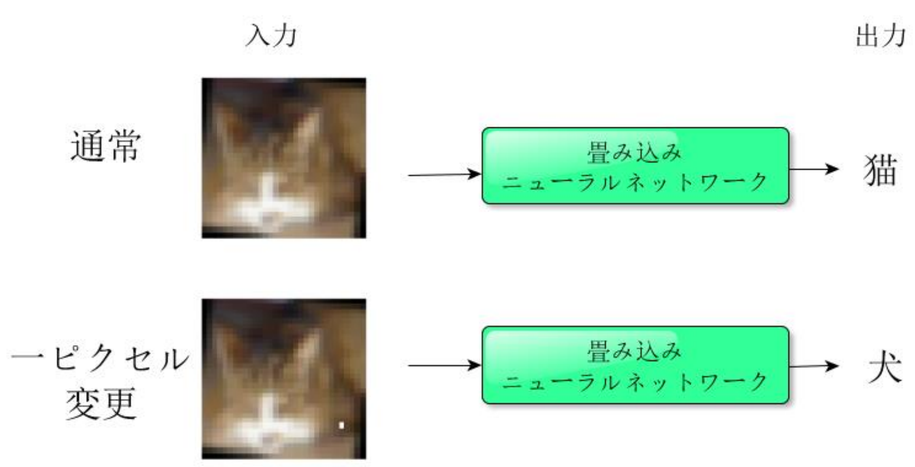
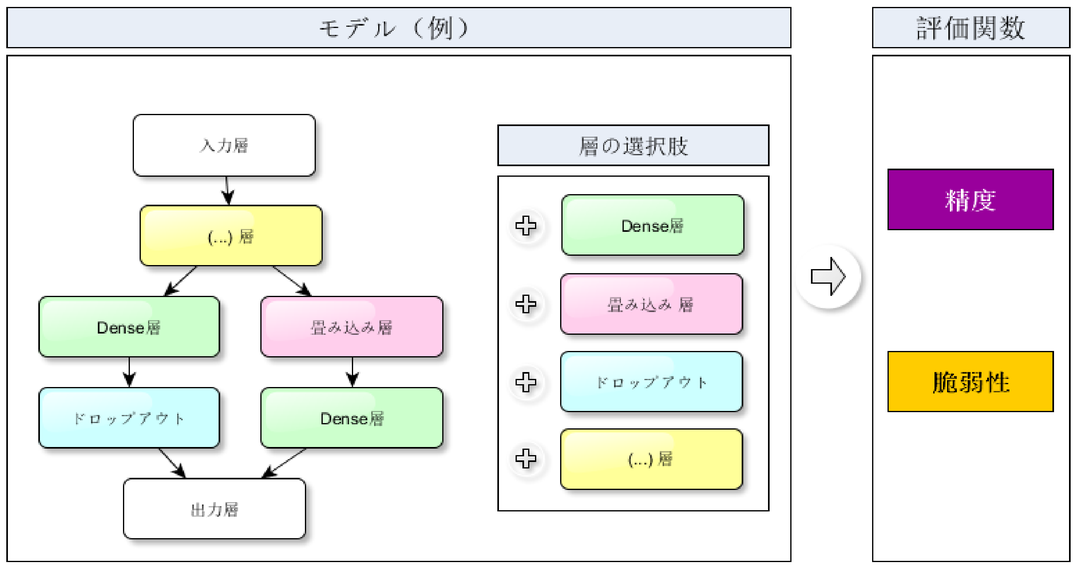
深層学習の安全な問題に関する研究

概要・アウトリーチ

本研究の代表者は、これまでの研究では、一つのピクセルを変えることでニュラールネットワークを誤魔化すことが可能と紹介した。その発見は畳みこみニュラールネットワークの脆弱性を表すことを実証している。この脆弱性の原因は畳みこみニュラールネットワークのモデルである。しかし、モデルの種類とパラメーターは複数あり、一番適切なモデルとパラメーターを見つけることは非常に時間がかかる。更に、深層学習のモデルはその問題を解決できない可能性もある。従って、本研究は最適化を利用し、自動的にロバスト深層学習を探索し、ロバスト性が高い深層学習の構造を発見した。



