



Adversarial Attacks on Neural Networks for Graph Data

Daniel Zügner Amir Akbarnejad Stephan Günnemann Technical University of Munich, Germany

https://scholar.google.com/scholar?as_q=&num=10&btnG=Search+Scholar&as_epq=Adversarial+Attacks+on+Neural+Networks+for+Graph+Data&as_oq=&as_eq=&as_occt=any&as_sauthors=Z%C3%BCgner

KDD/ 2018, Best Paper Award

Adversarial attacks deceive model of machine learning

Summary

By Yohei Kawakami

2019/01/23



Adversarial Attacks on Neural Networks for Graph Data

Daniel Zügner Amir Akbarnejad Stephan Günnemann Technical University of Munich, Germany

https://scholar.google.com/scholar?q=&num=10&btnG=Search+Scholar&as_epq=Adversarial+Attacks+on+Neural+Networks+for+Graph+Data&as_oq=&as_eq=&as_occt=any&as_sauthors=Z%C3%BCgner

- 1, Summary
- 2, What is Adversarial attacks ...??
- 3, Experiments, conclusion, and discussion

Adversarial attacks deceive model of machine learning

Summary

By Yohei Kawakami

2019/01/23





What is Adversarial attacks ...???

What is Adversarial attacks ...??:

ニューラルネットワークにおける敵対的攻撃とは？(先行研究)


1) 先行研究 [Szegedy et al. \(2013\)](#) 内での摂動による実験



Figure 5: Adversarial examples generated for AlexNet [9].(Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example. All images in the right column are predicted to be an “ostrich, *Struthio camelus*”. Average distortion based on 64 examples is 0.006508. Please refer to <http://goo.gl/huaGPb> for full resolution images. The examples are strictly randomly chosen. There is not any postselection involved.

2) 本研究における摂動の説明 <https://www.youtube.com/watch?v=9edSZzTg3co>

Adversarial Attacks on Neural Networks for Graph Data



Machine learning model

TUM D. Zügner, A. Akbarnejad, S. Günnemann
Technical University of Munich
www.kdd.in.tum.de/nettack

KDD 2018

Adversarial Attacks [Szegedy et al. \(2013\)](#)

機械学習の分類器に、加工した画像を学習させることによって、判定の精度を落とす攻撃のこと。

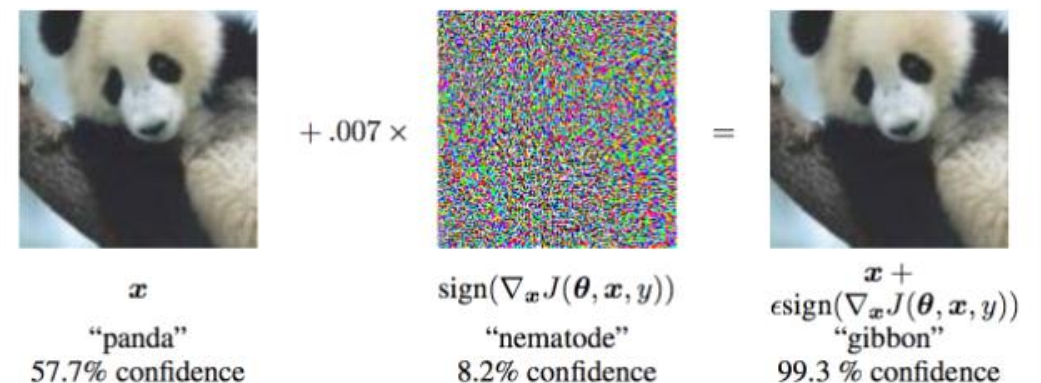
【概要】

誤認識させるための重ねる画像は、摂動(perturbation)と呼ばれ、その加工を、人間が目で判断することは極めて難しい。

【手法】

本来調整する勾配(重み)を固定し、逆に特徴量を動かして、最小値を得る特徴量を探る手法をとる。
(通常は勾配降下法で、勾配(重み)を更新し、入力を固定)

→その後の研究で、損失を最大化する方向に計算する手法も



Source: Goodfellow et al. (2014)

Conclusion: Adversarial Attackのstate-of-artな攻撃手法Nettackを開発できた。

Problem>>Farmer Work>>

(問題点/背景) 機械学習の脆弱性は、意外に知られていない。

Ex] 交通標識に敵対的攻撃が加えた場合、自動運転実用が難しくなる。機械学習を実装した防犯カメラの映像が信用できなくなる。

(各種先行研究) 攻撃側の研究と防御側の研究の両方が行われている。

Intriguing properties of neural networks (ChristianSzegedy.etc/2013)から使い古された互いに独立したデータを扱っている。

(本研究の意義) 攻撃側のstate-of-artな結果を提示する。

先行研究のモデルよりもより、分類器が誤認識するようにした。

What is this thesis for?

