Chinese Word Segmentation

Model

As shown in Fig. 1 the presented model consist in a stacked bidirectional LSTM which takes in input a set of strings and an embed matrix for both **unigram** and **bigram** chinese character and concatenate them before feeding the model. The dense layer outputs the class predictions through a softmax activation function and finally the model is trained using the Stochastic Gradient Descent as optmizer.

Input and Dataset

The dataset used is the *MSR* training set. From this dataset I obtained also the embedding matrix for unigram and bigram, after splitting the file in n-grams. Due to the huge size of the training dataset, the 10% of the above-mentioned file is also beeing used for the validation dataset.

HyperParameters

The hyperparameters used for this project have been gathered from the other papers mentioned in ours.

- n-Gram embedding size: 50
- Learning Rate: 0.04
- Batch Size:32
- DropOut Rate: 0.1

No Hyperparameters Tuning has been performed.

Results and Improvements

As shown in Fig. 2, Fig. 3 and Fig. 4 , after about 30 epochs, even if the train accuracy continue increasing, the validation accuracy starts decreasing, which means that the model is Overfitting the train dataset. For this reason an early stopping is performed at epoch 32, taking care of the Validation Accuracy.

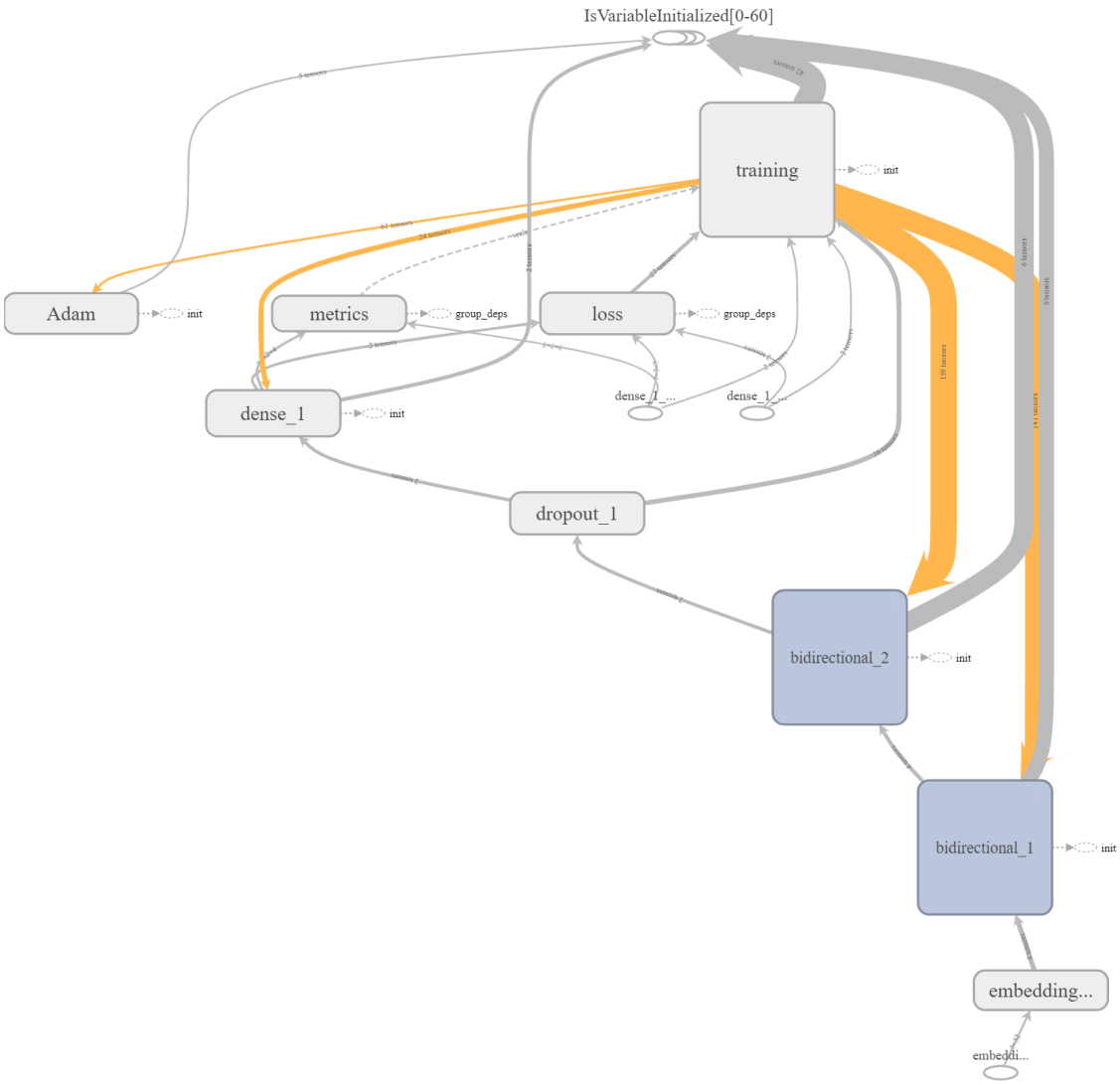

Figure 1: Model Architecture

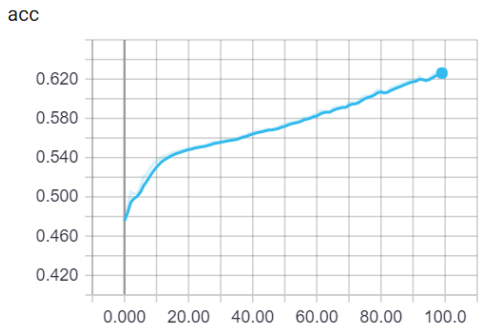

Figure 2: Train Accuracy

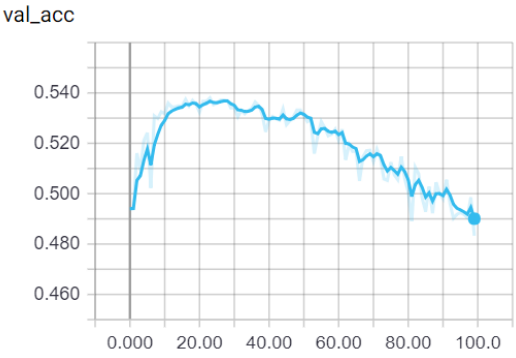

Figure 3: Validation Accuracy

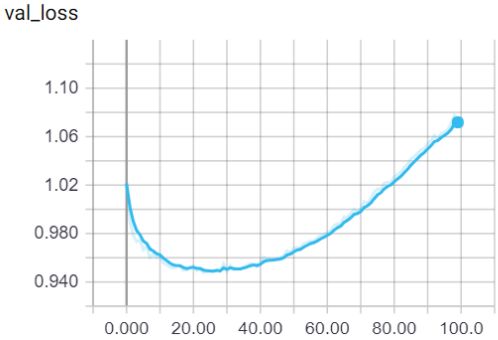

Figure 4: Validation Loss