Graph-Structured Representations for Visual Question Answering

Damien Teney, Lingqiao Liu, Anton van den Hengel

Overview

Definition of visual question answering
Input: image + text question
Output: text answer from predefined set of frequent ones (classification problem)

Contribution: inputs represented as graphs
- Scenes: structured description readily available
  - nodes = objects
  - edges = relative spatial relations (dense connectivity)
- Questions: leverage existing NLP tools for syntactic parsing
  - nodes = words
  - edges = syntactic relationships (sparse connectivity)

Technical details

Propagation of information over the graph from neighbours, over several iterations

\[
h_0 = 0
\]

\[
n_t = \text{pool}_j(e_{ij} \times x_j')
\]

\[
h_t = \text{GRU}(h_{t-1}, [x_j'; n_i]) \quad t \in [1, T].
\]

Matching the graphs of questions and image

- Attention weights

\[
a_{ij} = \sigma(W_5(x_i^Q \circ x_j^S) + b_i)
\]

- Weighted sum of the features

\[
y_{ij} = a_{ij} \cdot [x_i^Q ; x_j^S]
\]

Network architecture

Results

As seen on P/R curve: the model’s output (after softmax or sigmoid) is a good measure of its confidence/uncertainty, especially when trained with soft scores as targets.

Practically, this can be used to derive the answer I don’t know.