Towards Trustworthy Machine Learning Under Distribution Shifts

Jun Wu

Michigan State University
wujun4@msu.edu

Abstract

Transfer learning aims to transfer knowledge or information from a source domain to a relevant target domain. It involves two key challenges: distribution shifts and trustworthiness concerns. Having these challenges in mind, my research focuses on understanding transfer learning from the perspective of knowledge transferability (e.g., IID and non-IID learning tasks) and trustworthiness (e.g., adversarial robustness, data privacy, and performance fairness).

Introduction

Standard machine learning assumes that training and test samples follow the same data distribution. However, this assumption is often violated in real-world applications, especially when collecting samples from different sources/environments. The distribution shifts between training and testing data can impair the generalization performance of learning algorithms. To address this problem, my research investigates the problem of trustworthy transfer learning. The overall goal of trustworthy transfer learning is to transfer knowledge from source (training) data to relevant target (testing) data in a trustworthy and reliable manner. Specifically, there are two crucial components in understanding trustworthy transfer learning.

- **Part I: Knowledge Transferability:** My research quantitatively measures and enhances knowledge transferability across domains for IID and non-IID learning tasks.
  - *IID Scenarios:* It assumes that samples within each domain are independent and identically distributed (IID), e.g., cross-domain object recognition. My research (Wu et al. 2022) develops a distribution-informed neural network (DINO) to build the distribution-aware relationship of inputs and outputs from different domains. The connection between DINO and transferable Gaussian processes is theoretically analyzed.
  - *Non-IID Scenarios:* It involves a relaxed assumption that samples within each domain can be non-independent and identically distributed (non-IID), e.g., cross-network node classification. My work (Wu, He, and Ainsworth 2023) studies graph transfer learning by quantitatively measuring the distribution shifts between source and target graphs.
- **Part II: Knowledge Trustworthiness:** In addition to understanding knowledge transferability, my research also investigates the trustworthiness properties of deep transfer learning, including adversarial robustness, data privacy, and performance fairness.
  - *Adversarial Robustness:* My research (Wu and He 2021) demonstrates the adversarial vulnerability of unsupervised domain adaptation techniques. The crucial idea of the proposed I2Attack framework is to maximize the label-informed joint distribution discrepancy between raw and poisoned source domains under several constraints.
  - *Privacy and Fairness:* To protect data privacy among clients, federated learning frameworks have been proposed in past years. However, it is observed that some clients might suffer from negative transfer. To improve performance fairness and protect client privacy, FEDORA (Wu et al. 2023) is proposed based on adaptive parameter propagation and selective regularization.

In the future, my research will focus on further exploring the open questions in trustworthy transfer learning. To name a few, how can the decision boundary between positive and negative transfer be rigorously defined? How can we theoretically understand the generalization performance of transfer learning across different data modalities? What is the inherent trade-off between transfer accuracy and trustworthiness under different distribution shifts and data modalities?

References

Wu, J.; Bao, W.; Ainsworth, E.; and He, J. 2023. Personalized federated learning with parameter propagation. In *KDD*.

Wu, J.; and He, J. 2021. Indirect invisible poisoning attacks on domain adaptation. In *KDD*.

Wu, J.; He, J.; and Ainsworth, E. 2023. Non-IID transfer learning on graphs. In *AAAI*.

Wu, J.; He, J.; Wang, S.; Guan, K.; and Ainsworth, E. 2022. Distribution-informed neural networks for domain adaptation regression. *NeurIPS*.