A Attention Layer

Following Vaswani et al. (2017), the attention layer can be expressed in terms of key, query and value vectors, denoted as $\mathbf{k}_i$, $\mathbf{q}_i$ and $\mathbf{v}_i$ respectively, where the subscript i denotes the location in the sequence. Specifically, the attention layer in our models is defined as Equation 1.

$$w_{ij} = \frac{\exp \alpha_{ij}}{\sum_n \exp \alpha_{in}} \tag{1}$$

α_{ij} in Equation 1 is computed with Equation 2, where W_a and b are learnable parameters.

$$\alpha_{ij} = \tanh(\mathbf{q}_i \cdot W_a \cdot \mathbf{k}_j{}^T + b) \tag{2}$$

Here w_{ij} is the weight assigned to location j for location i. Then the output of the attention layer at location i is computed by taking the weighted sum of value vectors at all locations, i.e. $\mathbf{o}_i = \sum_n w_{in} \cdot \mathbf{v}_n$, where $\mathbf{o}_i$ denotes the output of attention layer at location i. Unless otherwise noted, throughout this paper $\mathbf{k}_i$, $\mathbf{q}_i$ and $\mathbf{v}_i$ are all equal to the hidden vector at position i from the previous layer $\mathbf{h}_i$.

B Experiment Details

Except for ELECTRA, the rest of the models are trained with Adam optimizer (Kingma and Ba, 2014) with learning rate of 0.001. Text is tokenized using pre-trained Fasttext embeddings (Bojanowski et al., 2017). All LSTM modules are bi-directional and have 3 layers, with hidden size of 512. Batch size is set to 64. We experimented with various choices of batch sizes, including 32, 64, 96 and 128, and noted only minimal differences. ELECTRA is trained with Adam optimizer with learning rate of 0.00002 and with batch size of 16.

C Additional Experiments Results

Figure 7 to Figure 8 show the top-10, and top-30 recalls on diagnosis prediction, respectively. Table 3 shows the complete performance of models on diagnosis prediction.

Model	**Validation performance**					
	Top-5 recall		**Top-10 recall**		**Top-30 recall**	
	Pre-trained	From scratch	Pre-trained	From scratch	Pre-trained	From scratch
LSTM	**26.20%**	15.49%	**40.00%**	26.33%	**63.57%**	45.78%
LSTM+SA	**28.08%**	15.43%	**41.75%**	26.33%	**65.15%**	46.33%
Electra	**28.63%**	28.08%	**42.35%**	41.74%	**65.64%**	65.37%
	Test performance					
LSTM	**26.94%**	15.67%	**40.59%**	25.97%	**65.49%**	45.15%
LSTM+SA	**27.47%**	15.93%	**41.24%**	25.97%	**65.86%**	45.67%
Electra	27.88%	**27.90%**	41.76%	**41.82%**	66.23%	**66.49%**

Table 3: Performance on diagnosis prediction[a][b]

[a] Note that, as discussed in Section 3, on our dataset the highest possible top-5, top-10 and top-30 recalls are 50.17%, 79.48% and 99.88% on validation set, and 49.75%, 79.23% and 99.79% on test set.

[b] Bold font indicates the training strategy (pre-trained or from scratch) that has higher accuracy.

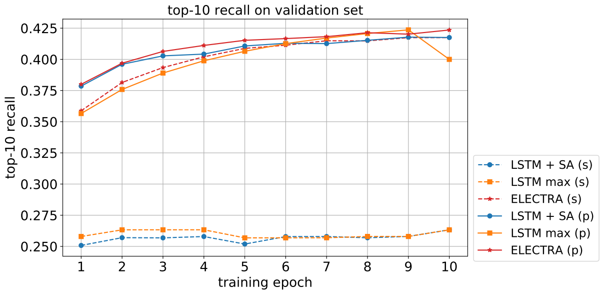

Figure 7: Top-10 recall on diagnosis prediction validation set. ‘SA’ stands for self attention layer. ‘max’ represents max-pooling output layer. ‘(s)’ and ‘(p)’ indicates whether the model is trained from scratch or pre-trained, respectively.

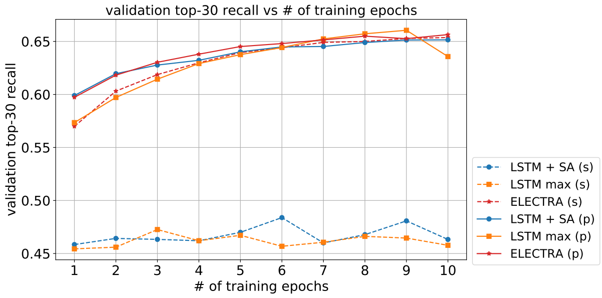

Figure 8: Top-30 recall on diagnosis prediction validation set. ‘SA’ stands for self attention layer. ‘max’ represents max-pooling output layer. ‘(s)’ and ‘(p)’ indicates whether the model is trained from scratch or pre-trained, respectively.