Exposé zur Masterarbeit

Adversarial Attacks on Recommender Systems

Philipp Normann (its103541)

9. Juni 2020

1 Motivation

Als einer der größten deutschen Onlinehändler, setzt auch OTTO (GmbH & Co KG) verstärkt auf lernende Empfehlungssysteme zur gezielten Kundenansprache. Jedoch sind genau diese lernenden Systeme, welche unsere täglichen Entscheidungen beeinflussen, Bedrohungen von feindlichen Akteuren ausgesetzt. Diese Akteure versuchen durch die gezielte Manipulation der Trainings- oder Inputdaten, die Empfehlungssysteme zu Ihren Gunsten auszunutzen (z.B. zur Promotion ihrer eignen Produkte oder zu Manipulation des öffentlichen Meinungsbildes). Da eine erfolgreiche Kompromittierung solcher Systeme, je nach Anwendungsgebiet, weitreichende Auswirkungen haben kann, ist ein besseres Verständnis möglicher Angriffe und Verteidigungen essentiell. Die allgemeine Anfälligkeit lernender Systeme für *feindliche Beispiele* ist seit 2013 bekannt [Big+13; Sze+14; GSS14] und hat sich mittlerweile als ein eigenes Forschungsgebiet (*adversarial learning*) etabliert. Bisherige Arbeiten im Bezug auf Empfehlungssysteme haben sich vor allem damit beschäftigt, Collaborative Filtering Methoden, durch das einschleusen von Fake-Profilen, zu beeinflussen [OHS02; CB19] und diese Fake-Profile durch Methoden der Anomalieerkennung aufzudecken [CNZ05; WM06]. Inwieweit auch Inhaltsbasierten Empfehlungen von gegnerischen Bedrohungen betroffen sind und welche Methoden es dazu gibt, um den Einfluss *feindlicher Beispiele* zu reduzieren, wurde bisher wenig Aufmerksamkeit geschenkt und soll daher im Rahmen dieser Arbeit genauer untersucht werden.

2 Zielsetzung

Das Ziel der Arbeit soll es sein, Modelle zur inhaltsbasierten Empfehlungen, auf ihre potenzielle Anfälligkeit gegen populäre Angriffe aus dem Bereich *Adversarial Learning* zu testen und geeignete Verteidigungsmaßnahmen zu implementieren und zu evaluieren.

3 Optionale Zusätze

Falls ich nach der Zielerreichung noch ausreichend Zeit habe, werde ich zusätzlich einen größeren Fokus auf die Methoden des kollaborativen Filterns setzten und dort ebenfalls Angriffe und Verteidigungen für diese Art von Systemen reproduzieren und evaluieren.

4 Herangehensweise

4.1 Grundlagen

Zuerst sollen die notwendigen theoretischen Grundlagen erarbeitet werden. Dazu gehört eine allgemeine Einführung in die Thematik der *Adverserial Attacks* gegen tiefe Neuronale Netzwerke und eine Aufschlüsselung der Ziele und der verscheidenden Arten von Empfehlungssystemen.

4.2 Bedrohungsanalyse

Bevor mit der Exploration möglicher Angriffe und Verteidigungen begonnen wird, soll jeweils eine strukturierte Bedrohungsanalyse für inhaltsbasierte und kollaborative Empfehlungssysteme vorgenommen werden, um ein besseres Verständnis zur Bedrohungslandschaft und zu schützender Assets zu erlangen. Die hierbei bestimmten Bedrohungen sollen als Grundlage für die Erarbeitung von Angriffen und Verteidigungen dienen.

4.3 Datengrundlage

Als Benchmark für alle folgenden Experimente sollen vorwiegend öffentlich zugängliche Datensätze verwendet werden, um eine Vergleichbarkeit und Reproduzierbarkeit zu gewährleisten. Dazu bietet sich z.B. der *MovieLens* Datensatz von der University of Minnesoata [HK15] oder der *Netflix* Datensatz [BL+07] an.

4.4 Angriffe

Es soll eine Auswahl populärer *white-box* und *black-box* Angriffe implementiert und evaluiert werden. Erst für das klassische Problem der Handschrifterkennung (MNIST) [LC10] und danach für verbreitete inhaltsbasierte Empfehlungssysteme.

Eine Auswahl möglicher Angriffe:

1. Projected Gradient Descent (PGD) [KGB16]
2. Fast Gradient Sign Method (FGSM) [GSS14]
3. Transfer Attacks [Pap+17]
4. Carlini & Wagner Attacks (CW) [CW17]
5. ...

4.5 Verteidigungen

Gegen die entwickelten Angriffe soll ebenfalls nach angemessenen Verteidigungen gesucht werden. Diese sollen dann ebenfalls zuerst auf MNIST und danach auf inhaltsbasierte Empfehlungssysteme angewandt und evaluiert werden.

Eine Auswahl möglicher Verteidigungen:

1. Adversarial Training [Tra+17]
2. Thermometer Encoding [Buc+18]
3. ...

4.6 Fazit

Nach der Implementierung und Evaluierung der Angriffe und Verteidigungen, soll ein Fazit bezüglich der Gefährdung von Empfehlungssystemen gezogen werden und ein Ausblick auf zukünftige Forschungsfragen gegeben werden.

