Encoding Prior Knowledge with Label Relation Graphs

Md Enayat Ullah, Anshul Goyal, Abheet Agarwal
Indian Institute of Technology Kanpur

Objectives

- An end-to-end module which automatically encode the label structure with a HEX graph, and tweaks the probabilistic classifier with the constraints imposed by the HEX graph.
- Tractable exact inference on the graphical model.
- Combine it with a Deep Neural Net and exploit the label structure to improve the performance measure.

Introduction

With the advent of deep learning, its plausible that we are able to unearth the underlying abstractions from raw data. However, a basic existing limitation is that they do not leverage the real world knowledge. In most multi-label classification tasks, the label space exhibits a rich structure. Traditional classification models either consider the labels mutually exclusive(softmax) or pairwise-independent(logistic regressions). However, a knowledge graph over the labels is ideal to exploit this rich structure among labels.

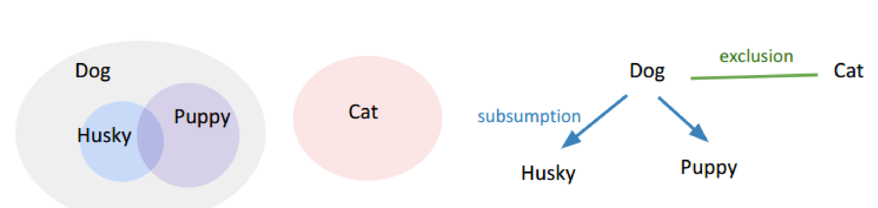

Figure 1: Heirarchy and Exclusion

Various line of works are dedicated to quantify these relations in different ways. Vedantam et al made use of human generated abstract scenes made from clipart for learning semantic visual information[1]. NEIL is another such powerful work which exploits the large-scale visual data to automatically extract commonsense relationships[2].
We draw inspiration from Deng et al, wherein they proposed a probabilistic classification model which combines powerful feature extraction with a graphical structure to encode prior beliefs[3]. They model the label constraints using a label relation graph (Hierarchical and Exclusion - HEX), and their method is shown to be at par with the state-of-the-art in a number of datasets.

Formalism

Hierarchical and Exclusion (HEX) Graph: A HEX graph $G = (V, E_h, Ee)$ is a graph consisting of a set of nodes $V = \{v_1, \cdots, v_n\}$, directed edges $E_h \subseteq V \times V$, and undirected/exclusion edges $E_e \subseteq V \times V$, such that the subgraph $G_h = (V, E_h)$ is a directed acyclic graph (DAG) and the subgraph $G_e = (V, E_e)$ has no self loop.
A HEX graph $G = (V, E_h, E_e)$ is consistent if and only if for any label $v_i \in V, E_e \cap (\alpha(v_i) \times \alpha(v_i)) = \phi$
The joint distribution of an assignment of all labels $y \in \{0, 1\}$

$$P(y|x) = \Pi_i e^{f_i(x;w)[y_i=1]} \Pi_{(v_i,v_j)\in E_h}[(yi, yj) \neq (0,1)]$$
$$\Pi_{(v_i,v_j)\in E_e}[(yi, yj) \neq (1,1)]$$

To compute the probability of a label, we need to marginalize over all labels.
We formulate the loss function of the model as the negative log likelihood using the marginal probability of the ground truth labels. The weights are estimated using Stochastic Gradient Descent(SGD).

$$L(D, w) = -\Sigma_l log Pr(y^{(l)}_{g(l)} | x^{(l)}; w)$$

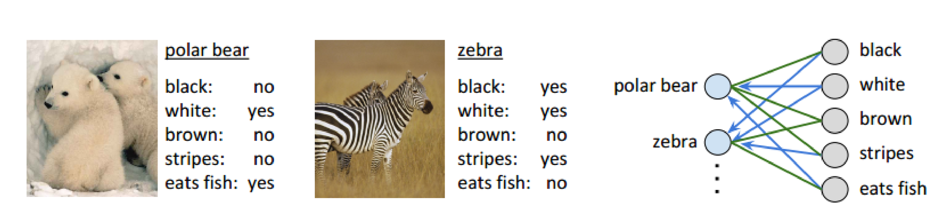

Figure 2: Label Structure

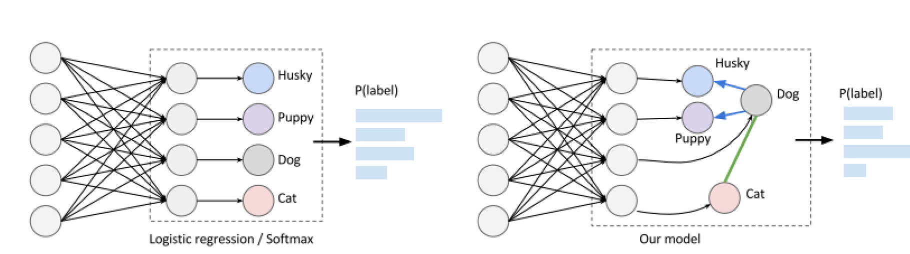

Figure 3: SoftMax vs HEX

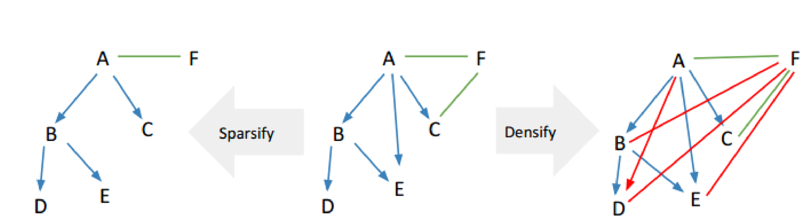

Figure 4: Sparsify and Densify

Performance

The HEX graph is constructed using the wordnet semantic hierarchy. "Exclusion wherever possible" principle is adopted to add the exclusion edges. This ensures that the graph is densely connected. Marginalizing over the labels by brute force incurs a time complexity exponential in the order of nodes. We resolve this using the modified junction tree algorithm presented in figure which runs in the order of size of the tree width.

