

Development and Improvement of White-box Attack Considering the Balance between Diversification and Intensification

山村, 圭一郎

<https://hdl.handle.net/2324/7329411>

出版情報 : Kyushu University, 2024, 博士 (数理学), 課程博士
バージョン :
権利関係 :



氏 名	山村 圭一郎			
論 文 名	Development and Improvement of White-box Attack Considering the Balance between Diversification and Intensification (多様化と集中化を考慮した敵対的攻撃手法の開発および改良)			
論文調査委員	主 査	九州大学	教授	藤澤 克樹
	副 査	九州大学	教授	鍛冶 静雄
	副 査	九州大学	教授	川村 秀一
	副 査	青山学院大学	准教授	小林 和博

論 文 審 査 の 結 果 の 要 旨

近年の生成 AI ブームに代表されるように、深層学習は必須の技術である。一方で敵対的例と呼ばれる入力に対して脆弱であることが知られている。敵対的例は通常、元の入力をわずかに摂動することで生成され、モデルの出力を大きく変化させることが知られている。このような入力の存在は、特に自動運転や顔認証などの安全性が極めて重要なアプリケーションにおいて致命的な結果を引き起こす可能性がある。したがって、深層学習モデルの堅牢性を向上させることが望まれる。堅牢性向上において有力とされる方法の一つに、敵対的学習と呼ばれる手法がある。敵対的学習は敵対的例を用いてモデルを訓練するため、通常の学習と比較して膨大な計算リソースを必要とする。このような背景を踏まえ、本研究では高速かつ大量に敵対的例を生成する、より強力な敵対的攻撃手法の開発を目指す。

敵対的攻撃は非凸な連続最適化問題として定式化される。この最適化問題における目的関数は複雑なニューラルネットワークを含むため、多峰性を持つことが予想される。多峰性関数の最適化では、異なる局所解の列挙（多様化）と局所解への素早い収束（集中化）が重要である。限られた計算時間の中で目的関数値を改善させるため、トレードオフの関係にある多様化と集中化を適切な割合で制御することが重要である。敵対的攻撃においては特に多様化の改善が攻撃性能の改善に寄与すると考えられている。本研究ではより適切な多様化を実現する、二つの異なるアプローチを提案する。一つ目は異なる目的関数と探索方向を適切に協働させるアプローチである。目的関数と探索方法を選択するためのアルゴリズム及び選択した組み合わせを適切に協働させるためのフレームワークを開発した。提案したフレームワークは既存の有力なアプローチと親和性が高く、併用することでさらに高い攻撃性能を期待することができる。

二つ目は高い多様化性能を持つ共役勾配法に基づく敵対的攻撃手法に注目し、更新式とステップサイズ制御方法を修正するアプローチである。更新式とステップサイズ制御方法の修正により、探索における多様化の改善に成功した。

提案手法の有効性は、主要なデータセットで学習された高いロバスト性を持つモデルを攻撃対象とした数値実験によって検証された。大規模で包括的な数値実験により、提案した二つのアプローチがともに探索の多様化度合いを改善し、既存手法と比較して高い攻撃性能を持つことを確認した。提案手法は特に ImageNet で学習されたモデルに対して高い有効性を示した。

以上の結果は、深層学習分野のジャーナルである Transactions on Machine Learning Research (TMLR) 及び国際会議 International Conference of Pattern Recognition and Artificial

Intelligence (ICPRAI)に採択されており、本分野において価値ある業績と認められる。
よって、本研究者は博士（数理学）の学位を受ける資格があるものと認める。