abstract

Machine learning has demonstrated great potential across a wide range of applications such as computer vision, robotics, speech recognition, drug discovery, material science, and physics simulation. Despite its current success, there are two major challenges for machine learning algorithms: limited robustness and generalizability.

The robustness of a neural network is defined as the influence that input perturbations have on its final prediction. It has been shown that neural networks are very sensitive to input perturbations. For convolutional neural networks, its prediction can be totally different for input images that are visually indistinguishable to human eyes. Based on such property, hackers can reversely engineer the input to trick machine learning systems in targeted ways. These adversarial attacks have shown to be surprisingly effective, which has raised serious concerns over safety-critical applications like autonomous driving. In the meantime, how to improve the robustness of neural networks is still an open question.

The generalizability of a neural network refers to its ability to be effective across a range of different inputs. On one hand, machine learning algorithms require a large number of samples from the data distribution in order to generalize well. It brings a big need for labeled data to perform supervised learning, and over-fitting on training data needs to be avoided for better generalization. On the other hand, machine learning models often fail to carry out reliable generalizations whenever there is a scarcity of supervised data. Many techniques have been proposed to improve the model generalizability; however, generalization with a few samples is still a challenging task.

In this dissertation, we are thus motivated to improve the robustness and generalizability of neural networks. Firstly, unlike traditional bottom-up classifiers, we use a pre-trained generative model to perform top-down reasoning and infer the label information. The proposed generative classifier has shown to be promising in handling challenging classification tasks like adversarial attacks and input distribution shifts. Secondly, we focus on improving the network robustness and propose an extension to adversarial training by considering the transformation invariance. Proposed method improves the robustness over state-of-the-art methods by 2.5\% on MNIST, 3.7\% on CIFAR-10, and 1.1\% on restricted ImageNet. Thirdly, we focus on designing networks that generalize well at predicting physics response. Our physics prior knowledge is used to guide the designing of the network architecture, which enables efficient learning and inference. Proposed network is able to generalize well even when it is trained with a single image pair.