NExT プログラム成果報告

第七期・後期 プログラム期間：2017 年 10 月 1 日～2018 年 3 月 31 日

「道路標識における Adversarial Examples の騙しやすさと影響度を考慮したリスク評価」

コンチネンタル・オートモーティブ（株）　J.DRIVE 水原 志暢

1. 背景

自動運転の技術の一つとして、道路標識認識システムがある。これにより、車は道路標識に従い安全な運転を可能とする。このシステムの基礎技術として、教師ありの深層学習の発展形である畳み込みニューラルネットワークが使われている。認識精度も非常に高いとされ、今後の応用が期待されている。一方で、学習器に対して高い確率で意図的に別の画像と誤認識させてしまう画像データ「Adversarial Examples」の脅威が存在する。現状、この脅威を根本的に解決する術はなく、評価に関しては検知度をもってして評価されてきたが、「安全性」の考慮がされてこなかった。

2. 目的

Adversarial Examples は、ある道路標識を別の道路標識と誤認識させることで、人命に関わる事故を起こさせる可能性がある。しかしながら、日本に存在する道路標識は約 200 種類以上存在し、すべてに対して対策を実施するのはコストの面で非常に困難である。従って、リスクの高さに応じて対策を行うために優先度が必要となってくる。そこで、本研究では、道路標識における Adversarial Examples の影響度と騙しやすさの両方を評価することで、Adversarial Examples のリスク評価の手法を提案する。

3. 方法

本研究では、安全に影響を与えるハザードを特定することができる FMEA (Failure Mode and Effects Analysis :故障モード影響解析)、を用いて評価を行った。FMEA では、リスク優先度(RPN) = 深刻度 × 発生度 × 検知度を計算し、リスクの程度の大きさを評価する。今回は、深刻度 : 事故を起こしたときの死亡率、発生度 : 違反をしたときの事故の発生率、検知度の代わりに脆弱度 : Adversarial Examples の誤認識率（高いほど騙され易い）と定義し評価計算を行った。

4. 結果

今回提案した評価手法では、交通違反やそれによって起きる事故、事故を起こした場所、事故の相手といった詳細な情報を用いることで、Adversarial Examples によって起こる車の動作から想定される事故の影響度を定量的に求めた。それによって、事故のリスク評価に妥当性を持たせ、さまざまな評価条件に対応できる評価手法を提案することができ、Adversarial Examples の騙しやすさと影響度を考慮したリスク評価が可能となった。