Fast and Accurate Entity Recognition with Iterated Dilated Convolutions

Emma Strubell, Patrick Verga, David Belanger & Andrew McCallum

College of Information and Computer Sciences, University of Massachusetts Amherst

UMassAmherst College of Information and Computer Sciences

Summary

- Want fast and accurate NER / tagging
- LSTMs attain highest accuracy, but $O(n)$ parallelized runtime; Feed-forward nets are $O(1)$, but less accurate
- We introduce ID-CNNs: distinct combination of network structure, parameter sharing and training enabling 14-20x speedups while retaining Bi-LSTM-CRF accuracy.
- ID-CNNs trained on entire documents are even more accurate while maintaining $8\times$ test time speeds.

Model

The first layer in the network is a dilation-1 convolution $D_1^{(0)}$:

$$\mathbf{i_t} = D_1^{(0)}\mathbf{x_t} \tag{1}$$

We apply L_c layers of exponentially increasing dilation width to $\mathbf{i_t}$. Beginning with $\mathbf{c_t}^{(0)} = \mathbf{i_t}$:

$$\mathbf{c_t}^{(j)} = r\left(D_{2^{L_c-1}}^{(j-1)}\mathbf{c_t}^{(j-1)}\right) \tag{2}$$

and add a final dilation-1 layer to the stack:

$$\mathbf{c_t}^{(L_c+1)} = r\left(D_1^{(L_c)}\mathbf{c_t}^{(L_c)}\right) \tag{3}$$

We iteratively apply this *block* $B(\cdot)$ of convolutions L_b times, starting with $\mathbf{b_t}^{(1)} = B\left(\mathbf{i_t}\right)$:

$$\mathbf{b_t}^{(k)} = B\left(\mathbf{b_t}^{(k-1)}\right) \tag{4}$$

Finally, we obtain per-class scores for each token $\mathbf{x_t}$:

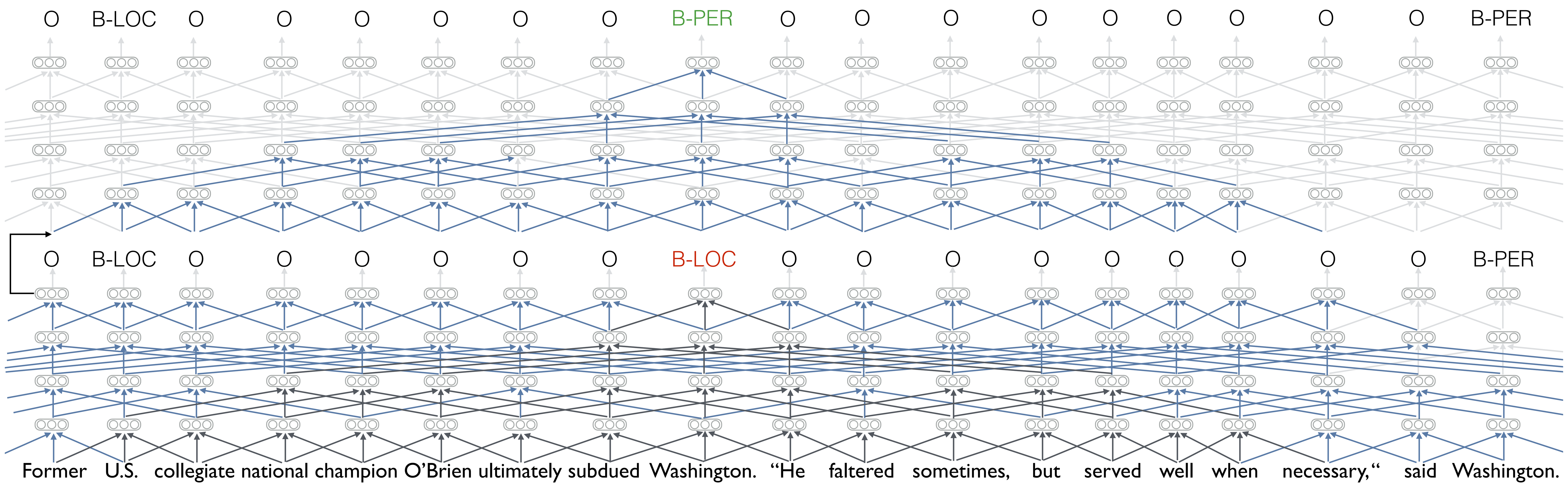
$$\mathbf{h_t}^{(L_b)} = W_o\mathbf{b_t}^{(L_b)} \tag{5}$$

Training

Let $\mathbf{h_t}^{(k)}$ be the result of applying W_o from Eqn. (5) to $\mathbf{b_t}^{(k)}$. We minimize the average of the losses for each application of the block:

$$\frac{1}{L_b}\sum_{k=1}^{L_b}\frac{1}{T}\sum_{t=1}^{T}\log P(y_t \mid \mathbf{h_t}^{(k)}) \tag{6}$$

Dilated Iterated CNN



Experimental results

English CoNLL 2003:

Model (sentence)	F1	Speed
Ratinov and Roth (2009)	86.82	
Collobert et al. (2011)	88.67	
Passos et al. (2014)	90.05	
Lample et al. (2016)	90.20	
Bi-LSTM-CRF (re-impl)	90.43 ± 0.12	1×
ID-CNN-CRF	90.54 ± 0.18	1.28×
Bi-LSTM	89.34 ± 0.28	9.92×
4-layer CNN	89.97 ± 0.20	18.40×
5-layer CNN	90.23 ± 0.16	12.38×
ID-CNN	90.32 ± 0.26	14.10×

Model (document)	F1	Speed
Bi-LSTM-CRF	90.60 ± 0.19	1×
Bi-LSTM	89.09 ± 0.19	4.60×
ID-CNN	90.65 ± 0.15	7.96×

Varying loss, parameter sharing:

Model	F1
ID-CNN noshare	89.81 ± 0.19
ID-CNN 1-loss	90.06 ± 0.19
ID-CNN	90.65 ± 0.15

English OntoNotes 5.0:

Model	F1	Speed
Ratinov and Roth (2009)[1]	83.45	
Durrett and Klein (2014)	84.04	
Chiu and Nichols (2016)	86.19 ± 0.25	
Bi-LSTM-CRF	86.99 ± 0.22	1×
Bi-LSTM-CRF-Doc	86.81 ± 0.18	1.32×
Bi-LSTM	83.76 ± 0.10	24.44×
ID-CNN-CRF (1 block)	86.84 ± 0.19	1.83×
ID-CNN-Doc (3 blocks)	85.76 ± 0.13	21.19×
ID-CNN (3 blocks)	85.27 ± 0.24	13.21×
ID-CNN (1 block)	84.28 ± 0.10	26.01×