Visual Word Sense Disambiguation using CLIP, BLIP2, and ALBEF

Lorenzo Ciarpaglini

January 10, 2025

1 Introduction

Visual Word Sense Disambiguation (VWSD) is a task in Natural Language Processing (NLP) focused on resolving semantic ambiguity associated with words in a visual context. The goal is to match a polysemous word (a word with multiple meanings) to the most relevant image within the context of surrounding sentences. This involves understanding the semantic context of the sentences and capturing meaningful relationships between text and images.

2 Models

2.1 CLIP

CLIP (Contrastive Language-Image Pre-training) is a model designed for understanding and connecting images and sentences. It accomplishes this by learning a shared representation space where the embeddings of images and corresponding textual descriptions are close to each other. CLIP employs a contrastive learning approach, encouraging the model to pull together positive pairs (matching image-text pairs) while pushing apart negative pairs (non-matching pairs).

2.2 BLIP2

BLIP2 (Bootstrapped Language Image Pretraining 2) is a model that extends the capabilities of its predecessors in understanding and connecting visual content with textual descriptions. This model leverages a large-scale pretraining approach, involving extensive datasets with image-text pairs.

2.3 ALBEF

ALBEF (Align before Fuse) is a model that emphasizes the alignment of visual and textual representations before their fusion. ALBEF's method involves separately processing the visual and textual inputs and then aligning them in a shared space, which allows for a more accurate and contextual interpretation of combined visual-textual data.

3 Experiments

In this section a description of the various experiments conducted on the 3 languages datasets (english, italian and farsi) with 2 different preprocessing strategies for the input sentences.

3.1 Preprocessing 1

In the first experiment, I augmented both the target word and the target phrase by creating two additional generic sentences aimed at providing additional context to enhance the learning of the text embedding. The prompts used are as follows:

- 'This is target word/target phrase'.
- 'This is an image that explains ¡target word/target phrase¿'

3.2 Preprocessing 2

In the second experiment, augmentation of the target word and target phrase involved using the definitions of synsets related to the target word for disambiguation. These synset definitions were added to the prompts generated in Experiment 1.

4 Conclusion

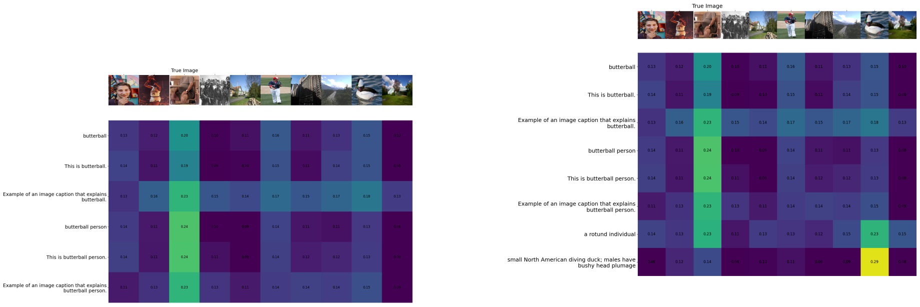
This section analyzes the results obtained by combining the three models (CLIP, BLIP2, ALBEF) with the two augmented datasets, applied across three languages: English, Italian, and Farsi. The results obtained are reported in Table 1 and Table 2. As shown the augmentation through Wordnet synsets definitions doesn't increase the performances, instead in many cases it reduces the accuracy on the predictions since it introuduces sentences that are more appropriate for other images. This is due to the impossibility to pass the golden phrase to Wordnet and in this way we are not taking the disambiguated sense definition causing a miss prediction. An example is reported in Figure 1 and Figure 2. A further implementation could be a Word Sense Disambiguation on the target word in order to get the right definition to increase the context and consequently the accuracy. Another approach could be to use a generative models to generate sentences relative to the target phrase.

Model	**English**	**Italian**	**Farsi**
CLIP	0.5	...	...
CLIPMultilingual	0.42	0.19	0.20
BLIP2	0.46	0.2	0.10
ALBEF	0.10	0.09	0.07

Table 1: Results of Experiment 1, augmenting using only the prompts sentences.

Model	**English**	**Italian**	**Farsi**
CLIP	0.45	...	...
CLIPMultilingual	0.41	0.19	0.20
BLIP2	0.45	0.2032	0.095
ALBEF	0.10	0.10	0.075

Table 2: Results of Experiment 2, augmenting with Wordnet synsets definitions.



(a) Prediction made only on prompts sentences

(b) Prediction made on prompts sentences and Wordnet synsets definitions.

Figure 1: Prediction error introduced by augmenting the input sentences by using Wordnet synsets definitions