Visual Question Answering (VQA) is a recently proposed multimodal task in the general area of machine learning. The input to this task consists of a single image and an associated natural language question, and the output is the answer to that question. In this thesis we propose two incremental modifications to an existing model which won the VQA Challenge in 2016 using multimodal compact bilinear pooling (MCB), a novel way of combining modalities. First, we added the language attention mechanism, and on top of that we introduce an image attention mechanism focusing on objects detected in the image (“region attention”). We also experiment with ways of combining these in a single end-to-end model. The thesis describes the MCB model and our extensions and their two different implementations, and evaluates them on the original VQA challenge dataset for direct comparison with the original work.