Research Project (3A): Mid-Term Report

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

Author: Ambroise LAROYE–LANGOUËT

Supervisor: Son Vu

son.vu@ensea.fr

January 29, 2025

Contents

1 Introduction **2**
1.1 Subject and methodology . . . 2
1.2 Context and motivation . . . 2
1.3 Definition . . . 3

2 Background **3**

3 Research question **4**

4 State of the art **5**
4.1 Spurious correlations . . . 5
4.2 Identifying spurious correlations in text classification . . . 5
4.3 Robustness to spurious correlation . . . 6
4.3.1 Pre-trained Language Models (PLMs) / Fine Tuning . . . 7
4.3.2 Data augmentation / Human annotation . . . 8
4.3.3 Generating Counterfactual . . . 8
4.3.4 Transfer Learning : Adversarial Learning / specific task (HANS, PAWS) . . . 8
4.3.5 Out of Distribution processing . . . 9
4.3.6 Latent space removal . . . 9
4.3.7 Regularization / overparametrization . . . 9
4.4 Optimizing method (distillation, adapter or pruning) . . . 10

5 Forgettable examples to enhance robustness against spurious correlations in pre-trained models **12**
5.1 Forgettable examples . . . 12
5.2 Tasks / Datasets . . . 13
5.3 PLM . . . 14
5.4 Results . . . 15

6 Conclusion **18**

1 Introduction

1.1 Subject and methodology

In this report, I have divided the discussion into two parts. In the first part, I will present and introduce the topic, define the key terms, and detail the state of the art. This will allow us to determine the direction we will take, identify the most promising avenues, and highlight those that can be further explored. In the second part, I will focus primarily on an experiment that I will have selected based on the state of the art. The objective is to implement a program to address a specific research question.

This project focuses on addressing the vulnerabilities of deep learning models caused by shortcut learning / spurious correlations and adversarial attacks. Shortcut learning refers to models relying on superficial patterns in data, such as background features or syntactic heuristics, which can lead to significant errors when tested on diverse or unexpected data. Adversarial attacks, exploit these vulnerabilities to manipulate model predictions or create deceptive content. The project aims to develop optimization-based learning techniques to mitigate these issues, improving the robustness and security of AI systems.

1.2 Context and motivation

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 rises, necessitating rigorous validation, robust methodologies, and critical thinking. By understanding the underlying causes and implications, researchers and practitioners can minimize the impact of these misleading relationships and ensure the integrity of their conclusions.

Let's consider the relationship from the chart below. Can we really assume that U.S. government subsidies in science truly influence the number of suicides? The answer is obviously no, and yet a model would tell us that the two are linked. Indeed, that relationship is casual; it is just a coincidence. There might be variables that perfectly predict the response, but purely by luck.

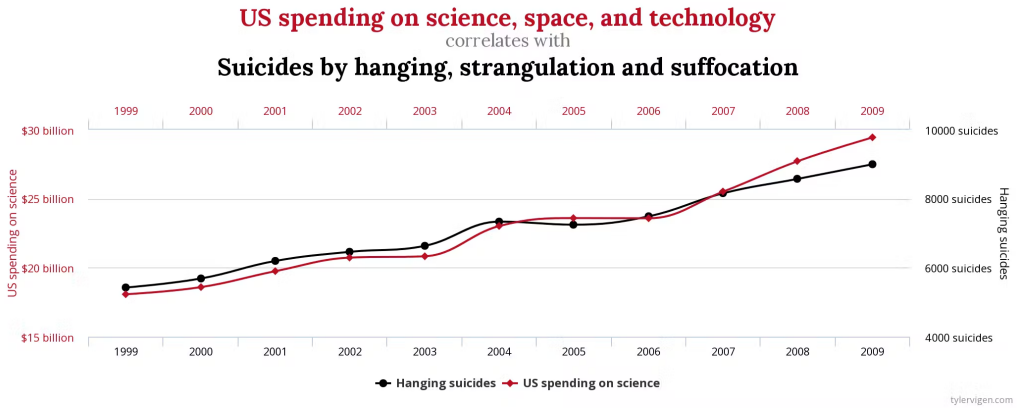

Figure 1: An example of a spurious correlation

In this work, we will focus on spurious correlations applied in Natural Language Processing (NLP). The issue of spurious correlations exists in both NLP and Computer Vision. However, I have primarily limited the scope to NLP for practical reasons and to delve deeper into a specific domain rather than spreading across multiple topics. That said, there is no drastic difference between NLP and Computer Vision in this context. The distinction lies only in the examples and concrete cases chosen within NLP. We will focus on the task of classification.

We first propose a brief review and state-of-the-art solutions to address spurious correlations and explore how we can enhance robustness against them.

1.3 Definition

We propose a short and simple definition of spurious correlation.
In statistics, a spurious correlation (or spuriousness) 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.