ニューラルネットワークにおける敵対的攻撃に関する最新技術(攻撃側)

Where is an important point compared to previous researches?

先行研究よりも機械学習モデルを誤認識させる精度があがった。
先行研究のモデルよりも複雑な設定で攻撃が可能である。
つまり、見た目にほぼ変化を加えずに攻撃する。

Where are the key points of technology and method?

- ①構造、または特徴量の摂動に対する敵対的攻撃のための、効率的、反復的、貪欲的なアルゴリズムを作成した。
- ②グラフ(化学結合)の畳み込みネットワークにおける線形化バージョンを使用した。(あらゆるノード分類モデルを攻撃できる)
- ③摂動による変化を閉形式で計算する方法を開発した。

How to verified whether it is valid?

下記二点を検証。

- ①当研究モデルが劇的に分類結果を悪化させることを統計的に示す。
- ②上記①が再構築されたモデルに転移され、攻撃されたデータの一部しか保持されない時でも、同様の結果を得ることができる。

Is there discussions?

既存のモデルの拡張を派生させて、再び堅牢にすることを目指す。(防御側)
さらに、ノード分類以外のタスクについても検討していく。(汎用性)

Which reserches should I read next?

AdverSzegedy et al. (2013), Intriguing properties of neural networks

Athalye and Sutskever (2017), Synthesizing Robust Adversarial Examples

Evtimov et al. (2017), Robust Physical-World Attacks on Deep Learning Models

敵対的攻撃の課題とは？(攻撃側)

【条件課題】

<https://www.youtube.com/watch?v=9edSZzTg3co>

①ポイズニングアタック(訓練データを不正に修正すること)は、一般的に扱いにくい二層問題を解決する必要がある。



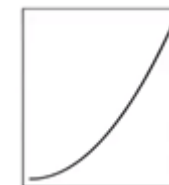
②ネットワーク内のインスタンスは互いに影響を与え合う。
あるノードへの摂動は、すべてのノードに影響を与える。



