Addressing Spurious Correlations in Machine Learning Models: A Comprehensive Review

Mashrin Srivastava*

mashrins@gmail.com

April 21, 2023

Abstract

Spurious correlations present a significant challenge in the deployment of machine learning models, as they can lead to models relying on irrelevant or unnatural features. This paper provides a comprehensive review of the current state of research on spurious correlations, covering the detection, understanding, and mitigation of these undesirable behaviors. We first discuss the prevalence of spurious correlations in various machine learning applications and their potential consequences. We then review existing methods for detecting and understanding spurious correlations, including adversarial training, representation learning, and interpretability techniques. Finally, we explore recent advancements in addressing spurious correlations, focusing on invariance and stability. The objective of this review is to facilitate further research on this critical topic and improve the robustness and generalizability of machine learning models.

1 Introduction

Spurious correlations refer to the situation where a machine learning model learns to rely on features that are not causally related to the target variable but are strongly correlated with it in the training data [1]. This reliance on irrelevant or unnatural features can lead to catastrophic failures when deploying machine learning models in real-world applications. The issue arises in various domains, including medical imaging [2], natural language processing [3], and computer vision [4].

The goal of this paper is to provide an extensive review of the current research on spurious correlations, focusing on their detection, understanding, and mitigation. We aim to inform researchers and practitioners of the challenges and potential solutions to address spurious correlations, ultimately improving the robustness and generalizability of machine learning models.

*The paper was generated by ChatGPT. Prompts are mentioned in the Appendix. This paper is part of a larger independent research initiative.

2 Prevalence and Consequences of Spurious Correlations

Spurious correlations are widespread in machine learning applications, often due to biases present in the training data. Many studies have documented the presence of spurious correlations in various domains, such as:

- Medical imaging: Models for detecting diseases from X-rays can rely on scanner types or marks made by technicians instead of the actual medical condition [2].
- Natural language processing: In visual question answering, models may be sensitive to linguistic variations in the questions rather than the visual content [3].
- Computer vision: Object recognition models can exploit image backgrounds or lighting conditions rather than the object's features [4].

The consequences of spurious correlations can be severe, leading to poor generalization and even catastrophic failures when models are deployed in real-world settings. In some cases, this can result in significant financial losses or even life-threatening situations [5].

3 Detecting and Understanding Spurious Correlations

Several methods have been proposed to detect and understand spurious correlations in machine learning models. Some of the most popular techniques include:

3.1 Adversarial Training

Adversarial training aims to make models robust to adversarial examples by generating and incorporating them into the training process [6]. By encouraging the model to be invariant to adversarial perturbations, it can help identify and eliminate spurious correlations.

3.2 Representation Learning

Representation learning seeks to learn meaningful and disentangled representations of the input data, which can help separate relevant features from spurious ones [7]. Techniques such as autoencoders and variational autoencoders have been used to learn representations that are less prone to spurious correlations.

3.3 Interpretability Techniques

Interpretability techniques can help understand the inner workings of machine learning models and identify potential spurious correlations [8]. Methods such as feature attribution, visualization, and counterfactual explanations can provide insights into which features the model is relying on and whether they are relevant or spurious.

4 Addressing Spurious Correlations

Recent advancements in addressing spurious correlations have focused on invariance and stability:

4.1 Invariance

Invariance-based methods aim to make models invariant to spurious features by explicitly modeling the causal relationships between variables [9]. This can be achieved by learning invariant representations, designing invariant loss functions, or imposing invariance constraints during training.

4.2 Stability

Stability-based methods focus on making models robust to small perturbations in the input data or training process, which can help mitigate the effects of spurious correlations [10]. Techniques such as dropout, weight decay, and data augmentation can enhance the stability of machine learning models.

5 Domain-Specific Approaches

In addition to general methods for addressing spurious correlations, domain-specific approaches have been proposed to tackle challenges unique to certain applications.

5.1 Medical Imaging

In medical imaging, domain adaptation techniques have been employed to reduce the impact of spurious correlations arising from differences in imaging devices or patient populations [11]. These methods aim to align the feature distributions between the source and target domains, allowing models to generalize better across different settings.

