Explaining Deep Neural Networks Through Fooling

Shpresim Sadiku
(Technische Universität Berlin & Zuse Institute Berlin)

The phenomenon of *adversarial attacks* reveals critical vulnerabilities in deep neural networks, as even well-trained models exhibit susceptibility to small input perturbations. For networks trained on data lying in low-dimensional subspaces, standard training methods often produce non-robust models with large gradients in directions orthogonal to the data subspace, making them highly prone to adversarial perturbations. Moreover, the demand for *explainable adversarial attacks* is growing, as they offer deeper theoretical insights into model behavior. While basic adversarial attacks resemble noise, advanced techniques, such as group-wise sparse attacks, produce structured, semantically meaningful perturbations that expose vulnerabilities in a more interpretable manner. Lastly, adversarial insights facilitate the generation of *counterfactual explanations* that lie within the data subspace. Leveraging efficient optimization techniques, such as accelerated proximal gradient methods, these counterfactuals align closely with the data subspace, providing plausible explanations for model decisions across a variety of classifiers.