High Order Regularization for Semi-Supervised Learning of Structured Output Problems

Yujia Li, Richard S. Zemel

University of Toronto

Structured Output Learning

- Data and their labels usually have structures that need to be taken into account when making predictions.

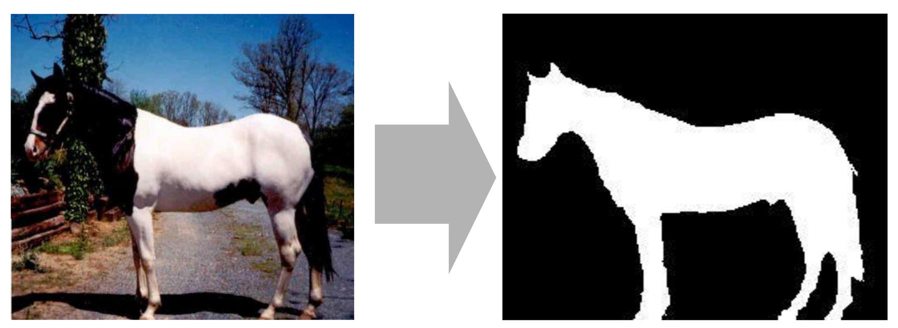


Image Segmentation

NNP VBZ DT JJ NN .

Beijing is a beautiful city .

Part of Speech Tagging

- Structured prediction models

$$\mathbf{y} = \mathrm{argmax}_{\mathbf{y}'} f(\mathbf{x}, \mathbf{y}', \mathbf{w})$$

  - Max-margin training - structured hinge loss

$$\mathcal{L} = \max_{\mathbf{y}} [f(\mathbf{x}, \mathbf{y}, \mathbf{w}) + \Delta(\mathbf{y}, \mathbf{y}^*)] - f(\mathbf{x}, \mathbf{y}^*, \mathbf{w})$$

  - Maximum likelihood training - negative log likelihood loss

$$\mathcal{L} = -\log p(\mathbf{y}^* | \mathbf{x}, \mathbf{w})$$

- Labeling structured data is very expensive, but plenty of unlabeled data is available.
  - Segmentation datasets: PASCAL VOC 2012 < 3k
  - Classification datasets: ImageNet > 1 million
  - Unlabeled images: almost infinite

Relation to Posterior Regularization

- Posterior Regularization: probabilistic models + regularizers defined on posterior distributions

$$\min_{\mathbf{w},q} \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) + \lambda R(q) + \mu \sum_{j=L+1}^{L+U} \mathrm{KL}(q_j(\mathbf{y}) || p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}_j))$$

- Temperature augmented formulation

$$p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}, T) = \frac{1}{Z_T^p} \exp\left(\frac{f(\mathbf{x}, \mathbf{y}, \mathbf{w})}{T}\right) \qquad q(\mathbf{y}, T) = \frac{1}{Z_T^q} \exp\left(\frac{g(\mathbf{y})}{T}\right)$$

$$\min_{\mathbf{w},q} \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}, T) + \lambda R(q_T) + \mu \sum_{j=L+1}^{L+U} \mathrm{TKL}(q_j(\mathbf{y}, T) || p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}_j, T))$$

- *T*-augmented PR equivalent to our formulation when *T*=0

$$\mathrm{TKL}(q_j(\mathbf{y}, T) || p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}_j, T)) \rightarrow \max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) - f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})$$

$$q_j(\mathbf{y} = 1) \rightarrow \mathbf{y}_j$$

$$\text{negative log} - \text{likelihood} \rightarrow \text{max} - \text{margin loss}$$

Semi-Supervised Learning for Structured Prediction Models

- L Labeled data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{L}$, U unlabeled data $\{\mathbf{x}_i\}_{i=L+1}^{L+U}$

$$\min_{\mathbf{w}} \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) + R\left(\{\mathbf{y}_j\}_{j=L+1}^{L+U}\right)$$

$$\text{s.t.} \quad \mathbf{y}_j = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}), \quad \forall j \geq L+1$$

- Regularizer R defined on predictions of the model on unlabeled data
  - Many expressive constraints / regularizers can be defined
  - Hard constraints make the optimization hard
- Relaxation: relax hard constraint to a (soft) penalty

$$\mathbf{y}_j = \mathrm{argmax}_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) \iff f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w}) = \max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w})$$

  - Relaxed objective ($\mathbf{Y}_U$ is a shorthand for the concatenation of all $\mathbf{y}_j$ for unlabeld images)

$$\min_{\mathbf{w}, \mathbf{Y}_U} \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) + R(\mathbf{Y}_U) + \mu \sum_{j=L+1}^{L+U} \left[\max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) - f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})\right]$$

  - Decoupled $\mathbf{Y}_U$ and $\mathbf{w}$, optimization made easy
- Alternating optimization (coordinate descent)

Step 1. Fix $\mathbf{w}$ and optimize over $\mathbf{Y}_U$

$$\min_{\mathbf{Y}_U} R(\mathbf{Y}_U) - \mu \sum_{j=L+1}^{L+U} f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})$$

  - Recent advances in optimization for models with high order potentials makes this step efficient.

