Hot Topics In NLP

Homework 3 - Visual-WSD

1. Introduction

Our task is that given a word and some limited textual context, the task is to select among a set of ten candidate images the one which corresponds to the intended meaning of the target word. The dataset encompasses three languages: English, Italian, and Farsi. Each input includes a word, its context, and ten distinct images associated with the word. The objective is to establish the output by jointly analyzing the word and context, pinpointing the appropriate image among the ten provided that accurately correlates with our given word and context. To validate the accuracy of the output, a golden_text file is employed, containing the correct image name from the ten images that are appropriately associated with our specified word and context.

2. Description

The assigned task offers numerous viable approaches, wherein individuals may undertake the training and development of their custom image and text-based models from the ground up, utilizing the provided dataset to achieve optimal accuracy. Alternatively, one may opt to fine-tune a pre-existing model tailored for image, text, or a combination of both. Regrettably, due to constraints imposed by the limitations of my machine's capacity, I am unable to conduct training or fine-tuning on the provided dataset. Consequently, I have adopted an alternative approach, as elaborated upon below.

2.1 Approach 1:

In my initial approach, I have adopted the baseline methodology outlined in reference [1]. This involves utilizing our test dataset exclusively with the pretrained open-source multi-lingual model known as CLIP (Contrastive Language-Image Pre-Training) developed by OpenAI. CLIP is a neural network that has undergone training on a diverse set of (image, text) pairs. It possesses the capability to comprehend natural language instructions to predict the most relavant text snippet corresponding to a given image.

Given the nature of our task, the suitability of CLIP becomes evident, as our test dataset comprises text data (specifically, target words and limited context) along with corresponding images. In this approach, we input this dataset into the aforementioned model to obtain image and text encodings for each instance. Subsequently, we compute the cosine similarity between the text encoding and the entire set of images at that specific instance. The image exhibiting the highest similarity is selected as the predicted image corresponding to the target word. And then accuracy was calculated.

2.2 Approach 2:

In my second approach to enhance task accuracy, I include a brief definition of the target word along with the provided context. I used WordNet to obtain the target word's definition, or use a neutral sentence or empty vector if definition is not available. After feeding this definition and context we get the combine text encoding where then is compared with the corresponding image encoding calculating cosine similarities.

Here we had to face a problem when working with Italian and Farsi dataset because wordnet only provides definitions of English words; To tackle this problem we used deep translator library to translate the target (Italian, Farsi) word into English and then feeding it to the wordnet to get the English definition and then again translating the definition back to the original language. Which is then feed along with the context to the multilingual pretrained model to get the text encodings.

3. Results and Conclusion

The outcomes from the aforementioned approaches did not meet my anticipated level of satisfaction, particularly for the Italian and Farsi languages. In the case of the English dataset, the accuracy without a definition stood at approximately 57%. Surprisingly, incorporating a definition along with context, which was expected to improve accuracy, resulted in a reduction to 46%. A similar trend was noted for the Italian language, where no positive changes in accuracy were observed even after incorporating the definition.

Accuracy		
Language	**Approach I**	**Approach II**
English	57.67%	46.00%
Italian	28.50%	27.21%
Farsi	28.0%	-

To conclude it is evident that the model work better on English language than other languages it could be due to the training data and resource used to trained the pretrained models there are significantly more data available for English than any other language.

The other methods that can be use to improve the accuracy of the model could be using a data augmentation for text , like adding similar words or generating additional context generation using GPT. Also omitting ambiguous words from the phrase or definition could help to highlight the main context.

4. Reference:

[1] GitHub - asahi417/visual-wsd-baseline