LEARNING APPROXIMATE INFERENCE NETWORKS FOR STRUCTURED PREDICTION

Lifu Tu **Kevin Gimpel**
Toyota Technological Institute at Chicago

ABSTRACT

- structured prediction is challenging due to exponentially-large output spaces
- How to speed up the inference time?
- Structured prediction energy networks (SPENs; Belanger & McCallum 2016): use neural networks to define structured energy functions
- Belanger & McCallum used gradient descent for inference with SPENs
- We replace this use of gradient descent with a neural network trained to approximate structured argmax inference.
- We develop large-margin training criteria for joint training of the structured energy function and inference network
  1. On multi-label classification, high accuracy and 10-60x speed-ups compared to Belanger et al. (2017)
  2. Sequence labeling: accuracies comparable to exact inference with faster inference speeds
  3. Improved accuracy by augmenting energy with "tag language model"
- We also show how inference networks can replace dynamic programming at test time for conditional random fields (see paper for details)

INFERENCE NETWORKS

We define an inference network $\mathbf{A}_\Psi(\boldsymbol{x})$ with the goal that

$$\mathbf{A}_\Psi(\boldsymbol{x}) \approx \underset{\boldsymbol{y}\in\mathcal{Y}_R(\boldsymbol{x})}{\operatorname{argmin}} E_\Theta(\boldsymbol{x},\boldsymbol{y})$$

SPENs TRAINING

SPENs are trained with the following SSVM loss:

$$\min_\Theta \sum_{\langle \boldsymbol{x}_i,\boldsymbol{y}_i\rangle\in\mathcal{D}} \left[\max_{\boldsymbol{y}\in\mathcal{Y}_R(\boldsymbol{x})} (\triangle(\boldsymbol{y},\boldsymbol{y}_i) - E_\Theta(\boldsymbol{x}_i,\boldsymbol{y}) + E_\Theta(\boldsymbol{x}_i,\boldsymbol{y}_i))\right]_+$$

Here $[.]_+ = \max(0,.)$, $\triangle(\boldsymbol{y},\boldsymbol{y}_i)$ is the error function between a prediction and the ground truth

However, the "cost-augmented" inference step is expensive.

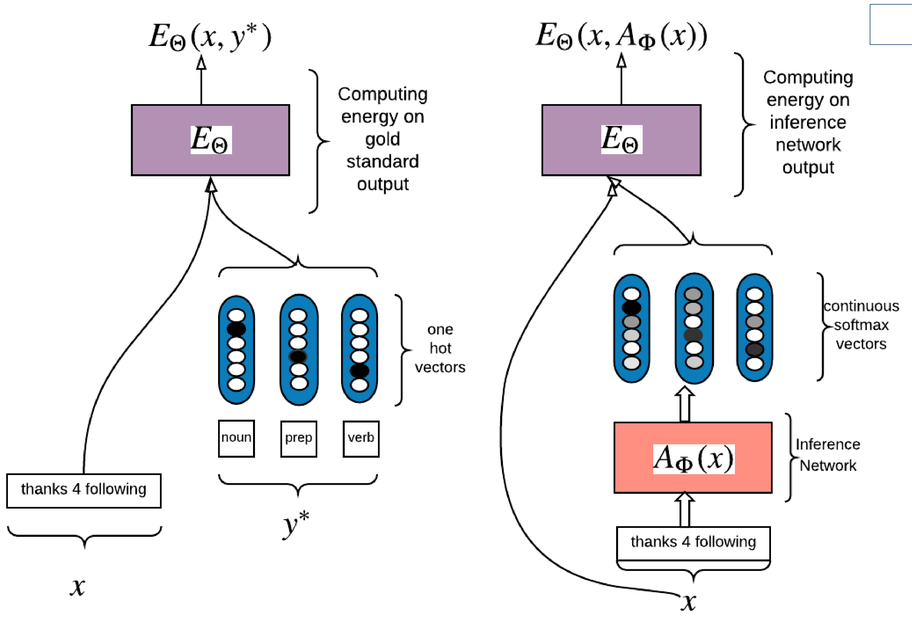
- In SPENs, this step uses gradient step. $\mathcal{Y}_R$ **:relaxed output space**
- However, hard to do exact optimization in the inner loop with gradient descent

ENERGIES FOR SEQUENCE LABELING

$$E_\Theta(\boldsymbol{x},\boldsymbol{y}) = -\left(\sum_t U_{y_t}^\top f(\boldsymbol{x},t) + \sum_t W_{y_{t-1},y_t}\right)$$ **CRF**

$$E_\Theta(\boldsymbol{x},\boldsymbol{y}) = -\left(\sum_t \sum_{\substack{i=1\\|\boldsymbol{y}|+1}}^{L} y_{t,i}\left(U_i^\top f(\boldsymbol{x},t)\right) + \sum_t y_{t-1}^\top W y_t\right)$$ **General**, **continuous label space**

$$E^{\mathrm{TLM}}(\boldsymbol{y}) = -\sum_{t=1} \log(y_t^\top \mathrm{TLM}(\langle y_0,\ldots,y_{t-1}\rangle))$$



ENERGIES FOR MULTI-LABEL CLASSIFICATION

$$E^{loc}(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{L} y_i b_i^\top F(\boldsymbol{x}) \qquad E^{lab}(\boldsymbol{y}) = c_2^\top g(C_1\boldsymbol{y})$$

$$E_\Theta(\boldsymbol{x},\boldsymbol{y}) = E^{loc}(\boldsymbol{x},\boldsymbol{y}) + E^{lab}(\boldsymbol{y})$$