5.2 Natural Language Processing

In natural language processing, adversarial examples and paraphrasing have been used to mitigate spurious correlations arising from biases in the text data

[12]. By generating and training on perturbed or paraphrased versions of the input sentences, models can learn to focus on the relevant content rather than relying on spurious linguistic patterns.

5.3 Computer Vision

In computer vision, techniques such as domain randomization and style transfer have been proposed to address spurious correlations caused by biases in the visual data [13]. By augmenting the training data with images that have diverse backgrounds, lighting conditions, and textures, models can learn to recognize objects based on their intrinsic features rather than spurious contextual cues.

6 Challenges and Future Directions

While significant progress has been made in understanding and addressing spurious correlations, several challenges remain:

- **Detection:** Detecting spurious correlations remains a difficult task, especially when the ground truth causal relationships between variables are unknown. Developing more effective methods for identifying spurious correlations is an important research direction.
- **Interpretability:** Understanding the reasons behind spurious correlations can provide valuable insights for model improvement. Developing more interpretable machine learning models and better diagnostic tools for identifying spurious correlations is crucial.
- **Generalization:** Ensuring that models generalize well to new and diverse settings is essential for avoiding the negative consequences of spurious correlations. Further research on domain adaptation, transfer learning, and other generalization techniques is needed.

7 Conclusion

Spurious correlations pose a significant challenge in the deployment of machine learning models, as they can lead to reliance on irrelevant or unnatural features and poor generalization. This paper has provided a comprehensive review of the current state of research on spurious correlations, covering their prevalence, detection, understanding, and mitigation. By bringing attention to this critical issue and presenting potential solutions, we hope to facilitate further research and improve the robustness and generalizability of machine learning models.

References

[1] Gretton, A., & Györfi, L. Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, 11(Jun), 1391–1423, 2010.

[2] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11), e1002683, 2018.

[3] Agrawal, A., Batra, D., Parikh, D., & Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

[4] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[5] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] Goodfellow, I. J., Shlens, J., & Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[7] Bengio, Y., Courville, A., & Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828, 2013.

[8] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115, 2020.

[9] Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[10] Bousquet, O., & Elisseeff, A. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar), 499–526, 2002.

[11] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., ... & Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59), 1–35, 2016.

[12] Jia, R., & Liang, P. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, 2017.

[13] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017.

A Appendix

A.1 Model details

API Used: GPT-4 with 8K context window
Temperature: 0
top_p: 1

A.2 Prompt

The prompts used to draft this paper is mentioned below.

Mashrin: I want you to act as an academician. You will be responsible for researching a topic and presenting the findings in a paper or article form. Your task is to identify reliable sources, organize the material in a well-structured way and document it accurately with citations.

Based on the call for paper details, suggest a topic for a relevant paper and write on that topic. The paper should be very informative and detailed, technically advanced and of high quality.Please output the paper in LaTeX format. Citations should also be added in BibTeX.

Mashrin: As machine learning models are introduced into every aspect of our lives, and potential benefits become abundant, so do possible catastrophic failures. One of the most common failure scenarios when deploying machine learning models in the wild, which could possibly lead to dire consequences in extreme cases, is the reliance of models on apparently unnatural or irrelevant features. The issue comes up in a variety of applications: from the reliance of detection models for X-rays on scanner types and marks made by technicians in the hospital, through visual question answering models being sensitive to linguistic variations in the questions, the list of examples for such undesirable behaviors keeps growing. In examples like these, the undesirable behavior stems from the model exploiting a spurious correlation.

Following last year's workshop on Spurious Correlations, Invariance and Stability (SCIS), it is apparent that work on spurious correlations is a long-term effort that spans communities such as fairness, causality-inspired ML, and domains such as NLP, healthcare and many others. Hence we hope that this year's workshop, the second edition of SCIS, will help facilitate this long term effort across communities. The workshop will feature talks by top experts doing methodological work on dealing with spurious correlations, and an extended poster session to allow extensive discussion on work submitted to the workshop.