Graph Structured Prediction Energy Networks (GSPEN)

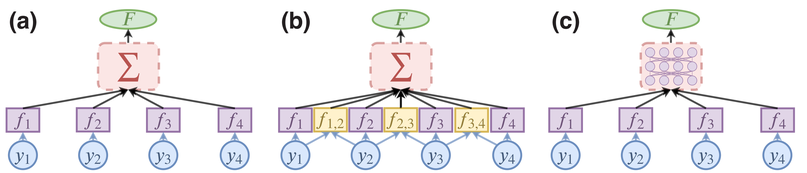
Colin Graber Alexander Schwing
University of Illinois at Urbana-Champaign

Motivation & Related Work

Objective:

Develop a model that allows known, local structure to be specified while also learning unknown global structure.

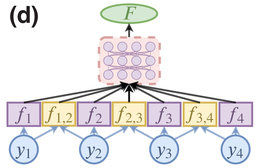
Approaches to Structured Prediction:



Unstructured Models (a):
- Independent score per variable

Classical Structured Models (b):
- Scoring functions factorize according to graphical model
- Modeling global structure is generally intractable

Structured Prediction Energy Networks [1] (c):
- Score is nonlinear function of joint variable prediction
- Cannot use known structured scores



Ours: Graph Structured Prediction Energy Networks (d):
- Score is nonlinear function of joint structured prediction
- Allows the use of any structured score function

Model Formulation

Notation:
- $x \in \mathcal{X}$: Input data (*e.g.*, images)
- $y = (y_1, \ldots, y_K)$: output space object
- $(y_1, \ldots, y_k) \in \mathcal{Y} := \prod_{k=1}^{K} \mathcal{Y}_k$: discrete output space
- $r \subseteq \{1, \ldots, K\}$, $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$: region graph
- $p_{\mathcal{R}} = (p_1, \ldots, p_r) \in \prod_{r \in \mathcal{R}} \Delta^{|\mathcal{Y}_r|}$: 1-hot structured representation of y

Score Functions: assign score to input-output pair (x, y)
- Unary (a): $F(x, y; w) = \sum_{k=1}^{k} f_k(x, y_k; w)$
  - Function f_k per output variable k; simple sum of scores
- Struct (b): $F(x, y; w) = \sum_{r \in \mathcal{R}} f_r(x, y_r; w)$
  - Function f_r per graph region r; simple sum of scores
- SPEN (c): $F(x, p_1, \ldots, p_k; w) = T(\bar{f}(x; w), p_1, \ldots, p_K; w)$
  - Global nonlinear function T applied to feature transformation $\bar{f}$ and one-hot output $p_1, \ldots, p_k$
- NLStruct [5]: $F(x, p_{\mathcal{R}}; w) = T(f_{\mathcal{R}}(x; w) \circ p_{\mathcal{R}}; w)$
  - Nonlinear re-weighting of terms of Struct scoring function
- GSPEN (d): $F(x, p_{\mathcal{R}}; w) = T(\bar{f}(x; w), p_{\mathcal{R}}; w)$
  - T contains both local structured scores and global joint score function
  - Includes NLStruct as special case

Example:

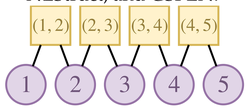
$\mathcal{X}$: Set of images with letters
$\mathcal{Y}$: ('A', 'B', ..., 'Y', 'Z')^5

x:
y: 'S' 'I' 'N' 'C' 'E'

Graph used by Unary and SPEN:

$\mathcal{R} = \{1, 2, 3, 4, 5\}$

Graph used by Struct, NLStruct, and GSPEN:



$\mathcal{R} = \{1, 2, 3, 4, 5, (1, 2), (2, 3), (3, 4)\}$

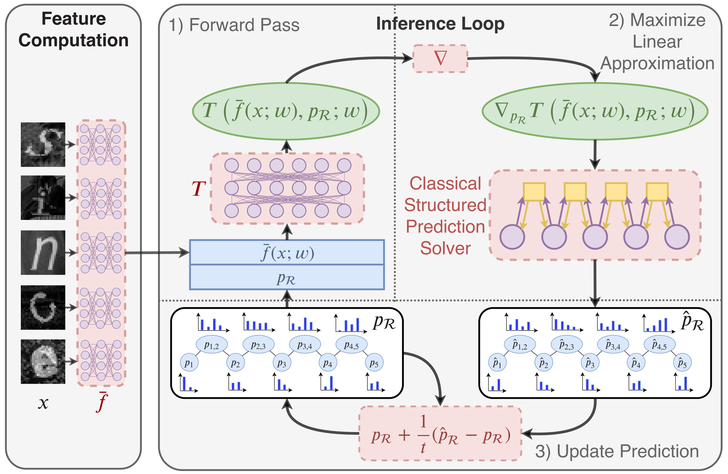
GSPEN Inference and Learning

Inference objective: maximize $F(x, p_{\mathcal{R}}; w)$ w.r.t. $p_{\mathcal{R}}$ such that $p_{\mathcal{R}} \in \mathcal{M}$ (marginal polytope)

- Solving inference requires using a constrained inference algorithm
- Because we have a non-linear objective subject to linear constraints, we can use Frank-Wolfe
- Inner maximization problem reduces to classical structured inference
- We also experiment with entropic mirror descent

