Hot topics in NLP Report: Visual WSD

Ludovico Comito
Sapienza University of Rome
comito.1837155@studenti.uniroma1.it

Abstract

This report describes the work involved in solving the third homework of the Hot topics in NLP class. The task proposed is Visual WSD, in which the model has to choose the most appropriate image given a certain context. The proposed work shows the advantages of utilizing a CLIP-based segmentation model against a standard CLIP baseline and tries to draw qualitative and quantitative explanations about the yield results.

1 Introduction

Visual Word Sense Disambiguation (VWSD) is a recently introduced task that arises from the latest developments in the field of Multimodal Natural Language Processing. Given a context sentence, a target word, and a series of candidate images, the goal of this task is to identify the image that most appropriately represents the sense of the target word for the given target word. The proposed baseline approach to tackle this problem is to utilize a pre-trained CLIP model to compute the similarities between the context and the candidate images, picking the image with the highest score. As an additional experiment, this work shows how utilizing a specialized version of CLIP for image segmentation (CLIPSeg) leads to a strong improvement in results. The intuition behind this choice is that the task of segmentation can be considered closer to a kind of disambiguation in the field of images, where the model has to properly focus on highlighting specific parts of the image and separate it from the surroundings. Finally, to deal with multilingual test data, input contexts are translated using specific translation models for Italian and Farsi.

2 Proposed architecture

2.1 CLIP

CLIP is among the most popular models in the field of multimodal learning. The model was trained on a large dataset consisting of image-text pairs, exploiting a contrastive learning based approach to maximize the cosine similarity between correct pairs against all the other possible candidates. As a result, CLIP has been shown strong capability in a variety of downstream tasks, particularly zero-shot classification. The specific version of CLIP used for this task is "*clip-vit-base-patch32*" from the Huggingface platform.

2.2 CLIPSeg

CLIPSeg is an image segmentation model that is obtained by adding a transformer-based decoder on top of the standard CLIP. The decoder takes as input the embeddings generated by CLIP and generates a binary segmentation. This model proved to be capable of being able to deal also with zero-shot and one-shot learning scenarios.

2.3 Proposed architecture

English test set In the English test set case, both context and images were fed to the models without any preprocessing except for the one performed by the default preprocessors from the Huggingface implementation of the models. Both models output logits for each context-image pair, which are the result of the cosine similarity between the respective embeddings.

Italian and Farsi test set As both CLIP and CLIPSeg are not natively trained on multilingual datasets, feeding data in their original language would lead to very poor results. For this reason, the proposed solution is to translate the data into English before passing it to the model. This is achieved by using specific text translation models. In the Italian case, the utilized model was Helsinki's "*opus-mt-it-en*". In the Farsi language case, the translation model of choice was mt5 base from PersianNLP. As in the English case, after being translated the text is fed to the model which will output the cosine similarities for the embeddings.

3 Methodology

3.1 Data pre-processing

The test dataset is provided in a txt format, which is parsed using the pandas library. Labels are created by converting each golden label filename to its corresponding index.

3.2 Model testing

The models were tested on the English, Italian and Farsi test set by computing their accuracy score using CLIP, CLIPSeg and CLIPSeg with Translation.

4 Results

4.1 Quantitative results

In order to establish a baseline, a first test was performed with the CLIP model on the English test dataset. This method yielded an accuracy score of 58.31%, which still proves the versatility of CLIP in zero-shot learning scenarios without any additional fine-tuning. Supporting the intuition described in the Introduction, the CLIPSeg model's accuracy reached 63.28%, significantly outperforming the baseline by +4.97%. Subsequently, a first attempt was made to test the CLIPSeg model on the Italian and Farsi datasets without any translation, yielding very poor results (18.03% for Italian and 9.5% for Farsi). As expected, the usage of translation models drastically increased the performances in both cases, scoring 50.49% for the Italian dataset and 32% for the Farsi one. Table 1 illustrates completely the obtained results.

5 Qualitative analysis

In order to further investigate the hypothesis that the CLIPSeg model has superior scene understanding capabilities, a qualitative analysis by attempting to interpret the comparison of saliency maps between both models and the segmentation results from CLIPSeg.

5.1 Saliency maps

Saliency maps allow us to interpret which regions of the input image were of particular interest with respect to the model's predictions. In this context, we want to interpret which regions of the image are closer to the text query. The idea is to compare samples where the CLIPSeg model predicted correctly and the baseline CLIP missed the prediction, to check whether this is due to CLIPSeg's better capabilities at recognizing relevant parts of the image. As can be appreciated from Figure 3 and Figure 4, CLIPSeg was better at attending to significant parts of the images that relate to the query (in particular Figure 3, where CLIP seems clueless given the query "animal tail").

