ABSTRACT

Despite superior performance in many situations, deep neural networks are often vulnerable to adversarial examples and distribution shifts, limiting model generalization ability in real-world applications. To alleviate these problems, recent approaches leverage distributional robustness optimization (DRO) to find the most challenging distribution, and then minimize loss function over this most challenging distribution.

Regardless of having achieved some improvements, these DRO approaches have some obvious limitations. First, they purely focus on local regularization to strengthen model robustness, missing a global regularization effect that is useful in many real-world applications (e.g., domain adaptation, domain generalization, and adversarial machine learning). Second, the loss functions in the existing DRO approaches operate in only the most challenging distribution, hence decouple with the original distribution, leading to a restrictive modeling capability.

In this thesis, we propose a novel regularization technique, following the veins of Wasserstein-based DRO framework. Specifically, we define a particular joint distribution and Wasserstein-based uncertainty, allowing us to couple the original and most challenging distributions for enhancing modeling capability and applying both local and global regularizations. Empirical studies on different learning problems demonstrate that our proposed approach significantly outperforms the existing regularization approaches in various domains.