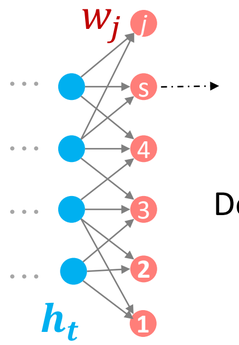
RECURRENT SVM FOR SPEECH RECOGNITION

Shi-Xiong Austin Zhang, Rui Zhao, Chaojun Liu, Jinyu Li and Yifan Gong
Microsoft, Redmond, WA, USA

1. Introduction

- RNNs use the softmax activation function in the last layer



$$P(s|\boldsymbol{x}_t) = \frac{\exp(w_s^{\mathrm{T}} \boldsymbol{h}_t)}{\sum_{j=1}^{N} \exp(w_j^{\mathrm{T}} \boldsymbol{h}_t)}$$

Denominator is constant for classification

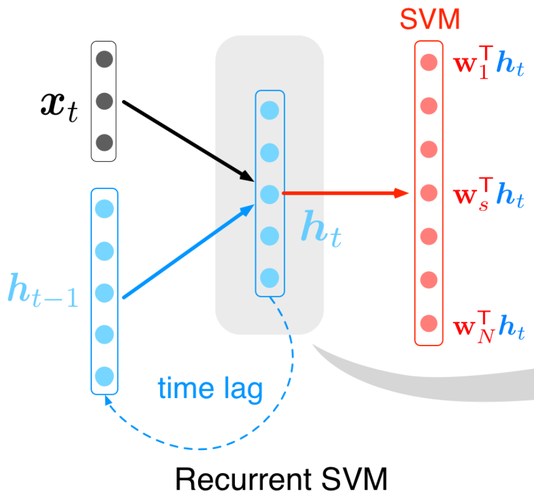
$$\arg\max_s P(s|\boldsymbol{x}_t) = \arg\max_s w_s^{\mathrm{T}} \boldsymbol{h}_t$$

- Multiclass SVM, same classification: $\arg\max_s w_s^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_t)$

Replace softmax layer in RNN with SVM ➔ Recurrent SVM

2. Recurrent SVM

Architecture: LSTM+SVM



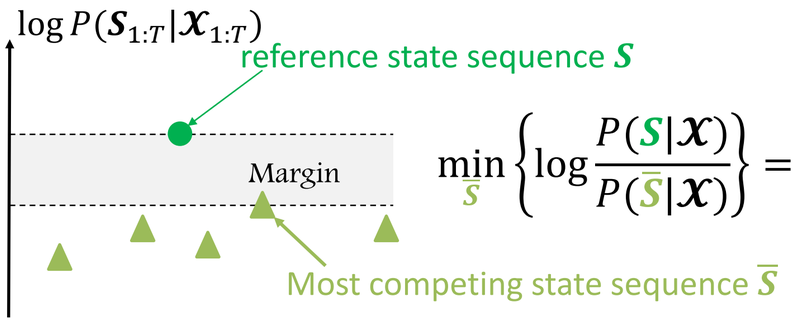
How to train them jointly?

1st-step: fixed LSTM, training SVM using the quadratic programing.

2nd-step: fixed SVM, training LSTM using the subgradient methods

3. Max-Margin Sequence Training

What is the Margin? Objective Function?



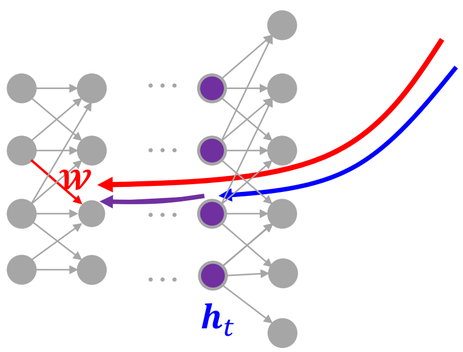
$$\min_{\bar{S}} \left\{ \log \frac{P(\boldsymbol{S}|\boldsymbol{X})}{P(\bar{\boldsymbol{S}}|\boldsymbol{X})} \right\} = \mathcal{F}(w^{\mathrm{SVM}}, w^{\mathrm{LSTM}})$$

How to train the last layer w^{SVM}?

1st-step: fixed w^{LSTM}, $\mathcal{F}(w^{\mathrm{SVM}}, w^{\mathrm{LSTM}})$ is convex, training w_{SVM} ⇔ the structured SVM

How to train the previous layers w^{LSTM}?

2nd-step: fixed w^{SVM}, $\mathcal{F}(w^{\mathrm{SVM}}, w^{\mathrm{LSTM}})$ is non-differentiable, training w^{LSTM} requires subgradients



$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{w}} = \underbrace{\left(\frac{\partial \mathcal{F}}{\partial \boldsymbol{h}_t}\right)^{\mathrm{T}}}_{\text{key}} \underbrace{\left(\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{w}}\right)}_{\text{Same as LSTM, Same BPTT ☺}}$$

$$\frac{\partial \mathcal{F}}{\partial \boldsymbol{h}_t} = 2C \sum_{t=1}^{T} \max\{\mathrm{Score}_{\bar{S}} - \mathrm{Score}_{\boldsymbol{S}}, 0\} \; (\mathbf{w}_{\bar{S}_t}^{\mathrm{SVM}} - \mathbf{w}_{S_t}^{\mathrm{SVM}})$$

Only the support vectors have gradients! ☺

4. Experiments & Conclusion

- **Training data**: 60 hours of US-English Windows Phone Short Message Dictation
- **Testing data**: 3 hours of data from same task

Model	WER (%)
6-layer DNN (MMI training)	21.1
4-layer LSTM (MMI training)	20.8
Recurrent SVM (Max Margin training)	19.8

- 4.8% WERR. More results in the paper.

Model	WER (%)
Recurrent SVM (only train last layer)	20.2
Recurrent SVM (+ previous layer)	19.8

- 65% gains are from updating the LSTM layers

References

[1] Schmidhuber, Jürgen, et al. Evolino for Recurrent Support Vector Machines. *Euro. Symp. on Artificial Neural Networks*, 2006.