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

Yujia Li and Richard Zemel

University of Toronto
Canadian Institute for Advanced Research

Structured Output Problems and Models

- Rich structure in data and labels

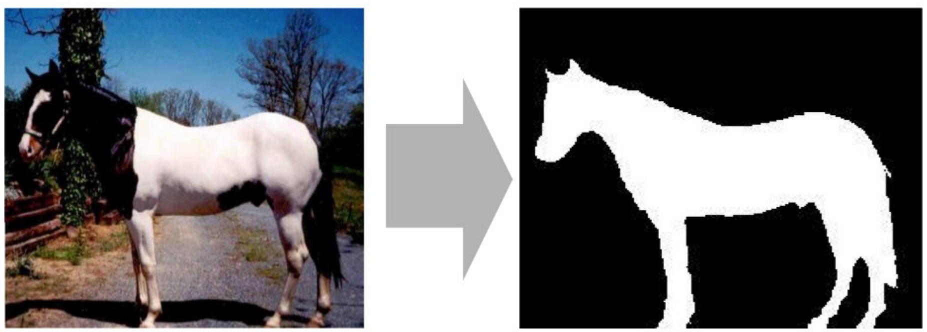


Image Segmentation

NNP VBZ DT JJ NN .

Beijing is a beautiful city .

Part of Speech Tagging

- Modeling such structure is beneficial
- Standard structured prediction models

$$\mathbf{y} = \underset{\mathbf{y}'}{\operatorname{argmax}} f(\mathbf{x}, \mathbf{y}', \mathbf{w})$$

Structured Output Learning and Challenges

- Supervised learning: loss minimization
  - Max-margin method $\mathcal{L} = \max_{\mathbf{y}} [f(\mathbf{x}, \mathbf{y}, \mathbf{w}) + \Delta(\mathbf{y}, \mathbf{y}^*)] - f(\mathbf{x}, \mathbf{y}^*, \mathbf{w})$
  - Probabilistic method $\mathcal{L} = -\log p(\mathbf{y}^* | \mathbf{x}, \mathbf{w})$
- Full labels needed for supervised learning – but they are expensive to obtain

Classification	Segmentation
ImageNet > 1M	PASCAL < 3k

  - Model capacity limited by small labeled data sets [Li et.al. 2013]
  - Semi-supervised learning is important

Regularization Based Framework of Semi-Supervised Learning

- *L* labeled examples $\{\mathbf{x}_i, \mathbf{y}_i\}$, *U* unlabeled examples $\{\mathbf{x}_j\}$

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{L} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w}) + R\left(\{\mathbf{y}_j\}_{j=L+1}^{L+U}\right)$$
$$\text{s.t.} \quad \mathbf{y}_j = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}), \quad \forall j \geq L+1$$

- Regularization based approach
  - Efficient at test time
  - Separation based methods and graph based methods fit in the framework

- Regularizer defined directly on model predictions
  - Lots of expressive regularizers can be used

Solving the Hard Optimization Problem

- The objective function is a complicated function of **w** due to the hard constraint

- Observation:

$$\mathbf{y}_j = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) \quad \Leftrightarrow \quad f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w}) = \max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w})$$

  - Constraint violation

$$\max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) - f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})$$

- Relaxed objective, $\mathbf{Y}_U = (\mathbf{y}_{L+1}, \ldots, \mathbf{y}_{L+U})$

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

Alternating Optimization

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

- Relaxation decouples *R* and **w**
- Alternating optimization:

Step 1: fix **w** solve for $\mathbf{Y}_U$ (MAP inference with high order potentials)

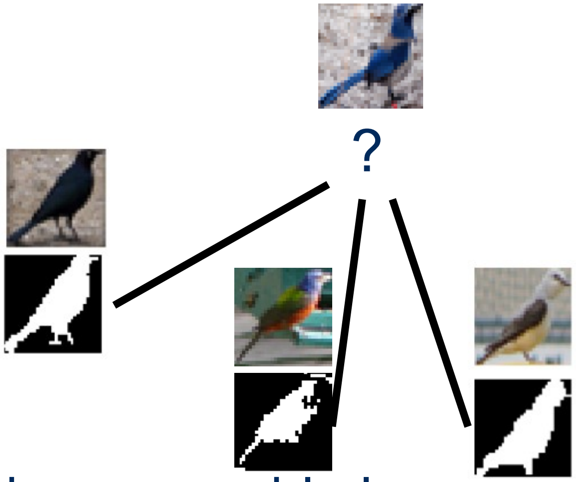
$$\min_{\mathbf{Y}_U} \quad R(\mathbf{Y}_U) - \mu \sum_{j=L+1}^{L+U} f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})$$

