The Falsification Challenge

Investigating Simplicity Bias in Neural Networks

Overview

In this challenge, you are not being asked to build the best model, reproduce a paper exactly, or obtain the highest accuracy. Instead, you will take the role of a skeptical AI researcher.

You will investigate the following broad claim:

> When multiple features can be used to solve a task, neural networks tend to preferentially learn simpler features over more complex ones.

Your objective is to critically examine this claim and design a research study that attempts to falsify, weaken, qualify, or better understand it.

The central question is not:

> How can I prove that simplicity bias exists?

Instead, it is:

> Under what conditions might this claim fail, and how would we know scientifically?

The Skeptical Research Mindset

Scientific research does not progress only by confirming existing ideas.

Suppose a paper makes the claim:

> Neural networks prefer simple features.

A weak research approach would be:

Read the paper.

Repeat its experiment.

Obtain similar results.

Conclude that the paper was correct.

A stronger approach is to ask:

- What assumptions does this claim depend on?
- What exactly does simple mean?
- Does the claim hold for all architectures?
- Does it depend on the optimizer?
- Does it depend on the dataset construction?
- Could another explanation produce the same observed result?
- Can we construct a situation where a more complex feature is learned first?
- Is the observed phenomenon actually a property of the network, or of the optimization process?

Your goal is to actively search for the boundaries of the claim.

Importantly:

> You are not expected to successfully disprove the claim.

A carefully designed experiment that fails to falsify the hypothesis can be more scientifically valuable than a poorly designed experiment that appears to contradict it.

Background: What Is Simplicity Bias?

Neural networks are highly expressive models capable of representing extremely complicated functions. Despite this, research has found that training procedures can exhibit a preference for simpler predictive patterns.

One influential formulation argues that neural networks may exhibit extreme simplicity bias, sometimes relying primarily on the simplest available predictive feature while ignoring other predictive features that may be more complex or even more robust. (arXiv)

For example, imagine a dataset containing two features that predict the same class.

Feature A: Simple

> Background color to Class

Feature B: Complex

> Object shape to Class

Even if both features are predictive, a neural network might primarily learn the simpler background correlation.

This raises an important question:

> Is the preference for simplicity universal, or does it depend on how we define simplicity and how the network is trained?

The difficulty begins here: “simple” is not a single universally agreed concept. It might refer to:

- Linear separability
- Low dimensionality
- Low-frequency structure
- Few parameters
- Low computational complexity
- Ease of optimization
- Short description length

This ambiguity is not a problem to avoid it is potentially one of the most interesting things to investigate.

Recommended Literature

You are not required to understand every mathematical detail of these papers. Focus on understanding:

What claim is being made?

How do the authors define or operationalize simplicity?

What experiments support the claim?

What assumptions might those experiments depend on?

Primary paper

The Pitfalls of Simplicity Bias in Neural Networks Shah et al.

The paper argues that neural networks trained with standard procedures can exhibit an extreme preference for simpler predictive features, potentially ignoring more complex features even when those features are useful or robust. It also emphasizes that the definition of simplicity is itself an important issue. (arXiv)

Primary paper: The Pitfalls of Simplicity Bias in Neural Networks

Supporting paper

Simplicity Bias in 1-Hidden Layer Neural Networks Morwani et al., NeurIPS 2023

This work gives a more formal treatment of simplicity bias and studies the idea that networks can effectively depend on a low-dimensional projection even when more complex predictive features are available. (NeurIPS Papers)

Simplicity Bias in 1-Hidden Layer Neural Networks

Optional background

On the Spectral Bias of Neural Networks Rahaman et al., ICML 2019

This paper investigates a related but distinct idea: neural networks can exhibit a learning bias toward lower-frequency functions, meaning that lower-frequency patterns may be learned faster than higher-frequency ones. It also shows that the geometry of the data manifold can affect this behavior. (Proceedings of Machine Learning Research)

On the Spectral Bias of Neural Networks

Your Research Task

Your task is to design a small research study investigating simplicity bias.

You should begin with the broad hypothesis:

Research Claim

> Neural networks preferentially learn simpler predictive features over more complex predictive features.

Your task is to challenge this claim.

You may investigate questions such as:

- Can a neural network learn a complex feature before a simple one?
- Does the phenomenon depend on the definition of simplicity?
- Does the preference disappear for certain architectures?
- Does optimization influence the observed simplicity bias?
- Does training duration change which feature the model ultimately relies upon?
- Can data correlations overpower feature simplicity?
- Is a supposedly simpler feature actually easier to optimize rather than intrinsically simpler?
- Can multiple notions of simplicity produce contradictory results?

These are examples, not restrictions. You are encouraged to formulate your own research question.

What We Primarily Want From You

The main deliverable is an experimental research proposal.

You are not required to run the experiments.

Instead, we want to see whether you can design a scientifically meaningful experiment suite.

Your submission should explain:

Your Research Question

Clearly state what specific part of the simplicity-bias claim you are investigating.

For example:

> Does simplicity bias persist when feature simplicity and feature-label correlation are independently controlled?

Your Hypothesis

State what you expect to happen and why.

Your Null Hypothesis

For example:

> H_0: Neural networks do not systematically prefer simpler features when multiple predictive features are available.

