Visual Word Sense Disambiguation

Francesco Citeroni 1758276

March 6, 2024

Introduction

In the ever-evolving landscape of natural language processing and computer vision, the task of associating words with corresponding images is of paramount importance. This report delves into three distinct approaches for solving this task, employing state-of-the-art models.

Clip

The first approach begins by tackling the task of mapping words to images through a combination of CLIP models. The 'sentence-transformers' library is employed to encode multilingual text into embeddings. The text embeddings are then compared to image embeddings generated by the `clip-ViT-B-32` model. The similarity scores are calculated using cosine similarity, allowing for the identification of the image with the highest similarity to the given text. In the context of this approach, it calculates the cosine similarity between the text embedding and each image embedding. The image with the highest similarity score is then selected.

WordNet+Lesk Algorithm

English approach

Inthis approach combining the Lesk algorithm for word sense disambiguation with text generation and image-text matching techniques. The Lesk algorithm is utilized to disambiguate word senses within given contexts, enhancing the descriptive quality of generated sentences. Text generation is facilitated by a transformer-based model, which synthesizes contextual information to produce coherent descriptions. Subsequently, a similarity comparison is conducted between the generated textual descriptions and a set of images using cosine similarity. This approach enables the model to associate textual descriptions with semantically relevant images, showcasing its efficacy in multimodal understanding tasks. The experimental evaluation demonstrates the effectiveness of the proposed technique, yielding an accuracy of 0,64. These findings underscore the potential of this approach in advancing research in natural language understanding and multimodal AI applications.

Multilingual approach

This study presents a comprehensive approach to multimodals semantic matching, focusing on Italian and Farsi language understanding and image association. The methodology involves translation, context generation, and similarity analysis to bridge textual descriptions with relevant images. Firstly, the Lesk algorithm is employed for word sense disambiguation in the Italian and Farsi context, enhancing the quality of generated descriptions. Subsequently, a transformer-based translation model synthesizes descriptive sentences, followed by context addition to enrich the textual input. The model then generates contextualized prompts for image association, leveraging a transformer-based image-text matching model to compute semantic similarity between textual and visual features. Experimental results demonstrate the effectiveness of the proposed approach.

Image Captioning

English approach

This approach presents a novel approach to image classification and captioning by leveraging sentence embeddings and cosine similarity metrics. The methodology involves iteratively comparing candidate images with a target context to identify the most semantically relevant image. Firstly, candidate images are processed to generate captions using a captioning model. Subsequently, sentence embeddings are computed for both the generated captions and the target context. By calculating the cosine similarity between these embeddings, the model identifies the candidate image that best matches the given context.

Multilingual approach

In this study, we proposed a method for cross-lingual image captioning evaluation, applied to both Italian and Farsi languages. The approach involves translating textual image descriptions from the source language to the target language using pre-trained translation models. For each translated text, we generate candidate captions for corresponding images and compute the similarity between these generated captions and the translated text. We utilize sentence embeddings and cosine similarity to measure the resemblance between the translated text and the generated captions. Through this method, we aim to assess the effectiveness of image captioning models across different languages. Experimental results demonstrate the applicability of the proposed approach in evaluating image captioning systems for Italian and Farsi languages, contributing to the advancement of cross-lingual image understanding.

Conclusion

In conclusion, all the approach offer viable solutions to the VWSD task, each with its unique strengths. Based on the tables below, it's evident that the WordNet+Lesk approach outperforms the other two methods consistently across all languages.

Approach	English	Italian	Farsi	Average accuracy
Clip	0.58	0.19	0.07	0.28
WordNet+Lesk	0.64	0.49	0.29	0.48
Image Captioning	0.28	0.40	0.05	0.25