5.2 Image segmentation

In another attempt to investigate CLIPSeg's capabilities at visual disambiguation, another experiment was performed by comparing different contexts corresponding to a certain ambiguous word and looking at how much the segmentation changes with the same word in different contexts for images that were correctly disambiguated. For example in Figure 1 we have the ambiguous word "madeira" and the images and contexts for "madeira river" and "madeira wine". As can be noticed, for example using the context "madeira wine" in the river image results in a much weaker segmentation that the respective correct context "madeira river". Another significant example can be seen in Figure 2 for the mercury statue. This suggests a correlation between the segmentation and disambiguation capabilities of the model.

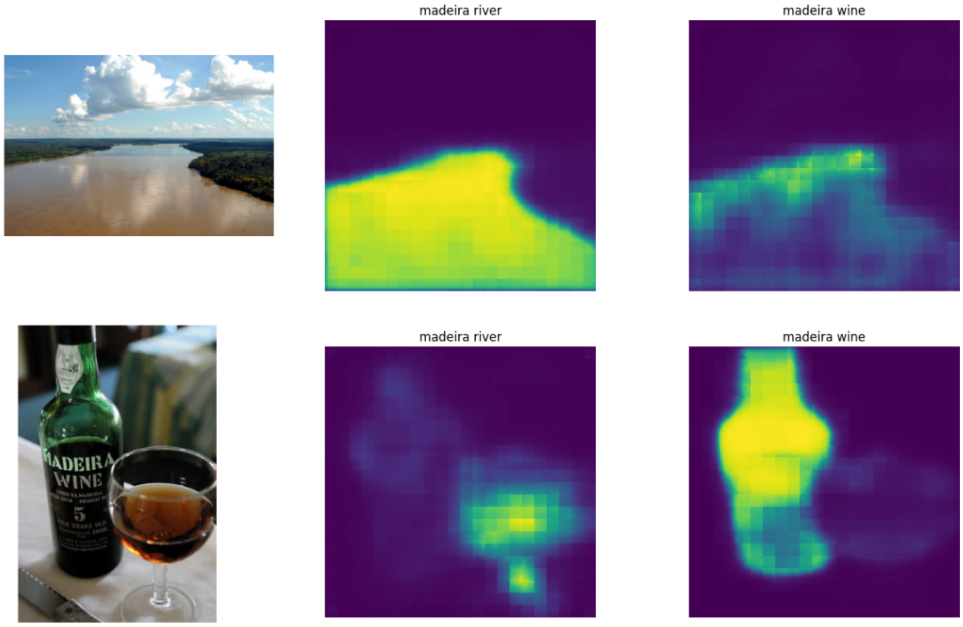

Figure 1: Segmentation results for the ambiguous word "madeira" given the contexts "madeira river" and "madeira wine" and the corresponding images.

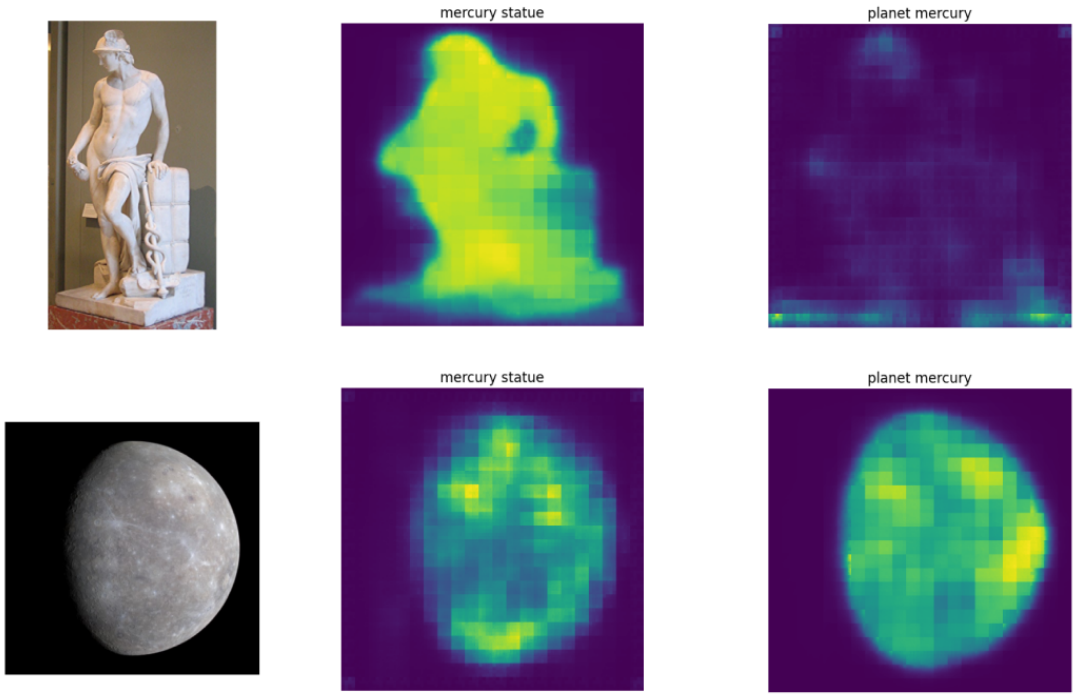

Figure 2: Segmentation results for the ambiguous word "mercury" given the contexts "mercury statue" and "planet mercury" and the corresponding images.