問題：一ピクセル変更だけでも深層学習を誤魔化される



解決手法：自動的に安全な人工知能の構造を探索手法

Architecture Search	Testing ER	ER on Adversarial Samples
DeepArchitect	25%	75%
Smash	23%	82%
Ours	18%	42%

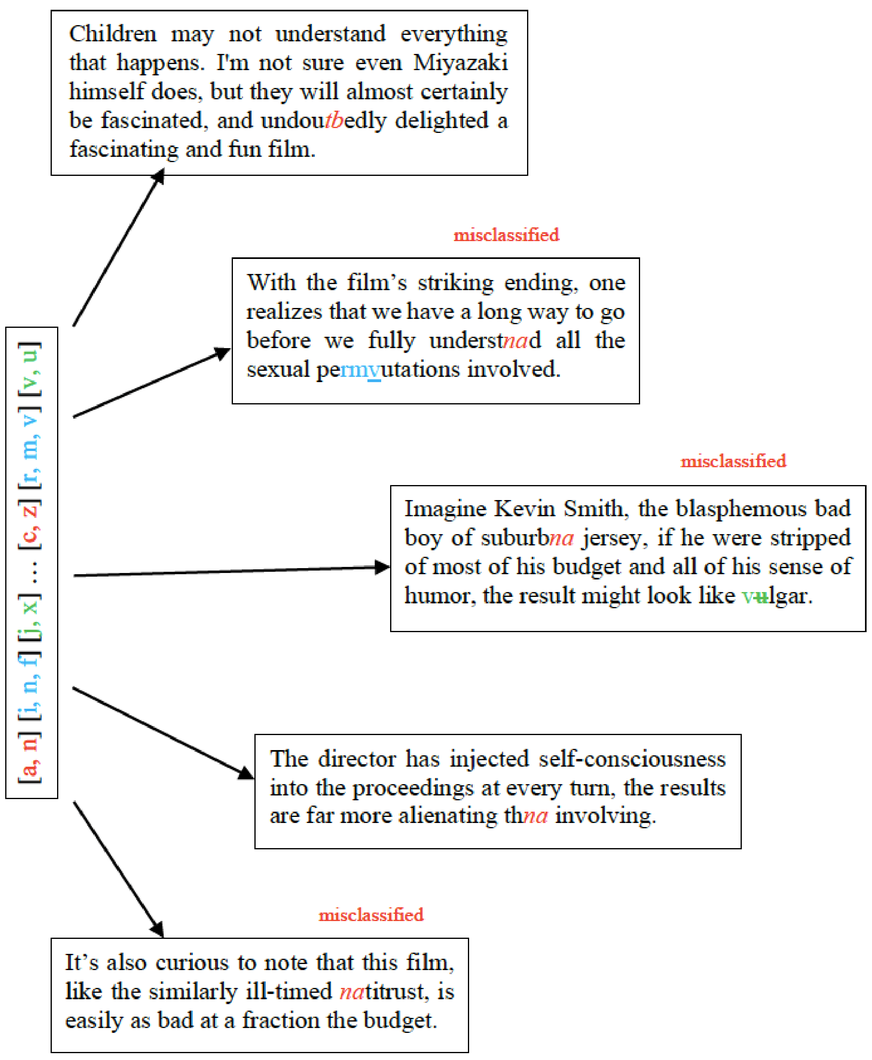
成果：より安全な深層学習の構造を発見した

本研究の成果では、以下のことが初めて明らかにしました：

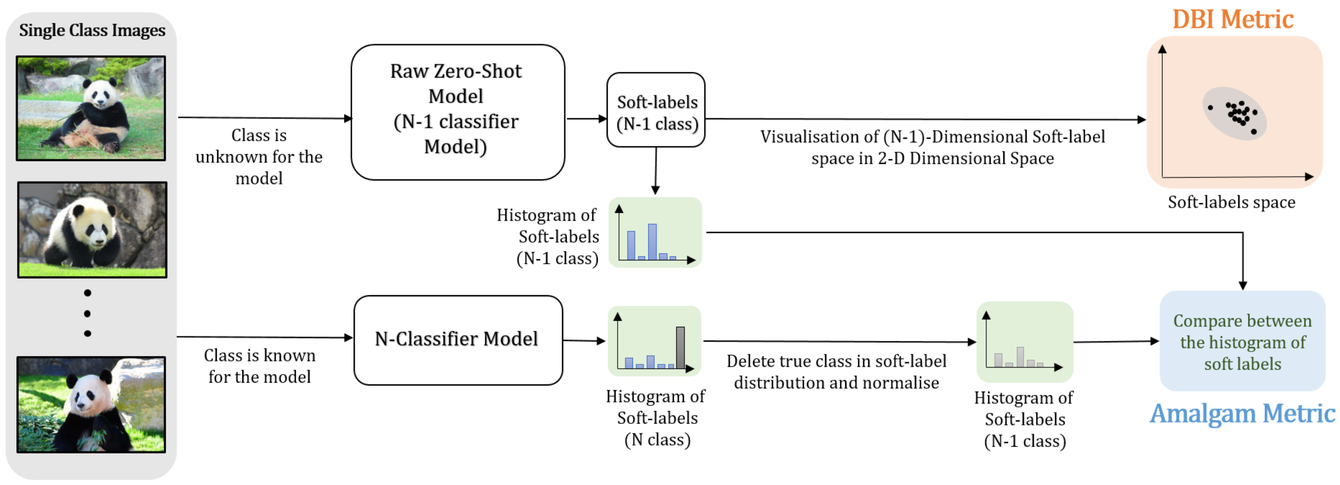
- ロバスト構造　ー　初めて、他の特別な訓練をせずにロバスト性を向上する構造が可能と紹介した。
- 文字列の脆弱性の発見　ー　ルールによって、探索をせずに、Adversarial Sampleに変更することが可能と紹介した。
