Toxicity Classification in Civil Comments

-Tirupal Rao Ravilla(trr321)

Problem:

Toxicity classification in Civil comments/ online conversations on a public platform, focused on news articles.

A NLP task where comment texts are processed to learn the toxicity in them, to classify whether a comment is toxic or not; where toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion.

As part of a competition on Kaggle, the Conversation AI team, a research initiative founded by Jigsaw and Google, released the data labeled for toxicity. Crowdsourcing was used for annotation.

The problem can be divided into two parts: building a model to learn to classify toxicity in the comments; and mitigating the unintended model bias for identity terms(e.g. muslim, gay. Background: here)

For this project, we focus primarily on building a deep stacked Bi-LSTM model to train for the toxicity classification.

Data:

1.8 million civil comments labeled for toxicity

Initialization Details:

0.75/0.25 train and test division used for this project.

300d GloVe word embeddings pretrained on Wikipedia and Gigaword are used for input feature representation.

Dropout of 0.2 used on each LSTM layer. Adam optimizer is used.

Random Uniform initialization for Embedding weights. Glorot Uniform initialization for all hidden kernel weights, and Orthogonal initialization for recurrent kernel weights in LSTM units.

-Tirupal Rao Ravilla(trr321)

Model:

Stacked bidirectional LSTM model, with 2 BiLSTM layers, a ReLu activated fully connected layer on top, and an sigmoidal output layer. Figure 1b. gives an overview of our model.

We used Keras CuDNNLSTM layer while constructing the model, which implements a standard LSTM layer. Fig 1a. gives an overview of the BiLSTM structure. The input gate, the forget gate, the output gate and the input cell state, which are represented in the LSTM cell in Fig. 1a, can be represented ad:

$f_t = \sigma_g (W_f x_t + U_f h_{t-1} + b_f)$

$i_t = \sigma_g (W_i x_t + U_i h_{t-1} + b_i)$

$o_t = \sigma_g (W_o x_t + U_o h_{t-1} + b_o)$

$\tilde{C}_t = \tanh(W_C x_t + U_C h_{t-1} + b_C)$

where Wf, Wi, Wo, and WC are the weight matrices mapping the hidden layer input to the three gates and the input cell state; while the Uf, Ui, Uo, and UC are the weight matrices connecting the previous cell output state to the three gates and the input cell state. The bf, bi, bo, and bC are four bias vectors. The σg is the gate activation function, and the tanh is the hyperbolic tangent function. From these, the cell output state and layer output state can be represented as:

$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

$h_t = o_t * \tanh(C_t)$

Finally, in the BiLSTM layer concatenates these operations are performed in both directions, forward and backward on the sequences, thus generating two sets of hidden outputs., which are concatenated to be passed as inputs to the next layer. Each hidden output can be represented as:

$y_t = \text{concatenate}(\overrightarrow{h}_t, \overleftarrow{h}_t)$

Minimal tuning done on the hyperparameters, and results shown are using the optimal parameters found during the process.

Model summary

Layer (type)	Output Shape	Param #
embedding_4 (Embedding)	(None, 150, 300)	92703600
bidirectional_7 (Bidirection	(None, 150, 200)	321600
dropout_7 (Dropout)	(None, 150, 200)	0
bidirectional_8 (Bidirection	(None, 200)	241600
dropout_8 (Dropout)	(None, 200)	0
dense_7 (Dense)	(None, 64)	12864
dense_8 (Dense)	(None, 1)	65

Total params: 93,279,729
Trainable params: 576,129
Non-trainable params: 92,703,600

Training Results

Train on 1353655 samples, validate on 451219 samples
Epoch 1/4
1353655/1353655 [==============================] - 495s 366us/step - loss: 0.1474 - acc: 0.9453 - val_loss: 0.1304 - val_acc: 0.9496

Epoch 2/4
1353655/1353655 [==============================] - 494s 365us/step - loss: 0.1250 - acc: 0.9512 - val_loss: 0.1259 - val_acc: 0.9506

Epoch 3/4
1353655/1353655 [==============================] - 494s 365us/step - loss: 0.1187 - acc: 0.9531 - val_loss: 0.1229 - val_acc: 0.9518

Epoch 4/4
1353655/1353655 [==============================] - 494s 365us/step - loss: 0.1134 - acc: 0.9546 - val_loss: 0.1228 - val_acc: 0.9522

-Tirupal Rao Ravilla(trr321)

