研究課題別事後評価結果

１．研究課題名：　Discouraging adversarial attacks through improving the adversarial training

２．個人研究者名

Zhang Jingfeng（理化学研究所革新知能統合研究センター　研究員）

３．事後評価結果

敵対的攻撃に対する安全性保障のための以下のような研究を展開した。(1)敵対的学習の頑健性を向上するために、ノイズを含んだラベルを加える”NoiLin”を提案、(2)アンサンブル学習において、正確なモデルが少数派に残る危険性を排除した「Synergy-of-Experts (SoE)」手法の提案、(3)深層学習による画像ノイズ除去に対する攻撃への頑健性向上のために、「Hybrid Adversarial Training (HAT)」と呼ばれる敵対的学習戦略を提案、(4)2 つのサンプルの生成分布推定 2 サンプルテストに対する攻撃に対する防御についての max-min 最適化手法を提案。

以上のように敵対的学習に関する多様な問題に対して優れたアイデアを提案し、トップ会議にて発表した実績は社会へ大きなインパクトを与えたと評価できる。