Sources of Irreproducibility in Machine Learning: A Review

Selina Cheng

Authors

Odd Erik Gundersen, Norwegian University of Science and Technology, Trondheim, Norway **(same first author as Paper 7)**

Kevin Coakley, Norwegian University of Science and Technology, Trondheim, Norway, San Diego Supercomputer Center, California, USA

Christine Kirkpatrick, San Diego Supercomputer Center, California, USA

Overview

Problem: Unreliable results in ML studies because of non-reproducibility

Goal: Identify and categorize root causes of irreproducibility in ML

Methods

- Literature Review
- Taxonomy of issues

Study design factors	HARKing, p-fishing
Algorithmic factors	Random weight initialization
Implementation factors	Ancillary software, compiler settings
Observation factors	Dataset bias, data preparation
Evaluation factors	Error estimation, selective reporting
Documentation factors	Readability, data, code, implementation

Experimental Design

- Review process
  - Impact of methodology decisions on reproducibility
- Case studies
  - Different experiment types should prioritize different sources of irreproducibility

Results

- Top Causes
  - Lack of code sharing
  - Failure to control for randomness
  - Non-standard evaluation metrics (for comparison)
- Recommendations
  - Documentation
  - Standardization
  - Validation

Key Takeaways

To ensure sustainable progress in ML research, controlling for reproducibility is critical.

Confirmatory Hypotheses vs *Hypothesis Generating*

- Confirmatory hypotheses
  - Should discuss controlling as many as possible
- Hypotheses generating
  - Can relax study design and observation factors
  - Lightly consider implementation and evaluation factors
  - Prioritize algorithmic (when relevant) and documentation factors