検出可能なAdversarial Exampleに関する研究

前田 拓海

概要

”Adversarial Example”とはAIを騙す技術のことです。

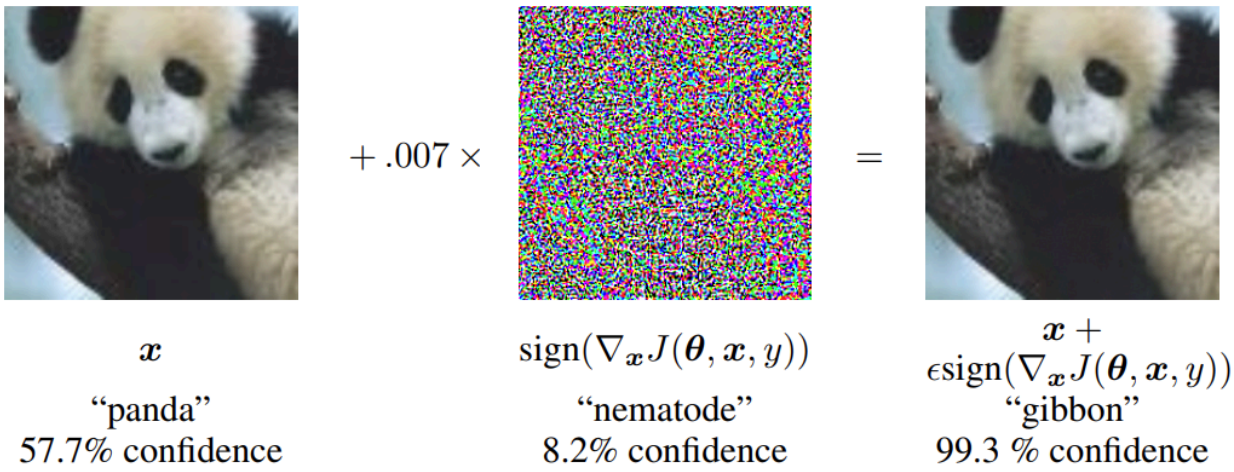

図1: パンダの画像に特殊な処理を施すことで、テナガザルと誤分類される例（Goodfellow et al. (2014)より）
左: 処理前のパンダの画像。正しくパンダと認識される。
中央: 特殊なノイズ。これを左の画像に加えることで右の画像になる。
右: 処理後のAdversarial Example画像。テナガザルと誤分類される。

例えば、動物の画像を分類するAIに、特殊な処理を施してあるパンダの画像を認識させると、テナガザルに誤分類されます（図1）。

処理する前の画像はもちろんパンダに正しく分類されます。
ですが、処理前の画像に僅かなノイズを加えるだけで、人間の目にはほとんど変化がないように見えても、パンダではなくテナガザルに誤分類されてしまう画像に変化してしまいます。
このような画像やデータなどは、Adversarial Exampleと呼ばれています。

このように、特殊な処理が施された画像やデータに騙されてしまう性質は、様々なAIが持っていることは今までの研究で分かっています。
（また、それに対する対策手法も、これまでの研究で幅広く数多く提案されてきています。）

ここまでの話だと、一見非常に危険な技術に思えます。
ですが、この技術は人間が必ず判断すべき事柄、つまりAIが判断するべきではない事柄に対しては使い方ができます。
この技術を使うことで、AIが正しく判断することが不可能になるため、人間が判断せざるを得なくなります。

ですが、このような使い方をするときに一つ問題があります。
AIは、Adversarial Exampleに騙されて正しく判断できなかったことを自覚できないため、判断結果を正しいものとして判断を進めてしまいます。
このような事態を防ぐためには、AI自身が騙されていることを認識できるような仕組みが必要です。

この研究では、Adversarial Exampleの生成方法を工夫することで、AI側に騙していることを伝える仕組みを提案しようと考えています。

アプローチ
AIの判断結果が、通常のデータ・Adversarial Exampleとは異なる、”変な”値になるAdversarial Exampleを生成します。
そして、このような変な値は、単純な方法で通常の判断結果と区別することができ、Adversarial Exampleに騙されていることをAI自身で認識することができるようになります。

現状の進捗

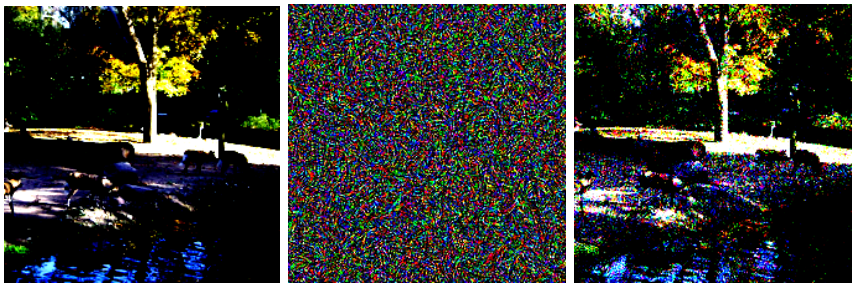

図2: これまでの予備実験で生成したAdversarial Example
左から、処理前の画像・ノイズ・処理後のAdversarial Example画像

同窓会からの支援を受けて購入した機材（パソコン）を用いて、本格的な実験に向けての下準備を進めています。

これまでに、下記のような作業を進めてきました。
・論文を通じて先行研究の調査
・実験に用いるAIの構築・プログラミング
・既存のAdversarial Example生成手法のプログラミング（図2）

これからは、下記のような作業を進めていく予定です。
・新たなAdversarial Exampleの生成手法の考案・プログラミング
・実験手法・内容の設定
・実験の実施
・プレゼンテーションや論文を通じた研究結果の発表

また、これからの研究はUTokyoGSC-Nextという東京大学主催のプログラム内で進めていきます。
同窓会からの支援が、このプログラム内での研究計画の作成の一助になりました。

謝辞
私の研究やUTokyoGSC-Nextへの挑戦を応援してくれた同窓会の方々に、心から感謝します。
本当にありがとうございました。