

A TRAINING LOSS

We demonstrate the training loss optimization process with $BERT_{base}$ model on HotpotQA dataset. The loss is logged every 200 updates. The regularization of BST objective on attention mechanism doesn't reduce the optimization speed.

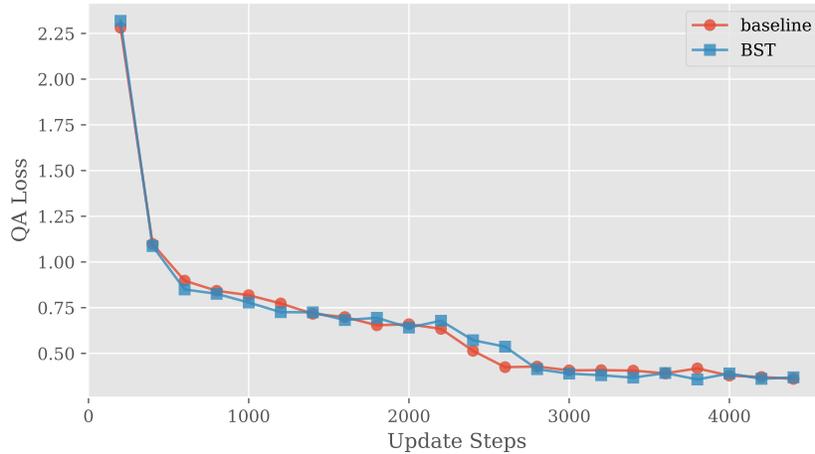


Figure 4: Training loss optimization process of baseline $BERT_{base}$ and with BST .

B BST CLASSIFIER F1 SCORE

Because we only show the BST classifier accuracy performance of layer 4 and middle layer in the main text for conciseness. Here we present the F1 score results of $BERT_{base}$ and $BERT_{large}$ results of all layers.

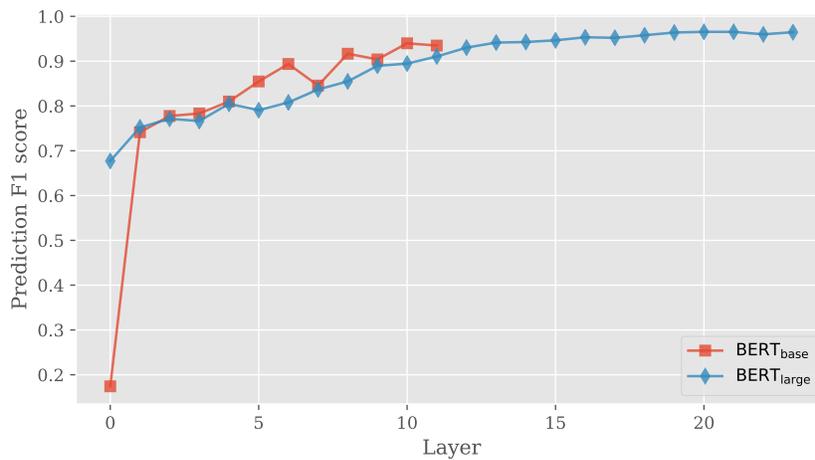


Figure 5: BST classifier F1 score of each layer of $BERT_{base}$ and $model_{large}$ on SQuAD dataset.