Abstract: In this talk, I will present our work on a multi-modal AI task called Visual Question Answering (VQA) – I will provide a brief overview of the VQA task, dataset and baseline models, highlight some of the problems with existing VQA models, and talk about our works on fixing some of these problems by proposing – 1) a new evaluation protocol, 2) a new model architecture, and 3) a novel objective function. Towards the end of the talk, I will also present some very recent work towards building agents that can generate diverse programs for scenes when conditioned on instructions and trained using reinforced adversarial learning.