Step 2. Fix $\mathbf{Y}_U$ and optimize over $\mathbf{w}$

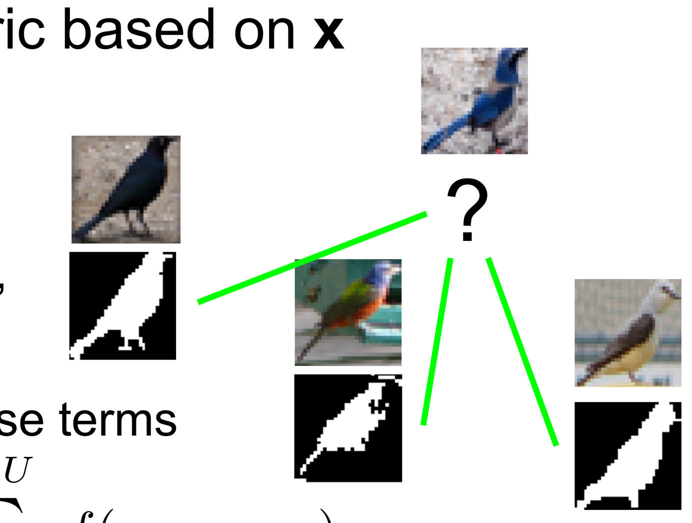
$$\min_{\mathbf{w}} \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) + \mu \sum_{j=L+1}^{L+U} \left[\max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) - f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})\right]$$

  - No harder than standard structured output learning

Example High Order Regularizers

- Graph regularizer given a similarity metric based on **x**

$$R_G(\mathbf{Y}_U) = \lambda \sum_{i,j:s_{ij}>0} s_{ij} \Delta(\mathbf{y}_i, \mathbf{y}_j)$$

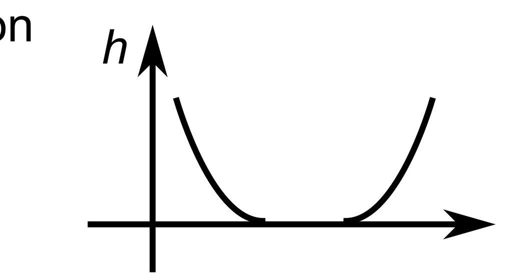


  - The more similar two examples are in **x** space, the more similar their **y** should be
  - Hamming loss - $R(\mathbf{Y}_U)$ decomposes into pairwise terms

$$\min_{\mathbf{Y}_U} \lambda \sum_{i,j:s_{ij}>0} s_{ij} \sum_{c} \Delta(y_{ic}, y_{jc}) - \mu \sum_{j=L+1}^{L+U} f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})$$

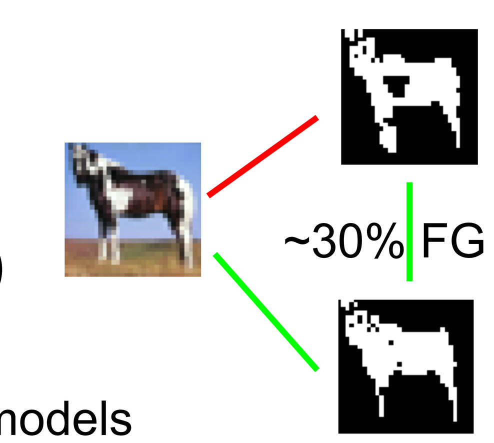
  Solvable using efficient graph-cuts based solvers for pairwise models

  - Non-decomposable loss - solvable with efficient high order loss optimization methods and dual decomposition
- Cardinality regularizer



$$R_C(\mathbf{Y}_U) = \gamma h\left(\textstyle\sum_{j,c} y_{jc}\right)$$

$$\min_{\mathbf{Y}_U} \gamma h\left(\textstyle\sum_{j,c} y_{jc}\right) - \mu \textstyle\sum_{j=L+1}^{L+U} f(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$$

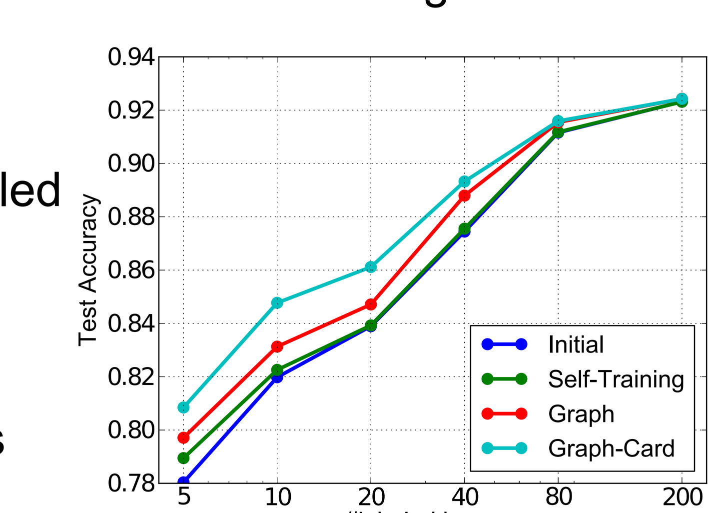


  - Efficient solver for unary only models by sorting
  - Decomposition methods for pairwise and higher order models
- Combining multiple high order regularizers
  - Dual decomposition inference

Experiments

- Settings
  - Horse: train, test on Weizmann horses unlabeled data from CIFAR-10
  - Bird: train on PASCAL, test on CUB, unlabeled data from CUB
  - See paper for a few more settings
  - Base model: pairwise CRF with NN unaries
  - Semi-supervised learning of NN parameters
- Models compared
  - Initial: pure supervised training
  - Self-Training: self-training baseline
  - Graph: SSL with graph regularizer R_G
  - Graph-Card: SSL with both graph and cardinality regularizer R_G+R_C

Horse Segmentation



Bird Transfer Learning