③グラフ(化学結合)は離散的な性質のため、
攻撃モデルはデータに小さな値の変更を加えることはできず、
勾配情報を頼りに作成することができない。



④可能摂動数が、はなので、徹底的な解析は難しい。



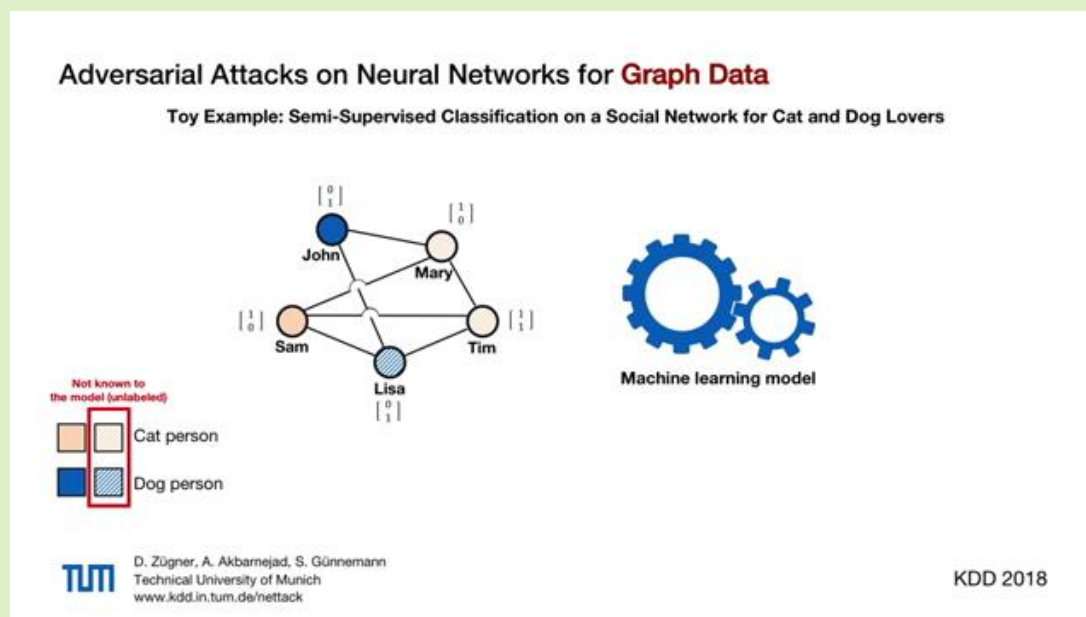
⑤摂動は、かすかで気づけないものであるべきである。



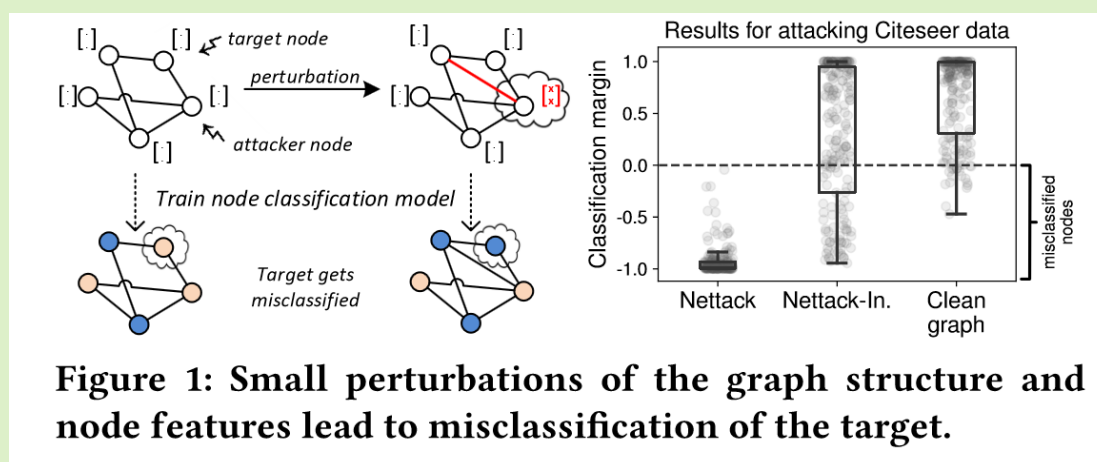
What is Nettack...??:

Nettack とは？(本研究)

3) ノードへの攻撃



4) Figure 1



NETTACK/ Adversarial Attacks

グラフデータ(化学結合)内のノードの行列を書き換える。

①構造、または特徴量の摂動に対する敵対的攻撃のための、効率的、反復的、貪欲的なアルゴリズムを作成した。

②グラフ(化学結合)の畳み込みネットワークにおける線形化バージョンを使用した。(あらゆるノード分類モデルを攻撃できる)

③摂動による変化を閉形式で計算する方法を開発した。

<https://www.youtube.com/watch?v=9edSZzTg3co>

What is Adversarial attacks ...??:

ニューラルネットワークにおける敵対的攻撃とは？(先行研究)


1) 先行研究 [Szegedy et al. \(2013\)](#) 内での摂動による実験



Figure 5: Adversarial examples generated for AlexNet [9].(Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example. All images in the right column are predicted to be an “ostrich, *Struthio camelus*”. Average distortion based on 64 examples is 0.006508. Please refer to <http://goo.gl/huaGPb> for full resolution images. The examples are strictly randomly chosen. There is not any postselection involved.

2) 本研究における摂動の説明 <https://www.youtube.com/watch?v=9edSZzTg3co>

Adversarial Attacks on Neural Networks for Graph Data



Machine learning model

TUM D. Zügner, A. Akbarnejad, S. Günnemann
Technical University of Munich
www.kdd.in.tum.de/nettack

KDD 2018

Adversarial Attacks [Szegedy et al. \(2013\)](#)

