Authorship Verification

Made by Zihao Chen and Zhimeng Liu

Data Preprocessing

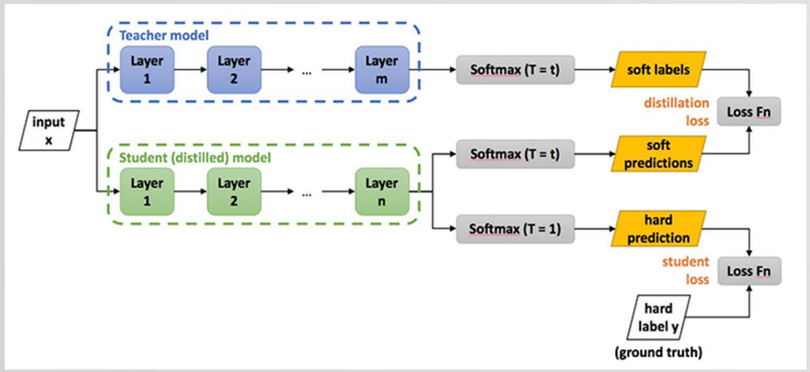
Training Set: 30k pairs of the text. Validation Set: 6k pairs of the text. Data Cleaning: Exclude pairs with shorter than 5 characters for each sequence.

DistilBERT Tokenizer: Two model instances both employed the DistilBERT tokenizer.

Data Augmentaion:Using [UNK] as "blank nosing" to prevent overfitting following the Stanford University research [1].

DistilBERT

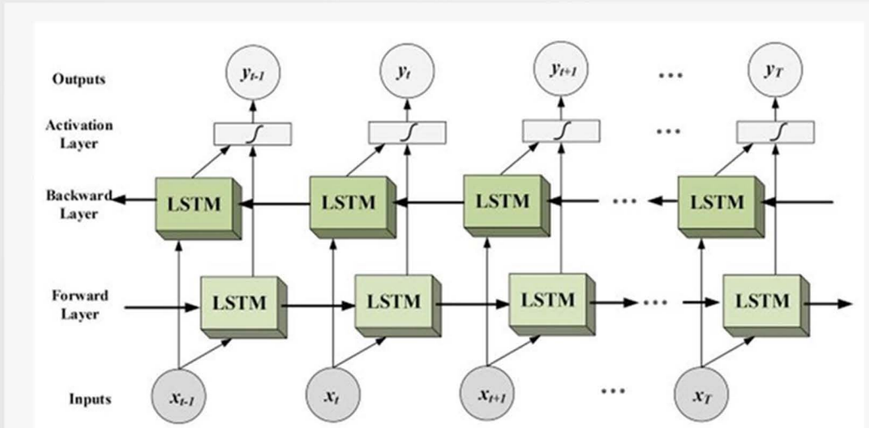
DistilBERT utilized knowledge distillation[2]. It firstly pre-trained a Teacher model using dynamic Masked Language Modeling and deriving soft labels, with the temperature-scaled softmax. Student model (DistilBERT) was also trained by MLM using the same temperature-scaled softmax and deriving soft predictions with fewer encoder blocks. Soft labels of Teacher model and soft predictions of Student model were combined to calculate distillation losses. Simultaneously, Student model (DistilBERT) used the original softmax to predict hard labels with student losses. Cosine losses combined both distillation losses and student lossese were also considered.



[3]

Bi-LSTM

Bi-LSTM combined with the DistilBERT embedding, was trained. The output along the sequential direction was averaged as the whole sequence feature. There were three Bi-LSTM layers.



[4]

Training Process

Sequence Comparison	Evaluation	Loss Function: Cross-Entropy	Optimizer: Adam	Hyperparameter
By Cosine similarity	F1-Score, Accuracy, Precision, Recall	When calculating, divide the cosine similarity by 0.1 to prevent the potential gradient explosion.	High computational efficiency, not sensitive to learning rate, suitable for fine-tuning pre-trained language models.	Seed = 68 Batch Size: 16 Learning Rate: 1e-5 Epoch: 10 Max Token Length: 256

Result on Validation Set

DistilBERT	F1-Score	Accuracy	Precision	Recall
Result	73.65%	76.32%	83.52%	65.87%

Bi-LSTM	F1-Score	Accuracy	Precision	Recall
Result	65.04%	69.94%	78.25%	55.64%

[1] Ziang Xie, Sida I Wang, Jiwei Li, Daniel L´evy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. Data noising as smoothing in neural network language models. arXiv preprint arXiv:1703.02573, 2017.
[2] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[3] Devopedia. 2020. "Knowledge Distillation." Version 11, July 24. Accessed 2024-4-16. https://devopedia.org/knowledge-distillation
[4] Verma Y. 2021. "Complete Guide To Bidirectional LSTM with Python Codes." Accessed 2024-4-16. https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/