From Universal Language Model to Downstream Task: Improving RoBERTa-Based Vietnamese Hate Speech Detection

Quang Huu Pham[1] Viet-Anh Nguyen[1] Linh Bao Doan[1]
Ngoc N. Tran [1] Ta Minh Thanh[2]

[1]R&D Department
Sun-Asterisk Inc.

[2]Faculty of Computer Science
Le Quy Don Technical University

The 12th IEEE International Conference on Knowledge and Systems Engineering

Sun*

Table of Contents

Sun*

Table of Contents

Motivation and objective

- Motivation
  - Pre-trained language models are extremely useful for downstream tasks.
  - Fine-tuning a well-trained model on a task-specific dataset needs to be carefully handled.

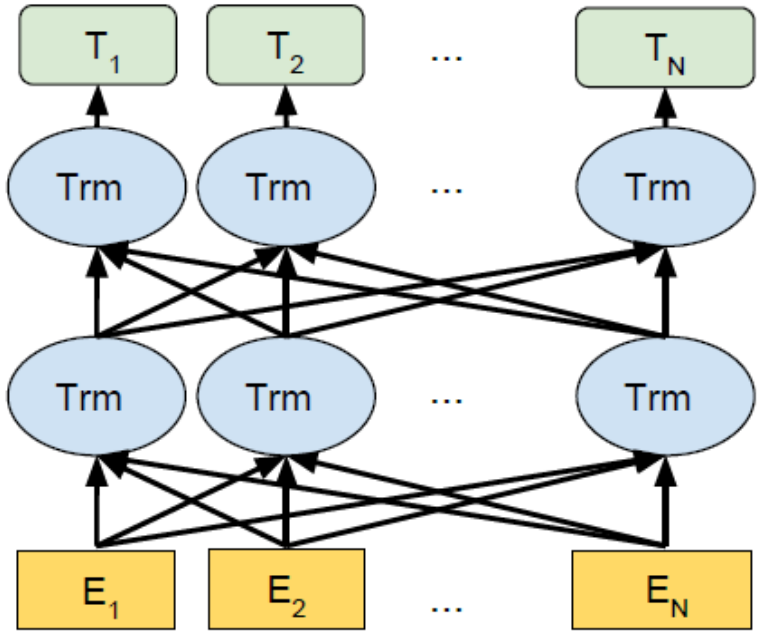
Motivation and objective

- Motivation
  - Pre-trained language models are extremely useful for downstream tasks.
  - Fine-tuning a well-trained model on a task-specific dataset needs to be carefully handled.
- Objective
  - Propose a general tuning strategy for language models on downstream tasks.
  - Apply the pipeline with a RoBERTa-based (i.e. PhoBERT) architecture to solve Vietnamese Hate Speech Detection task.

Sun*

Background

- RoBERTa language model
  - The original by Facebook[1]: BERT without NSP, trained on 160GB text.
  - Vietnamese version PhoBERT by VinAI[2]: Trained on 20GB texts (1GB Wikipedia, 19GB news)