5 Verwandte Arbeiten

- Adversarial attacks on an oblivious recommender [CB19]
- Adversarial Recommendation: Attack of the Learned Fake Users [CB18]
- Poisoning attacks to graph-based recommender systems [Fan+18]
- Preventing shilling attacks in online recommender systems [CNZ05]

Literatur

[Big+13] Battista Biggio u. a. "Evasion attacks against machine learning at test time". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2013, S. 387–402. URL: https://link.springer.com/content/pdf/10.1007/978-3-642-40994-3_25.pdf (besucht am 01. 04. 2020).

[BL+07] James Bennett, Stan Lanning u. a. "The netflix prize". In: *Proceedings of KDD cup and workshop*. Bd. 2007. Citeseer. 2007, S. 35. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.8094&rep=rep1&type=pdf (besucht am 01. 04. 2020).

[Buc+18] Jacob Buckman u. a. "Thermometer encoding: One hot way to resist adversarial examples". In: (2018). URL: https://openreview.net/pdf?id=S18Su--CW (besucht am 01. 04. 2020).

[CB18] Konstantina Christakopoulou und Arindam Banerjee. “Adversarial recommendation: Attack of the learned fake users”. In: *arXiv preprint arXiv:1809.08336* (2018). URL: https://arxiv.org/pdf/1809.08336.pdf (besucht am 01. 04. 2020).

[CB19] Konstantina Christakopoulou und Arindam Banerjee. “Adversarial attacks on an oblivious recommender”. In: (2019), S. 322–330. URL: https://www-users.cs.umn.edu/~baner029/papers/19/adv_attack.pdf (besucht am 01. 04. 2020).

[CNZ05] Paul-Alexandru Chirita, Wolfgang Nejdl und Cristian Zamfir. “Preventing shilling attacks in online recommender systems”. In: *Proceedings of the 7th annual ACM international workshop on Web information and data management*. 2005, S. 67–74. URL: https://www.researchgate.net/profile/Cristian_Zamfir/publication/220759092_Preventing_shilling_attacks_in_online_recommender_systems/links/00b49536abca48b9cf000000.pdf (besucht am 01. 04. 2020).

[CW17] Nicholas Carlini und David Wagner. “Towards evaluating the robustness of neural networks”. In: *2017 ieee symposium on security and privacy (sp)*. IEEE. 2017, S. 39–57. URL: https://arxiv.org/pdf/1608.04644.pdf (besucht am 01. 04. 2020).

[Fan+18] Minghong Fang u. a. “Poisoning attacks to graph-based recommender systems”. In: *Proceedings of the 34th Annual Computer Security Applications Conference*. 2018, S. 381–392. URL: https://dl.acm.org/doi/pdf/10.1145/3274694.3274706.

[GSS14] Ian J Goodfellow, Jonathon Shlens und Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014). URL: https://arxiv.org/pdf/1412.6572.pdf) (besucht am 01. 04. 2020).

[HK15] F Maxwell Harper und Joseph A Konstan. “The movielens datasets: History and context”. In: *Acm transactions on interactive intelligent systems (tiis)* 5.4 (2015), S. 1–19. URL: https://dl.acm.org/doi/pdf/10.1145/2827872 (besucht am 01. 04. 2020).

[KGB16] Alexey Kurakin, Ian Goodfellow und Samy Bengio. “Adversarial machine learning at scale”. In: *arXiv preprint arXiv:1611.01236* (2016). URL: https://arxiv.org/pdf/1611.01236 (besucht am 01. 04. 2020).

[LC10] Yann LeCun und Corinna Cortes. “MNIST handwritten digit database”. In: (2010). URL: http://yann.lecun.com/exdb/mnist/ (besucht am 01. 04. 2020).

[OHS02] Michael P O’Mahony, Neil J Hurley und Guenole CM Silvestre. “Promoting recommendations: An attack on collaborative filtering”. In: *International Conference on Database and Expert Systems Applications*. Springer. 2002, S. 494–503. URL: http://ftp10.us.freebsd.org/users/azhang/disc/springer/0558/papers/2453/24530494.pdf (besucht am 01. 04. 2020).

[Pap+17] Nicolas Papernot u. a. “Practical black-box attacks against machine learning”. In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017, S. 506–519. URL: https://dl.acm.org/doi/pdf/10.1145/3052973.3053009 (besucht am 01. 04. 2020).

[Sze+14] Christian Szegedy u. a. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2014). URL: https://arxiv.org/pdf/1312.6199.pdf (besucht am 01. 04. 2020).

[Tra+17] Florian Tramèr u. a. “Ensemble adversarial training: Attacks and defenses”. In: *arXiv preprint arXiv:1705.07204* (2017). URL: https://arxiv.org/pdf/1705.07204.pdf) (besucht am 01. 04. 2020).

[WM06] Chad Williams und Bamshad Mobasher. “Profile injection attack detection for securing collaborative recommender systems”. In: *DePaul University CTI Technical Report* (2006), S. 1–47. URL: https://www.researchgate.net/profile/Bamshad_Mobasher/publication/228815801_Thesis_Profile_Injection_Attack_Detection_for_Securing_Collaborative_Recommender_Systems/links/0fcfd507477e7b5f5f000000.pdf (besucht am 01. 04. 2020).