2 Background

Various causes of Spurious Correlation arise :

- **Confounding Variables**; A third variable, not accounted for, may influence both variables of interest, creating an illusion of correlation. For instance, ice cream sales and drowning incidents may appear correlated, but both are influenced by temperature.

- **Random Coincidence**; Large datasets or multiple hypothesis testing increase the likelihood of observing correlations purely by chance.
- **Sampling Bias**; Non-representative or biased samples can introduce artificial relationships. For example, analyzing a dataset with limited demographic diversity may reveal patterns that do not generalize to the broader population.
- **Data Transformations and Preprocessing**: Improper handling of data, such as scaling, normalizing, or filtering, can introduce artificial relationships between variables.

The consequences of spurious correlations are significant in both research and practical applications. False correlations can lead to incorrect theories or hypotheses, wasting resources and time. Models trained on datasets with spurious correlations may overfit, resulting in poor generalization to unseen data. Decisions based on spurious relationships can also lead to ineffective or harmful policies.

Despite the remarkable performance of deep learning models, recent studies have highlighted their vulnerability to shortcut learning, also known as spurious correlations, dataset biases, group robustness issues, and simplicity bias. In this work, we focus on spurious correlations, which refer to dependencies between observed features and class labels that only hold for certain groups of training data. In statistics, spurious correlation describes a situation where two variables appear related but are coincidental or influenced by an external variable.

For example, in sentiment classification of movie reviews, the term *Spielberg* may be correlated with positive sentiment because many of his movies receive positive reviews. However, the term itself does not inherently indicate positivity. Similarly, in natural language processing (NLP), models often rely on specific words or syntactic heuristics when predicting the sentiment of a sentence or the relationship between a pair of sentences.

In NLP, spurious correlations are often defined in multiple ways. One conceptual definition, referred to as *ingenuine* [Wang and Culotta, 2020a] [Rogers, 2021], describes a feature correlated with some output label for no apparent reason. Such features often result from the annotation process, known as annotation artifacts [Gururangan et al., 2018]. For instance, Gururangan et al. [2018] showed that the words *cat* and *sleeping* are correlated with contradictions in the SNLI dataset [Bowman et al., 2015].

3 Research question

How can the use of forgettable examples enhance robustness against spurious correlations in pre-trained models?

4 State of the art

4.1 Spurious correlations

Ye et al. [2024], provides a comprehensive taxonomy of methods addressing spurious correlations, categorized into, **Data-Centric Approaches**, Balanced datasets and adversarial data augmentation ; **Model-Centric Approaches**, Regularization and debiasing techniques ; **Evaluation Tools**, Robustness metrics and out-of-distribution benchmarks. This survey offer a broad framework and highlights the need for scalable, domain-agnostic strategies to enhance the robustness of machine learning systems. It emphasizes applications in computer vision, NLP, and healthcare, and highlights challenges such as model brittleness under domain shifts.
Du et al. [2023b] focusing on NLP, this work identifies "biased words" that disproportionately influence predictions. The proposed **Less Learn-Shortcut (LLS)** strategy mitigates over-reliance on such biases by downweighting biased examples during training. Experiments on tasks like NLI and sentiment analysis show improved adversarial robustness without sacrificing in-domain performance. LLS is task-agnostic and transferable across NLP tasks. Robust evaluation frameworks are essential for benchmarking progress in mitigating spurious correlations.
These results and papers lead us to several reflections. First, we must ask ourselves, before starting, how to identify spurious correlations. Then, we will focus on methods to reduce shortcuts. Ye et al. [2024] proposes numerous methods applied to computer vision and NLP, while we will focus more on text classification. ... specifically presents several methods to address spurious correlations, including data manipulation, representation learning, and learning strategies.

4.2 Identifying spurious correlations in text classification

Wang and Culotta [2020a] paper proposes a method to distinguish between spurious and genuine correlations in text classification tasks. The approach treats this as a supervised classification problem, using features derived from treatment effect estimators to isolate the influence of individual words on the class label, while controlling for the context in which they appear. Key contributions of this paper include, Supervised Classification for Spurious Correlation Detection, by leveraging treatment effect estimation techniques, the authors propose a classifier that can identify words contributing to spurious correlations, even with a small number of labeled examples. The classifier is robust to domain shifts, as the features used are generic and can be transferred across domains. Howeover, this work need the human action, to annotate data, which made it particularly inconvenient because we need to know the spurious correlations in advance, which is not always the case, on the contrary, as in the case of backdoors, for example.
Wang et al. [2022] presents a framework for identifying spurious correlations in NLP models automatically and at scale. It emphasizes the need for proactive detection of spurious

correlations in models to improve their robustness across multiple domains. The main innovations in this paper are the **automatic identification** of Spurious Correlations: The framework utilizes interpretability methods such as attention scores and integrated gradients to extract tokens that influence the model's decision-making. It then distinguishes "genuine" from "spurious" tokens through cross-dataset analysis and knowledge-aware perturbation. Genuine tokens are those that consistently influence model predictions across different domains, while spurious tokens tend to have domain-specific or unstable influences. The framework then tests the stability of model predictions by perturbing the extracted tokens with semantically similar alternatives. If a model's prediction changes significantly when a token is replaced with a similar one, this suggests that the token was a spurious correlation rather than a genuine feature of the task.

