Research Project (3A): Experiment

Avoiding Shortcut Learning in Deep Models for Enhanced Defense Against Adversarial Attacks

Author: Ambroise LAROYE–LANGOUËT

ambroise.laroye--langouet@ensea.fr

Supervisor: Son Vu

son.vu@ensea.fr

January 30, 2025

Contents

Github repository : `https://github.com/Ambroise012/robustness_to_spurious_correlation`

Key words:
Pre-trained Large Models, Fine-tuning, Classification, Spurious Correlation, Forgettable Examples, Robustness, Attention Score, Local Intrinsic Dimensionality (LID).

1 Research question

How can a two-step fine tuning enhance robustness against spurious correlations of pre-trained models? And what is the best method to extract spurious correlation ?

2 Methodology

We consider 3 parts in our methodology :

1. Identify Spurious correlation through forgettable example, feature extraction or LID or a combinaision of these methods
2. First fine tuning PLM on MNLI task
3. Second fine tuning on spurious examples

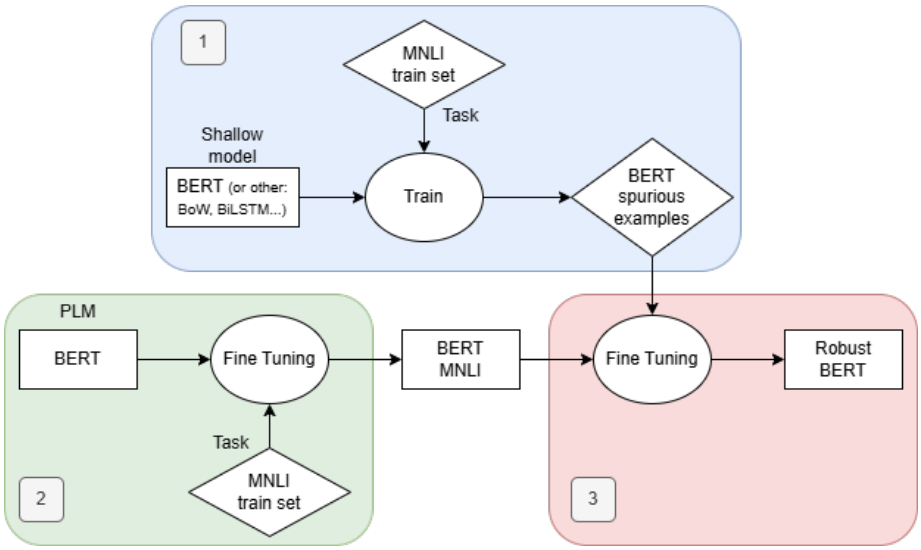


Figure 1: Methodology followed in this study

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 $\rightarrow$ “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 = ”forget”
2. Spurious = Important features through attention score = ”important”
3. Spurious = Important features through LID score = call ”LID” for simplicity
4. Spurious = combination of both important features and forgettable or ”LID”

Other cases have been studied in previous papers Yaghoobzadeh et al. [2021], 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 other GLUE tasks.

3 Spurious correlation extraction

3.1 Forgettable examples

This method was studied in the paper Yaghoobzadeh et al. [2021].
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.
If an example is forgotten at least once or is never learned during training it is dubbed forgettable.
For the extraction of spurious correlations, we consider a balanced approach by class. This means that we equalize the spurious correlations across all classes, thus avoiding disproportionality or overfitting of any single class.

3.2 Important example & attention score

In the previously discussed method, which involves using forgettable examples, we observe some limitations. Indeed, this method does not account for all spurious correlations. A

spurious correlation may consistently be misclassified without its classification ever changing. This reveals the need for a more effective method. One of my intuitions is to use the attention score. With this approach, we can determine the extent to which an example influences the model. Specifically, it helps identify whether an example is out-of-distribution within the dataset. By selecting a certain threshold, we can determine the attention score above which an example is classified as spurious.
Why use the attention score? Attention scores indicate the relative importance of different parts of an input in the model's prediction. These scores can help identify whether an example relies on a spurious correlation by analyzing where the model "focuses" when making its decisions.
For example, in text classification:
Input: "This product is horrible; I do not recommend it."
Spurious correlation: The model focuses only on the word "recommend" to predict a positive sentiment, ignoring the negative context.
Attention scores: If the scores show strong attention on "recommend" but little on "horrible," it indicates that the model is relying on a spurious correlation between "recommend" and positive sentiment.

