Semi-supervised Multitask Learning for Sequence Labeling

Marek Rei
The ALTA Institute
Computer Laboratory
University of Cambridge
United Kingdom

Motivation

- Due to the Sparseness of labels, many availabel data can't be utilized effectively.
- Due to the limitation of the single task, many features can't be learnt effectively.
- Due to the limitation of distribution of data of the single task, it can occurs over-fitting easily.

Solution-Multitask

- Multitask can incorporate new supervise signals so that learn the data representation better.
- Multitask can learn the feature representation not easy learnt in task A by incorporating task B.
- Multitask can reduce the over-fitting occurrence by incorporating multi-tasks learning.

Model

Neural Sequence Labeling

Bi-LSTM encoding

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}})$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}})$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

Non-Linear Transformation

$$d_t = tanh(W_d h_t)$$

Soft-Max

$$
\begin{aligned}
P(y_t|d_t) &= softmax(W_o d_t) \\
&= \frac{e^{W_{o,k} d_t}}{\sum_{\tilde{k} \in K} e^{W_{o,\tilde{k}} d_t}}
\end{aligned}
$$

Optimization Object

$$
E = -\sum_{t=1}^{T} log(P(y_t|d_t))
$$

Language Modeling Objective

Bi-Prediction

$$\overrightarrow{m_t} = tanh(\overrightarrow{W}_m \overrightarrow{h_t})$$

$$\overleftarrow{m_t} = tanh(\overleftarrow{W}_m \overleftarrow{h_t})$$

$$P(w_{t+1}|\overrightarrow{m_t}) = softmax(\overrightarrow{W}_q \overrightarrow{m_t})$$

$$P(w_{t-1}|\overleftarrow{m_t}) = softmax(\overleftarrow{W}_q \overleftarrow{m_t})$$

Optimization Object

$$\overrightarrow{E} = -\sum_{t=1}^{T-1} log(P(w_{t+1}|\overrightarrow{m_t}))$$

$$\overleftarrow{E} = -\sum_{t=2}^{T} log(P(w_{t-1}|\overleftarrow{m_t}))$$

The Over-Whole Option Object

$$\widetilde{E} = E + \gamma(\overrightarrow{E} + \overleftarrow{E})$$