JOINT ADVERSARIAL TRAINING OF SPENS AND INFERENCE NETWORKS

In SSVM loss, replacing expensive "cost-augmented" inference step with $\mathbf{A}_\Phi(\boldsymbol{x})$, then the new training objective is to minimize:

$$\min_\Theta \max_\Phi \sum_{\langle \boldsymbol{x}_i,\boldsymbol{y}_i\rangle\in\mathcal{D}} [\triangle(\mathbf{A}_\Phi(\boldsymbol{x}_i),\boldsymbol{y}_i) - E_\Theta(\boldsymbol{x}_i,\mathbf{A}_\Phi(\boldsymbol{x}_i)) + E_\Theta(\boldsymbol{x}_i,\boldsymbol{y}_i)]_+$$

We optimize the objective with the following two steps:
The objective for the cost-augmented inference network is:

$$\hat{\Phi} \leftarrow \underset{\Phi}{\operatorname{argmax}} [\triangle(\mathbf{A}_\Phi(\boldsymbol{x}_i),\boldsymbol{y}_i) - E_\Theta(\boldsymbol{x}_i,\mathbf{A}_\Phi(\boldsymbol{x})_i) + E_\Theta(\boldsymbol{x}_i,\boldsymbol{y}_i)]_+$$ **+ Reg.**

Regularization term

The objective for the energy function is:

$$\hat{\Theta} \leftarrow \underset{\Theta}{\operatorname{argmin}} [\triangle(\mathbf{A}_\Phi(\boldsymbol{x}_i),\boldsymbol{y}_i) - E_\Theta(\boldsymbol{x}_i,\mathbf{A}_\Phi(\boldsymbol{x}_i)) + E_\Theta(\boldsymbol{x}_i,\boldsymbol{y}_i)]_+ + \lambda\|\Theta\|_2^2$$

Training iterates between updating Φ and Θ

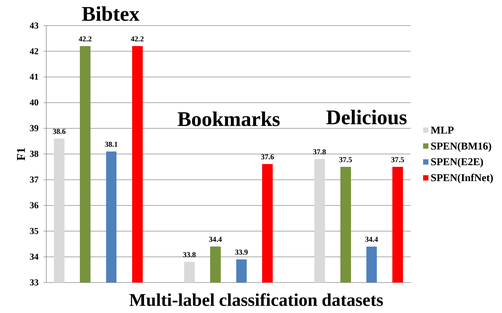
The final inference network with retuning(Ψ is initialized with Θ):

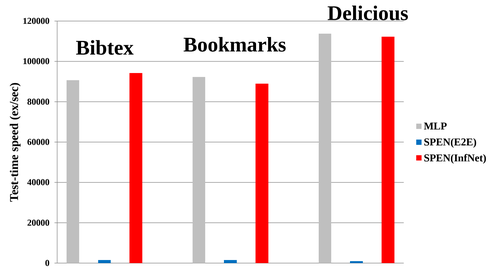
$$\hat{\Psi} \leftarrow \underset{\Psi}{\operatorname{argmin}} \sum_{\boldsymbol{x}\in X} E_\Theta(\boldsymbol{x},\mathbf{A}_\Psi(\boldsymbol{x}))$$ **+ Reg.**

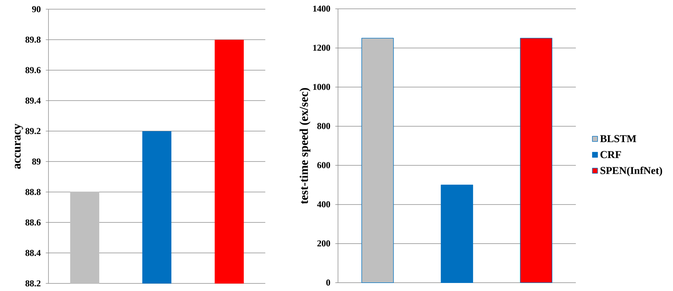
Regularization term

TEST-TIME INFERENCE

EXPERIMENTS







Twitter part-of-speech tagging

	test accuracy(%)
-TLM	89.6
+TLM	90.2

Twitter POS test accuracies when adding tag language model(TLM)

#	tweet (target word in bold)	-TLM	+TLM
1	... that's a t-17 , technically . does **that** count as top-25 ?	determiner	pronoun
2	... lol you know im down **like** 4 flats on a cadillac ... lol ...	adjective	preposition
3	... them who he is : he wants her to **like** him for his pers ...	preposition	verb
4	Cut my hair , **gag** and bore me	noun	verb
5	I wonder when Nic Cage is going to **film** " Another Something Something Las Vegas " .	noun	verb
6	... they had their fun , we **hd** ours ! ;) lmaooo	proper noun	verb
7	lmao I am not a sheep who listens to it **cos** everyone else does ...	verb	preposition
8	Noo its not cuss you have swag **andd** you wont look dumb ! ...	noun	coord. conj.

Examples of improvements when using tag language model

References

1. Belanger , David and McCallum., Andrew Structured Prediction Energy Networks. ICML2016
2. LeCun, Yann and Chopra, Sumit and Hadsell, Raia and Ranzato . A Tutorial on Energy-Based Learning . MIT Press 2006
3. Belanger, D. Yang, B., McCallum, A. "End-to-End Learning for Structured Prediction Energy Networks." Arxiv Preprint