A Study of Combining Transfer LearningMethods for Text Classification with RoBERTa

Huy Quang Dao, Tam Minh Nguyen

VNU University of Engineering and Technology , SUN Asterisk Inc.

Introduction

This paper reports our approach and the results of our experiments for W-NUT task 2: Identification of Informative COVID-19 English Tweets [Nguyen et al., 2020] . Our main contributions are:

- We perform numerous experiments to support our hypothesis. Which is fine-tuning state-of-the-art pre-trained language models such as RoBERTa [Liu et al., 2019] is more beneficial to the Identification of Informative COVID-19 English Tweets task than several training-from-scratch models.
- We combine many fine-tuning techniques to form our best model and experiments with the effect of each technique by gradually stacking them onto our base model then observe their effects.

Our best model results in a high F1-score of 89.89 on the task's test dataset and that of 90.96 on the public validation set without ensembling multiple models and additional data.

Model Architecture

Our base model is designed to test out the effectiveness of our choice of backbone, custom head, and training techniques:

- Choice of backbone: RoBERTa Base or RoBERTa Large.
- The backbone extracting all information of an input sequence into the [CLS] token. Those features are then linearly projected onto 2D, followed by a Softmax activation to predict the probability of the label being "Informative" or "Uninformative" given the input.
- We use cyclical learning rate and label smoothing to all model's settings in our experiment including the base model's settings.
- More complex model settings use a more complex custom head and are gradually added advanced training techniques.

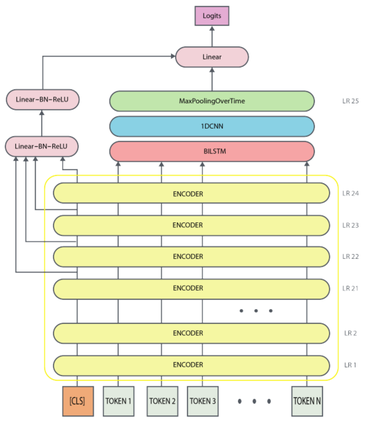

Figure 1: Our Proposed Architecture.

Custom head design:

- Using CLS token's features of 4 last backbones' layer.
- Combine with features of the rest of tokens at the last backbone layers, by passing them through Bidirectional LSTM, 1D Convolutional Neural Network, and Max Pooling Over Time, then concatenate with the transformed CLS token's features.

Transfer Learning Techniques

ULMfit [Howard and Ruder, 2018] introduced several fine-tuning methods which boost upmost NLP downstream task's performance, consists of three techniques.

- **Learning rate discrimination**: setting the learning rate differently for each layer, the lower layer has a small learning rate while higher ones update their weights at a higher rate.
- **Cyclical learning rate**: warm up learning rate by a gradual increase, followed by a gradual decrease.
- **Gradual unfreezing**: apply multi-phase training, each unfreezes one layer, from top to bottom gradually.

We also investigate the effect of these techniques in our proposed framework.

Traditional Training From Scratch Models

In order to emphasize the significant contribution of the pre-training method using the SOTA language model, we also train used-to-be state of the art models in text classification including:

- Hierarchical Attention Networks [Yang et al., 2016] combined with Weight-Dropped GRU [Merity et al., 2017].
- Bidirectional Long Short-Term Memory Networks with Two Dimensional CNN [Zhou et al., 2016].
- Non-Static Convolutional Neural Network [Kim, 2014].

Result

Model	Precision	Recall	F1	Accuracy
BiLSTM + 2DCNN (*)	77.89	83.31	80.51	80.90
Non Static CNN (*)	81.63	80.38	81.00	82.18
HAN + WD-GRU (*)	82.65	82.08	82.36	83.40
RoBERTa Base	87.60	92.97	90.21	90.46
RoBERTa Large	**88.40**	**92.84**	**90.57**	**90.86**

Table 1: Comparing pretraining using RoberTa and training from scratch with previous SOTA models on text classification. (*) denotes for our implementation.

Model	Precision	Recall	F1	Accuracy
RoBERTa Base	87.60	92.97	90.21	90.46
RoBERTa Base + DLr	87.84	93.09	90.39	90.64
RoBERTa Base + DLr + Head	88.66	92.33	90.46	90.80
RoBERTa Base + DLr + Head + GU	88.40	93.22	90.75	91.02
RoBERTa Large	88.40	92.84	90.57	90.86
RoBERTa Large + DLr	87.84	93.74	90.69	90.90
RoBERTa Large + DLr + Head	**88.75**	92.84	90.75	91.06
RoBERTa Large + DLr + Head + GU	88.15	**93.96**	**90.96**	**91.14**

Table 2: Comparing models using RoBERTa Base and RoBERTa Large as backbone. DLr denotes Discrimination Learning Rate, Head denotes Custom Head and GU denotes Gradual Unfreezing.

Table 1 compared the performance of several architectures training from scratch and pre-training method using RoBERTa in our base settings. These traditional architecture are Hierarchical Attention Networks combined with Weight-Dropped GRU (HAN + WD-GRU), Bidirectional Long Short-Term Memory Networks with Two Dimensional Convolutional Neural Network (BiLSTM + 2DCNN) and Non Static Convolutional Neural Network (Non-static CNN). As far as we expect, the pre-training method significantly out-performs training from scratch with traditional architecture results in a gain of around 8.0 F1-score.

Table 2 compared the performance of 4 models using RoBERTa Base and 4 models using RoBERTa Large as the backbone. Two main observation:

- Firstly, RoBERTa Large outperforms RoBERta Base in all model settings. Which meets our expectations.
- Secondly, gradually adding up training technique and custom head leads to a gradual increase in F1-score. This emphasizes the positive effect of all techniques and custom heads on this task's performance. Overall, our winning model results in an 89.89 F1-score.

References

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

[Kim, 2014] Kim, Y. (2014).
Convolutional neural networks for sentence 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.

[Merity et al., 2017] Merity, S., Keskar, N. S., and Socher, R. (2017).
Regularizing and optimizing lstm language models.

[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*.

[Yang et al., 2016] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016).
Hierarchical attention networks for document classification.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

[Zhou et al., 2016] Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016).
Text classification improved by integrating bidirectional lstm with two-dimensional max pooling.