Limited Query Black-box Adversarial Attacks in the Real World

Todorov, Hristo (School: High School of Mathematics and Natural Sciences "Professor Emanuil Ivanov")

Despite the success of machine learning models in an expansive range of tasks, they are still surprisingly brittle to seemingly inconspicuous input manipulations, called adversarial examples, which is a major security concern for their deployment. We study the creation of physical adversarial examples, which are robust to real-world transformations, using a limited number of queries to the target black-box neural networks, which provide very little information. We observed that robust models tend to be especially susceptible to perturbations to the foreground of a given image, which motivated our novel Foreground attack, which uses region priors. We demonstrated that gradient priors are a useful component that could be used during different attacks to improve their efficiencies and therefore introduced an improved version of the popular SimBA. We also proposed an algorithm for ensemble-based transferable attacks that selects the most similar surrogates to the target model. Both our black-box attacks outperform the current state-of-the-art approaches they are based on and support our belief that some deep learning models share a lot of similarities between themselves and that knowledge could be leveraged to build strong attacks in a limited-information setting.

Awards Won:

National Taiwan Science Education Center: Taiwan International Science Fair Special Award is a trip to participate in the Taiwan International Science Fair

Association for Computing Machinery: Fourth Award of $500