Deep Contextualized word representations

2018年3月30日 16:37

Pre-trained word representations **challenge**:

- Complex characteristics of word use
- How these uses vary across linguistic contexts

Approach: Embedding from Language Models:Functions of the entire input sentence

Procedure: computed as two-layer biLMs with character convolutions as a linear function of the internal network states.

1. **Bidirectional language models** N tokens $(t_1, t_2, \dots, t_N)$
   - Forward language model:computes the probability of sequence by modeling the probability of token t_k given the history $(t_1, \dots, t_{k-1})$

     $P(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \dots, t_{k\text{-}1})$
   - Backward LM: similar to a forward LM, but it runs over the sequence in reverse

     $P(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, \dots, t_N)$
   - Joinly maximizes the log likelihood of the forward and backward directions:

     $\sum_{k=1}^{N} (\log p(t_k | t_1, \dots, t_{k-1}); \Theta_x, \Theta_{LSTM,right}, \Theta_s) + \log p(t_k | t_{k+1}, \dots, t_N; \Theta_x, \Theta_{LSTM,left}, \Theta_s))$

     Token representation Θ_x Softmax layer Θ_s,
2. **ELMo**: linear combination of the intermediate layer representations in the biLM
   - **FOR each token t_k, L-layer biLSTM computes a set of** 2L+1 representations.

     $R_k = \left\{ x_k^{LM}, h_{k,j}^{LM_{right}}, h_{k,j}^{LM_{left}} \middle| j = 1, \dots, L \right\} = \{ h_{k,j}^{LM} | j = 0, \dots, L \}$

     $h_{k,0}^{LM}$ is token layer

     $ELMo_k = E(R_k; \Theta_e)$ collapses all layers in R into single vector.

Usage Example: in supervised NLP tasks

First : run the biLM and record all of the layer representations for each word.

Then: let the end task model learn a linear combination of these representations.

How to add ELMo to the supervised model:

1. Freeze the weights of the biLM
2. Concatenate the ELMo vector $ELMo_k^{task}$ with x_k and pass the EMLo enhanced representation $[x_k; ELMo_k^{task}]$ Into task RNN.

Pre-trained bidirectional LM

L=2 biLSTM layers with 4096 units and 512 dimension projections and a residual connection from the first to second layer