Abstract

Recent advancements in machine learning have led to a state in which achieving more accurate models is often no longer the primary aim. The focus is rapidly shifting towards designing models that can demonstrate diverse but valuable qualities in addition to excellent performance. One of the current challenges in deep learning is the lack of robustness — poor accuracy on out-of-distribution data, which opens the door for deliberate attacks on models, against which they fail to defend. Another concern is the insufficient level of transparency and explainability of deep neural networks. During the interaction with people, the ability to express the model's reasoning is vital, as it leads to greater trust and promotes the deployment of machine learning models. In this work, we aspire to address these two concerns.

One of the key concepts we leverage throughout most of this work is attention. Attention mechanisms in machine learning have brought huge success in the past years. Nevertheless, in our opinion, it is not a fully explored area, and still provides a solid basis for further development. With all this in mind, we focus on three parallel research lines in this work. First, we analyze a group of malicious inputs called adversarial examples. We utilize them in standard deep learning models and propose ways to investigate their distinctions from in-distribution data. Second, in the hope of mitigating the effects of adversarial examples on image classification, we propose and examine an attention-based model, RecViT. Third, we design multiple approaches to build upon the state-of-the-art in the task of addressee classification in the human-robot interaction scenario. By leveraging the attention modules in our model design, we are able to craft human-readable explanations during the addressee estimation. Hopefully, this facilitates smoother and more trustworthy human-robot interaction.

Keywords: attention, explainability, robustness, adversarial examples