- ロバスト性を持つための特徴　ー　様々な研究成果では特別な特徴を持つことでロバスト性が向上することを明らかにした。
- Adversarial Trainingにバイアスがあるため、訓練を変えることで解決できない可能性が高いと紹介した。

ロバスト性の問題を把握する研究

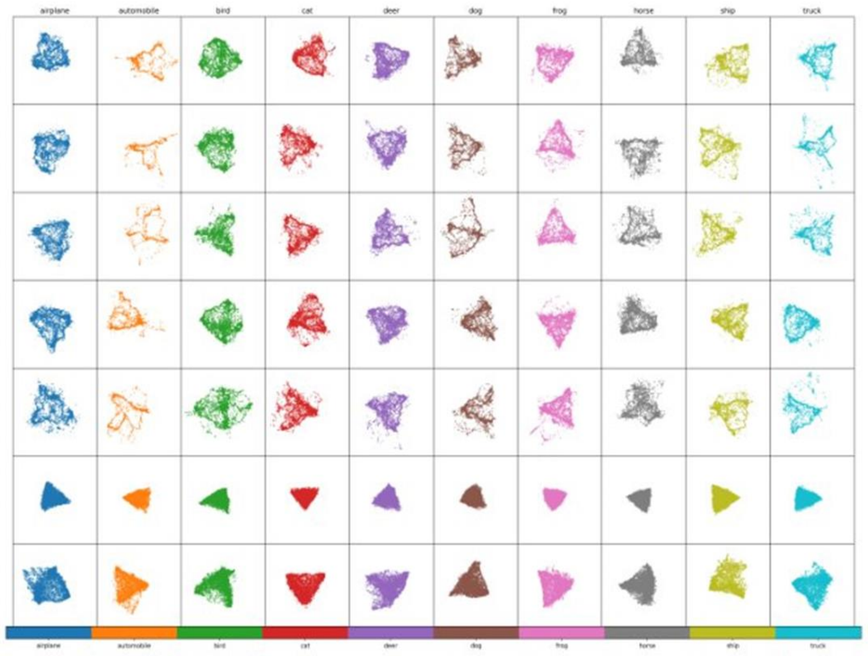
現在、深層学習のロバスト性の問題を多く理解されていない。より把握するため、ロバスト性と表現力の関連性と文字列のロバスト性を調査した。



文字列を処理する人工知能の脆弱性の発見



深層学習の表現を評価するインデクスとその理論



深層学習の出力空間を比較することで、表現力とロバスト性の関連性を明らかにした

Danilo Vasconcellos Vargas　ダニロ　ヴァスコンセロス　ヴァルガス
九州大学大学院システム情報科学研究院 情報学部門　准教授

国立研究開発法人 科学技術振興機構 Japan Science and Technology Agency