4.3 Robustness to spurious correlation

We have several possible solutions: adding information to avoid shortcut learning, removing problematic data, or using PLMs. The diagram below illustrates what we will develop in the following section with our different cases.

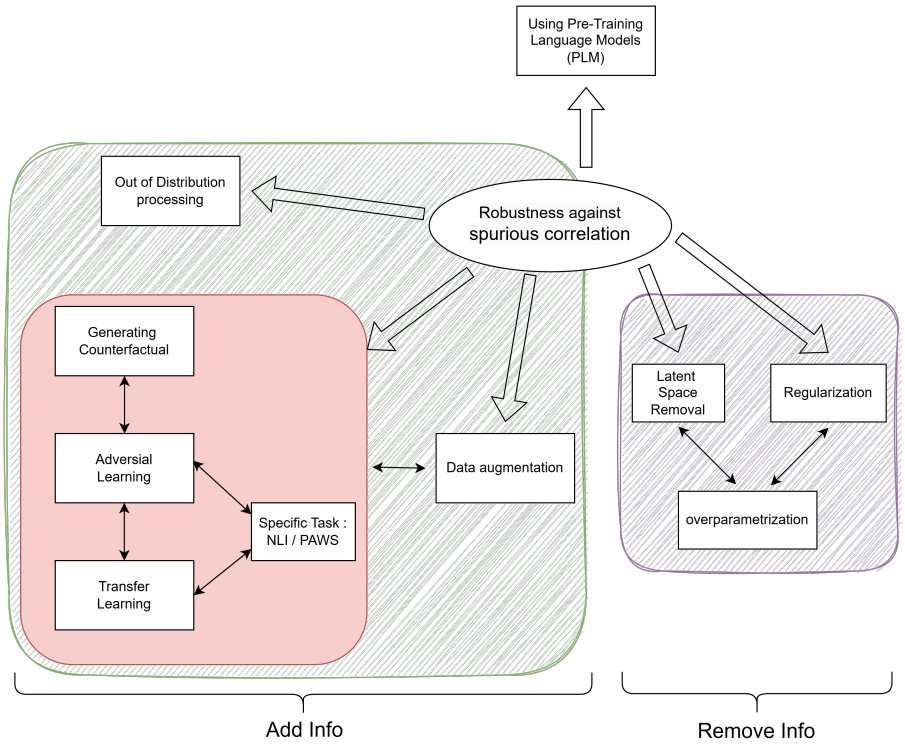

Figure 2: Robustness to spurious correlation

4.3.1 Pre-trained Language Models (PLMs) / Fine Tuning

One of solution that arise is using **Pre-trained Language Models** to increase robustness to spurious correlation. Recent work has shown that pre-trained language models improve robustness to spurious correlation in the dataset [Tu et al., 2020], [Hendrycks and Gimpel, 2018]. And specially Hendrycks and Gimpel [2018] show that although pre-training may not improve performance on traditional classification metrics, it improves model robustness and uncertainty estimates.
In a series of ablation studies, Huh et al. [2016] show that the benefits of pre-training are robust to significant variation in the dataset used for pre-training, including the removal of classes related to the target task.
Others study argue that pre-training does reduce overfitting when also measuring model calibration. Showing that pre-training can improve model robustness to label corruption, class imbalance, and adversarial attacks. [noa] [Sukhbaatar et al., 2015] [Szegedy et al., 2014].
Pre-trained models like BERT are more robust to spurious correlations because they can generalize from a minority of training examples that counter the spurious pattern [Tu et al., 2020]. Tu et al. [2020] propose multi-task learning (MTL) as a solution for extreme minority cases, demonstrating that MTL with carefully chosen auxiliary tasks improves performance on challenging datasets without compromising in-distribution accuracy. The study emphasizes the importance of data diversity and shows that pre-training and MTL both enhance generalization from minority examples, aiding out-of-distribution robustness.

Schwartz and Stanovsky [2022] proposed alternative ways for mitigating spurious correlations, including adding richer contexts to textual corpora, and allowing models to abstain or interact in cases of uncertainty. And then concluded by suggesting to reconsider the practice of **fine-tuning** pretrained models on large-scale training sets.
Yu et al. [2024] how fine-tuned pre-trained language models aid generalizability from single-domain scenarios under mild assumptions, targeting more general and practical real-world scenarios we introduced a method for constructing robust causal representations leveraging PLMs. Through a series of semi-synthetic and real-world experiments, we demonstrated the promising performance of our approach in OOD scenarios compared to standard PLM fine-tuning. The proposed method introduces a causal front-door adjustment to derive robust representations, leveraging PLMs for data augmentation under single-domain settings. Through synthetic and real-world experiments, the approach demonstrates improved robustness and generalizability compared to existing techniques, advancing the understanding of domain generalization by linking fine-tuning with causal mechanisms in representation learning.

