Hierarchical Attention Networks for Document Classification

By Jiawei

Abstract

- We propose a hierarchical attention network for document classification. Our model has two distinctive characteristics: (i) it has a hierarchical structure that mirrors the hierarchical structure of documents; (ii) it has two levels of attention mechanisms applied at the wordand sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation. Experiments conducted on six large scale text classification tasks demonstrate that the proposed architecture outperform previous methods by a substantial margin. Visualization of the attention layers illustrates that the model selects qualitatively informative words and sentences.

Hierarchical Attention Networks

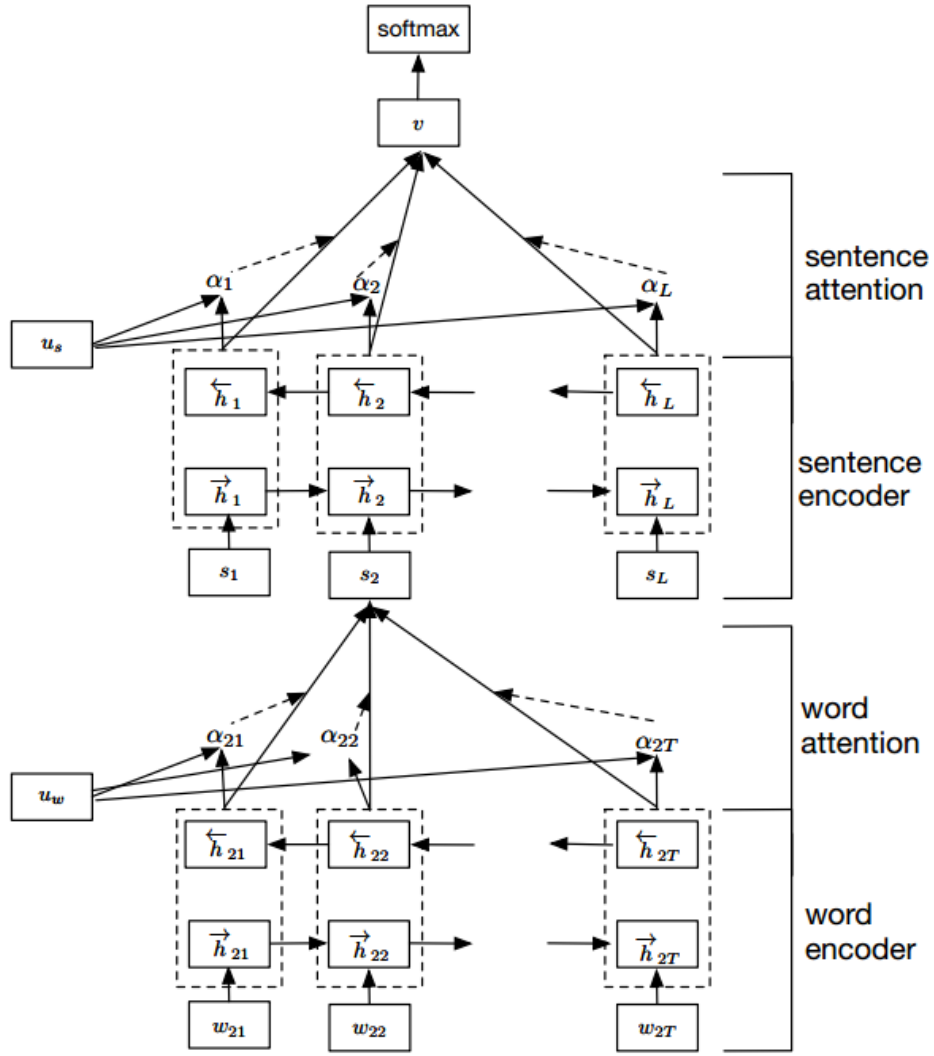

Figure 2: Hierarchical Attention Network.

GRU-based sequence encoder

•

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t. \quad (1)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \quad (2)$$

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h), \quad (3)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (4)$$

Hierarchical Attention

- Word Encoder

$$
\begin{aligned}
x_{it} &= W_e w_{it}, t \in [1, T], \\
\overrightarrow{h}_{it} &= \overrightarrow{\mathrm{GRU}}(x_{it}), t \in [1, T], \\
\overleftarrow{h}_{it} &= \overleftarrow{\mathrm{GRU}}(x_{it}), t \in [T, 1].
\end{aligned}
$$

- Word Attention

$$u_{it} = \tanh(W_w h_{it} + b_w)$$

$$\alpha_{it} = \frac{\exp(u_{it}^\top u_w)}{\sum_t \exp(u_{it}^\top u_w)}$$

$$s_i = \sum_t \alpha_{it} h_{it}.$$

Document Classification

$$p = \text{softmax}(W_c v + b_c).$$

Lost Function

$$L = -\sum_{d} \log p_{dj},$$

Experiments

- We evaluate the effectiveness of our model on six large scale document classification data sets. These data sets can be categorized into two types of document classification tasks: sentiment estimation and topic classification. The statistics of the data sets are summarized in Table 1. We use 80% of the data for training, 10% for validation, and the remaining 10% for test, unless stated otherwise

Tang et al., 2015	Paragraph Vector	57.7	59.2	60.5	34.1	-	-
	CNN-word	59.7	61.0	61.5	37.6	-	-
	Conv-GRNN	63.7	65.5	66.0	42.5	-	-
	LSTM-GRNN	65.1	67.1	67.6	45.3	-	-
This paper	HN-AVE	67.0	69.3	69.9	47.8	75.2	62.9
	HN-MAX	66.9	69.3	70.1	48.2	75.2	62.9
	HN-ATT	**68.2**	**70.5**	**71.0**	**49.4**	**75.8**	**63.6**