Robustness to Spurious Correlation using Pre-trained Language Models and two-step Fine Tuning

Ambroise LAROYE (ambroise.laroye--langouet@ensea.fr)
Supervisor : Son Vu (son.vu@ensea.fr)

Abstract

NLP models tend to rely on spurious correlation between labels and input features. Shortcut learning or spurious correlation refer to "biased words" that disproportionately influence predictions.
This work first proposes several identification methods and then a two-step fine tuning on BERT model, using classification task.
This study compares two identifying methods based on forgettable example and important feature.

Introduction

In statistics, a spurious correlation refers to a connection between two variables that appears to be causal but is not. With spurious correlation, any observed dependencies between variables are merely due to chance or related to some unseen confounder.
Recent studies highlighted the potential of leveraging pre-trained language models (PLMs), incorporating forgettable examples, and employing fine-tuning or transferring learning techniques.
The idea of this study is to combine these methods to increase the robustness of a model.
Recognizing and addressing spurious correlations is essential for producing reliable and actionable insights. As data analysis becomes increasingly automated and complex, the risk of spurious correlations and backdoor attacks rises, necessitating rigorous validation, robust methodologies.

Theoretical Framework

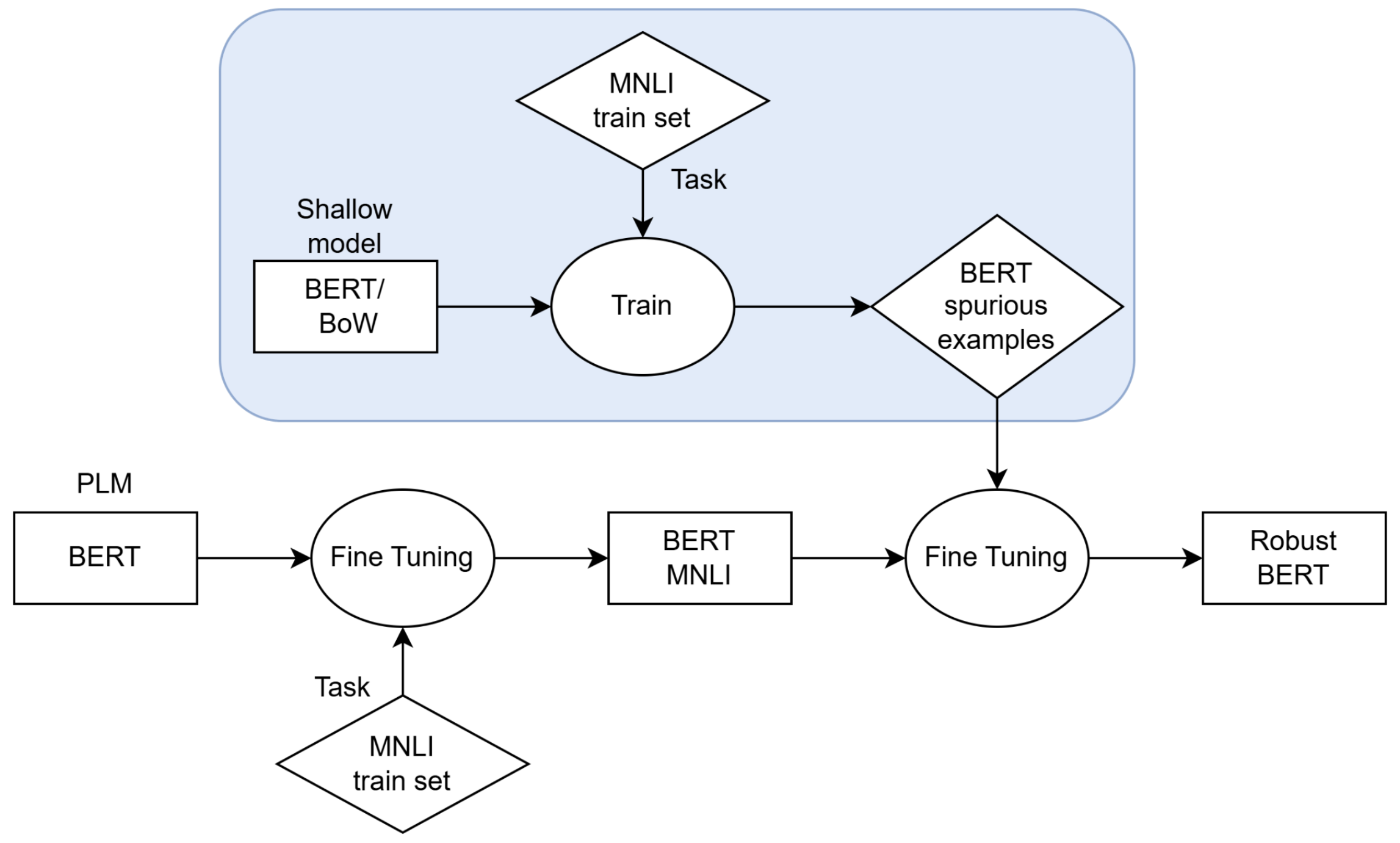
Forgettable example :
An example is forgotten if it goes from being correctly to incorrectly classified during training (each such occurrence is called a forgetting event). This happens due to the stochastic nature of gradient descent, in which gradient updates performed on certain examples can hurt performance on others.
Attention score / important feature:

1. Apply the trained BERT model to each sentence s_i to obtain the probabilities p_i^{pos} and p_i^{neg} for the positive and negative labels, respectively.
2. Extract attention scores $\{a_{i1}, a_{i2}, \dots, a_{im}\}$ for tokens $\{t_{i1}, t_{i2}, \dots, t_{im}\}$ in sentence s_i, (m : length of the sentence)
3. Update the attention scores :

$$\tilde{a}_j^i = \begin{cases} a_j^i \cdot p_i^{\text{pos}}, & \text{if } p_i^{\text{pos}} > p_i^{\text{neg}}, \\ -a_j^i \cdot p_i^{\text{neg}}, & \text{otherwise.} \end{cases}$$

Methodology

We consider 3 parts in our methodology :
1. Identify Spurious correlation through forgettable example or attention score
2. First fine tuning PLM on MNLI task
3. Second fine tuning on extract examples (spurious examples)



PLM :
BERT, is a Transformer, which follow the encoder-decoder framework using stacked multi-head self-attention and fully-connected layers for both the encoder and decoder.

Task :
MNLI, consists of sentence pairs annotated with labels indicating their relationship in terms of textual entailment: these labels include entailment, contradiction, or neutral.
HANS, is composed of both entailment and contradiction examples that have high word-overlap between hypothesis and premise(e.g. "The president advised the doctor" X → "The doctor advised the president"). Allowing to evaluate the robustness of a model. The evaluation on the HANS task is the primary focus here, with the objective being to maximize it. However, this must be achieved without a drop in accuracy on MNLI. Therefore, a trade-off must be made between these two metrics.

The study investigates 3 cases on BERT model on the task MNLI.
1. Spurious = Forgettable example only
2. Spurious = Important features
3. Spurious = combination of both forgettable and important features

Other cases have been studied in previous papers, where the shallow model used to extract spurious correlations differs from the base model. For example: BoW and BiLSTM.

A variation of this study involves using a dataset different from MNLI. For example: QQP, FEVER, or GLUE tasks.

Results

In this section, the goal is to assess whether our methodology is effective against spurious correlations. Specifically, we aim to evaluate whether our models perform satisfactorily on the HANS task compared to the initial BERT base model trained over 4 epochs.

Model	MNLI	HANS	Threshold
BERT	**84.2**	62.9	/
$BERT + Forget_{BERT}$	82.9	69.0	/
$BERT + Important_{BERT}$	82.4	73.3	0.08
$BERT + Important_{BERT}$	82.7	**73.6**	0.1
$BERT + Important_{BERT}$	82.5	73.4	0.3

Table 1: Accuracy after 4 epochs on BERT ; BERT + Forgettable examples ; BERT + Important samples

The notation "BERT + Forget$_{BERT}$" refers to an initial fine-tuning of BERT, followed by a second fine-tuning on the forgettable examples of BERT. First, the results are satisfactory, using forgettable examples and the two-stage fine-tuning process effectively enhances the robustness of the BERT model. Although there is a slight decrease in accuracy on MNLI, this difference is negligible compared to the improvement observed on the HANS dataset. The same applies to important samples. Additionally, we varied the threshold corresponding to the number of features selected. The most satisfactory result on HANS was achieved using important features with a threshold of 0.1.

Model	MNLI	HANS
BERT	**84.2**	69.5
$BERT + Forget_{BERT}$	82.9	69.0
$BERT + Forget_{BILSTM}$	82.9	**70.4**
$BERT + Forget_{BoW}$	**84.3**	**70.5**
$BERT + Forget_{HANS}$	83.9	69.5

Table 2: Accuracy after 4 epochs on BERT ; BERT + Forgettable examples

In this second table, we examine the shallow models used to detect spurious correlations. This is more of a side note within the context of forgettable examples, as our primary focus is on the BERT model. Nonetheless, note that the best results are achieved with BoW and BiLSTM.

Model	MNLI	HANS	Epoch
BERT	**84.2**	69.5	4
$BERT + Forget \cap Important_{BERT}$	83.63	72.34	1
	83.29	72.66	2
	83.18	72.73	3
	83.05	72.77	4

Table 3: Accuracy on BERT ; BERT + forgettable examples combine with important samples ; threshold = 1

This table presents the results for a combination of forgettable examples and important features. Interestingly, this approach does not significantly improve the accuracy on HANS, as the best result remains that achieved with important samples (Table 1). However, the accuracy on MNLI is preserved, which is an advantage not observed with the previous methods. Additionally, the accuracy on HANS reaches 72.34% after the first epoch, which is relatively high for the first epoch compared to the previous model.

Conclusion

This study highlighted the use of PLMs, a two-step fine-tuning approach (transfer learning), and the identification of spurious correlations through the methods of important features and/or forgettable examples to enhance a model's robustness.
The model used was BERT, and robustness evaluation was conducted using the HANS task.

The proposed method proved effective, achieving its best results with important features and a threshold of 0,1.
A trade-off was made between the accuracy of MNLI and that of HANS, prioritizing the latter while tolerating a 2% decrease in MNLI accuracy.
However, to maintain or even improve accuracy on both MNLI and HANS, it is possible to combine forgettable examples and important features.

References

[1]: Increasing Robustness to Spurious Correlations using Forget-table Examples (Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni). https://aclanthology.org/2021.eacl-main.291
[2]: Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models (Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang) http://arxiv.org/abs/2110.07736

Acknowledgements

I sincerely thank my supervisor, Son Vu, for his expertise, guidance, and feedback throughout this project. I am also grateful to Magalie Topczynski for her help in writing this poster.

Additional Information

CODE

More Of Me