4.3.2 Data augmentation / Human annotation

Izmailov et al. [2022] conclude that a strong regularization is not necessary for robustness to spurious correlations, but **appropriate data augmentation** and weight decay can provide a small improvement. Min et al. [2020] highlights the potential of targeted data augmentation to address syntactic weaknesses in pre-trained models. This method reach best result on subject/object inversion, increasing BERT's accuracy from 0.28 to 0.73. Some studies propose complementary dataset, like using AND-rules with negation [Yadav et al., 2022].
Other studies address the issue of data augmentation by adding **human annotation** in the dataset [Srivastava et al., 2020]. Making models robust to spurious correlation by leveraging humans' common sense knowledge of causality.

4.3.3 Generating Counterfactual

An other method to fight against spurious correlation is by generating counterfactuals. Counterfactuals are hypothetical scenarios that help assess the causal relationship between variables by imagining alternative outcomes had certain conditions been different. This approach allows us to isolate the true causal effects by testing how changes in specific variables impact the outcome while holding others constant. These methods require collection of counterfactual labeled data that can be used to regularize a classifier. [Bansal and Sharma, 2023][Kaushik et al., 2020][Veitch et al., 2021] [Wang and Culotta, 2020b]

4.3.4 Transfer Learning : Adversarial Learning / specific task (HANS, PAWS)

Transfer learning [Pan and Yang, 2010] in the context of robustness to spurious correlations refers to leveraging pre-trained models or knowledge learned from one task to improve performance on a target task, while addressing the challenge of spurious correlations. . It facilitates the learning of robust and generalizable representations for the target conditional probability distribution. Techniques like Bayesian optimization can assist in fine-tuning hyperparameters to improve transfer learning performance, though data selection is typically handled through other methods [Ruder and Plank, 2017].
Studies shows that transfer learning leads to better robustness without affecting LLM's performances [noa, 2024]

Adversarial learning involves training models to be robust against adversarial examples—inputs deliberately crafted to fool the model. Thus improving model robustness and generalization by exposing the model to challenging inputs.
The HANS (Heuristic Analysis for NLI Systems) dataset [McCoy et al., 2019] is designed to test models' reliance on shallow heuristics in natural language inference (NLI) tasks. A model relying exclusively on the word-overlap feature would not have a higher than chance

classification accuracy on HANS. Training models on HANS forces models to confront examples that exploit their heuristic biases, thus encouraging the development of deeper, more robust language understanding.
Papers introduce large-scale NLI benchmark dataset, collected via an iterative, adversarial human-and-model-in-the-loop procedure. And show that training models on this new dataset leads to better performance on popular NLI benchmarks, while raising a more difficult challenge with its new test set. [Nie et al., 2020] [Zhou and Bansal, 2020]

4.3.5 Out of Distribution processing

Out-of-distribution (OOD) detection is the task of identifying inputs that fall outside the training distribution. Out-of-distribution detection improves robustness to spurious correlations by enables a system to recognize when an input does not belong to the data distribution it was trained on, thereby avoiding unreliable predictions in unfamiliar contexts [Zhang and Ranganath, 2023]. OOD detection mitigates the risks of over-reliance on spurious correlations by helping systems identify when their learned patterns are not applicable, thereby enhancing robustness and reliability. [Lee et al., 2022]

4.3.6 Latent space removal

Prior work has introduced latent adversarial training (LAT) as a way to improve robustness. Sheshadri et al. [2024] reach out better result by using latent adversarial training (LAT) to make LLMs more robust to exhibiting persistent unwanted behaviors. In contrast to adversarial training with perturbations to the model's inputs, they train the model with perturbations to its hidden latent representations.

4.3.7 Regularization / overparametrization

In this part we investigate the the contrary of what we previously looking for. Indeed, previously the goal was to identify spurious correlations in order to eliminate them. However, now we will see that, contrary to what we might have thought, in certain cases, removing spurious correlations can decrease accuracy. Khani and Liang [2020] show that removal of spurious feature can decrease the accuracy even in balanced dataset where spurious features co-occur equally with all targets. This is due to the inductive biases of overparameterized models, which favor fitting spurious features to minimize training loss. Thus removing spurious features can make models more susceptible to other spurious features, further complicating robustness efforts. This work propose an alternative approach, robust self-training. Common mitigation strategies, like distributionally robust optimization (DRO) or upweighting minority groups, can improve worst-group error in strongly regularized models but fail in overparameterized models capable of achieving zero training error [Sagawa et al., 2020b].
Sagawa et al. [2020a] explores the application of distributionally robust optimization (DRO)

to overparameterized neural networks, which often fail on atypical data groups due to reliance on spurious correlations. The study highlights the critical role of regularization in achieving robust generalization for worst-case groups, even when unnecessary for average performance.

