Improving BERT-Based Noisy TextClassification with Knowledge of the Data domain

Linh Bao Doan, Viet-Anh Nguyen, Quang Huu Pham

SUN Asterisk Inc.

Introduction

Since the outbreak of COVID-19 pandemic, frequently updated information becomes a huge problem of concern. Social media platforms consequently become real-time sources for news about flare-up data. In any case, the flare-up has been spreading quickly, we observe a monstrous amount of information on social networks, for example around 4 million COVID-19 English Tweets every day on Twitter, in which most of these Tweets are uninformative. Therefore, it is crucial to collect the informative ones (for example Corona Virus Tweets identified with new cases or dubious cases) for downstream applications. In any case, manual ways to deal with recognizing useful Tweets require critical human endeavors, and hence are expensive.

Based on the dataset provided in WNUT-2020 Task 2: Identification of informative COVID-19 English Tweets [Nguyen et al., 2020], we propose a fine-tuning strategy to adopt the universal language model RoBERTa [Liu et al., 2019] as an backbone model for text classification purposes. We also conduct several experiments in varied fine-tuning architectures on the pre-trained RoBERTa. Our best model results in a high F1-score of **0.9005** on the task's private test dataset and that of **0.9218** on the public validation set with Multilayer Perceptron Head.

Model Architecture

Taking advantage of RoBERTa as a backbone, we propose a customized network with appreciably modifications. Figure 1 illustrates our proposed architecture. The "base" version of RoBERTa is used. It has 12 Transformer blocks, each block outputs a 768-D vector for each token. Since the output of different Transformer blocks represent different semantic levels for the inputs, in our experiments we combine outputs of those Transformer blocks by concatenation. This combination is fed to a classification head. We propose two types of the head:

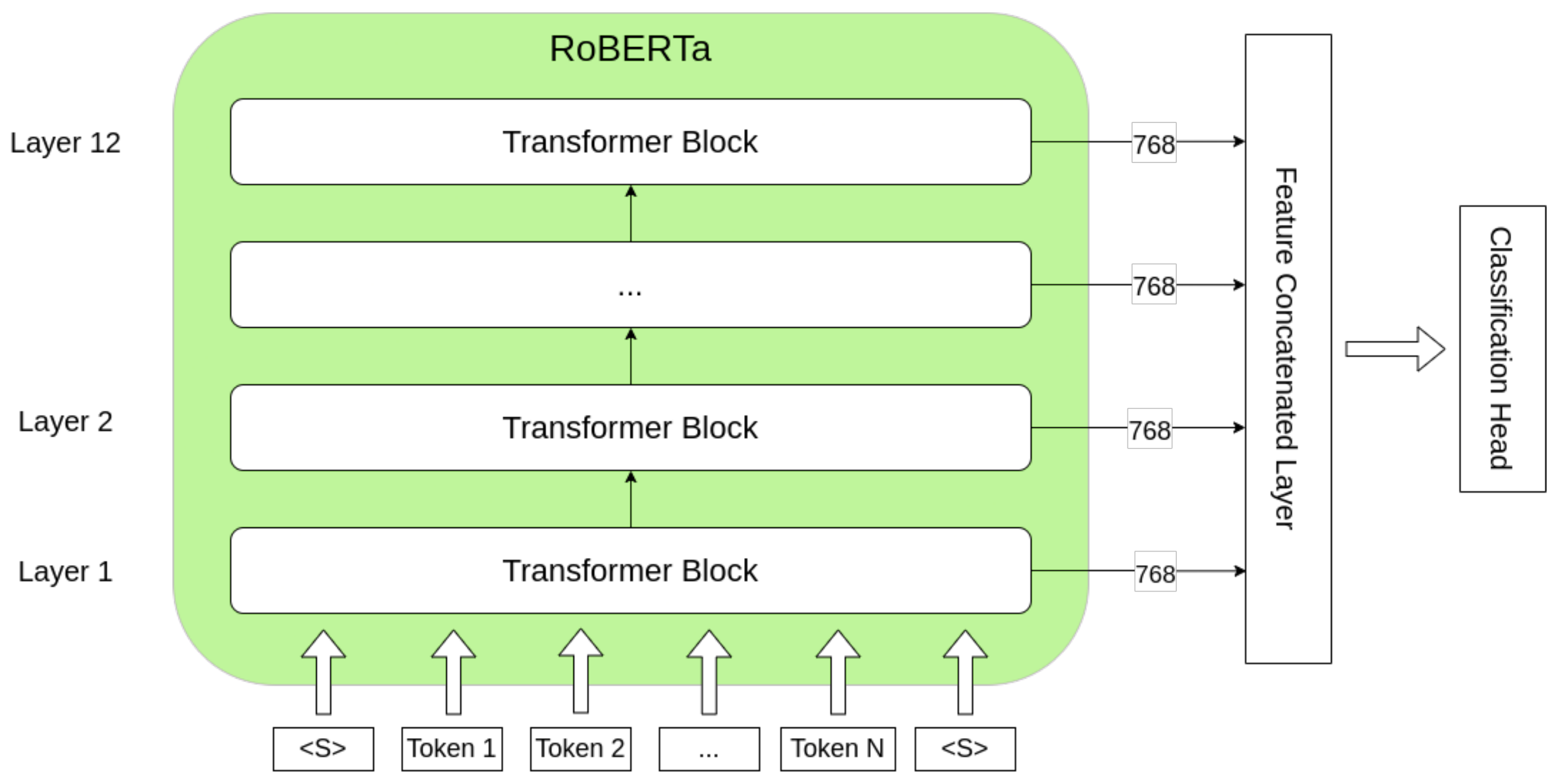

Figure 1: RoBERTa network for Text Classification Task

- **MLP Head:** A simple feed forward network with one hidden layer. This head takes the last token embedding as its input.
- **BiLSTM Head:** A recurrent neural network with one Bidirectional LSTM layer. This network takes embeddings of all tokens.

The input is tokenized into a sequence of BPE tokens. RoBERTa, the "base" version, takes this sequence and propagates it through 12 Transformer layers. By concatenating outputs from these 12 layers, we form a long sentence representation for the follow-up classification head, which is a simple Multi-layer Perceptron/Long Short-Term Memory network

Tuning

RoBERTa apparently is an excellent language model since it was trained on a huge dataset in a broad domain. However, the general domain is also a drawback when it comes to downstream tasks with completely different domains such as classifying users' tweets on Twitter. Therefore, in order to produce high-quality outputs from the model, there is a need of fine-tuning MLM task on the task dataset for RoBERTa. This adapts the universal language model into our narrow domain, giving it prior knowledge for later classification training.

Choosing learning rate is the key factor for the convergence. If learning rate is too small, the model may converge too slow causing harder to fit to new data distribution. On the other hand, large learning rate can lead to the problem of useful feature forgetting. Hence, we employ warm-up learning rate scheduler [Howard and Ruder, 2018] to help the model converge faster while preserving its good initialization.

Methods

We assume fine-tuning only on the dataset might cause overfitting on the chosen dataset only. Hence, we propose a hierarchical fine-tuning strategy for RoBERTa: the first phase we train with custom domain COVID Tweets dataset for domain adaptation, then the second phase is a fine-tuning process with WNUT Task 2 dataset for task adaptation. Our custom COVID Tweets dataset is gathered from Twitter platform, including unlabeled 1 million posts in general COVID domain, which has the hashtag of **#Covid**, **#Covid19**, and **#Coronavirus**. We expect this model to generalize better on different distributed dataset in the same field of COVID Tweets.

Figure 2 illustrates our process. For MLM tuning we propose hierarchical tuning process that consists of two steps: Domain adaptation using extra COVID data and Task adaptation using the given training data. After MLM Tuning, we utilize different training techniques for text classification such as back translation, warm-up learning rate, layer freezing and layer-wise learning rates. This section provides details of this pipeline.