機械学習の分類器に、加工した画像を学習させることによって、判定の精度を落とす攻撃のこと。

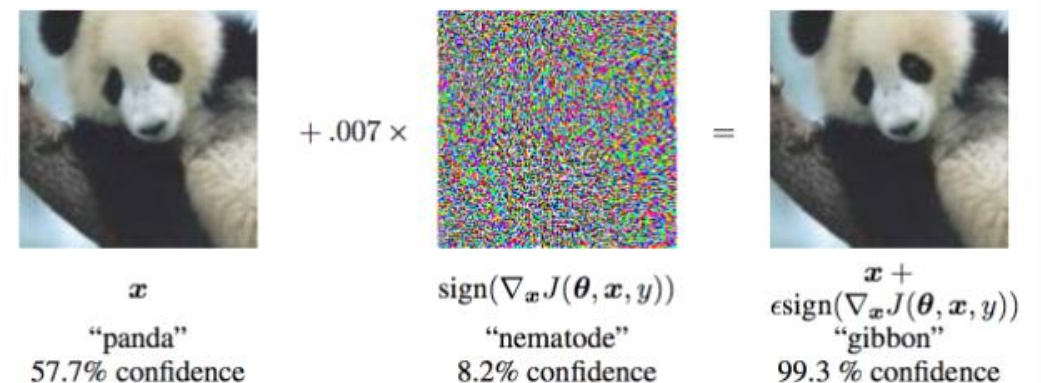
【概要】

誤認識させるための重ねる画像は、摂動(perturbation)と呼ばれ、その加工を、人間が目で判断することは極めて難しい。

【手法】

本来調整する勾配(重み)を固定し、逆に特徴量を動かして、最小値を得る特徴量を探る手法をとる。
(通常は勾配降下法で、勾配(重み)を更新し、入力を固定)

→その後の研究で、損失を最大化す方向に計算する手法も



Source: Goodfellow et al. (2014)



Two Experiments, Conclusion, Discussion

Experiment

Nettackモデルの検証：摂動により攻撃した多次元の特徴量

【使用するデータセット】

- Cora-ML
- Citeseerネットワーク
- POL. BLOGs

Dataset	N _{LCC}	E _{LCC}
CORA-ML [22]	2,810	7,981
CITESEER [29]	2,110	3,757
POL. BLOGS [1]	1,222	16,714

Table 1: Dataset statistics. We only consider the largest connected component (LCC).

ネットワークをラベル付きノード（20％）ラベルなしノード（80％）に分割。

【検証】

ラベル付けされたノードを均等部分のトレーニングと検証セットに分割して、代理モデルをトレーニングし、

- 攻撃なしデータ
- 先行研究手法による攻撃したデータ
- Nettack-In(間接的)で攻撃したデータ
- Nettackモデル(直接的)で攻撃したデータ

で正しく分類されるかどうかを検証した。

【摂動により攻撃された特徴量TOP10】

Class: neural networks				Class: theory				Class: probabilistic models			
constrained	unconstrained			constrained	unconstrained			constrained	unconstrained		
probabilistic	25	efforts	2	driven	3	designer	0	difference	2	calls	1
probability	38	david	0	increase	8	assist	0	solve	3	chemical	0
bayesian	28	averages	2	heuristic	4	disjunctive	7	previously	12	unseen	1
inference	27	accomplished	3	approach	56	interface	1	control	16	corporation	3
probabilities	20	generality	1	describes	20	driven	3	reported	1	fourier	1
observations	9	expectation	10	performing	7	refinement	0	represents	8	expressed	2
estimation	35	specifications	0	allow	11	refines	0	steps	5	robots	0
distributions	21	family	10	functional	2	starts	1	allowing	7	achieving	0
independence	5	uncertain	3	11	3	restrict	0	task	17	difference	2
variant	9	observations	9	acquisition	1	management	0	expressed	2	requirement	1