4.4 Optimizing method (distillation, adapter or pruning)

This part is based on my intuition, as there isn't much research on the subject.

During my second-year internship, I had the opportunity to work on optimization methods such as pruning, distillation, and replacing fine-tuning with adapters based tuning in NLP. Specifically, I also wrote an article. One of my questions has therefore been toward robust pruning. Can pruning make a model robust, or does it have the opposite effect? [Du et al., 2023a] and [Li et al., 2024]. Li et al. [2024] suggest comparing the robustness of a model based on pruning. Results demonstrate that their pruning strategy significantly enhances the robustness of language models, even under sparse architectures. This improvement is attributed to the regularization effects of sparsity. Furthermore, the method achieves superior robustness with fewer parameters, emphasizing its efficiency and effectiveness. While accuracy may be slightly lower at lower sparsity levels, the substantial gains in robustness more than compensate for this trade-off.
However, current research does not entirely address the same problem, as it mainly focuses on determining whether pruning can help maintain a model's robustness. In contrary, we aim to explore whether carefully selected pruning can improve a model's robustness.
Why do I think this? Because we've previously seen that there are methods to add information to a model, such as data augmentation, generating counterfactuals, and transfer learning. Similarly, there are methods to remove weights or connections from a model that act as shortcuts.
During pruning, the goal is precisely to remove weights deemed unnecessary, which only add useless complexity to the model. Therefore, it is reasonable to hypothesize that by applying pruning and combining it with adapter-based tuning (since this combination maintains the model's performance, unlike pruning alone), we could enhance robustness.

Structured pruning is widely used to improve performance and reduce the size of models in NLP, enabling compression that allows models to run on edge devices, such as mobile phones. The idea is to remove most of the weight that do not interfere with the performance of the model.

Adapters-based tuning has emerged as an alternative to fine-tuning Houlsby et al. [2019]. Adapters are new modules inserted between the layers of a pre-trained network. During training on a downstream task, only the adapter parameters are updated, adding just a few trainable parameters per new task and allowing for a high degree of parameter

sharing. Adapter-based tuning requires significantly less computational resources compared to fine-tuning.

Moreover, the issue of pruning and distillation becomes necessary as model sizes increase, requiring more energy resources, becoming more expensive, and increasing inference time. Energy costs are particularly significant, as a request on large pre-trained models (PLMs) like GPT-4 can be 10 times more expensive than making a reserch on Google.

5 Forgettable examples to enhance robustness against spurious correlations in pre-trained models

In the following section, I will delve into *Increasing Robustness to Spurious Correlations using Forgettable Examples* [Yaghoobzadeh et al., 2021]. I will provide more details about the experiment and the article.
But first, I would like to justify the choice of this article. Indeed, in the previous state-of-the-art review, we explored numerous methods to increase robustness against shortcut learning, out-of-distribution data, etc. A potential idea would be to combine several methods. For instance, it could be possible to combine the use of PLMs with adversarial learning on a specific task such as HANS. We observed that these are among the most reliable and straightforward methods to use.

In the work of Yaghoobzadeh et al. [2021], the principle is similar, if not better. Why? Because the forgettable example method allows for the identification of spurious correlations. The idea is to isolate these elements and then train them on specific tasks. This approach involves two fine-tuning steps: one is global, applied to the entire pre-trained model, and the other is focused solely on the minority elements. In doing so, we significantly contribute to the robustness of our model, at least theoretically. Now, let us see if the experiment confirms this.

5.1 Forgettable examples

Yaghoobzadeh et al. [2021] introduces a systematic, data-driven approach to identify and leverage "forgettable examples" to address these challenges and improve model robustness.

The notion of example forgetting, introduced by Toneva et al. [2019], forms the foundation of this approach. Forgettable examples are defined as instances that are correctly classified at some point and later misclassified, reflecting a "forgetting event." Or never correctly classified throughout training.

These examples often highlight minority cases or hard-to-learn patterns, making them essential for understanding spurious correlations in datasets. Deep models, due to their limited memorization capacity, have been shown to be particularly effective for identifying forgettable examples, as they focus less on memorizing noise and more on learning core patterns.

This work propose a method to identify minority examples through forgettable analysis, which does not require prior knowledge of the dataset's spurious correlations. The identified examples are then used to fine-tune the model, reducing its reliance on spurious correlations and enhancing its generalization capabilities. The proposed method employs

a two-step fine-tuning strategy to improve robustness:

1. Initial Fine-Tuning: The model is first trained on the full dataset to establish baseline performance.
2. Second Fine-Tuning: The model is subsequently fine-tuned on the subset of minority examples identified through forgettable analysis.