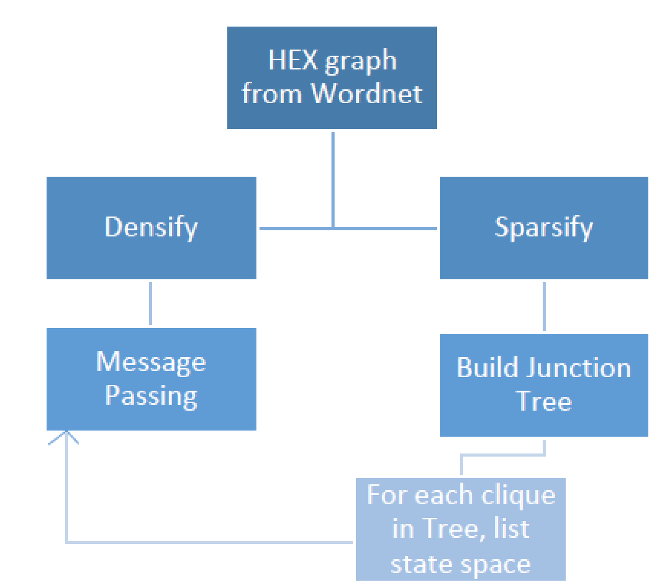

Figure 5: HEX Pipeline

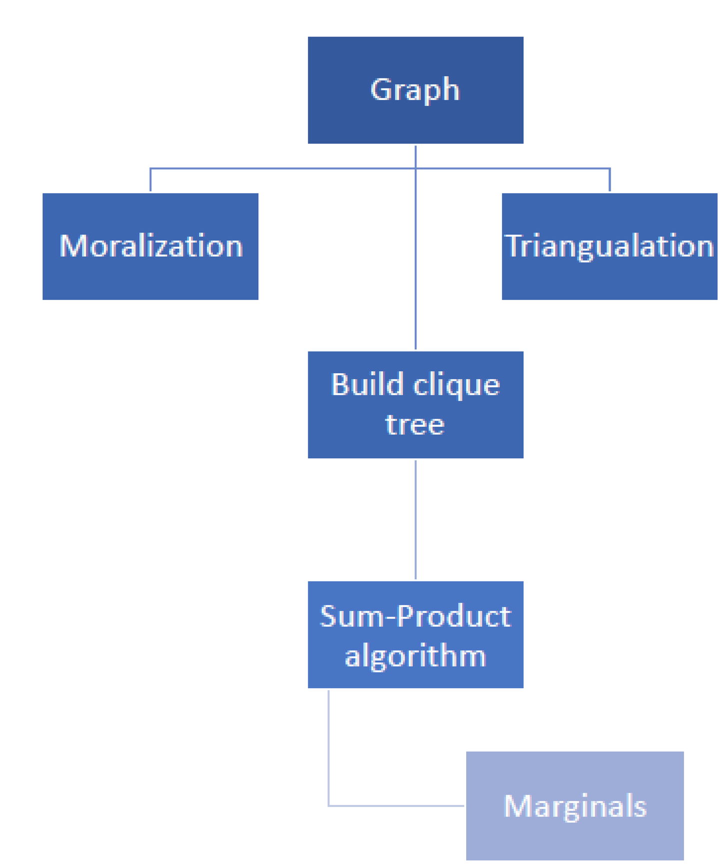

Figure 6: Junction Tree Pipeline

Dataset	SoftMax	HEX
Zoo	82.43	85.56
Forest	72.39	71.11

Table 1: Table caption

Conclusion

Based on the theoretical and experimental evidence, it is apparent that the classifier leverages on the additional constraints when the label space exhibits a rich structure. However, in cases where the labels are not semantically co-related, the performance is at par with the traditional softmax, owing to the HEX's property to reduce to a softmax when the graph consists of all mutually exclusive relations.
There are a couple of extensions to what we have achieved till now. We started with the objective of creating an end-to-end module which given the feature vectors and labels replaces the existing softmax with the automatically constructed HEX. We are currently operating on simple linear models, however the results are bound to improve if we put it over excellent feature extractors like Deep Neural nets.

References

[1] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh.
Learning common sense through visual abstraction.
In *Proceedings of the IEEE International Conference on Computer Vision*.

[2] Xinlei Chen, Ashish Shrivastava, and Arpan Gupta.
Neil: Extracting visual knowledge from web data.
In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013.

[3] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam.
Large-scale object classification using label relation graphs.
In *Computer Vision–ECCV 2014*.

[4] Nan Ding, Jia Deng, Kevin Murphy, and Hartmut Neven.
Probabilistic label relation graphs with ising models.
arXiv preprint arXiv:1503.01428, 2015.