審 査 の 結 果 の 要 旨

氏　　名　チャウダリー　シュボジト

本論文は、「Understanding Generalization in Neural Networks for Robustness against Adversarial Attacks (敵対的攻撃に対するニューラルネットワークの汎化性能の理解)」と題し、英文より書かれており6章よりなる。近年、深層学習による物体認識の性能向上は目覚ましいが、一方でごく僅かなノイズを画像に付与することで認識性能が劇的に劣化する場合があることが知られている。本論文ではそのような敵対的攻撃と呼ばれる問題に対して、学習時における敵対的攻撃手法、最適化手法による敵対的攻撃手法への耐性の違い、医用画像認識における敵対的攻撃、敵対的手法に対する効果的な防御手法について網羅的に論じている。

第1章は「Introduction (序論)」であり、敵対的攻撃及び攻撃に対する防御手法についてこれまでなされてきた研究の体系化を行っている。

第2章は「Adversarial Training Time Attack against Discriminative and Generative Convolutional Models (学習時の敵対的攻撃手法及び認識・生成ネットワークへの攻撃)」と題し、学習データにわずか数ピクセルに対してノイズを混入させただけで推論時に大きく性能が下がる現象を示し、遺伝的アルゴリズムによる効果的な攻撃パターン生成手法について論じている。また、一般的に言われる認識問題だけでなく、ニューラルネットワークを用いた画像生成においても攻撃が成立し、意図した画像が生成されなくなることを実験的に示した。

第3章は「Robustness of Neural Network Optimization under Training Perturbations (学習時の敵対的攻撃におけるニューラルネットワーク最適化のロバスト性)」と題し、確率的勾配法ベースの手法と適応的勾配法ベースの手法の敵対的攻撃に対する安定性について論じている。なぜ最適化手法によって安定性に差が出るのかを簡素なモデルで証明したあと、実験的に様々な最適化手法の敵対的攻撃に対する安定性を検証している。

第4章は「Deep Learning for Medical Images under Adversarial Training Attacks (学習時の敵対的攻撃の医用画像認識への影響)」と題し、医用画像という学習データにノイズが混入しやすい状況下での安定性について論じている。医用画像認識という特殊な画像ドメインにおいても第2章、第3章で議論した内容がそのまま成り立つことを実験的に示している。

第5章は「Adversarial Robustness of Convolutional Models Learned in the Frequency Domain (周波数領域を利用した敵対的攻撃への耐性獲得)」と題し、一般的な敵対的攻撃に対する防御手法として周波数領域を用いた学習を提案している。敵対的攻撃で用いられる信号は一般的にノイズ的であり高周波領域までスペクトルが広がった信号になる

一方で、自然画像のスペクトルは低周波領域に集中することに注目し、画像を周波数領域に変換したあと学習を行うことで敵対的攻撃への強い耐性を得られることを実験的に示している。

第6章は「Conclusions (結論)」であり、本論文の成果と残された課題をまとめている。

以上これを要するに、本論文は、深層学習で問題となっている画像の敵対的攻撃に対し、新たな学習時攻撃手法、攻撃に対する最適化関数の安定性の違い、周波数領域を用いた効果的な防御手法について網羅的に議論したものであり、電子情報学上貢献するところが少なくない。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。