Paper Summaries Table of Contents

[**Adversarial Deep Learning Models with Multiple Adversaries - Chivukula, Liu 2**](#_oavhvn33iuix)

[**Learning Adversarially Fair and Transferable Representations – Madras, Creager, Pitassi , Zemel 4**](#_3nvvdqs1ur7l)

# Adversarial Deep Learning Models with Multiple Adversaries - Janapriya, Anuradha, Srilakshimi

Synopsis/ Notes

This paper focuses on the using adversarial machine learning that generates deceptive data. Showing a weakness in current models and how adversarial attacks can exploit vulnerabilities.This paper focuses on improving adversarial learning algorithm for convolutional neural networks using game theory and evolutionary computation.

Introduction

ML models are used in high stakes applications such as finance, healthcare and cybersecurity. Adversarial learning shows that the vulnerabilities in models exposing the models sensitivity to sensitive inputs that can mislead the model. This shows that it is very important to develop the ML model to withstand adversarial attacks using adaptive learning.

Related Work

* Battista Biggo Adversarial attacks in spam filtering and pattern classification security
  + Challenges in securing ML models against adversarial threats focusing on spam detection and classifier .
  + Useful to read through if we want to use topics on adversarial inputs and creating defenses.

Game Formulation

This paper’s adversarial model architecture is based on deep learning algorithms such as CNNs, RNNs and FRNNs; as these models are used in disease prediction and facial recognition. This paper spoke about using a CNN that is resistant to adversarial modifications structured like [multiplayer stochastic game](https://en.wikipedia.org/wiki/Stochastic_game#:~:text=In%20game%20theory%2C%20a%20stochastic,or%20infinite%20number%20of%20stages.). Each step in this paper focuses on simulating real world condition where adversaries attempt to manipulate data while the model adapts to counter the attacks.

Stochastic Game Algorithm and Sequential Game ALgorithm

Algorithm Model

* Identifies similarities between input data and adversarial examples by selecting images and altering through mutation
* Change data attributes and transform positive classes into negative ones to mislead model.
* Maintain adversarial example relevance to help model learn to identify and mitigate threats.

Experiments

This paper expands their adversarial framework to use video data using augmented methods to anticipate adversarial manipulations across different data types.

# Learning Adversarially Fair and Transferable Representations – Madras, Creager, Pitassi , Zemel

Synopsis/Notes

This paper focuses on learning representations to ensure fairness in machine learning tasks when they are used by third parties with unknown objectives. This paper also focuses on linking group fairness such as demographic parity, equalized odds, equal opportunity to adversarial objectives.

Introduction

ML systems are built by two parties; the data owner who collects and represents data and the prediction vendor who uses the data to train models for [downstream tasks](https://www.baeldung.com/cs/downstream-tasks#:~:text=A%20downstream%20task%20depends%20on,once%20completing%20the%20previous%20one.). Vendors focus on prioritizing accuracy leading to biased predictions because of past discrimination. This paper shows that the data owner as a learning representation ensures that there is fairness. This helps with linking fairness to adversarial objectives, demonstrating fairness in transfer learning and experimentally validating naive classifiers trained on representing fairness.

Background

This paper focuses on Fairness in Classification and Adversarial Learning. In terms of Fairness in Classification; fairness ensures that predictions are unbiased with respect to sensitive attributes, there are 3 fairness metrics. These metrics include demographic Parity, Equalized odds and Equal opportunity. In terms of Adversarial Learning it is introduces by Goodfellow et al and involves a generator and discriminator optimizing opposing objectives. The framework is then adapted to ensure representations are fair by penalizing unfairness through adversary.

Adversarially Fair Representation

This paper proposing a model that learns a representation Z which reconstructs data X, predicts labels Y and protects the sensitive attributes A from adversary. This leads; to the Encoder f(X) mapping data to represent Z, classifier g(Z) prediction Y, ADversary h(Z) attempting to predict A and Decoder k(Z,A) reconstructing X from Z to A. This shows that the learning process alternate between minimizing classification dn reconstructing loss which adversarial loss is being maximized.

Experiments

This paper focuses on 2 Experiments, Fair Classification and Transfer Learning. The model (LAFTR) in this paper is tested on the UCI adult dataset to predict income which is the sensitive attribute. The results of the model showed that LAFTR got the best accuracy fairness trade off compared to baselines. It also showed that adversarial objectives tailored to specific fairness metrics such as DP, EO, and EOpp. For transfer learning the heritage health dataset where a task is used to predict health outcomes while ensuring fairness across age groups while the transfer task is used to learn representations to predict new outcomes. This resulted in LAFTR ensuring fairness on transfer task better than the baselines with minimal loss in accuracy while the adversarial objectives are directly accessing representations outperforming the ones that have access to classifier outputs only.

Conclusion

This paper shows the LAFTR which is a method for learning representations that ensure fairness in both classification and transfer learning. This results in theoretical guarantees which links fairness metrics to adversarial objectives which demonstrate practical benefits through experiments.

# Adversarial Deep Learning Models with Multiple Adversaries - Chivukula, Liu