Parsing Natural Scenes and Natural Language with Recursive Neural Networks

Review by

Shubham Gupta (10699)

Introduction

This paper basically introduces recursive neural networks for predicting recursive structures in multiple modalities such as the case of natural scene images or natural language sentences. This algorithm also helps to, aside from identification, understand the interaction of different sections of images to form a whole scene.

Basic Methodology

Images are divided into small regions to form several segments with features (like color, texture etc.) associated with them [1], [2]. These features are mapped into a semantic space using a neural network. Using these semantic region representations and adjacency matrix as input RNN computes

(i) A score that support neighboring segments being merged to form larger segments.
(ii) Modified semantic space and adjacency matrix for the new merged region.
(iii) Label of each node (defines the object categories such as building or street based on the training results).

The model is trained, using Max margin estimators [3], [4] and greedy approach, so that the score is high when neighboring regions have the same class label. After regions with the same object label are merged, neighboring objects are merged to form the full scene image. These merging decisions can be defined as a tree structure in which each node has associated with it the labels and scores, and higher nodes represent increasingly larger elements of the image. The tree with highest cumulative score is taken as the representation of the input image.

The same algorithm is used to parse natural language sentences. Words are used in place of image segments[5][6]. They are merged into phrases in a syntactically and semantically meaningful order. The class labels are phrase types such as noun phrase (NP) or verb phrase (VP).

Contribution

This paper has introduced a RNN to successfully merge image segments or words using recursively learned parse representations. It is the first recursive learning method to achieve state-of-the-art results on segmentation and annotation of complex scenes. This RNN architecture outperforms other methods that are based on conditional random fields or combinations of other methods. This method with a success rate of 78.1% in scene understanding outperforms previous other results (maximum 77.5%) reported on same Stanford data. With an accuracy of 88.1%, this process outperforms the Gist descriptors (Aude & Torralba, 2001), for scene categorization, which obtain only 84.0%. This algorithm is general in nature and can also parse natural language sentences obtaining competitive performance with the widely used Berkeley parser on the Wall Street Journal dataset.

References

Original Paper:
Parsing Natural Scenes and Natural Language with Recursive Neural Networks
Richard Socher Cliff Chiung-Yu Lin Andrew Y. Ng Christopher D. Manning

[1]Comaniciu, D. and Meer, P. Mean shift: a robust approach toward feature space analysis. IEEE PAMI, 24(5):603– 619, May 2002.
[2]Gould, S., Fulton, R., and Koller, D. Decomposing a Scene into Geometric and Semantically Consistent Regions. In ICCV, 2009.
[3]Taskar, B., Klein, D., Collins, M., Koller, D., and Manning, C. Max-margin parsing. In EMNLP, 2004.
[4]Manning, C. D. and Schu¨tze, H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, 1999.
[5]Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. A neural probabilistic language model. JMLR, 3, 2003.
[6]Collobert, R. and Weston, J. A unified architecture for natural language processing: deep neural networks with multitask learning. In ICML, 2008.