UNIVERSITY OF TRENTO

UNIVERSITY OF MALTA

Abstract

Ionut-Teodor Sorodoc

Combining visual understanding and natural language is one of the most important reseach problems in artificial intelligence. Recent research involving image captioning has shown very good results in jointly learning from images and text using models such as convolutional neural networks for object recognition and word embeddings extracted from huge amounts of text. Visual Question Answering (VQA) adds an extra layer of complexity because it has to pay attention to details, not just to give a vague description and it is a challenging testbed for multimodal systems. On one hand, neural networks have been shown to obtain excellent results on the task, using relatively simple systems which learn correlations between linguistic and visual features. On the other hand, the same systems have struggled with questions that require deeper reasoning. In particular, 'number questions' have been associated with low performance, and have only been studied superficially.

In this thesis, we focus on questions which may be answered with a quantifier (e.g., *Which proportion of dogs are black? Some/most/all of them*, etc.) relative to an image. We show that in order to learn to quantify, a multimodal model has to obtain a genuine understanding of linguistic and visual inputs and of their interaction. We build a dataset which is suitable for this task and we propose a model that extracts a fuzzy representation of the set of the queried objects (e.g., *dogs*) and of the queried property in relation to that set (e.g., *black* with respect to *dogs*), outputting the appropriate quantifier for that relation. Our model outperforms a state-of-the-art VQA system on this challenging task.