset and their similarity scores with the context "andromeda tree", using the *all-mpnet-base-v2* Sentence-BERT model.

which the encoding phase consists in adding noise to the image while the decoding step learns to remove the noise. The decoding is guided by the text given as input to the model ("prompt"), which conditions the final output thanks to a cross-attention mechanism in place between a text transformer and the denoising module.

In the remainder of this article we describe first the initial experiments we carried out with the trial data and subsequently the models we used for the submission to the Task, with some analysis of the errors we identified.

2 Methodology

Our first idea was to use textual embeddings obtained from the captions of the images via Sentence-BERT (SBERT) (Thakur et al., 2021) and compare them to the SBERT embeddings of the context sentence. As it can be noted from Figure 1, even if the text "andromeda tree" is not present from the captions, it could be possible to match "andromeda tree" with the right image by taking the one with the highest similarity score. We tested this idea on the trial collection, applying the ViT-GPT2 image captioning model to the trial candidate images, and subsequently applying the SBERT MiniLM-L6-v2 model to transform the image caption and the context query into embeddings. With this setup, we obtained an MRR of 0.575, inferior to the CLIP baseline (MRR 0.734). Unfortunately, the results are affected both by errors in the captioning and the fact that SBERT model sometimes provides lower scores for the right images as the contexts are too short.

The second experiment was to use the Stable Diffusion model on the query to produce an image and compare this image to the candidate images. To obtain an embedding for the images we applied the same CLIP model and calculated similarity using cosine measure. With this setup, we obtained a MRR of 0.580.

Given these preliminary results, it was clear that these 0-shot approaches from pre-trained models were inferior to CLIP alone. With the availability of training data, we planned to build a neural network to fine-tune the SBERT and Stable Diffusion models on them, also in combination with CLIP. Inspired by the model used for Sentence BERT, we built a siamese network where the vectors coming from the target sentence and the candidate images are encoded in the same way, following a fully connected network with a first layer of 256 units and a second one of 512 units. A dropout with probability 0.2 is in place between the two fully connected layers. We tested in the preliminary phase different activation functions, in particular LeakyReLU and tanh, without any significant differences. The loss function for the siamese network is a cosine embedding loss with margin 0.8 (we tested margins of $0.2, 0.5$ and 0.8 on trial data, with the last one achieving the best results). The cosine embedding loss is defined as:

$$L(x, y, \theta) = \begin{cases} \cos(x_i, y_i) & \text{if} \theta_i = 1 \\ \max(0, \mu - \cos(x_i, y_i)) & \text{otherwise} \end{cases}$$

where x and y are tensors of the same size representing embeddings for pairs of samples, θ is a tensor of binary labels indicating whether the pairs are similar (1) or dissimilar (-1), N is the number of samples in the batch, $\cos(x_i, y_i)$ is the cosine similarity between the i-th pair of embeddings, and

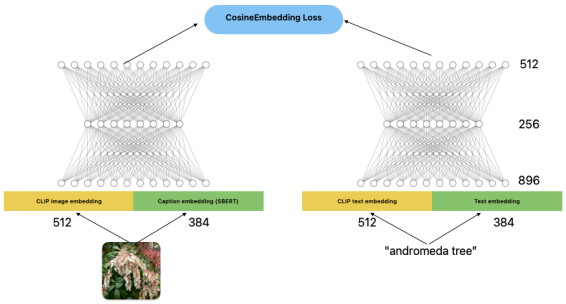

Figure 2: captions+CLIP model schema

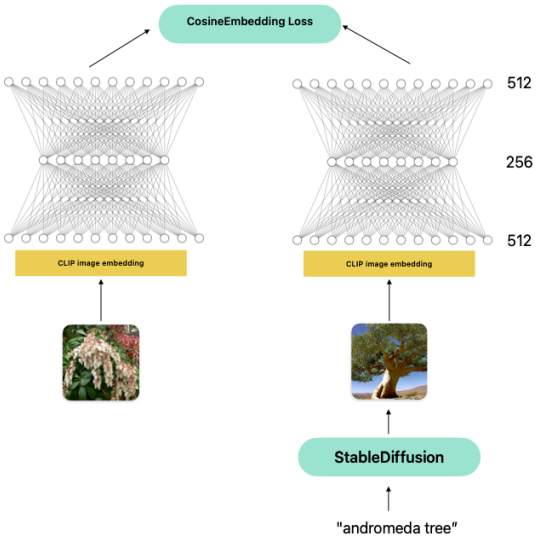

Figure 3: CLIPDiffusion model schema

μ is a hyperparameter that controls the degree of separation between the similar and dissimilar pairs.

Since we had only two possible choices, we focused on two models: the first one (captions+CLIP) uses in input a concatenation of CLIP and SBERT embeddings, resulting in a vector of size 896 (512 from CLIP and 384 from SBERT), for both the target sentence and the candidate images. The second one (CLIPDiffusion) uses the CLIP embedding for the image generated from the target sentence using Stable Diffusion: v1-4 model for English (Rombach et al., 2022) and the AltCLIP-m9 bilingual model for Italian (Chen et al., 2022). We disabled the NSFW filter in the diffusion model to avoid generating black images. Since diffusion can yield different results, we kept only the first generated image for each target text and we stored it to avoid to re-generate obtaining a different result. An overview of the models can be seen in Figures 2 and 3.

