CS-260D - Final Project

Investigating spurious correlations for self-supervised learning (feature suppression)

Phillip Kwan pkwan930@gmail.com
Joseph Lin josephlin95@g.ucla.edu
Terry Pederson terryap@g.ucla.edu
Tong Wu wutong0218@ucla.edu
Department of Computer Science, UCLA, Los Angeles, CA, 90024.

Abstract

1. Summarize key objectives, methodology, findings, and significance.
2. Keep it brief and enticing for readers to explore the full content.

1 Introduction

Investigating spurious correlations in the context of self-supervised learning, particularly for feature suppression, involves understanding how models may inadvertently learn to associate unrelated features or patterns in data, leading to biased or inaccurate outcomes. Here's a breakdown of key aspects to consider:

Understanding Spurious Correlations: In machine learning, a spurious correlation refers to a relationship where two variables appear to be connected, but their association is either coincidental or due to an unseen third variable. In self-supervised learning, where models are trained to understand data without explicit labels, they might pick up such misleading correlations.

Self-Supervised Learning: This learning paradigm relies on algorithms that generate their own supervisory signal from the input data. It's useful for scenarios where labeled data is scarce or expensive to obtain. The model learns to predict some parts of its input from other parts, gaining a form of understanding about the data structure.

Feature Suppression in Self-Supervised Learning: Feature suppression involves identifying and minimizing the impact of less relevant or misleading features on the model's learning process. It's crucial for improving the robustness and accuracy of the model. In the context of spurious correlations, this would mean identifying and reducing the influence of features that are correlated but not causally related to the task at hand.

Methods to Address Spurious Correlations:

- **Data Augmentation:** Altering data in ways that break spurious correlations without affecting real, causal relationships can help. For instance, adding noise or modifying aspects of the input that are related to spurious features.
- **Regularization Techniques:** Regularization methods, like dropout or weight decay, can prevent the model from overly relying on any small set of features, which might include spurious ones.
- **Contrastive Learning:** In self-supervised learning, contrastive methods that encourage the model to focus on similarities and differences across varied instances can reduce the impact of spurious features.
- **Interventional Techniques:** Methods like causal inference can help in identifying and controlling for spurious variables, leading to more robust feature representations.

Preprint. Under review.

Evaluating and Validating Models: Rigorous testing, including out-of-distribution evaluation, is vital. This ensures that the model's performance is not overly reliant on spurious correlations that might not hold in real-world scenarios or across different datasets.

Ethical Considerations and Bias: It's crucial to be aware of and address biases that might arise from spurious correlations, especially when these models are applied in critical areas like healthcare, finance, or law enforcement.

2 Related Work

1. Review existing literature relevant to your study.
2. Highlight key findings and methodologies from previous research.
3. Identify gaps or areas where your study contributes.

3 Problem Formulation

1. Clearly define the problem or question your research aims to address.
2. Try to be formal here (use mathematical notation as far as possible)
3. Clearly state the specific setting you are exploring rather than the overarching problem

4 Method

1. Describe the research design, approach, and methodology.
2. Detail the procedures and tools used for data collection and analysis.
3. Ensure clarity and reproducibility for readers.

5 Experiments

1. Present the design and execution of your experiments.
2. Report on the data collected and any statistical or analytical methods used.
3. Provide sufficient detail for others to replicate your experiments.

6 Conclusion

1. Summarize the key findings and their implications.
2. Reflect on how your results contribute to the broader context.
3. Address any limitations and suggest avenues for future research.

References

A Appendix 1