3.3 LID (Local Intrinsic Dimensionality)

In this section, we primarily address the fact that spurious correlations often arise from irrelevant or noisy features.
To tackle this, we use Local Intrinsic Dimensionality (LID) [Savić et al., 2023].
LID quantifies the perceived dimensionality of neighboring points in a space, essentially measuring the local complexity around a point. Specifically, it assesses how many dimensions are required to explain the variability of data in the vicinity of a given point.
Spurious correlations introduce additional variability into the data, which can increase the local dispersion of points in the representation space. Consequently, the LID may be higher in such cases.

However, certain limitations become apparent. In some cases, spurious correlations can lead to a low LID. For instance, if the model relies on a simple cue or an isolated feature that captures low local dimensionality (e.g., recognizing a camel based on the presence of sand in the background). There is no causal relationship here: a high or low LID alone does not prove that a correlation is spurious; it merely highlights areas where local relationships are either simple or complex.

Given our coarse assumption that spurious correlations typically exhibit a higher LID, we propose combining this score with previous methods, such as forgettable examples and/or attention scores.

4 Experiment

This study focuses exclusively on the BERT model as the pre-trained language model (PLM). Spurious correlations are extracted following the previously described methodology, specifically from the MNLI training set, using the shallow BERT model. We focuses only on BERT as we investigate solution to increase model robustness and compare these methods. We do not aim to find the perfect model. But in future work we could try the best method on few models. Or experiment with other dataset on BERT (eg: QQP)

4.1 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.
In Table 1, we present a comparison between three models: the BERT base model, the BERT model fine-tuned a second time on forgetting events, and the BERT model fine-tuned a second time using important features extracted based on attention scores. For the model fine tune with important example we include variations in the threshold, which selects examples of varying importance.

Model	MNLI	HANS	Threshold
BERT	**84.2**	62.9	/
$BERT + \mathscr{F}_{forget}$	82.9	69.0	/
$BERT + \mathscr{F}_{important}$	82.4	73.3	0.08
$BERT + \mathscr{F}_{important}$	82.7	**73.6**	0.1
$BERT + \mathscr{F}_{important}$	82.5	73.4	0.3

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

The notation "$BERT + \mathscr{F}_{forget}$" refers to an initial fine-tuning of BERT on MNLI, 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	Epoch
BERT	**84.2**	69.5	4
$BERT$	83.63	72.34	1
$+$	83.29	72.66	2
$\mathscr{F}_{important \cup forget}$	83.18	72.73	3
	83.05	72.77	4

Table 2: Accuracy on BERT ; BERT + forgettable examples combine with important samples ; threshold = 1

This table presents the results for the combination of forgetting events and important features. To do so, we consider the union between important examples and forgetting events, thus removing the duplicate examples. Interestingly, this approach does not significantly improve the accuracy on HANS, as the best result remains 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. Because for a comparison, HANS accuracy for the epoch 1 and 2 on $BERT + \mathscr{F}_{important}$ are respectively 69.5 and 70.1. We examine this point in the Discussion part.

Model	MNLI	HANS
$BERT$	**84.2**	62.9
$BERT + \mathscr{F}_{important}$	82.7	**73.6**
$BERT + \mathscr{F}_{important \cup LID}$	82.5	73.4
$BERT + \mathscr{F}_{LID}$	82.4	73.2
$BERT + \mathscr{F}_{important \cup forget}$	83.1	72.8
$BERT + \mathscr{F}_{forget}$	82.9	69.0

Table 3: Comparative table of models accuracies across various extraction methods

This table compares all the studied cases. The rows are sorted in descending order of HANS accuracy. We observe that extraction using attention scores remains the best approach. It also does not add significant computational complexity. Combinations are quite effective, and the use of LID seems to work well, outperforming forgetting events. Although initially skeptical about this method, as it does not clearly highlight spurious correlations but rather noise and harder-to-learn features, it has shown promising results.

