Title:	Depth based Graph Network for VQA Counting
Author(s):	Srivastava, Deepankar
Supervisor(s):	Venkatesh, K S Namboodiri, Vinay P
Keyword(s):	Vision Language VQA Deep Learning Graph Network Counting
Subject(s):	Vision and Language Deep Learning
Abstract:	Humans make use of various cues for solving semantic tasks. One of the cues is based on depth. In this work, we leverage the use of depth estimation for solving the task of Visual Question Answering (VQA). Our approach relies on a Multimodal Graph Attention network that combines the cues obtained from image based responses with depth based responses to answer queries more accurately than other methods in Visual Question Answering. The graphs of image and depth modalities provides an e ective way to combine depth information. Our work is focused on the "Counting Questions" of VQA. We show that the performance for the highly challenging count based queries in VQA task are particularly improved. The results are validated on standard VQA datasets. We also provide graph-based visualization technique and validate our model using Human Attention Dataset.

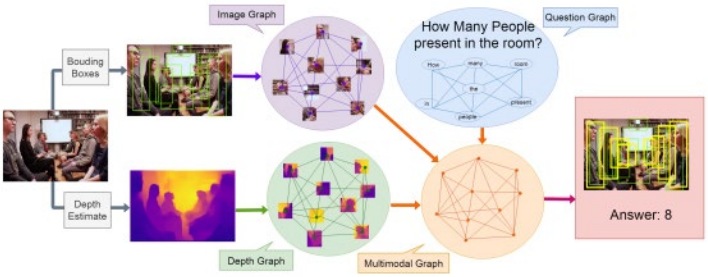


Figure 1.1: Overview of using Depth Maps and Graph Network for VQA Counting

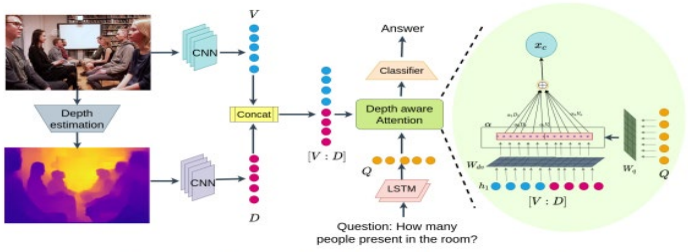


Figure 4.1: Overview of Depth-aware Attention Network

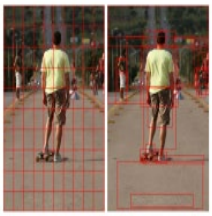


Figure 2.1: Grid vs object level attention. Image courtesy: paper [1]

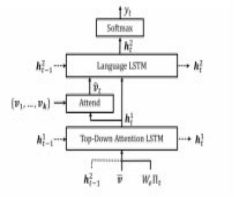


Figure 2.2: Overview of Bottom-up Top-down attention. Image courtesy: paper [1]

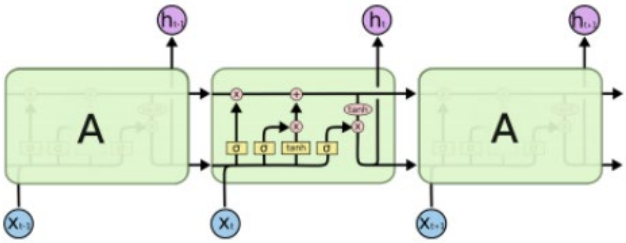
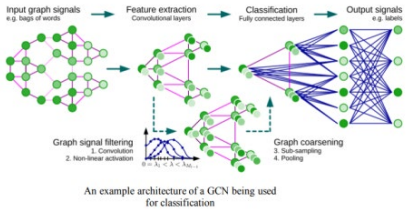


Figure 2.3: LSTM. Image courtesy: http://colah.github.io/posts/2015-08-Understanding-LSTMs/



An example architecture of a GCN being used for classification