This dual fine-tuning approach improves the model's out-of-distribution performance while maintaining strong in-distribution accuracy.

Forgettable examples contain a higher proportion of minority cases compared to random subsets of equivalent size. However, training exclusively on forgettables leads to poor overall performance, emphasizing the need for balanced training. Identifying forgettables using a shallower network than the primary model enhances effectiveness, as shallow networks are less prone to overfitting and memorization.

Larger pre-trained language models (PLMs) exhibit greater robustness to spurious correlations. Fine-tuning these models on forgettable examples further enhances their out-of-distribution performance.
The techniques "Reg-conf" Utama et al. [2020], which aim to improve OOD robustness without introducing additional hyperparameters, are complementary to the forgettable example approach, which directly identifies challenging examples for targeted fine-tuning.

This work makes the key contributions by introducing a novel method to identify minority examples without prior knowledge of spurious correlations using example forgetting. Proposes a two-step fine-tuning strategy that enhances model robustness to spurious correlations.
By leveraging forgettable examples, this approach systematically addresses the limitations of existing methods, offering a scalable and generalizable solution to improve robustness in pre-trained language models.

5.2 Tasks / Datasets

In this project we aim to evaluate robustness and performance in out of distribution data. We thus need to find specific tasks to evaluate model's performances.

MNLI [Williams et al., 2018], consists of sentence pairs annotated with labels indicating their relationship in terms of textual entailment: these labels include entailment, contradiction, or neutral.

- Entailment: The second sentence logically follows from the first.
- Contradiction: The second sentence contradicts the first.
- Neutral: The second sentence is neither entailed nor contradicted by the first, meaning the relationship is neither directly supported nor directly refuted.

With a corresponding dataset : **HANS** [McCoy et al., 2019] 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"). A model relying exclusively on the word-overlap feature would not have a higher than chance classification accuracy on HANS.

QQP [Jia and Liang, 2017]. The goal is to determine whether two questions are semantically equivalent. Given a pair of questions, the task involves binary classification: identifying if the questions are duplicates (at least have the same meaning) or not.
With the dataset : **PAWS** [Zhang et al., 2019]. Paraphrase Adversaries from Word Scrambling dataset is designed to evaluate models for paraphrase identification. It contains pairs of sentences that are highly similar in word order but differ in meaning. PAWS is particularly challenging because it tests a model's ability to go beyond surface-level similarities and understand deeper semantic differences.

The task of fact verification [Schuster et al., 2019]. It aims to verify a claim given an evidence. The labels are classify between Supported or Refuted or Not Enough Information.
With the dataset : **FEVER** [Thorne et al., 2018]. Fact Extraction and Verification dataset It consists of claims paired with evidence from Wikipedia, and the task is to classify each claim as Supported, Refuted, or Not Enough Information based on the evidence.

And the evaluation metrics used is the accuracy.

5.3 PLM

We are interested in the robustness of large PLMs. In this work, we focus on **BERT** model [Vaswani et al., 2023] (more specifically BERT base Large). 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. We used the pre-trained **google-bert/bert-base-cased** model with 12 layers and 12 attention heads.
The encoder consists of N layers, each containning a multi-head self-attention (MHA) layer and a feed-forward (FFN) layer.

An MHA are N_h heads that takes an input $x \in \mathbb{R}^d$ and ouput :

$$MHA(X) = \sum_{h=1}^{N_h} \text{Att}(W_Q^{(h,l)}, W_K^{(h)}, W_V^{(h,l)}, W_O^{(h,l)}, X)$$

The attention head h in layer l is parametrized by the matrices $W_Q^{(h,l)}, W_K^{(h,l)}, W_V^{(h,l)} \in \mathbb{R}^{d_h \times d}$ and $W_O^{(i)} \in \mathbb{R}^{d \times d_h}$ and where :

$$Att(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

And the feed-forward layer is composed of 2 projection layers, up and down, parameterized by $W_U \in \mathbb{R}^{d \times d_f}$ and $W_D \in \mathbb{R}^{d_f \times d}$:

$$FFN(X) = ReLU(XW_U)W_D$$

5.4 Results

The aim is to increase the accuracy on HANS dataset, which means that the model is more robust.

