Object Detection Grammars

Pedro Felzenszwalb
University of Chicago
pff@cs.uchicago.edu
people.cs.uchicago.edu/p̃ff/

David McAllester*
TTI-Chciago
mcallester@ttic.edu
ttic.edu/mcallester

February 11, 2010

We formulate a general grammar model motivated by the problem of object detection in computer vision. We focus on four aspects of modeling objects for the purpose of object detection. First, we are interested in modeling objects as having parts which are themselves (recursively) objects. For example a person can be represented as being composed of a face, a trunk, arms, and legs where a face is composed of eyes, a nose and a mouth. Second, we are interested modeling object (and part) categories as being composed of subcategories or subtypes. For example we might distinguish sitting people from standing people and smiling faces from frowning faces. Third, we are interested in modeling the relative positions of the parts that make up an object. For example, in a person, the position of the hand is related to the position of the lower arm which is related to the position of the upper arm which is related to the position of the torso. Fourth, we are interested in modeling the appearance of objects so that we can find them in images. For example, a pattern of edges in a particular location of an image might give evidence for, or against, the presence of a part at that location. These four aspects of models — parts, subtypes, positions, and appearance — can be represented in a single grammar formalism that we call an object detection grammar.

We can define an object detection grammar as a set of grammar production schemas of the form

$$X(\omega_0) \xrightarrow{\alpha} \{Y_1(\omega_1), \ldots, Y_n(\omega_n)\} \tag{1}$$

where X, Y_1, …, Y_n are nonterminal symbols and ω_0, …, ω_n are "placements" and α is a score

[*] Presenting Author

(or cost) for using this production. This general formalism covers the mixture of star models that we have been using in the PASCAL competition but is more general. Productions of the form $X(\omega) \overset{\alpha}{\to} Y(\omega)$ allow the model X to select between different types of submodels. Productions of the form (1) with more than one right hand nonterminal allow a model to be broken down into part models. Productions of the form $X(\omega_0) \overset{d(\omega_0,\omega_1)}{\to} Y(\omega_1)$ can be used to model deformations where a model is found at a place ω_1 somewhat different from the nominal position ω_0. Terminal symbols compute a score from a patch of image determined by the placement of the terminal symbol.

Given an object detection grammar, detection can be done by dynamic programming analogous to the CKY algorithm for context free parsing. Each chart entry in this algorithm can be viewed as a node in a neural network. Nodes for nonterminals compute a max over the possible productions from that nonterminal and nodes for productions compute a sum over the score of the nonterminals on the right hand of the production. This interpretation gives is a conceptual analogy between a detection grammar and a max/sum convolutional neural network.

A very simple instance of a detection grammar is a mixture of pictorial structure models which has achieved considerable success in the PASCAL object detection challenge [1]. Formalizing and implementing a general class of grammar models allows direct experimentation with a much larger class of models. In many application areas involving statistical methods there is history of sophisticated models being outperformed by simple models such as n-gram language models or bag-of-word image models. However, recently there has been a general trend toward somewhat more sophisticated models. Our objective is to continue to improve performance through a steady increase in the level of model sophistication. We will describe our experiences in this direction.

References

[1] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. preprint.

Topic: Visual Processing and Pattern Recognition

Preference: oral