BrainTrain: Aligning Deep Neural Networks to Human Behavior to Improve Robustness and Generalization

Anand, Bharath (School: West Lafayette Junior/Senior High School)

The remarkable success of deep neural networks (DNNs) has led to great interest in their adoption in critical applications such as healthcare, autonomous vehicles, and law enforcement. DNNs are known to produce erroneous results under real-world noisy inputs, presenting a major bottleneck to their use in applications where lives, safety, or significant resources are at stake. Prior efforts to address this problem significantly increase training time and produce improvements only on specific types of noise. My research is motivated by the observation that humans are often resilient to inputs that are challenging for ANNs – in fact, humans may barely perceive perturbations that cause ANNs to fail spectacularly. I hypothesize that statistically aligning DNNs to human behavior during training can cause them to inherit desirable robustness traits. This led me to propose BrainTrain, a framework to create more robust DNNs through human behavior alignment. BrainTrain consists of (i) a cross-platform mobile application that enables the collection of human behavioral data at larger scales than previously feasible, (ii) a novel training method that uses a composite loss function to co-optimize accuracy and human behavior alignment during stochastic gradient descent (SGD) based training, and (iii) an evaluation framework to compare BrainTrain-ed DNN models with conventional models. I implemented BrainTrain using open-source software frameworks and applied it to state-of-the-art Residual Networks. BrainTrain-ed DNNs showed up to 26% higher accuracy under a wide range of noisy inputs and 16 times lower calibration error with no increase in training time. My work offers a pathway to addressing a key challenge facing DNNs and can enable their adoption in critical applications.