HOT TOPICS IN NLP: VISUAL WORD SENSE DISAMBIGUATION

Elective in Artificial Intelligence

Author
Lorenzo Cirillo 1895955
Sapienza University of Rome
February 16, 2024

Contents

1 Dataset

To solve the Visual Word Sense Disambiguation (Visual WSD) task, we use the test data proposed by the organizers of [1]. In few words, each sample consists of a target word, a full phrase (i.e., a sentence containing the target word with some limited context), a list of ten candidate images and the gold image (i.e., the ground truth). These test data are available in English, Italian and Farsi.

2 Baseline

The baseline solution for the Visual WSD task consists of using a CLIP model [2] to compute the similarity between the full phrase and the ten candidate images. More in detail, I compare two different models: clip-vit-base-patch16 and clip-vit-base-patch32. The main difference between such models is the patch size (respectively 16 and 32). From now on, I will refer to them respectively with Base16 and Base32.

After choosing the model, I iterate over all the samples of the selected dataset and give the full phrase with the ten candidate images to the model itself. By tokenizing the full phrase and encoding the images, the obtained output contains a tensor (e.g. *logits_per_image*) with ten elements, each one representing the similarity between the full phrase and the corresponding candidate image. At this stage, I select the image with the highest similarity through the *ArgMax* function, which represents our prediction. Finally, by comparing each prediction with the corresponding gold image, I compute the accuracy on each of the three datasets for each model, reported in Table 1.

3 Proposed Solution

What I have discussed so far is considered as a baseline. To improve its performances, I decided to translate the italian and farsi full phrases and target words to English. Indeed, the CLIP models work better with english sentences but, in contrast, some context can be lost during the translation due to the fact that each language attributes a specific meaning to a word in a sentence. The accuracy values are reported in Table 2.

Trying to improve further the performances, my personal proposal is to consider another phrase (called added phrase) in addition to the initial full phrase. By doing so, I have another comparison term by confronting the added phrase with each of the candidate images. Therefore, to get the predicted image, I make the prediction by comparing the initial full phrase with the candidate images. If the prediction is not equal to the corresponding gold image (i.e., the ground truth), I compute again the prediction but considering now the added phrase. The latter is chosen by following this flow: find the synsets of the target word through Wordnet [3] and compute the cosine similarity between the full phrase and the definition of each synset. The added phrase is the definition of the synset with the highest similarity.

In this way, we have more probability to get an added phrase which has the same meaning of the full phrase thanks to the highest similarity between the definition of the synsets and the full phrase. Moreover, all the elements of the dataset are involved since the synsets are taken by starting from the target word, the best synset definition found

Table 1: Accuracy values of base16 and base 32 on the three datasets (baseline approach).

Model	English	Italian	Farsi
Base16	**0.63**	**0.22**	**0.12**
Base32	0.58	0.18	0.10

Table 2: Accuracy values of base16 and base 32 on the three datasets after translating all the full phrases and target words to English.

Model	English	Italian	Farsi
Base16	**0.63**	**0.42**	**0.42**
Base32	0.58	0.41	**0.42**

through a comparison with the full phrase and all the synsets definitions, and the prediction by comparing both the full phrase and the added phrase with the candidate images.

Regarding the italian and farsi data, I first translate them to English and successively perform the mentioned proposed method.

The results of this approach are reported in Table 3.

3.1 Results

I report in Table 1 the accuracy values obtained by achieving the Visual WSD task with the three different mentioned models tested on the baseline approach. As It can be appreciated, the Base16 performs better than the other on all the three datasets. This is due to the architecture of the models and the dataset conformation, meaning that a smaller patch size is well suited.

We can find in Table 2 the accuracy values regarding the case where the italian and farsi target words and full phrases are translated to English. It can be appreciated that there are higher values both for italian and farsi test data after the translation. As said, this is due to the fact that CLIP models handle better english sentences. The very high improvements regarding the farsi sentences is explained by the fact that Farsi is managed worse than Italian by the CLIP models.

Even better results are reported in Table 3, in which we can appreciate not only the improvement of the accuracy values on the Italian and Farsi dataset, but also on the English dataset. This means that, in some cases, the added phrase is able to provide a well grounded similarity with the candidate images and, as a consequence, the prediction is more reliable. By the way, these are the best results among the three approaches (i.e., baseline, translation, added phrase + translation). Indeed, with respect to the baseline, there is an improvement of +15.9% and +22.4% respectively for Base16 and Base32 on the english dataset.

Table 3: Accuracy values of base16 and base 32 on the three datasets after translating all the target words and full phrases to English and considering the added phrase.

Model	English	Italian	Farsi
Base16	**0.73**	**0.52**	**0.53**
Base32	0.71	0.50	**0.53**

4 Conclusion

I solved the Visual WSD task proposed in [1] by adopting three different approaches. Regarding the first one, it is simply the baseline in which a CLIP model is used to predict the most similar image to the full phrase. The second one provides a translation of italian and farsi data before give them as input to the model. The results are good, but they are improved by the third approach, consisting in considering an added phrase to have a possible more reliable comparison with the candidate images. The added phrase is the definition of the synset (synsets are taken by using Wordnet on the target word) with the highest similarity with the full phrase, in order to not lose the general meaning. Its results, reported in Table 3, are the best ones considering all the three approaches, confirming the fact that the added phrase can be more valid with respect to the full phrase.

References

[1] A. Raganato, I. Calixto, A. Ushio, J. Camacho-Collados, and M. T. Pilehvar, "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, Jul. 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[3] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.