Abstract

Recent years have seen a tremendous up flux of data. In NLP, the massive amount of data growth became the motivation for the development of text processing systems which can effectively process data. However, the data on the internet tends to be more versatile where multiple modalities like images, audio, video and text are intertwined. Hence there is a need to develop systems that can process the different modalities together.
Visual entailment is a task involving multiple modalities. It is inspired by the textual entailment task in linguistics which aims to classify the relation between a text premise and text hypothesis. The Visual Entailment task aims to classify the relation between an image premise and a text hypothesis. This study aims to analyse the use of the VE task to observe how the cross-media relations of image-caption pairs in news articles differ across categories of news. This is a challenging task as it involves both computer vision and natural language processing. Image-caption pairs along with article text are scraped using links to news articles provided in the Kaggle News Category data set. The One-For-All framework is utilised to caption images reducing the VE task to textual entailment. CLIP is applied to rank the actual and generated captions with the intent to draw conclusions depending based upon which type of caption ranks higher. Clustering and statistical analyses are performed on the CLIP generated ranks. BERT is used to find the semantic similarity between actual and generated captions.
It is found that captioning systems like OFA cannot be employed for reducing VE to TE for news articles. While the cross-media relations do not differ across the categories of news, there is significant interaction between the CLIP ranks for the actual and generated image captions and the entailment relations. A survey of different existing datasets is provided followed by a new dataset of 40K news articles containing image-caption pairs from the HuffPost website. It is hoped that this research work can act as a reference for future work in this area.