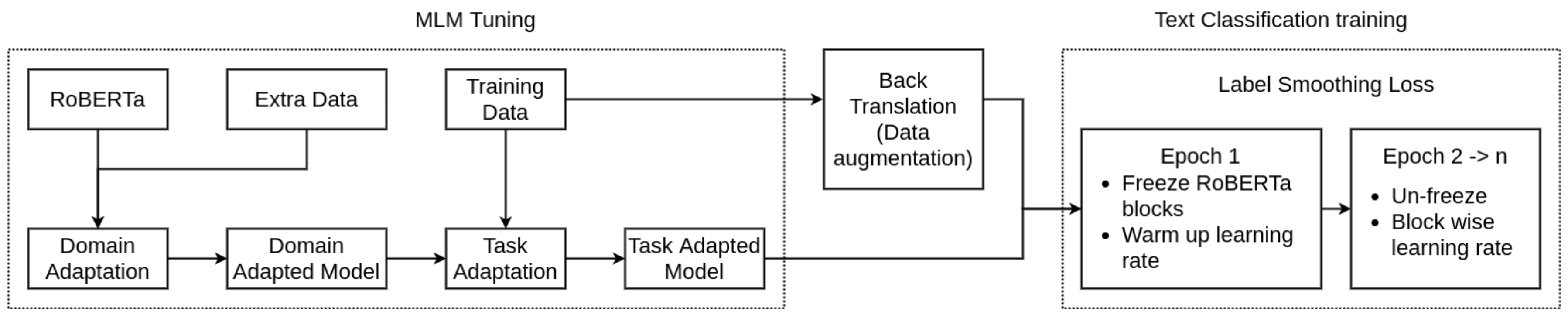

Figure 2: Pipeline for hierarchical MLM tuning and main task training.

Back translation [Xie et al., 2019, Edunov et al., 2018] can be used as a potential form of data augmentation in Text Classification. In our experiment, 25% of the data samples is back-translated into Vietnamese, the same amount goes for Italian and French, and the rest 25% is kept unchanged. This assures the languages contribute equally to the overall dataset. Totally, the dataset size is increased by 75%.

Result

Model	Precision	Recall	F1	Accuracy
RoBERTa + MLP Head	0.9407	0.8740	0.9061	0.9080
RoBERTa + BiLSTM Head	0.9322	0.8853	0.9082	0.9110
Direct tuning + MLP Head + Label smoothing	**0.9492**	0.8960	**0.9218**	**0.9240**
Direct tuning + BiLSTM Head + Label smoothing	0.9364	**0.8983**	0.9170	0.9200
Direct tuning + MLP Head + Back translation + Label smoothing	0.9343	0.8909	0.9121	0.9150
Hierarchical tuning + MLP Head + Back translation + Label smoothing	0.9449	0.8745	0.9084	0.9100

Table 1: Comparison of different tuning and training techniques on the public validation set.

To facilitate the outcome, table 1 compares the performance of multiple trial architectures training with pre-trained method using RoBERTa in our base settings. The original RoBERTa with MLP Head shows the better result than LSTM head, but the difference is not really noticeable (0.9082 vs. 0.9061). When applying direct tuning MLM and label smoothing, the gap has been widened, specifically, 0.9218 for MLP Head and 0.9170 for LSTM Head.

References

[Edunov et al., 2018] Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018).
Understanding back-translation at scale.

[Howard and Ruder, 2018] Howard, J. and Ruder, S. (2018).
Universal language model fine-tuning for text classification.

[Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019).
Roberta: A robustly optimized bert pretraining approach.

[Nguyen et al., 2020] Nguyen, D. Q., Vu, T., Rahimi, A., Dao, M. H., Nguyen, L. T., and Doan, L. (2020).
WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets.
In *Proceedings of the 6th Workshop on Noisy User-generated Text.*

[Xie et al., 2019] Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., and Le, Q. V. (2019).
Unsupervised data augmentation for consistency training.

R&D Lab, SUN* Inc. **November 16, 2020**