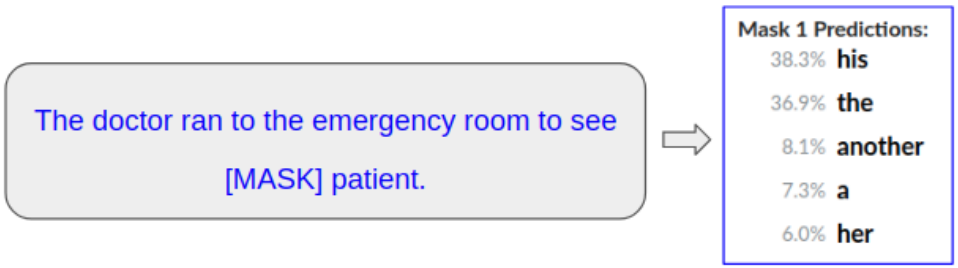


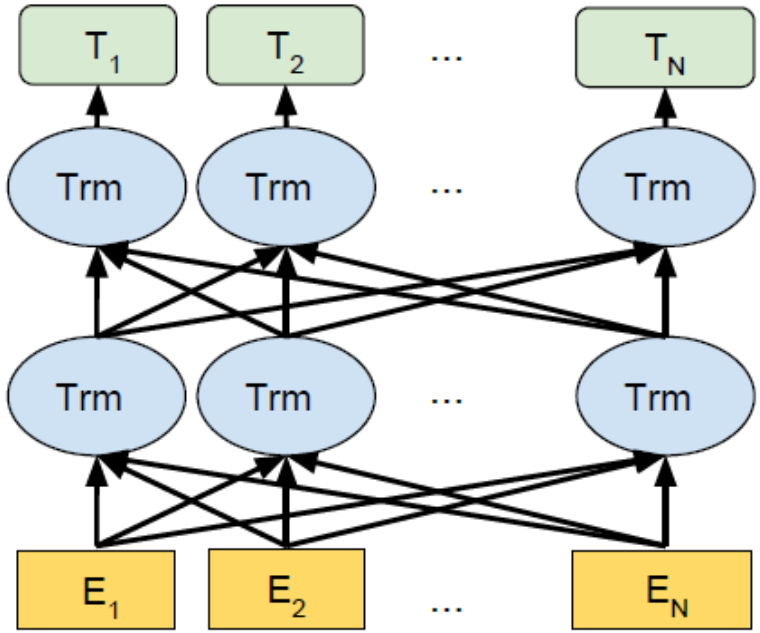
[1] Yinhan Liu et. al. *Roberta: A robustly optimized bert pretraining approach, 2019.*
[2] N.Q.Dat and N. A. Tuan. *Phobert: Pre-trained language models for Vietnamese, 2020.*

Sun*

Background

- RoBERTa language model
  - The original by Facebook[1]: BERT without NSP, trained on 160GB text.
  - Vietnamese version PhoBERT by VinAI[2]: Trained on 20GB texts (1GB Wikipedia, 19GB news)
- Masked language modeling: fill-in-the-blank task





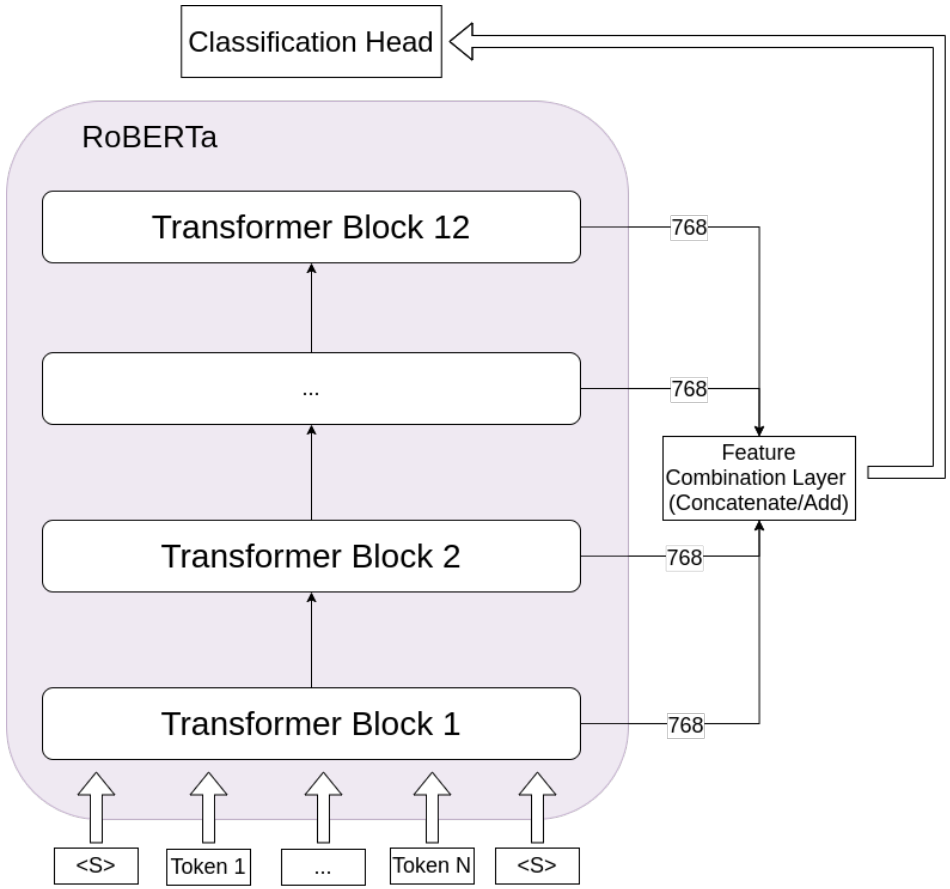
[1] Yinhan Liu et. al. *Roberta: A robustly optimized bert pretraining approach, 2019.*
[2] N.Q.Dat and N. A. Tuan. *Phobert: Pre-trained language models for Vietnamese, 2020.*