Algorithm 1 Frank-Wolfe Inference for GSPEN

Input: Initial set of predictions $p_{\mathcal{R}}$; Input x; Factor graph $\mathcal{R}$
for $t = 1 \ldots T$ **do**
  $g \Leftarrow \nabla_{p_{\mathcal{R}}} F(x, p_{\mathcal{R}}; w)$
  $\hat{p}_{\mathcal{R}} \Leftarrow \max_{\hat{p}_{\mathcal{R}} \in \mathcal{M}_{\mathcal{R}}} \sum_{r \in \mathcal{R}, y_r \in \mathcal{Y}_r} \hat{p}_r(y_r) g_r(y_r)$
  $p_{\mathcal{R}} \Leftarrow p_{\mathcal{R}} + \frac{1}{t}(\hat{p}_{\mathcal{R}} - p_{\mathcal{R}})$
end for
Return: $p_{\mathcal{R}}$

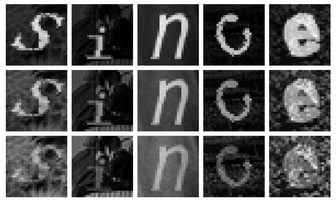


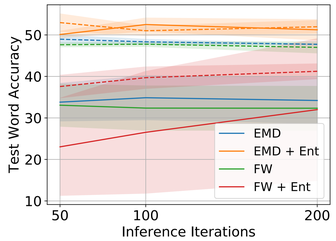
Learning Objective: SSVM objective with loss-augmented inference

$$\min_{w} \sum_{(x^{(i)}, p_{\mathcal{R}}^{(i)})} \left[\max_{\hat{p}_{\mathcal{R}} \in \mathcal{M}} \left\{ T\left(\bar{f}(x; w), \hat{p}_{\mathcal{R}}; w\right) + L\left(\hat{p}_{\mathcal{R}}, p_{\mathcal{R}}^{(i)}\right) \right\} - T\left(\bar{f}\left(x^{(i)}; w\right), p_{\mathcal{R}}^{(i)}; w\right) \right]_{+}$$

Experiments

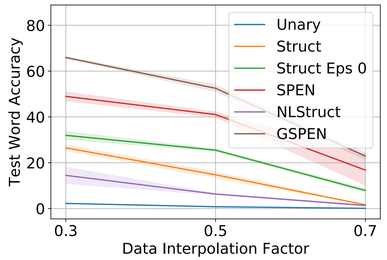
Optical Character Recognition
- Characters rendered on high-noise background patches with different interpolation factors
- Examined different inference procedures during learning (below)





Experiments (Continued)

OCR (Continued)



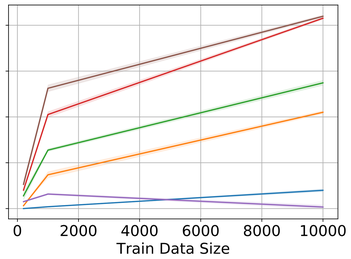
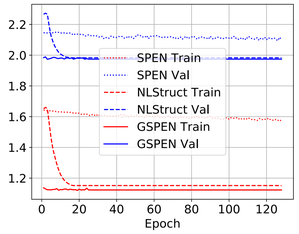


Image Tagging
- Problem: label images with subset of 24 possible tags
- Reported metric: Hamming Loss
- GSPEN trained using pre-trained Struct model from [5]
- Test results:
  - SPEN: 2.158
  - NLStruct: 2.037
  - GSPEN: 2.029

animals
flowers
plant life



Multilabel Classification
- Problem: given binary input feature vector, assign it to a subset of labels
- Two datasets: Bibtex (159 labels) and Bookmarks (208 labels)
- Reported metric: F1
- Top two rows taken from [3]

	Bibtex		Bookmarks	
	Val.	Test	Val.	Test
SPEN	–	42.2	–	34.4
DVN	–	44.7	–	37.1
Unary	43.3	44.1	38.4	37.4
Struct	45.8	46.1	39.7	38.9
SPEN	46.6	46.5	40.2	39.2
GSPEN	**47.5**	**48.6**	**41.2**	**40.7**

Named Entity Recognition

B-ORG	O	B-PER	I-PER	O	O	O	O	O	O
Celtic	's	Jackie	McNamara	,	who	did	well	with	last	...

- Problem: label each word in a sentence with tags indicating whether they are part of a named entity
- Metric: F1
- Trained GSPEN models initialized from two different Struct models

	Avg. Val.	Avg. Test
Struct [2]	94.88 ± 0.18	91.37 ± 0.04
+ GSPEN	$\mathbf{94.97} \pm 0.16$	$\mathbf{91.51} \pm 0.17$
Struct [4]	95.88 ± 0.10	$\mathbf{92.79} \pm 0.08$
+ GSPEN	$\mathbf{95.96} \pm 0.08$	92.69 ± 0.17

References

[1] D. Belanger, A. McCallum. Structured Prediction Energy Networks. In *ICML 2016*
[2] X. Ma, E. Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *ACL 2016*
[3] M. Gygli, M. Norouzi, A. Angelova. Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs. In *ICML 2017*
[4] A. Akbik, D. Blythe, R. Vollgraf. Contextual String Embeddings for Sequence Labelin. In *COLING 2018*
[5] C. Graber, O. Meshi, A. Schwing. Deep Structured Prediction with Nonlinear Output Transformations. In *NeurIPS 2018*