Graph-Structured Representations for Visual Question Answering(2)

Cheng Guan

May 23, 2018

1 Introduction

Challenge The questions in the clip-art dataset vary greatly in their complexity. Some can be directly answered from observations of visual elements, e.g. Is there a dog in the room ?, or Is the weather good ?. Others require relating multiple facts or understanding complex actions, e.g. Is the boy going to catch the ball?, or Is it winter?. An additional challenge, which affects all VQA datasets, is the sparsity of the training data. Even a large number of training questions (almost 25,000 for the clip art scenes of [1] cannot possibly cover the combinatorial diversity of possible objects and concepts. Adding to this challenge, most methods for VQA process the question through a recurrent neural network (such as an LSTM) trained from scratch solely on the training questions.

Language Representation The above reasons motivate us to take advantage of the extensive existing work in the natural language community to aid processing the questions. First, we identify the syntactic structure of the question using a dependency parser [2]. This produces a graph representation of the question in which each node represents a word and each edge a particular type of dependency (e.g. determiner, nominal subject, direct object, etc.). Second, we associate each word (node) with a vector embedding pretrained on large corpora of text data [3]. This embedding maps the words to a space in which distances are semantically meaningful. Consequently, this essentially regularizes the remainder of the network to share learned concepts among related words and synonyms. This particularly helps in dealing with rare words, and also allows questions to include words absent from the training questions/answers. Note that this pretraining and ad hoc processing of the language part mimics a practice common for the image part, in which visual features are usually obtained from a ?xed CNN, itself pretrained on a larger dataset and with a different (supervised classi?cation) objective.

Scene Representation Each object in the scene corresponds to a node in the scene graph, which has an associated feature vector describing its appearance. The graph is fully connected, with each edge representing the relative position of the objects in the image.

Applying Neural Networks to graphs The two graph representations feed into a deep neural network that we will describe in Section 4. The advantage of this approach with text- and scene-graphs, rather than more typical representations, is that the graphs can capture relationships between words and between objects which are of semantic signi?cance. This enables the GNN to exploit (1) the unordered nature of scene elements (the objects in particular) and (2) the semantic relationships between elements (and the grammatical relationships between words in particular). This contrasts with the typical approach of representing the image with CNN activations (which are sensitive to individual object locations but less so to relative position) and the processing words of the question serially with an RNN (despite the fact that grammatical structure is very non-linear). The graph representation ignores the order in which elements are processed, but instead represents the relationships between different elements using different edge types. Our network uses multiple layers that iterate over the features associated with every node, then ultimately identi?es a soft matching between nodes from the two graphs. This matching re?ects the correspondences between the words in the question and the objects in the image. The features of the matched nodes then feed into a classi?er to infer the answer to the question Fig. 1.

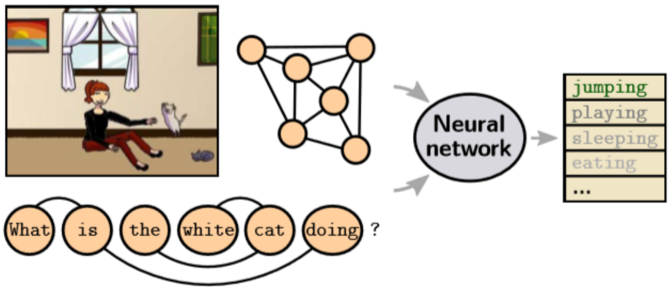

Figure 1: We encode the input scene as a graph representing the objects and their spatial arrangement, and the input question as a graph representing words and their syntactic dependencies. A neural network is trained to reason over these representations, and to produce a suitable answer as a prediction over an output vocabulary.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, and D. Parikh. VQA: Visual Question Answering. *In Proc. IEEE Int. Conf. Comp. Vis.*, 2015.

[2] M. Marneffe and C.D. Manning. The stanford typed dependencies representation. *In COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, 2008.

[3] J. Pennington, R. Socher, and C. Manning. Glove: Global Vectors for Word Representation. *In Conference on Empirical Methods in Natural Language Processing*, 2014.