ICVSS 2016
Sicily ~ 17 - 23 July
International Computer Vision Summer School

Visual Question Answering (VQA)

Devi Parikh
Virginia Tech, USA

Abstract

I will present models, datasets, and open research questions in free-form and open-ended Visual Question Answering (VQA).

Given an image and a natural language question about the image (e.g., “What kind of store is this?”, “How many people are waiting in the queue?”, “Is it safe to cross the street?”), the machine’s task is to automatically produce an accurate natural language answer (“bakery”, “5”, “Yes”).

Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Answering any possible question about an image is one of the ‘holy grails’ of AI requiring integration of vision, language, and reasoning.

Keywords
Vision and language, deep learning, common sense reasoning, semantic scene understanding