Large Scale Structured Prediction with Hidden Variables

Alexander Schwing
ETH Zurich

Tamir Hazan
TTI Chicago

Marc Pollefeys
ETH Zurich

Raquel Urtasun
TTI Chicago

Unlike standard learning problems which involve simple scalar outputs, structured prediction deals with structured outputs such as sequences, grids, or more general graphs. Ideally, one would want to make joint predictions on the structured labels instead of simply predicting each element independently, as this additionally accounts for the statistical correlations between label elements, as well as between training examples and their labels. These properties make structured prediction appealing for a wide range of applications such as scene understanding, image segmentation, computational biology and natural language parsing.

In many structured prediction tasks, there is useful modeling information that is not available as part of the training data. For example, in object detection, one is typically given a training image with a bounding box around the object, but not the locations of the object parts. Similarly, in machine translation, one may be given the translation of a training sentence, but not the hidden linguistic structure. This missing information is important for learning a good model, and incorporating it may result in better prediction.

Structured prediction models with hidden variables have been recently proposed, including log-likelihood models such as hidden conditional random fields (HCRFs, [1]), and latent structured support vector machines (SVMs) [3]. For HCRFs, learning is done by performing gradient descent to minimize a smooth non-convex function. Learning latent structured SVMs is done by the convex-concave procedure, iteratively minimizing convex regularized structured hinge loss.

More formally, in structured prediction the goal is to learn the weighting of the features such that it best explains the training data. The weighting parameters $\boldsymbol{\theta}$ are typically learned by minimizing the norm-dependent loss

$$\sum_{(x,y)\in\mathcal{S}} \bar{\ell}(\boldsymbol{\theta}, x, y) + \frac{C}{p}\|\boldsymbol{\theta}\|_p^p,$$

defined over a training set $\mathcal{S}$. Latent structured SVMs use the structured hinge loss [3]:

$$\bar{\ell}_{hinge}(\boldsymbol{\theta}, x, y) = \max_{\hat{y}\in\mathcal{Y}, h\in\mathcal{H}} \left\{\ell(y,\hat{y}) + \boldsymbol{\theta}^\top \Phi(x,\hat{y},h)\right\} - \max_{h\in\mathcal{H}} \left\{\boldsymbol{\theta}^\top \Phi(x,y,h)\right\}.$$

HCRFs are log-linear models that use the negative log-likelihood loss [1] :

$$\bar{\ell}_{log}(\boldsymbol{\theta}, x, y) = \ln \frac{1}{\sum_h p_{x,y}(y, h; \boldsymbol{\theta})}, \qquad p_{x,y}(\hat{y}, h; \boldsymbol{\theta}) \propto \exp\left(\boldsymbol{\theta}^\top \Phi(x, \hat{y}, h)\right)$$

Many structured prediction applications with hidden information need to consider exponentially many configurations of the hidden states. For example, in scene understanding, every pixel corresponds to a binary hidden variable which indicates if it is part of the clutter or the holistic scene. The existence of many hidden configurations affects the behavior of the learning algorithm, since it implies there are many local minima.

In this work we suggest to constrain the space of hidden states by enforcing a graphical model on the hidden variables. For example, in computer vision applications, the graph adjacencies are usually inherited from the pixels or super-pixels adjacencies in the image. Thus for every feature function we relate a graphical model over the output variables $y \in \mathcal{Y}$ and the hidden variables $h \in \mathcal{H}$:

$$\phi_r(x, y, h) = \sum_{v \in V_{r,x}} \phi_{r,v}(x, y_v) + \sum_{v \in V_{r,x}} \phi_{r,v}(x, h_v) + \sum_{\alpha \in E_{r,x}} \phi_{r,\alpha}(x, y_\alpha, h_\alpha).$$

The computational efficiency depends on the learning algorithm we use. We derive a message-passing algorithm that integrates the updates of $\boldsymbol{\theta}$ with the inference of h, y. This allow us to deal with large scale problems.

We demonstrate the effectiveness of our approach on scene understanding, focusing on finding the most likely box that describes the layout of the room [2]. Intuitively, we take into account the lines in the room explained by the three dominant vanishing points, neglecting those which belong to clutter. The clutter, which is not labeled in the training images, is captured by the hidden variables. Once the correct hidden information is learned, finding the correct box can be done by considering the lines which belong to the global scene.

References

[1] A. Quattoni, S. Wang, L.P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852, 2007.

[2] H. Wang, S. Gould, and D. Koller. Discriminative Learning with Latent Variables for Cluttered Indoor Scene Understanding. *Computer Vision–ECCV 2010*, pages 435–449, 2010.

[3] C.N.J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.

Topic: Graphical Models, Structured Prediction
Preference: None