4.2 To extend this work

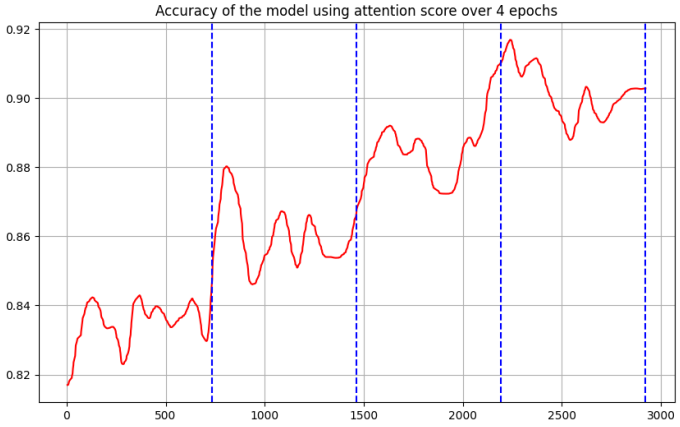
Model	MNLI	HANS
BERT	**84.2**	69.5
$BERT + Forget_{BERT}$	82.9	69.0
$BERT + Forget_{BILSTM}$	82.9	**70.4**
$BERT + Forget_{BoW}$	**84.3**	**70.5**
$BERT + Forget_{HANS}$	83.9	69.5

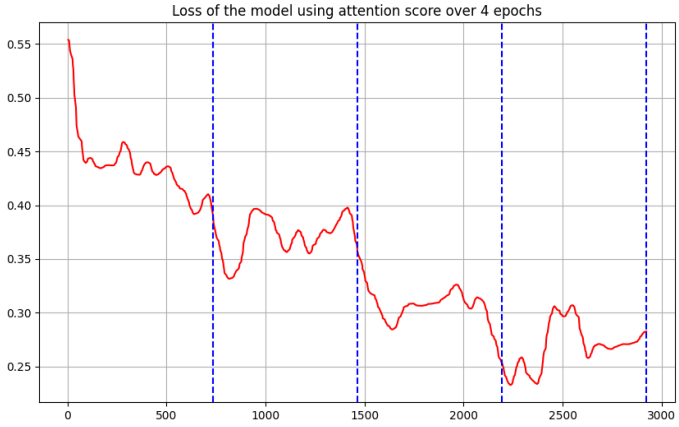
Table 4: Accuracy after 4 epochs on BERT ; BERT + Forgettable examples

In this Table 4, 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.

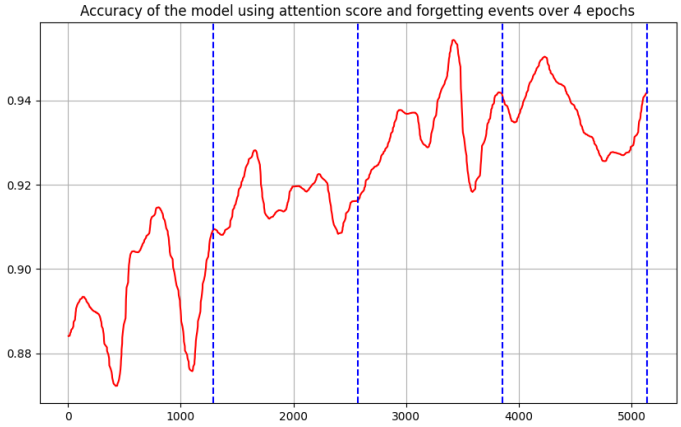
4.3 Discussion

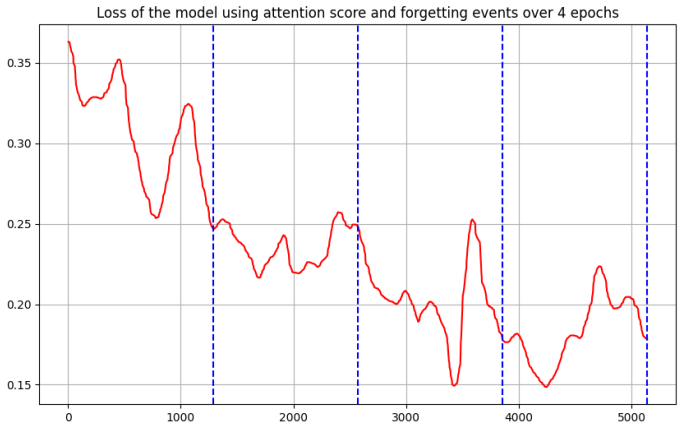
Previously we have seen that accuracy of the model combining forgettable examples and important feature converges faster. We propose to explore metrics as accuracy and loss during the second fine tuning of these two models. Observation have been made during 4 epochs of this fine tuning. (epochs are represented with the vertical line)
Important examples