Result on BERT

	mnli.acc	hans.acc
0	0.3193071828833418	0.5
1	0.8202750891492613	0.5039
2	0.8405501782985226	0.5756333333333333
3	0.8398369842078451	0.6099
4	0.8410596026490066	0.6271666666666667

Figure 3: Accuracy after BERT base first fine tuning for each epoch

	mnli.acc	hans.acc
0	0.8409577177789098	0.6272
1	0.8353540499235863	0.6758666666666666
2	0.8308711156393276	0.6913
3	0.8302598064187469	0.6897333333333333
4	0.8293428425878757	0.6901333333333334

Figure 4: Accuracy after BERT base second fine tuning on forgettable example for each epoch

Accuracy on HANS increase, which means that our algorithm is working

Model	**MNLI**	**HANS**	**Avg.**
BERT	**84.4**$_{\pm 0.1}$	62.9$_{\pm 1.5}$	73.7$_{\pm 0.8}$
BERT+ $\mathcal{F}_{\text{BERT}}$	83.0$_{\pm 0.4}$	68.9$_{\pm 1.4}$	75.9$_{\pm 0.7}$
BERT+ $\mathcal{F}_{\text{BiLSTM}}$	82.9$_{\pm 0.4}$	70.4$_{\pm 0.9}$	76.7$_{\pm 0.5}$
BERT+ $\mathcal{F}_{\text{BoW}}$	83.1$_{\pm 0.3}$	**70.5**$_{\pm 0.7}$	**76.8**$_{\pm 0.4}$
BERT + Rand$_{63,390}$	84.3	63.6	73.9
BERT+ $\mathcal{F}_{\text{HANS}}$	83.9$_{\pm 0.4}$	69.5$_{\pm 0.9}$	76.7$_{\pm 0.5}$

Figure 5: Accuracy result on comparativ shallow model : BERT, BoW, BiLSTM

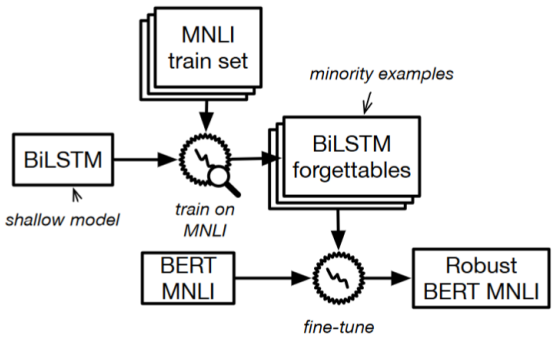

Figure 6: Diagram on an example of the algorithm, BiLSTM

The results confirm that tuning the model towards minority examples improves robustness with a slight drop in MNLI accuracy. The best model is obtained by fine-tuning on BoW. Fine-tuning on BiLSTM is comparable to finetuning on BoW, which demonstrates that both BoW and BiLSTM models learn similar spurious correlations.

6 Conclusion

In this first project report, we conducted a comprehensive review of the state of the art regarding various identification methods and strategies to manage spurious correlations. We concluded this review by recommending the use of multiple complementary approaches to enhance robustness against spurious correlations. Specifically, we highlighted the potential of leveraging pre-trained language models (PLMs), incorporating forgettable examples, and employing fine-tuning or transfer learning techniques.

In the second part of this report, we focused on implementing this combination of methods to evaluate its effectiveness. The experimental framework for this implementation was inspired by the work of Yaghoobzadeh et al. [2021].

For future research, it would be valuable to delve deeper into the experimental setup by varying key parameters, adjusting evaluation metrics, and exploring the impact of different proportions of forgettable examples. This would provide further insights into optimizing the proposed approach and broadening its applicability.

References

[2004.07780] Shortcut Learning in Deep Neural Networks. URL `https://arxiv.org/abs/2004.07780`.

(PDF) On Adversarial Robustness of Language Models in Transfer Learning. In *ResearchGate*, December 2024. URL `https://www.researchgate.net/publication/387130633_On_Adversarial_Robustness_of_Language_Models_in_Transfer_Learning`.

Parikshit Bansal and Amit Sharma. Controlling Learned Effects to Reduce Spurious Correlations in Text Classifiers, June 2023. URL `http://arxiv.org/abs/2305.16863`. arXiv:2305.16863 [cs].

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Llu\'ı s Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075/`.

Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and Ahmed Hassan Awadallah. Robustness Challenges in Model Distillation and Pruning for Natural

Language Understanding, February 2023a. URL http://arxiv.org/abs/2110.08419. arXiv:2110.08419 [cs].

Yanrui Du, Jing Yan, Yan Chen, Jing Liu, Sendong Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Bing Qin. Less Learn Shortcut: Analyzing and Mitigating Learning of Spurious Feature-Label Correlation, June 2023b. URL http://arxiv.org/abs/2205.12593. arXiv:2205.12593 [cs].

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Dan Hendrycks and Kevin Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, October 2018. URL http://arxiv.org/abs/1610.02136. arXiv:1610.02136.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP, June 2019. URL http://arxiv.org/abs/1902.00751. arXiv:1902.00751 [cs, stat].

Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes ImageNet good for transfer learning?, December 2016. URL http://arxiv.org/abs/1608.08614. arXiv:1608.08614 [cs].

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On Feature Learning in the Presence of Spurious Correlations, October 2022. URL http://arxiv.org/abs/2210.11369. arXiv:2210.11369 [cs].

Robin Jia and Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1215. URL https://aclanthology.org/D17-1215/.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data, February 2020. URL http://arxiv.org/abs/1909.12434. arXiv:1909.12434 [cs].