Your null hypothesis should be appropriate for your particular experiment.

Alternative Explanation

Explain what else might explain the results reported in the literature.

For example:

> Perhaps the network is not selecting the simpler feature because it is simpler. Perhaps that feature is simply easier to optimize.

This is an extremely important part of the challenge.

Experimental Design

Describe exactly how your experiment would work.

Include:

- Dataset design or selection
- Features being controlled
- What changes between experiments
- What remains fixed
- Model architecture(s)
- Training procedure
- Evaluation methodology

Controls and Baselines

A research experiment must distinguish between competing explanations.

For example, if you claim that feature complexity affects learning, you should ask:

> What experiment would distinguish complexity from correlation strength?

Explain what controls you would use.

Expected Outcomes

Describe possible outcomes.

For example:

Possible ResultInterpretation

Simple feature is consistently learned Supports the original claim

Complex feature is consistently learned Challenges the claim

Behavior changes by optimizer Suggests optimization may be responsible

Behavior changes by definition of simplicity Suggests the claim requires qualification

You do not need to predict that your experiment will falsify the claim. In fact, you should discuss multiple possible outcomes.

Limitations

Every experiment has limitations.

Explain:

- What your experiment cannot establish
- What variables may still be confounded
- Whether your definition of simplicity is incomplete
- How future experiments could improve the investigation

Implementation Is Optional

The primary objective of this challenge is:

> Design a rigorous experiment or experiment suite.

You are not required to implement or run your proposed experiments.

However, if your proposed experiment is sufficiently simple and computationally feasible, we strongly encourage you to implement it.

For example, if you design a synthetic dataset with:

- One simple feature
- One complex feature
- A small neural network

there may be no reason to leave the experiment entirely theoretical. Running it could provide useful evidence.

Therefore:

> Implementation and experimental results are optional, but will receive additional credit when they meaningfully strengthen the investigation.

Do not implement experiments merely for the sake of producing graphs.

A small, carefully controlled experiment is more valuable than a large amount of uncontrolled computation.

The Principle of Appropriate Complexity

This challenge is not asking you to design a massive research project.

You do not need:

- Large-scale GPU training
- Large language models
- Massive datasets
- Complex infrastructure
- Extensive hyperparameter searches

A good experiment could potentially involve:

Synthetic Dataset+Small Neural Network+One Carefully Controlled Variable\text{Synthetic Dataset} + \text{Small Neural Network} + \text{One Carefully Controlled Variable}

For example:

> Construct two predictive features, systematically change their complexity and predictive strength, train a small network, and measure which feature the network relies upon.

The simplicity of an experiment is not a weakness.

In research, a simple experiment that isolates a variable clearly can be significantly more valuable than a complicated experiment where many factors change simultaneously.

Suggested Submission Structure

Title

A concise title describing your investigation.

Introduction

What claim are you investigating and why?

Literature Review

Briefly discuss the provided literature and any additional work you find.

Focus on:

- The central claims
- Experimental evidence
- Definitions of simplicity
- Open questions

Research Question

State your specific question clearly.

Hypothesis and Null Hypothesis

Define both precisely.

Proposed Experiment Suite

Explain your experiments in sufficient detail that another researcher could implement them.

Controls and Confounding Variables

Explain how your design distinguishes between competing explanations.

Expected Outcomes and Interpretation

Describe how different results would affect your conclusions.

Optional Experimental Results

If you implemented experiments, present:

- Methodology
- Results
- Analysis

Limitations

Discuss weaknesses honestly.

Conclusion

State what your proposed investigation could potentially tell us about the original claim.

How You Will Be Evaluated

You will not primarily be evaluated on whether you successfully falsify the paper.

Instead, we will evaluate:

CategoryWhat We Look For

Research Question Is the question meaningful and precise?

Skeptical Thinking Did you genuinely challenge the original assumptions?

Hypothesis Is it falsifiable?

Experimental Design Does the experiment isolate the relevant variables?

Controls Have alternative explanations been considered?

Originality Does the investigation show independent thinking?

Scientific Reasoning Are conclusions proportional to the evidence?

Limitations Does the candidate understand what their study cannot prove?

Optional Implementation Does running experiments meaningfully strengthen the study?

Important

A participant who says:

> “I attempted to falsify the claim, but my experiments were unable to do so under the conditions I tested.”

can receive a very high score.

A participant who claims:

> “I disproved the paper”

without strong controls and evidence will not.

Final Challenge Statement

> ## The Falsification Challenge
>
> Neural networks are often claimed to exhibit simplicity bias a tendency to preferentially learn simpler predictive features when multiple features can solve the same task. Your task is not to accept this claim or reproduce it blindly. Instead, act as a skeptical researcher: identify assumptions behind the claim, formulate a falsifiable research question, propose alternative explanations, and design a controlled experiment or experiment suite capable of challenging the claim. Your primary deliverable is a rigorous research proposal. Implementation is optional and will receive additional credit where it meaningfully strengthens the investigation. We are not looking for the largest model or the highest accuracy we are looking for evidence of scientific thinking, experimental reasoning, skepticism, and the ability to turn an interesting AI claim into a testable research study.

I think this structure is especially strong because it lets even someone with limited compute design a genuinely meaningful research investigation, while stronger candidates can go further and actually run their proposed experiments.