Step 2: fix $\mathbf{Y}_U$ update **w** (no harder than standard structured output learning)

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

Example High Order Regularizers

- Graph regularizer

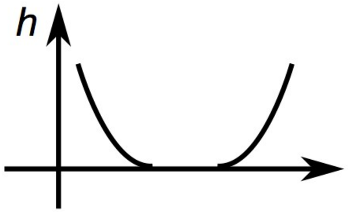


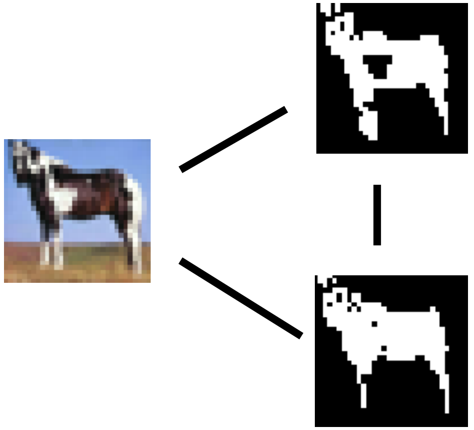
$$R_G(\mathbf{Y}_U) = \lambda \sum_{i,j:s_{ij}>0} s_{ij} \Delta(\mathbf{y}_i, \mathbf{y}_j)$$

  - Decomposable for Hamming distance
  - Efficient high order loss optimization for non-decomposable losses

- Cardinality regularizer

$$R_C(\mathbf{Y}_U) = \gamma \, h(1^\top \mathbf{Y}_U)$$





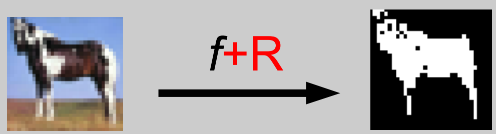
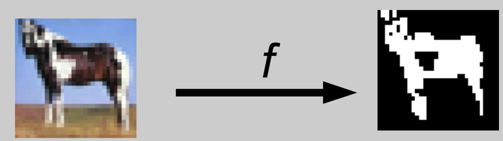
  - Efficient inference for unary models by sorting
  - Decomposition methods for pairwise models

- Combining multiple regularizers
  - Dual decomposition inference

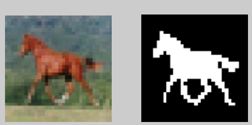
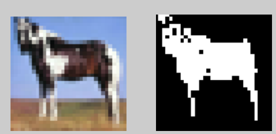
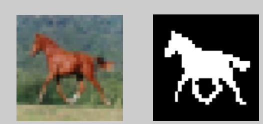
Illustration of the Learning Process

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

Step 1, fix **w** solve for $\mathbf{Y}_U$

Our Method	Self-Training
 $\xrightarrow{f+R}$	 $\xrightarrow{f}$

Step 2, fix $\mathbf{Y}_U$ update **w**

Our Method		Self-Training	
Labeled Data	Unlabeled Data	Labeled Data	Unlabeled Data
			
$\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$	$\boxed{\max_{\mathbf{y}} f(\mathbf{x}_j, \mathbf{y}, \mathbf{w}) - f(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})}$	$\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$	$\boxed{\mathcal{L}(\mathbf{x}_j, \mathbf{y}_j, \mathbf{w})}$
Update **w**		Update **w**	

Relation to Posterior Regularization

- PR [Ganchev, 2010] is a framework for probabilistic models
  - Regularizers defined on posterior distributions
  - Auxiliary distribution *q* and KL penalty

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

- Temperature-augmented formulation

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

- Equivalent to our max-margin formulation when *T*=0

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

$$q_j(\mathbf{y} = 1, T) \to \mathbf{y}_j$$

Negative log-likelihood $\rightarrow$ Max-margin loss

Experiment Settings

- Binary segmentation tasks

	Train	Test	Unlabeled
Horse	Weizmann	Weizmann	CIFAR-10
Bird	PASCAL	CUB	CUB

- Images resized to 32x32 as all images in CIFAR-10 are of this size
- Base model is a pairwise CRF, with neural network unary potentials
- Semi-supervised learning of NN parameters
- See paper for a few more settings

