機械学習システム
セキュリティガイドライン

「付録：攻撃検知技術の概要」

Version 0.8

2022 年 4 月 4 日

機械学習システムセキュリティガイドライン策定委員会

機械学習システムセキュリティ・セーフティワーキンググループ

日本ソフトウェア科学会 機械学習工学研究会

目次

1. はじめに

機械学習システムには、判断を誤らせたり情報を奪ったりするような、機械学習特有の攻撃が存在するため、従来のセキュリティ対策だけでなく、機械学習特有の攻撃への対策も必要となる。対策として、攻撃に堅牢なシステムを構築することに加え、攻撃を検知することも重要となる。

機械学習への攻撃を検知する手法は多く提案されているが、特定のデータやタスクに対してのみ有効な場合も多く、決定的な検知手法はまだ確立されていない。Carlini らは敵対的サンプルの検知について既存の代表的な手法を評価し、検知手法を知った上での攻撃に対しては十分な検知性能が得られないことを示している [1]。また、Kumar らは機械学習セキュリティの課題の一つに攻撃検知を挙げており、セキュリティ分析者の間で検知に関する知見の共有ができていないことについて言及している [2]。したがって現状では、機械学習セキュリティの専門知識が無ければ、適切な検知手法を実装することは容易ではないと言える。

一般的なサイバー攻撃対策においては、攻撃戦略や手法を体系化した MITRE ATT&CK [3]などのフレームワークがあり、検知手法を検討する際にも活用されている [4]。機械学習システムへの攻撃については、MITRE ATLAS [5]において機械学習特有の攻撃戦略（攻撃に必要な情報収集などの偵察活動、攻撃の実行など）が整理されており、検知対象とする攻撃の選定の参考になる。一方で、著者が知る限り、機械学習特有の攻撃を検知する手法を攻撃戦略に沿ってまとめた文献はない。そこで本付録では、攻撃戦略を「偵察活動や攻撃のための下準備」（攻撃の前兆）と、「実際に誤判定を引き起こさせたり情報を奪ったりする攻撃」（攻撃の兆候）の 2 段階とし、それぞれを検知する手法を前兆検知と兆候検知に分けて整理した。また、検知に使用するデータが機械学習システムにおいて取得可能かどうかも検知手法選定時には重要な情報となるため、使用するデータ（訓練データ、訓練済みモデルなど）の観点からも検知手法を整理した。

本付録は、開発者やセキュリティ分析者の検知手法選定を支援することを目的とし、攻撃検知に関する文献を攻撃戦略と使用するデータの観点から整理した。なお、対象とするシステムは画像分類システムとし、対象とする攻撃は回避攻撃、ポイズニング攻撃、モデル抽出攻撃とする（対象範囲は今後拡張していく予定）。

2. 攻撃戦略毎の検知手法について

本付録では、攻撃戦略を「偵察活動や攻撃のための下準備」と、「実際に誤判定を引き起こさせたり情報を奪ったりする攻撃」の 2 段階に分けて考える。前者を攻撃の前兆、後者を攻撃の兆候と呼ぶこととし、それぞれの検知を前兆検知と兆候検知とする。本節では回避攻撃、ポイズニング攻撃、モデル抽出攻撃の検知手法を前兆検知と兆候検知という 2 つの検知目的に分けて整理した（整理した一覧表は表 1 に示す）。

2.1. 前兆検知

2.1.1. 回避攻撃の検知

画像分類を行う機械学習システムの場合、攻撃者は入力画像に対して摂動と呼ばれる小さなノイズを加えることで、機械学習モデルを誤認識させることができる [6]。このような摂動を加えて誤認識させるようなデータを敵対的サンプルと呼ぶ（敵対的サンプルに関する論文リスト（arXiv）が Nicholas Carlini

によってまとめられている［7]）。

敵対的サンプルの作成方法は攻撃者が有している知識によっても異なる。例えば、攻撃対象のシステムに対する知識を持っていない場合は、システムに対して一連のクエリを実行して敵対的サンプルを作成する方法がある［8]。一方、攻撃対象システムに対する知識を有している場合は、攻撃者の手元で攻撃対象モデルを再現し、攻撃対象システムへのクエリを実行せずに敵対的サンプルを作成できる［9]。前者のようにクエリ実行を通して敵対的サンプルを作成しようとする活動は、攻撃の前兆として攻撃が成功（敵対的サンプル作成）する前に検知することが望ましい。検知手法について以下に示す。

■　敵対的サンプルを作成しようとする操作を検知する

攻撃者が敵対的サンプルを作成するためには、複数の類似するデータを入力する可能性が高いため、攻撃者が入力する一連のクエリが正常ユーザとは異なる分布になることを利用する検知手法が提案されている［10, 11]。また、このような攻撃はクエリ数が正常ユーザと比べて多くなる場合や、出力ラベルに偏りが出る場合が多いため、単位時間当たりのクエリ頻度や出力ラベルの分布をモニタリングする等、単純な方法で検知できる可能性もある。

2.1.2.　ポイズニング攻撃の検知

ポイズニング攻撃には、訓練データを汚染することによりモデルの推論精度を意図的に劣化させる攻撃［12]と、訓練データにバックドアデータを仕込み、特定の入力データを攻撃者が意図したクラスに誤分類させるようなバックドア攻撃がある［13]。訓練データの準備やモデル作成のアウトソーシング［14]、信頼できない Web サイトからのデータ収集、連合学習［15]や転移学習［16]によってデータやモデルが汚染される恐れがある［17]。