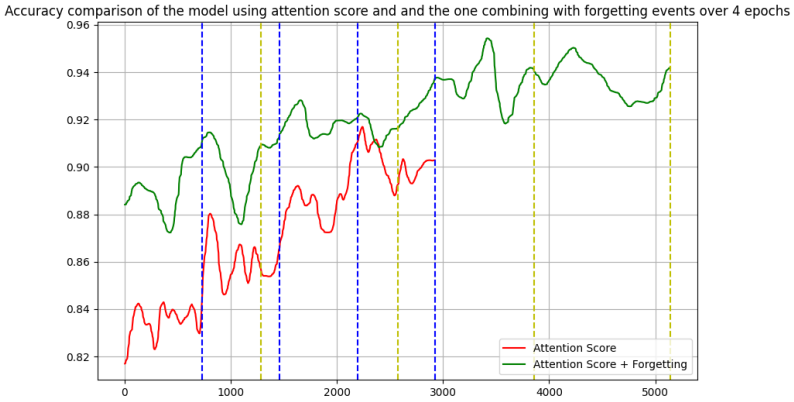


Forgettable examples





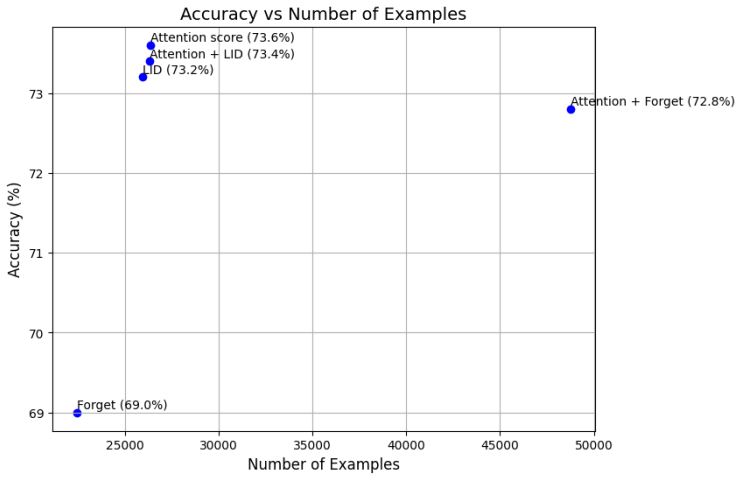
Comparison



In the figure above, we highlight that the number of data, meaning the number of spurious correlations extracted, is not the same.
One potential criterion that could explain the faster convergence of one model compared to another during fine-tuning is the number of spurious correlations extracted. In other words, the more parameters there are, the less the model needs a large number of epochs. The inequality in the convergence of epochs is solely due to the inequality in the number of examples that are re-fine-tuned.
We then propose to compare whether there is a correlation between the number of examples extracted and the accuracy.

extraction method	number of examples	accuracy
Forget	22 452	69.0
Attention score	26 364	73.6
LID	25 946	73.2
Attention + Forget	48 749	72.8
Attention + LID	26 323	73.4

Table 5: Number of examples train in the second fine tuning



We could conclude that there is a correlation between the number of examples and the performance. But a important number of examples extracted does not imply a good accuracy.

5 Conclusion

In this project report, we have detailed our experience and presented our results. The comparison of several methods for extracting spurious correlations revealed that using attention scores is quite effective, significantly improving accuracy on HANS results that had never been achieved before with the BERT model. The combination of these extraction methods could also be a promising solution, as some, like forgettable joins with important features, help maintain accuracy on MNLI.

For future research, it would be interesting to explore a different type of task. Previously, certain limitations of this study had been highlighted, particularly regarding the evaluation on HANS. We may wonder whether the sole use of HANS is sufficient to assert that a model is robust. To be more rigorous, we should test on another task. In the state-of-the-art review conducted earlier, we highlighted another task and its associated dataset to evaluate robustness, namely the QQP task and the PAWS dataset.

References

Miloš Savić, Vladimir Kurbalija, and Miloš Radovanović. Local Intrinsic Dimensionality Measures for Graphs, with Applications to Graph Embeddings, July 2023. URL `http://arxiv.org/abs/2208.11986`. arXiv:2208.11986 [cs].

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. Increasing Robustness to Spurious Correlations using Forgettable Examples. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.291. URL `https://aclanthology.org/2021.eacl-main.291`.