Models Compared

Initial: base model trained without using unlabeled data

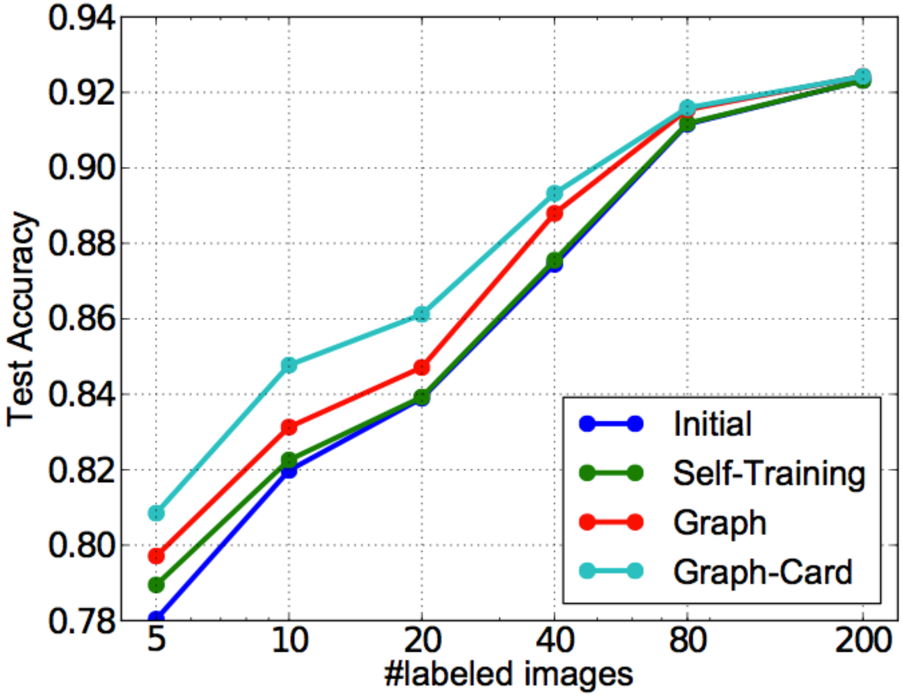
Self-Training: self-training baseline

Graph: SSL with graph regularizer R_G

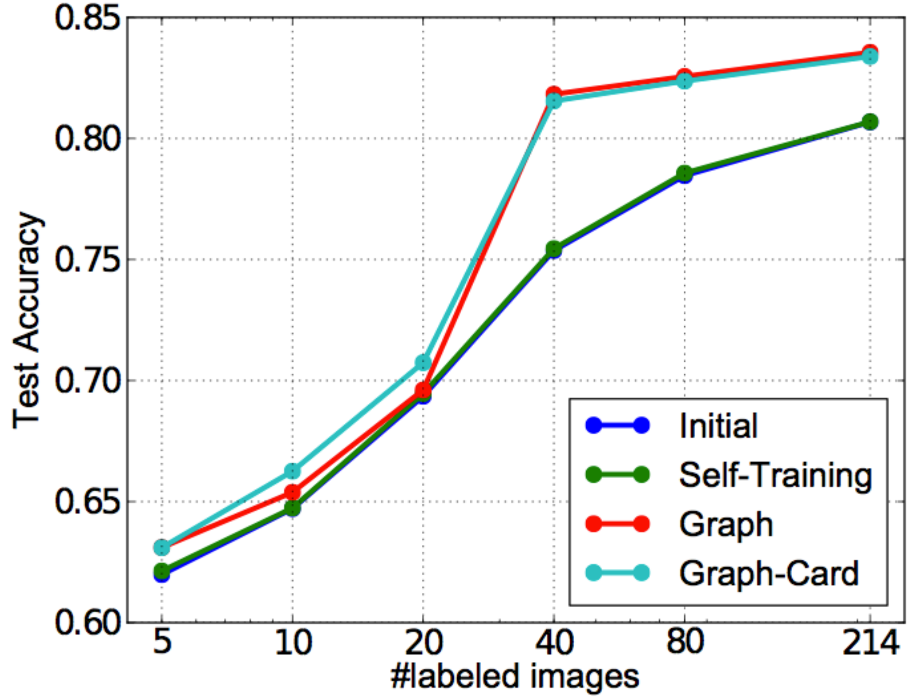
Graph-Card: SSL with both graph and cardinality regularizers $R_G + R_C$