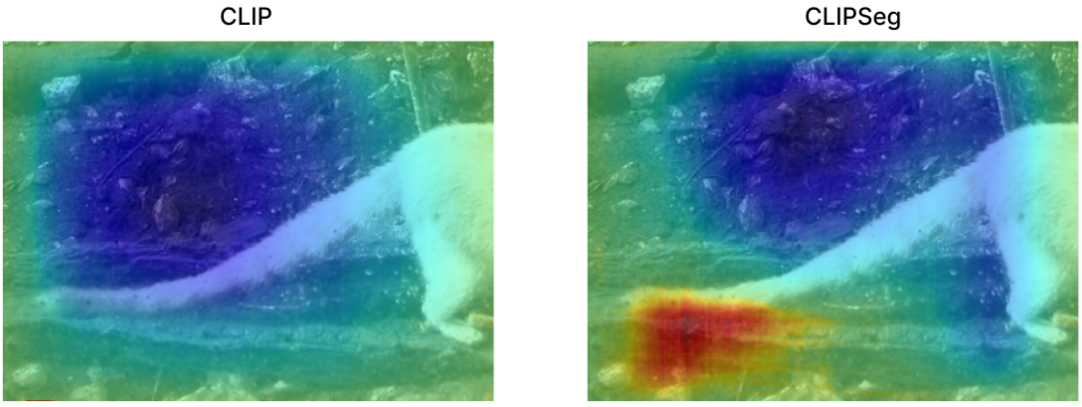

Figure 3: Comparison between saliency maps from CLIP and CLIPSeg given the query: "animal tail".

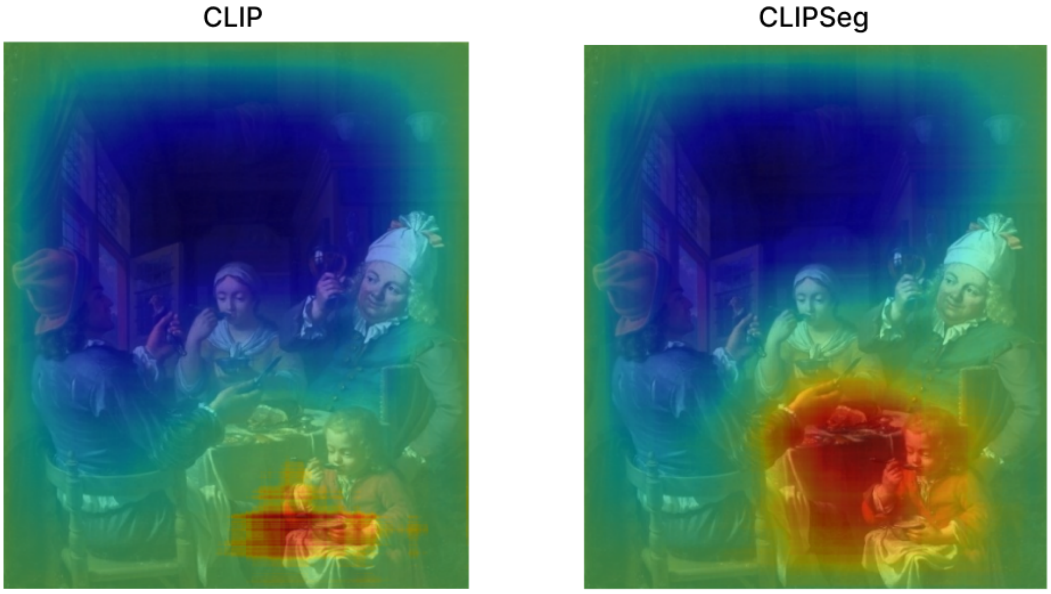

Figure 4: Comparison between saliency maps from CLIP and CLIPSeg given the query: "glutton hungry".

	English	Italian	Farsi
CLIP	58.31%	-	-
CLIPSeg	**63.28%**	18.03%	9.5%
CLIPSeg + Translation	-	**50.49%**	**32%**

Table 1: Performances results for the English, Italian and Farsi test sets.