Model	English	Italian
captions+CLIP	0.625	**0.559***
CLIPDiffusion	0.590	0.459
CLIP baseline	**0.738**	0.426

Table 1: MRR (Mean Reciprokal Rank) obtained in the task by the two models, compared to the CLIP baseline. $*$-result not submitted to the challenge.

For the training, we randomly split the training data into train and validation sets, respectively with 90% and 10% of the training data. The models were trained for 5 epochs (when the validation loss started to increase), with batch size 16.

3 Results and Discussion

We submitted two runs with each of the models above for English and we submitted only the Diffusion model run for Italian. This was due to the fact that, given that no reliable captioning model was available at the time of participation, the only possibility to apply the captions+CLIP model was to translate the contexts into English. This option seemed to potentially introduce more noise into the model and we discarded it.

The obtained results are shown in Table 1, including the result of the model that we discarded. The results are compared with the CLIP only baseline. In general, the results were disappointing in particular for English, and even more in comparison with other participants who were able to outperform the baseline. However, it is interesting to notice that in the case of Italian the Stable Diffusion-based model obtained a better score than the baseline.

We took a more detailed look into what happened with the diffusion model in the Italian dataset. We were able to identify some situations in which Stable Diffusion was not able to produce a good representation of the original text input. Please refer to Figure 4 for the visual examples mentioned in the list.

1. Strange results of the diffusion model, apparently unrelated from the original text input (ex. “gomma per smacchiare" (eraser): the generated image represents an animal);

2. Insufficient training data for the diffusion model. For instance, “asino gioco di carte" (Donkey card game) is almost unknown even in Italy, there’s high probability that not even an image of this game was used in the model

training. The image produced by the diffusion is made by pictues of donkey-like animals arranged as cards in a 3x3 square;

3. The image produced by the model is affected by the ambiguity: for example, for the context "alfiere in diagonale" (bishop moves diagonally) the bishop is a human bishop, and for "colonna missione" (mission column) the column is interpreted not in military sense but in the architecture one;

4. The choice for the produced image is not particularly representative of the expressed concept (for instance, the image generated for the text "adamo ed eva" is representing what would seem a couple in the late XIX century);

5. The produced image is affected by the tokenization of the original sentence (for instance, the image generated for "box per infanti" (children playpen) seems to be actually produced for "boxe per infanti" (children boxing)

Some of these results depend on the way Stable Diffusion is trained. In particular, the self-attention mechanism may have been misled by giving too much weight to the ambiguous parts of the context. For instance, "diagonal" is what is important to focus on to disambiguate "alfiere" (bishop). Instead, it looks like it's defaulting to the most common sense of the target word. The last error (no.5 in the above list) seems a problem of the transfomer tokenizer which is unable to differentiate "box" from "boxe", the term commonly used in Italian to refer to the sport.

4 Conclusions

Although our participation was unsuccessful in terms of results, we were able to learn something interesting regarding the diffusion models, which seem particularly affected by the ambiguity problem, yielding bad example images even when the context in the input text should be enough. Probably a good strategy would have been to use the training data to improve the diffusion models instead of fine-tuning the CLIP embeddings obtained from the images. The code used for this participation is available at the following address: https://github.com/dbuscaldi/VisualWSD23. The automatically generated data (captions, generated images) are stored on Zenodo at the following address: https://doi.org/10.5281/zenodo.7860213.

Limitations

The main limitation of this work is given by the fact that we cannot indicate the parameters to produce the same images using the stable diffusion model. We will store the generated images in a repository to allow other researchers to inspect the results of our generation process. Another important limitation is that these results refer exclusively to the chosen models (in particular AltCLIP-m9 and Stable Diffusion v1-4). The results may vary if other models are used to produce the images. Another limitation of this work consists in the fact that we used the *all-MiniLM-L6-v2* model for SBERT, which is faster than the others; however, from a small experiment that we carried out, it looks like larger models such as *all-mpnet-base-v2* would yield better results (for instance, the scores in Figure 1 with the MiniLM model would be higher for the third (wrong) image. Unfortunately we didn't have time to retrain our models with this configuration.

Ethics Statement

We disabled the NSFW filter in the generation models, thus possibly generating disturbing content. The input texts were also containing possibly NSFW prompts so we considered that the filter removal was necessary to participate in this task.

References

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities.

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*.

Ankur Kumar. 2022. The illustrated image captioning using transformers. *ankur3107.github.io*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

“Gomma per smacchiare"
(eraser)

“Asino gioco di carte"
(Donkey card game)

“Box per infanti"
(children playpen)

“Alfiere si muove in diagonale"
(bishop moves diagonally)

“Colonna missione"
(mission column)

“Adamo ed Eva"
(Adam and Eve)

Figure 4: Examples of generated images from the Stable Diffusion model with the generating prompt.

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

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: a survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.