Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering
Medhini Narasimhan, Svetlana Lazebnik, Alexander Schwing

Overview

- **Objective**: To answer open ended questions about an image using facts from an external knowledge base.
- We use the FVQA Dataset containing image – question pairs and the corresponding FVQA Knowledge Base of facts. [1]
- We develop a model that reasons using message passing across multiple relevant facts before arriving at an answer.

Learning Knowledge Base Retrieval

1. **Retrieval of Relevant Facts**
   - Fact consists of (visual concept, relation, attribute), e.g., \((\text{Orange}, \text{IsA}, \text{Fruit})\)
   - 100 relevant facts retrieved based on GloVe similarity of the fact with the question and visual concepts in image
   - One relation out of 13 possible is obtained from the question by using an LSTM, proposed in [2]
   - Top 100 facts further reduced by filtering according to the predicted relation, e.g., IsA
   - **Entity Embedding**
     - Each entity, (visual concept, attribute) in the fact is embedded using an LSTM

2. **Question and Visual Concept Embedding**
   - **Question**: Embedding of dimension 100 learned using an LSTM
   - **Visual Concepts**: Objects, scenes, and actions detected using pre-trained models

3. **Node Embedding and Graph Construction**
   - The visual concept, question, and entity embeddings are concatenated to form an embedding of a node
   - The nodes of the graph are connected based on the relations connecting the entities

4. **Answer Prediction from the Graph**
   - A 2-layer graph convolution network (GCN) performs a joint assessment of the nodes in the graph
   - Each hidden layer of the GCN is a non-linear function given by,
     \[
     H^{(i)}(\mathbf{X}) = \sigma(D^{-1/2}A D^{-1/2}H^{(i-1)}W^{(i-1)}) \quad \forall i \in [1, \ldots, L]
     \]
   - The output of the GCN is passed through an MLP which predicts the answer

Learning

1. **Relation Prediction**
   - The LSTM is trained using ground truth question-relation pairs and standard cross-entropy loss

2. **Answer Prediction**
   - The answer predictor’s parameters consist of the question and entity embedding, the layers of the GCN and MLP
   - The LSTMs and the GCN and the MLP are trained end-to-end using the ground truth answer and binary cross-entropy loss

References