Experiment Results

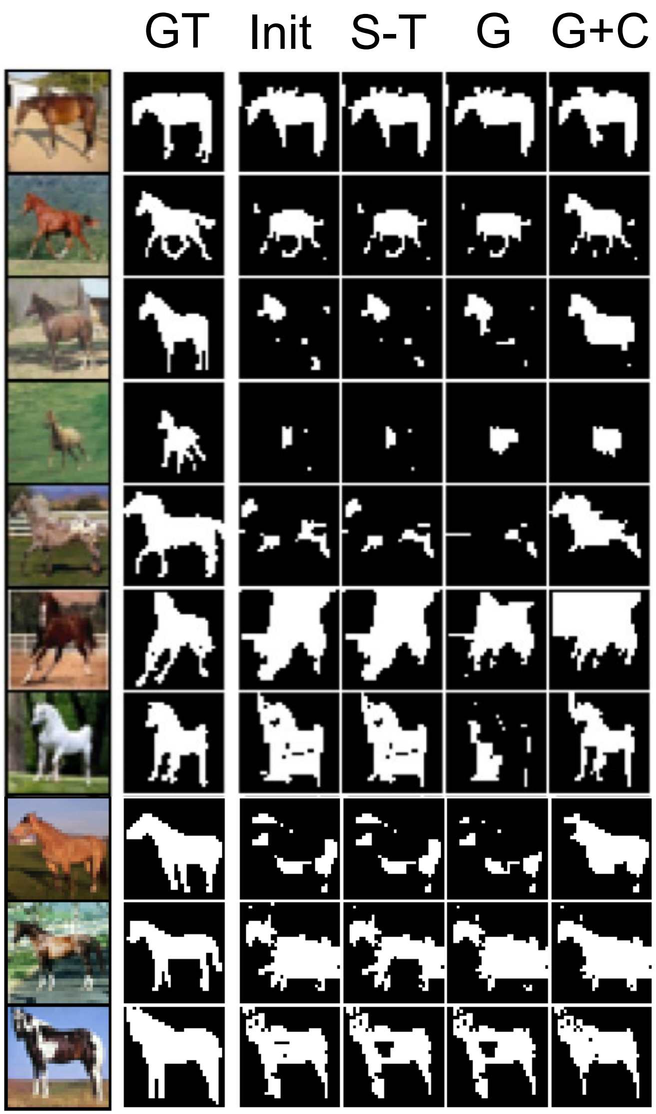
Semi-supervised learning (Horse)

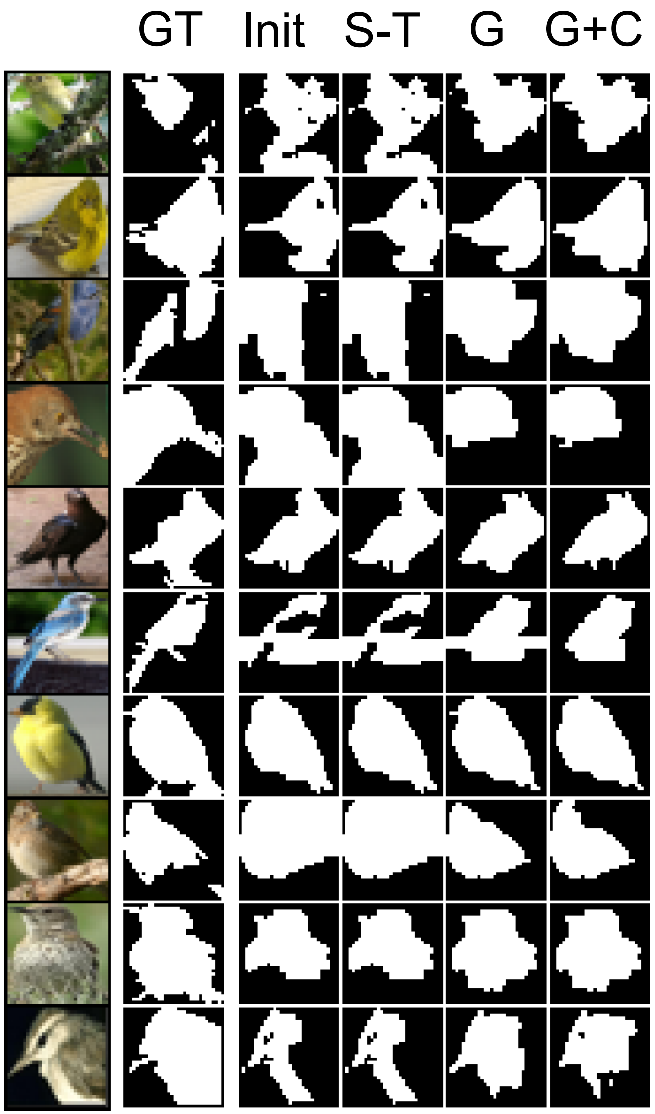


Transfer learning (Bird)



Segmentation Examples





GT: ground truth. Init: Initial. S-T: Self-Training. G: Graph. G+C: Graph-Card.

Q & A

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

Yujia Li and Richard Zemel

University of Toronto
Canadian Institute for Advanced Research