Sun*

Table of Contents

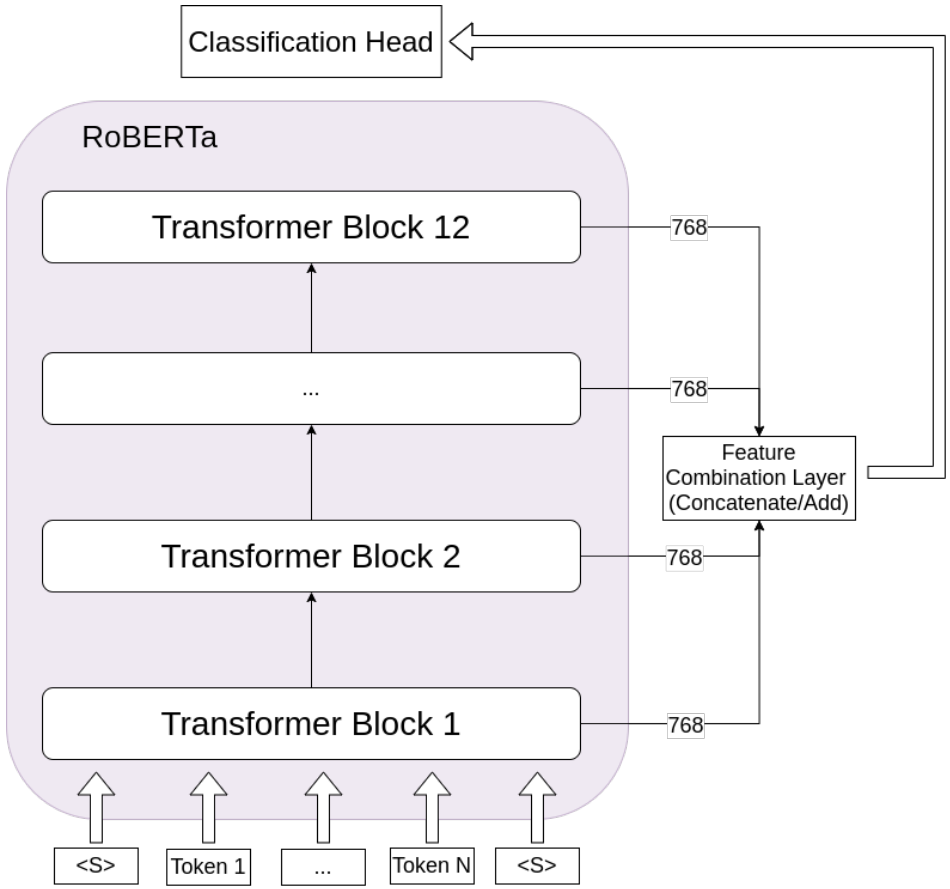
Classification architecture

- RoBERTa-base (PhoBERT's weights) as backbone network



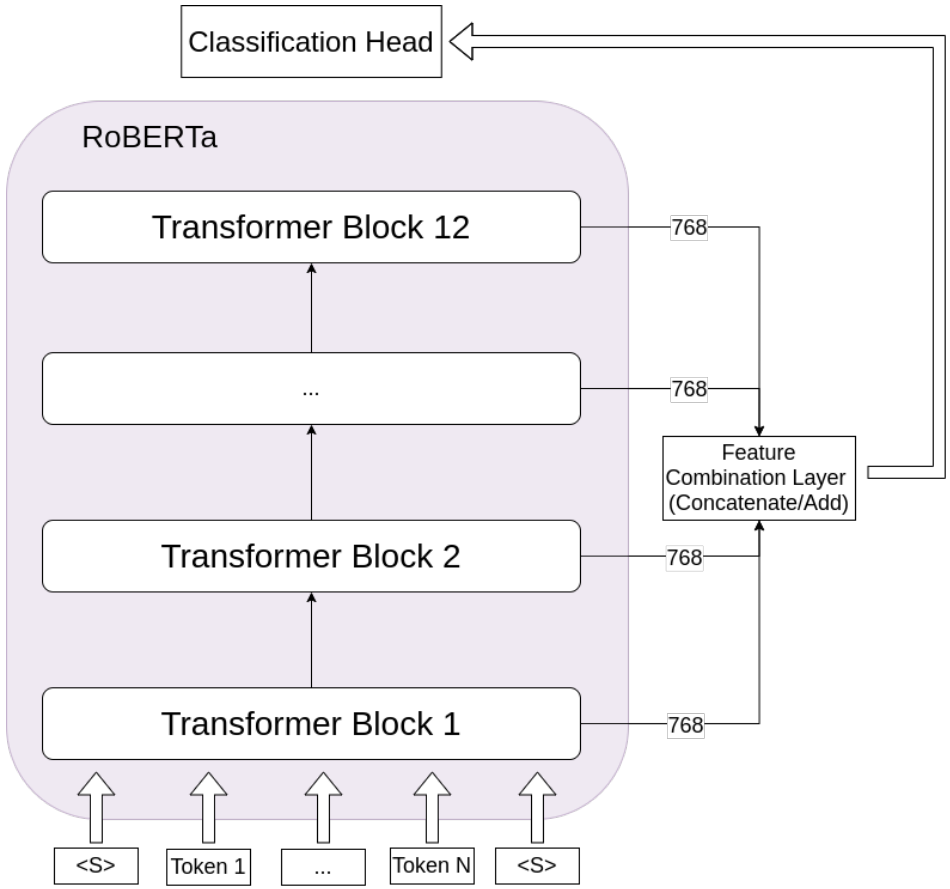
Classification architecture

- RoBERTa-base (PhoBERT's weights) as backbone network
- Combination of different layer embeddings