Fereshte Khani and Percy Liang. Removing Spurious Features can Hurt Accuracy and Affect Groups Disproportionately, December 2020. URL http://arxiv.org/abs/2012.04104. arXiv:2012.04104 [cs].

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and Disambiguate: Out-of-Distribution Robustness via Disagreement. September 2022. URL https://openreview.net/forum?id=RVTOp3MwT3n.

Jianwei Li, Qi Lei, Wei Cheng, and Dongkuan Xu. Towards Robust Pruning: An Adaptive Knowledge-Retention Pruning Strategy for Language Models, January 2024. URL http://arxiv.org/abs/2310.13191. arXiv:2310.13191 [cs].

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, June 2019. URL http://arxiv.org/abs/1902.01007. arXiv:1902.01007 [cs].

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.212. URL https://aclanthology.org/2020.acl-main.212/.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding, May 2020. URL http://arxiv.org/abs/1910.14599. arXiv:1910.14599 [cs].

Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010. ISSN 1558-2191. doi: 10.1109/TKDE.2009.191. URL https://ieeexplore.ieee.org/document/5288526. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Anna Rogers. Changing the World by Changing the Data. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.170. URL https://aclanthology.org/2021.acl-long.170/.

Sebastian Ruder and Barbara Plank. Learning to select data for transfer learning with Bayesian Optimization, July 2017. URL http://arxiv.org/abs/1707.05246. arXiv:1707.05246 [cs].

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020a. URL http://arxiv.org/abs/1911.08731. arXiv:1911.08731 [cs].

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations, August 2020b. URL http://arxiv.org/abs/2005.04345. arXiv:2005.04345 [cs].

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards Debiasing Fact Verification Models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1341. URL https://aclanthology.org/D19-1341/.

Roy Schwartz and Gabriel Stanovsky. On the Limitations of Dataset Balancing: The Lost Battle Against Spurious Correlations, April 2022. URL http://arxiv.org/abs/2204.12708. arXiv:2204.12708 [cs].

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs, August 2024. URL http://arxiv.org/abs/2407.15549. arXiv:2407.15549 [cs].

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to Spurious Correlations via Human Annotations, August 2020. URL http://arxiv.org/abs/2007.06661. arXiv:2007.06661 [cs].

Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training Convolutional Networks with Noisy Labels, April 2015. URL http://arxiv.org/abs/1406.2080. arXiv:1406.2080 [cs].

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. URL http://arxiv.org/abs/1312.6199. arXiv:1312.6199 [cs].

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June

2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning, November 2019. URL http://arxiv.org/abs/1812.05159. arXiv:1812.05159 [cs].

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *Transactions of the Association for Computational Linguistics*, 8:621–633, 2020. doi: 10.1162/tacl_a_00335. URL https://aclanthology.org/2020.tacl-1.40/. Place: Cambridge, MA Publisher: MIT Press.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards Debiasing NLU Models from Unknown Biases. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.613. URL https://aclanthology.org/2020.emnlp-main.613.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests, November 2021. URL http://arxiv.org/abs/2106.00545. arXiv:2106.00545 [cs].

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models, May 2022. URL http://arxiv.org/abs/2110.07736. arXiv:2110.07736 [cs].

Zhao Wang and Aron Culotta. Identifying Spurious Correlations for Robust Text Classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3431–3440, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.308. URL https://aclanthology.org/2020.findings-emnlp.308/.

Zhao Wang and Aron Culotta. Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals, December 2020b. URL http://arxiv.org/abs/2012.10040. arXiv:2012.10040 [cs].

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, February 2018. URL http://arxiv.org/abs/1704.05426. arXiv:1704.05426 [cs].

Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. Robust Interpretable Text Classification against Spurious Correlations Using AND-rules with Negation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 4439–4446, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3. doi: 10.24963/ijcai.2022/616. URL https://www.ijcai.org/proceedings/2022/616.

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.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. Spurious Correlations in Machine Learning: A Survey, May 2024. URL http://arxiv.org/abs/2402.12715. arXiv:2402.12715 [cs].

Jialin Yu, Yuxiang Zhou, Yulan He, Nevin L. Zhang, and Ricardo Silva. Fine-Tuning Pretrained Language Models for Robust Causal Representation Learning, October 2024. URL http://arxiv.org/abs/2410.14375. arXiv:2410.14375 [cs].

Lily H. Zhang and Rajesh Ranganath. Robustness to Spurious Correlations Improves Semantic Out-of-Distribution Detection, February 2023. URL http://arxiv.org/abs/2302.04132. arXiv:2302.04132 [cs].

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase Adversaries from Word Scrambling. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131/.

Xiang Zhou and Mohit Bansal. Towards Robustifying NLI Models Against Lexical Dataset Biases, May 2020. URL http://arxiv.org/abs/2005.04732. arXiv:2005.04732 [cs].