Recent advances in computer vision, natural language processing and related areas has led to a renewed interest in artificial intelligence applications spanning multiple domains. Specifically, the generation of natural human-like captions for images has seen an extraordinary increase in interest. I will describe approaches that combine state-of-the-art computer vision techniques and language models to produce descriptions of visual content with surprisingly high quality. Related methods have also led to significant progress in answering questions about images, and differentiating multiple object instances in images. The limitations of current approaches and the challenges that lie ahead will both be emphasized.