The Effect of Corrupted Data on Modern ML Models

Supervisor: Ljubomir Rokvić

January 4th, 2023

1 Project Overview & Goal Description

Due to the rapid development of machine learning, models have become very large and difficult to train. As models get more, and more powerful the group of problems that they can solve increases. This leads to datasets becoming much larger both in number of entities and number of features. Furthermore, models have been continuously growing in both efficiency and performance. Some of the most popular models include ResNet [2] and Visual Image Transformers [1]. All real world problems suffer from inaccuracies, which can sometimes not be avoided. Sometimes these inaccuracies are caused due to poor sensors, other times due to human error, and sometimes even due to malicious intent. It is unclear how different types of data corruption affect the outcome of the model.

The goal of this project is to identify and select a set of realistic corruption techniques. Following that the student needs to select a subset of popular classification models and implement them [4]. Furthermore, after evaluating the model performance on non-corrupt and corrupt data, the student has to do the same for a distributed (Federated Learning) .

2 Project Implementation

1. Get acquainted with the current state of the art ML models.
2. Implement the most common models. [4]
3. Select different data corruption techniques, and show how they impact said models.
4. Compare all the results and analyze how effect each data corruption method is.
5. Expand this further into the domain of Federated Learning, and analyze the impact of these techniques on FedAvg. [3]

3 Required Skills

- Good programming knowledge and skills, preferably python.
- General knowledge on machine learning.
- Decent academic and research skills.
- Being passionate about the topic and good English skills are a must.

References

[1] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* 2020. DOI: `10.48550/ARXIV.2010.11929`. URL: `https://arxiv.org/abs/2010.11929`.

[2] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: `10.48550/ARXIV.1512.03385`. URL: `https://arxiv.org/abs/1512.03385`.

[3] H. Brendan McMahan et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In: (2016). DOI: `10.48550/ARXIV.1602.05629`. URL: `https://arxiv.org/abs/1602.05629`.

[4] Shagun Sharma and Kalpna Guleria. "Deep learning models for image classification: comparison and applications". In: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE. 2022, pp. 1733–1738.