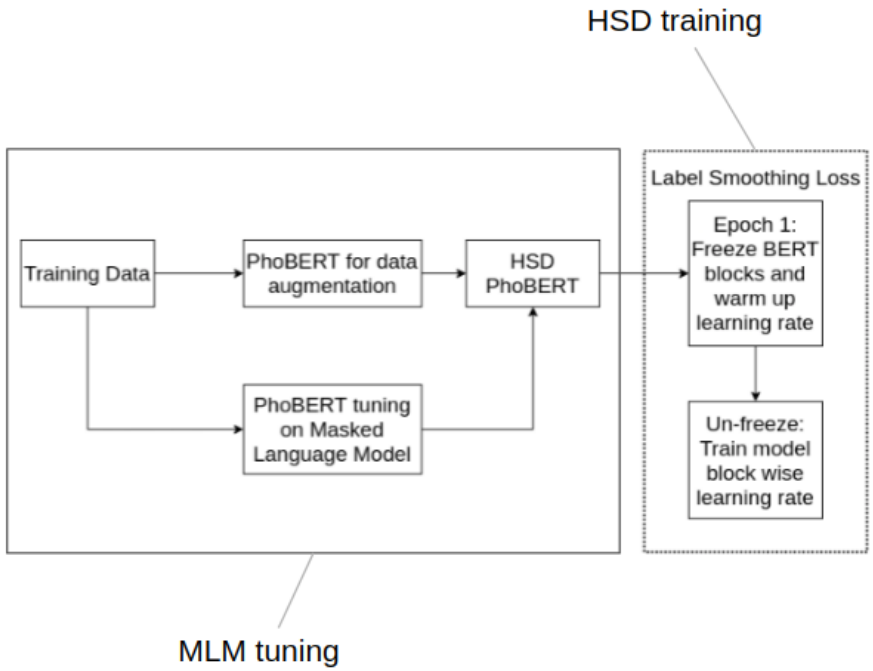
Classification architecture

- RoBERTa-base (PhoBERT's weights) as backbone network
- Combination of different layer embeddings
- Classification head: Multi-layer perceptron



Fine-tuning pipeline

- MLM Tuning:
  - Randomly replace 5 tokens using PhoBERT
  - Tune the language model on training data



