MASTER THESIS PROPOSAL

Exploiting Model Multiplicity in Machine Learning

Student:
Maksim Fediushkin
makcfd@gmail.com

Supervisors:
Florian Lemmerich
florian.lemmerich@uni-passau.de

Applied Machine Learning
May 21, 2023

Contents

1 Motivation

Machine learning (ML) models have gained widespread use in many applications, ranging from finance to healthcare. However, the increasing reliance on these models has also raised concerns about potential risks and vulnerabilities. One aspect of these models that has recently drawn attention is their multiplicity, or the differences in predictions that arise due to various factors such as initializations, architectures, and training data. This research will focus on exploiting machine learning model multiplicity, specifically on understanding the factors that contribute to model exploitation and developing strategies to mitigate the associated risks. The work will exclude data poisoning, as this topic has been extensively covered in existing literature (Steinhardt et al., 2017).

2 Research Questions

The proposed research will address the following questions:

1. How can ML model multiplicity be exploited for personal or group benefit, and what are the potential risks associated with this exploitation?
2. What methods can be implemented to mitigate the risks arising from the exploitation of ML model multiplicity?

3 Literature Review

A literature review will be conducted to identify the current state of knowledge in the field of machine learning model exploitation and risk mitigation. Preliminary works to be reviewed:

1. "Predictive Multiplicity in Classification" (Wu et al., **2021**). This study analyzes the causes of multiplicity in classification tasks and provides insights into the role of initialization, architecture, and data in creating multiple solutions.
2. "Disentangling Model Multiplicity in Deep Learning" (Fort et al., **2021**). This work investigates the multiplicity of deep learning models and proposes a framework for understanding the factors contributing to it.

3. "Torch.manual seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision" (Engstrom et al., **2021**). This paper examines the impact of random seed selection on the performance and stability of deep learning models in computer vision.

4. "Intriguing properties of neural networks. This paper explores the adversarial examples in neural networks and their implications on the stability of the model predictions" (Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. **2013**). Although this paper not directly focusing on the exploitation of prediction instability, it discusses the underlying causes of such instability in deep learning.

5. "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning" (Gal, Y., Ghahramani, Z. **2016**). The authors propose a Bayesian interpretation of dropout, a popular regularization technique, to quantify the uncertainty in the predictions of deep learning models. This work can provide insights into how uncertainty estimation can be exploited for specific observations.

6. "Taking human out of learning applications: A survey on automated machine learning" (Yao, Q., Wang, M., Chen, Y., Dai, W., Hu, Q., Li, Y. **2019**). This survey paper provides a comprehensive overview of automated machine learning (AutoML) techniques, some of which may explore and exploit the instability of machine learning predictions in order to optimize model performance.

7. "Train longer, generalize better: closing the generalization gap in large batch training of neural networks" (Hoffer, E., Hubara, I., Soudry, D. **2020**). This paper investigates the relationship between large-batch training and the generalization gap in neural networks. It suggests that increasing the training time can improve the generalization performance of large-batch training.

4 Methodology

This study will employ a mixed-methods approach, combining quantitative and qualitative research methods, to address the research questions. The

main proposed research phases are:

1. Run an analysis of the factors affecting model multiplicity based on the literature review.

2. Setting up the baseline of the research: To establish a robust foundation for this research, a baseline model will be defined, selection of the appropriate dataset, and selection of specific data observations on which to conduct experimental analysis of model multiplicity exploitation will be executed.

3. Investigate exploitation and risks: Exploitation of ML model multiplicity for personal or group benefit will be studied using both theoretical and empirical approaches. The potential risks associated with this exploitation will be analyzed and discussed. This step involves quantifying the effects from experiments on irrelevant parameters of the model training will be taken place (not limited):

   (a) Changing random seed and parameters initialisation

   (b) "Early stopping" when exploitation of a particular instance has been reached

   (c) Changing architecture of NN, for example:

      i. Change number of neurons in the layers near input and/or output

      ii. Change number of layers of NN

   (d) Changing batch ordering

   (e) Changing batch sizes

4. Develop mitigation strategies: Based on the findings from previous steps, a set of strategies for mitigating the risks arising from the exploitation of ML model multiplicity will be proposed. These strategies will be evaluated using experimental setups to test their effectiveness.

5 Expected Outcomes

By the end of this research, the following outcomes are expected:

1. An understanding of the factors contributing to ML model multiplicity and their quantification.
2. Identification of potential exploitation strategies and associated risks in leveraging ML model multiplicity for personal or group benefit.
3. Proposal and evaluation of mitigation strategies for managing the risks arising from the exploitation of ML model multiplicity.

6 Timeline

The proposed timeline for this work is as follows:

1. Months 1-2: Literature review and analysis of factors contributing to ML model multiplicity.
2. Months 3-4: Investigation of exploitation strategies and risks.
3. Months 5-6: Development and evaluation of mitigation strategies. Writing and submitting the thesis

7 Miscellaneous

Pytorch

Immatruclation Transcripion of records Form with stated thesis title

Sign the form

Karin Brezel secretary Thursday.

first step simple model small dataset change random seed run 100 experiments and compare the output instability measurements metrics

It will be in two weeks "Survey of measure the stability of NN".