Table 2: Top-10 feature perturbations per class on Cora

Result

State-of-artな結果(攻撃)を出力した。
 先行研究と比べても、明確に識別力を低下させることができた。

<https://www.youtube.com/watch?v=9edSZzTg3co>

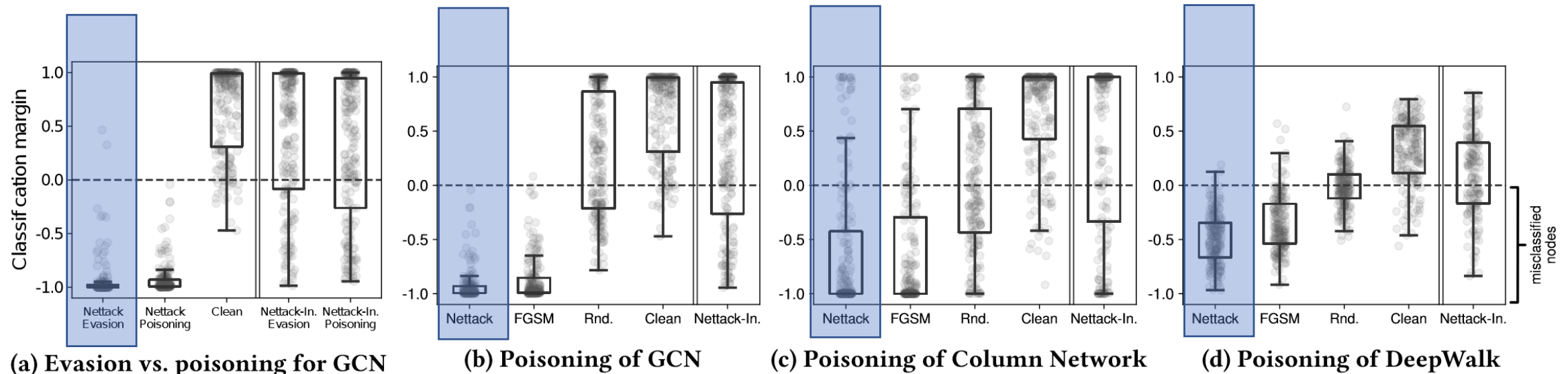


Figure 6: Results on Cora data using different attack algorithms. Clean indicates the original data. Lower scores are better.

Attack method	Cora			Citeseer			Polblogs		
	GCN	CLN	DW	GCN	CLN	DW	GCN	CLN	DW
Clean	0.90	0.84	0.82	0.88	0.76	0.71	0.93	0.92	0.63
NETTACK	0.01	0.17	0.02	0.02	0.20	0.01	0.06	0.47	0.06
FGSM	0.03	0.18	0.10	0.07	0.23	0.05	0.41	0.55	0.37
RND	0.61	0.52	0.46	0.60	0.52	0.38	0.36	0.56	0.30
NETTACK-IN	0.67	0.68	0.59	0.62	0.54	0.48	0.86	0.62	0.91

Table 3: Overview of results. Smaller is better.

Conclusion

Adversarial Attackのstate-of-artな攻撃手法を開発することができた。

Discussion

(防御側)

既存のモデルの拡張を派生させて、再び敵対的攻撃から防ぐモデルを開発する。

(汎用性)

ノード分類以外のタスクについても検討していく。

Related Work

Breaking Linear Classifiers on ImageNet

Attacking Machine Learning with Adversarial Examples

CS231n, Stanford University, Lecture 16, Ian Goodfellow

Goodfellow et al. (2014), Explaining and Harnessing Adversarial Examples

Cadieu et al. (2014), Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition

Szegedy et al. (2013), Intriguing properties of neural networks

Carlini and Wagner (2016), Towards Evaluating the Robustness of Neural Networks

Kurakin et al. (2017), Adversarial examples in the physical world

Papernot et al. (2016a), Practical Black-Box Attacks against Machine Learning

Papernot et al. (2016b), Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Papernot et al. (2016c), Towards the Science of Security and Privacy in Machine Learning

Papernot et al. (2015a), Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks

Papernot et al. (2015b), The Limitations of Deep Learning in Adversarial Settings

Papernot and McDaniel (2017), Extending Defensive Distillation

Hinton et al. (2015), Distilling the Knowledge in a Neural Network

Carlini and Wagner (2016), Defensive Distillation is Not Robust to Adversarial Examples

Huang et al. (2017), Adversarial Attacks on Neural Network Policies

Lu et al. (2017a), NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles

Lu et al. (2017b), Standard detectors aren't (currently) fooled by physical adversarial stop signs

Szegedy et al. (2013), Intriguing properties of neural networks

Athalye and Sutskever (2017), Synthesizing Robust Adversarial Examples

Evtimov et al. (2017), Robust Physical-World Attacks on Deep Learning Models

Adversarial Attacks on Neural Networks for Graph Data

Daniel Zügner Amir Akbarnejad Stephan Günnemann Technical University of Munich, Germany

https://scholar.google.com/scholar?as_q=&num=10&btnG=Search+Scholar&as_epq=Adversarial+Attacks+on+Neural+Networks+for+Graph+Data&as_oq=&as_eq=&as_occt=any&as_sauthors=Z%C3%BCgner

KDD/ 2018, Best Paper Award

Adversarial attacks deceive model of machine learning

Summary

By Yohei Kawakami

2019/01/23

