Abstract

Natural Language Inference which is regarded as the basic step towards Natural Language Understanding is extremely challenging due to the natural complexity of human languages. In this thesis, we frame the NLI task with a generation approach and propose to extend it by employing visually augmented data. The aim of the proposed model is to generate a sentence (hypothesis) given an image and its description (premise) as the input. The generated hypothesis is entailed by the input premise. For this purpose, we develop a multimodal entailment generation model which is based on the encoder-decoder architecture which is used as the base-line framework. We also develop a modified version of image captioning model in which the goal is generating the hypotheses given the visual data corresponding to the premises. For unimodal and multimodal models, SNLI and multimodal-SNLI datasets were used respectively. Multimodal-SNLI was created by mapping the SNLI pairs with their corresponding image from Flickr30k dataset. Experiments in the current work show a marginal improvement in the generation process in multimodal models compared to the textual-only framework. This can imply the usefulness of incorporating visual information in NLI tasks.