Grounding Natural Language Inference on Images

Hoa Trong VU

July 20, 2018

Abstract

Despite the surge of research interest in problems involving linguistic and visual information, exploring multimodal data for Natural Language Inference remains unexplored. Natural Language Inference, regarded as the basic step towards Natural Language Understanding, is extremely challenging due to the natural complexity of human languages. However, we believe this issue can be alleviated by using multimodal data. Given an image and its description, our proposed task is to determined whether a natural language hypothesis contradicts, entails or is neutral with regards to the image and its description. To address this problem, we develop a multimodal framework based on the Bilateral Multi-perspective Matching framework. Data is collected by mapping the SNLI dataset with the image dataset Flickr30k. The result dataset, made publicly available, has more than 565k instances. Experiments on this dataset show that the multimodal model outperforms the state-of-the-art textual model.

References