Fine-tuning pipeline

- MLM Tuning:
  - Randomly replace 5 tokens using PhoBERT
  - Tune the language model on training data
- HSD training:
  - The first epoch: Freeze transformer encoders, train MLP head with warm-up learning rate
  - The rest: Unfreeze and train all encoders with block-wise learning rates
  - Label smoothing loss

$$y'_k = y_k(1-\alpha) + \alpha/K$$

K classes

one-hot label

smoothing parameter

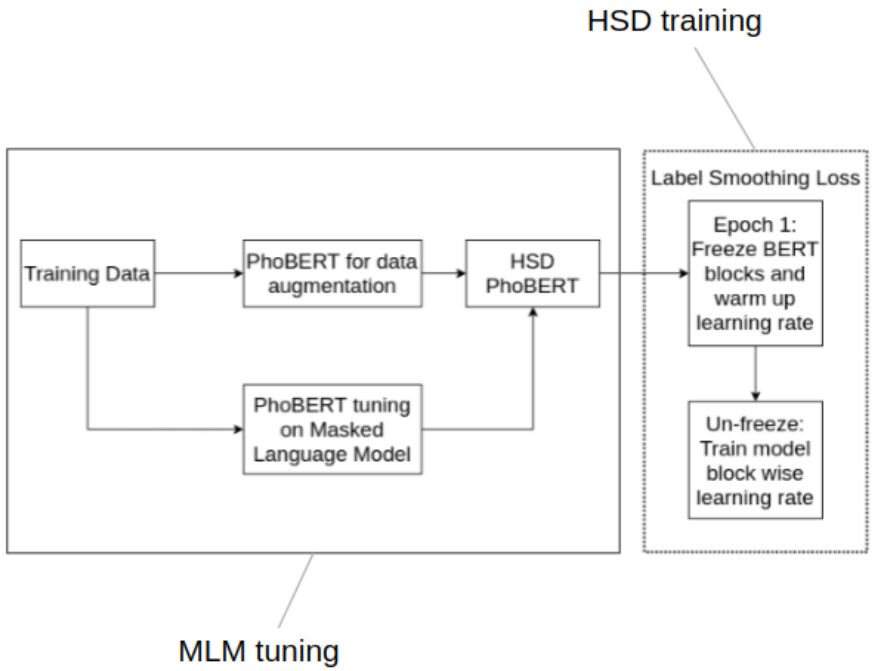


Table of Contents

Experiments

- Experiment with different combinations of embeddings from 12 layers.

	HATE	OFFENSIVE	CLEAN
Number of sample	709	1,022	18,614

Noise: abbreviations, emoji, special characters, foreign language, teen code, typing errors.

Crawled from SNS's

Experiments

- Experiment with different combinations of embeddings from 12 layers.
- Investigate effectiveness of each individual and all fine-tuning techniques.

	HATE	OFFENSIVE	CLEAN
Number of sample	709	1,022	18,614

Noise: abbreviations, emoji, special characters, foreign language, teen code, typing errors.

Crawled from SNS's

Results

Table: Mean of Macro $F1$ score on Stratified K-fold with k = 10 of difference blocks

Feature blocks	Mean of F1 score
Layer 6 (only single block)	0.6854
Layer 12 (only single block)	0.6978
Layer 3-6 (4 middle blocks)	0.6855
Layer 9-12 (4 last blocks)	**0.6989**
Layer 1-6 (6 first blocks)	0.6905
Layer 7-12 (6 last blocks)	**0.6989**
Layer 1-12 (all blocks)	0.6979

Table: Mean of Macro $F1$ score on Stratified K-fold with k = 10 with concatenate of layers 6-12 and our training approach

Proposed training approach	Mean of F1 score
Cross entropy loss	0.6922
Label Smoothing loss	0.7005
Non warm-up learning rate	0.6989
Warm-up learning rate	0.7062
Non Fine-tune MLM	0.6989
Fine-tune MLM	0.7162
Non Block wise learning rate	0.7051
Block wise learning rate	0.7079
Combine all the methods	**0.7211**

Sun*

Table of Contents

Conclusion

- What we have done:
  - Proposed a pipeline for adapting a universal language model to downstream tasks
  - Applied the pipeline into Hate Speech Detection task, achieved top 1 on the leaderboard.

Conclusion

- What we have done:
  - Proposed a pipeline for adapting a universal language model to downstream tasks
  - Applied the pipeline into Hate Speech Detection task, achieved top 1 on the leaderboard.
- Future work:
  - Design more complex classification head
  - Try employing the model and pipeline on different languages.

Sun*

Thank you for listening!