READING ASSIGNMENT 5
16785-Integrated intelligence in robotics

[redacted]

1. Deep Compositional Question Answering with Neural Module Networks

This paper bases its approach on semantic parsing of questions to tackle the visual question answering problem. It dynamically combines learning neural module networks and integrating them into deeper networks which are then trained. It performed exceedingly well on existing datasets for visual question answering especially pertaining to objects and attributes. In addition, the authors also demonstrated above average performance on a dataset curated by them containing highly compositional questions.

A filtered convolutional layer (attention map) is used to detect different instances in a sentence. This is then used as an input to a fully connected relu layer which serves as a measurement for attributes like existence. Stacking of two or more attention maps can also occur to combine two sentences. A standard single layer LSTM in combination with module network outputs are combined and re-weighted. Their average gives the final answer. Both semantic and syntactic irregularities are accounted for.

It targets a compositional approach to visual question answering and hence it will outperform most existing algorithms when it comes to fairly complex never heard of questions. Experiments should be performed demonstrating efficiency in the same. It would be wise to extend this framework's functionality to applications such as signal processing. Improvements could be made on the encoder that targets syntactic irregularities so that it generalizes better.

2. Modeling Relationships in Referential Expressions with Compositional Modular Networks

This paper targets referential expression in natural language parsing and semantic perception. Prior work has dealt with this problem by grounding entire referential information to a single scene or localizing relationships categorically. This approach is built on combining two types of modular networks which explore locally as well as define relationships between them.

This model inputs a given image and its corresponding query. A number of bounding boxes are created on the image and a score is calculated for each box. The highest score is the localization result. The score is not a simple result of local features, it also combines other bounding box regions in space. It is evaluated on a synthetic dataset and then tested on images from the Visual Genome dataset and Google-Ref dataset.

In conclusion, the relationship module of the model accounts only for spatial features. Hence disambiguation is unnecessary. Most times it will output a high score even if the image makes no sense. It only handles three types of referential expressions: subject, relationship and object. It might wise to consider other expressions. During experimentation, sometimes poor labels were assigned and ambiguous relationships were predicted.