Digital Genius

Named Entity Recognition with Parallel Recurrent Neural Networks

Andrej Zukov-Gregoric
Yoram Bachrach
Sam Coope

1. Introduction

We present a new architecture for named entity recognition. Our model employs multiple independent bidirectional LSTM units across the same input and promotes diversity among them by employing an inter-model regularization term. By distributing computation across multiple smaller LSTMs we find a reduction in the total number of parameters. We find our architecture achieves state-of-the-art performance on the CoNLL 2003 NER dataset.

2. LSTM and complexity

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\boldsymbol{W}_i\mathbf{h}_{t-1} + \boldsymbol{U}_i\mathbf{x}_t) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_f\mathbf{h}_{t-1} + \boldsymbol{U}_f\mathbf{x}_t) \\
\mathbf{o}_t &= \sigma(\boldsymbol{W}_o\mathbf{h}_{t-1} + \boldsymbol{U}_o\mathbf{x}_t) \\
\tilde{\mathbf{c}}_t &= \tanh(\boldsymbol{W}_c\mathbf{h}_{t-1} + \boldsymbol{U}_c\mathbf{x}_t) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
\end{aligned}$$

One way of measuring the complexity of a model is through its total number of parameters. Looking at the above, we note there are two parameter matrices, $\mathbf{W}$ and $\mathbf{U}$, for each of the three input gates and during cell update. If we let $\mathbf{W} \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{n\times m}$ then the total number of parameters in the model (excluding the bias terms) is $4(nm+n^2)$ which grows quadratically as n grows. Thus, increases in LSTM size can substantially increase the number of parameters.

3. LSTM definition without biases:

To reduce the total number of parameters we split a single LSTM into multiple equally-sized smaller ones:

$$h_{k,t} = \text{LSTM}_k(h_{k,t-1}, \mathbf{x})$$

where $k \in \{1, ..., K\}$. This has the effect of dividing the total number of parameters by a constant factor. The final hidden state h_t is then a concatenation of the hidden states of the smaller LSTMS:

$$h_t = [h_{1,t}; h_{2,t}; ...; h_{K,t}]$$

4. Promoting Diversity Between LSTMs

We take the cell update recurrence parameters $\mathbf{W}_i$ across LSTMs (we omit the c in the subscript for brevity; the index i runs across the smaller LSTMs) and for any pair we wish the following to be true:

$$\langle \text{vec}(W_c^{(i)}), \text{vec}(W_c^{(j)}) \rangle \approx 0$$

.

To achieve this we pack the vectorized parameters into a matrix:

$$\Phi = \begin{pmatrix} \text{vec}(W_c^{(1)}) \\ \text{vec}(W_c^{(2)}) \\ \vdots \\ \text{vec}(W_c^{(N)}) \end{pmatrix}$$

and apply the following regularization term to our final loss:

$$\lambda \sum_i \|\Phi\Phi^\top - I\|_F^2$$

5. Results

Model	F1
(Chieu and Ng, 2002)	88.31
(Florian et al., 2003)	88.76
(Ando and Zhang, 2005)	89.31
(Collobert et al., 2011)[‡]	89.59
(Huang et al., 2015)[‡]	90.10
(Chiu and Nichols, 2015)[‡]	90.77
(Ratinov and Roth, 2009)	90.80
(Lin and Wu, 2009)	90.90
(Passos et al., 2014)[‡*]	90.90
(Lample et al., 2016)[‡]	90.94
(Luo et al., 2015)[‡]	91.20
(Ma and Hovy, 2016)[‡]	91.21
(Sato et al., 2017)	91.28
(Chiu and Nichols, 2015)[‡*]	91.62
(Peters et al., 2017)[‡*]	91.93
This paper[‡]	**91.48** ±0.22

6. Architecture Choices & Ablations

# RNN units	Unit size	F_1
1	1024	87.54
2	512	91.25
4	256	91.29
8	128	91.31
16	64	91.48 ±0.22
32	32	90.60
64	16	90.79
128	8	90.41

Table 3: Performance of our model with various unit sizes resulting in a fixed final output size h_t. Single runs apart from 16 unit.

Unit size	F_1
8	89.78
16	89.77
32	90.26
64	91.48 ±0.22
128	89.28

Table 4: Performance as a function of the unit size for our best performing model (16 biLSTM units). Single runs apart from with size 64.

Component	F_1
No character embeddings	90.39
No orthogonal regularization	90.79
No Xavier initialization	91.09
No variational dropout	91.03
Mean pool instead of concat	90.49

Table 5: Impact of various architectural decisions on our best performing model (16 biLSTM units, 64 unit size). Single runs.