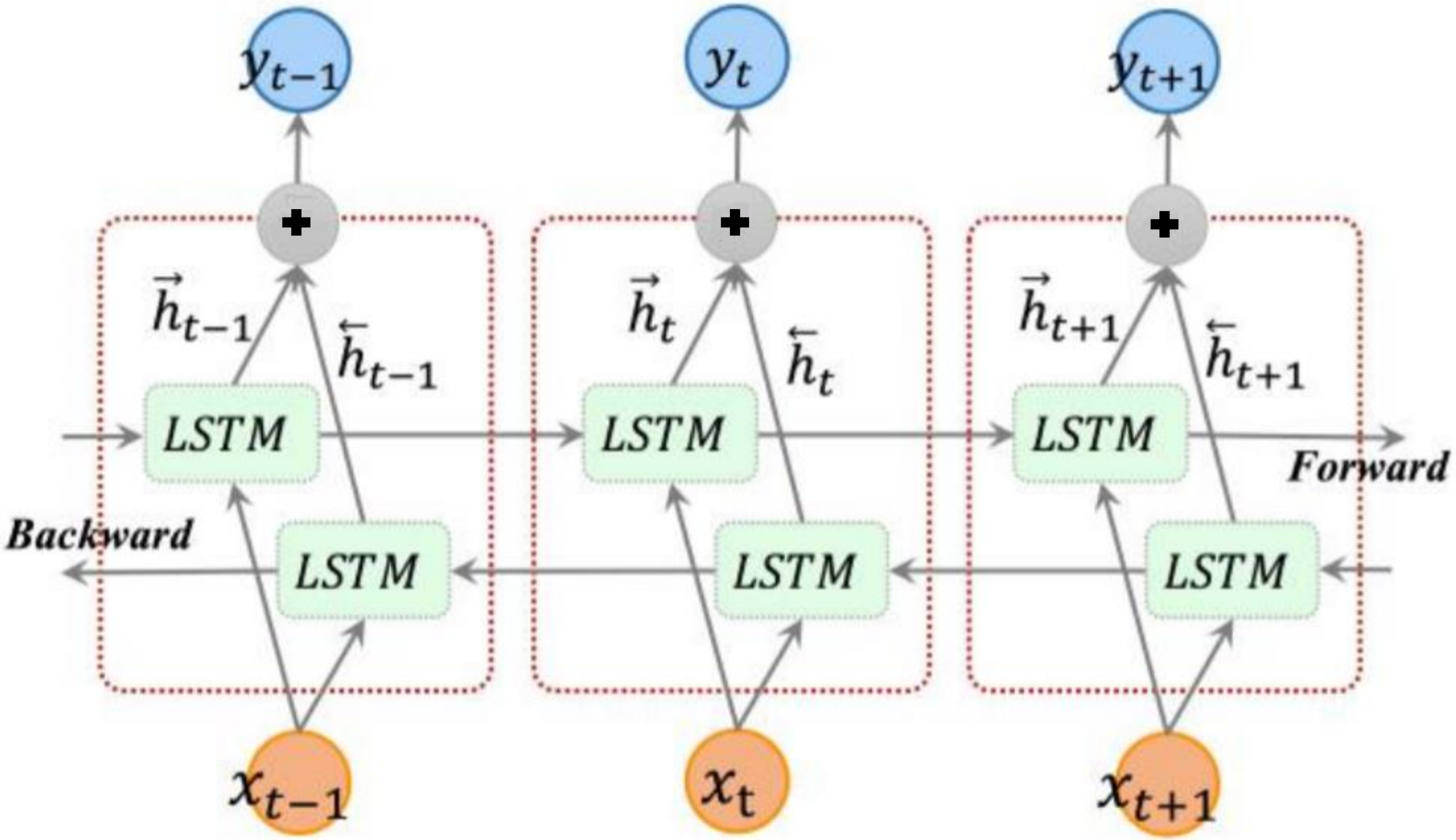

Figure 1a. BiLSTM internal architecture (courtesy)

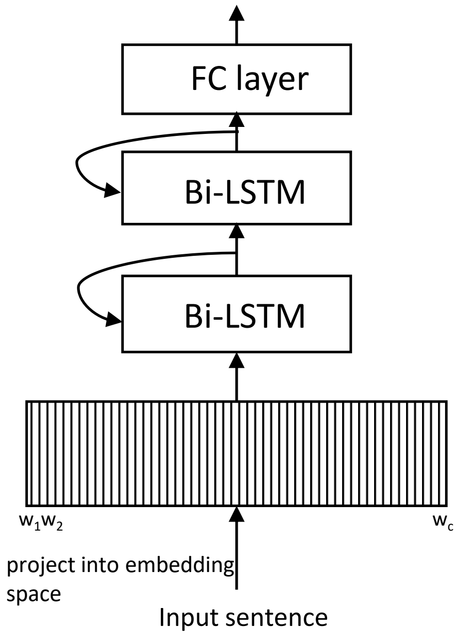

Figure 1b. Stacked BiLSTM