これらは攻撃の前兆ととらえることができ、テスト段階、もしくはそれよりも早期に検知することが望ましい。運用中の入力データを訓練データに加えて再学習するような場合も、入力されたデータが汚染されていないかを確認してから再学習する必要がある。検知手法について以下に示す。

■　精度劣化を引き起こすような訓練データセットの汚染を検知する

訓練用に収集したデータに、精度劣化を引き起こすようなデータが含まれていないかを検知する手法が提案されている［18, 19]。

■　訓練データセットのバックドアを検知する

訓練データセットに含まれるバックドアデータを検知する手法が提案されている［20, 21]。文献［17]はバックドア攻撃やその対策に関する包括的なレビューであり、検知手法に関しても整理されているため参照されたい。

■　訓練済みモデルが汚染されていることを検知する

外部から取得した訓練済みモデル等が汚染されていないかを検知する手法が提案されている［22, 23]。

2.2. 兆候検知

2.2.1. 回避攻撃の検知

■ 敵対的サンプルの入力を検知する

前述の通り、攻撃者が攻撃対象システムに対する知識を有している場合などは、何らかの方法で敵対的サンプルを作成し、システムに入力してくる可能性がある。このような攻撃を検知するためには、入力データ、モデルの出力結果、中間層出力のデータを分析することが多い [24-32]。ただし、敵対的サンプルの入力検知は一般的に難しいタスクと言われており [1]、あらゆるシステムに有効な検知手法はないため、可能な限り複数の検知手法を適用することを推奨する。

敵対的サンプルの検知手法に関する包括的なレビューは文献 [33]を参照すること.

2.2.2. ポイズニング攻撃の検知

■ 入力データにバックドアのトリガーが含まれるかを検知する

運用時にトリガーが含まれる画像が入力されていないかを検知する手法が提案されている [34]。

2.2.3. モデル抽出攻撃の検知

モデル抽出攻撃は、システムへの複数クエリをもとにモデルの入出力のペアを取得し、取得した情報をもとにシステムで使用されているモデルと似たふるまいをするモデルを作成する攻撃である。攻撃者にとってモデル抽出自体が目的である場合もあれば、他の攻撃に用いるためにモデル抽出攻撃を行う場合もある [35]（後者の場合は攻撃の前兆とも言える)。モデル抽出攻撃を検知する手法を以下に示す。

■ モデル抽出攻撃を目的とした異常なクエリを検知する

モデル抽出攻撃を実施する際は正常ユーザとは異なるログになる可能性が高いため、それを利用して検知する手法が提案されている [36-39]。また、このような攻撃はクエリ数が正常ユーザと比べて多くなる場合があるため、単位時間当たりのクエリ頻度をモニタリングする等、単純な方法で検知できる可能性もある。

3. 検知に使用するデータについて

「検知手法を実装する際に使用するデータ」の観点から既存の検知手法を整理する。ここでいうデータは、訓練データ、訓練済みモデル、運用中の入力データ、モデルの出力データの 4 つを考える。また、これらのデータは、検知手法によって、「検知・分析の対象となるデータ」と、「攻撃の痕跡はないが、検知・分析する際に必要なデータ」に分けることができる。前者は日時やアカウント等と紐づけて管理する必要があり（前者の例：敵対的サンプルの摂動の痕跡が残っている可能性がある入力画像)、後者は検知手法を実装する際に必要となるため適切に管理する必要がある（後者の例：攻撃判定に用いる閾値を算出するために用いる訓練データ)。上記の観点から検知手法を整理したものを表 1 に示す。なお、出力データはモデルが出力する確信度や中間層出力等を指し、検知手法によって異なるため、詳細については各文献を参照されたい。

表 1. 検知に使用するデータについて

●：検知・分析の対象となるデータ

□：攻撃の痕跡はないが、検知・分析する際に必要なデータ

検知目的	検知対象攻撃	検知手法	開発		運用	
			訓練データ	訓練済モデル	入力データ	出力データ
前兆検知	敵対的サンプル作成の検知	Chen et al. [10]	□		●	
		Li et al. [11]	□		●	
	データ汚染（精度劣化）の検知	Müller et al. [18]	●	□		
		Tavallali et al. [19]	●			
	バックドアの検知	Chen et al. [20]	●	□		
		Hayase et al. [21]	●	□		
		Dong et al. [22]		●		
		Huster et al. [23]		●		
兆候検知	敵対的サンプル入力の検知	Hendrycks et al. [24]	□		●	
		Meng et al. [25]	□		●	
		Grosse et al. [26]	□		●	
		Gong et al. [27]	□	□	●	
		Lu et al. [28]	□	□		●
		Feinman et al. [29]	□	□		●
		Aigrainet al. [30]	□	□		●
		Xu et al. [31]	□	□	●	●
		Monteiro et al. [32]	□	□	●	●
	バックドアトリガーの検知	Kiourti et al. [34]	□	□	●	●
	モデル抽出攻撃の検知	Juuti et al. [36]			●	
		Pal et al. [37]	□		●	
		Atli et al. [38]	□		●	
		Sadeghzadeh et al. [39]	□		●	

4. まとめ

本付録では、開発者やセキュリティ分析者の検知手法選定を支援することを目的とし、回避攻撃、ポイズニング攻撃、モデル抽出攻撃の検知に関する文献を攻撃戦略と使用するデータの観点から整理した。攻撃戦略に従って検知手法を整理することによって、一連の攻撃をその前兆と兆候で多段階に検知する場合や、可能な限り早い段階で攻撃を検知したい場合等の参考になると考えられる。また、使用するデータの観点から検知手法を整理することによって、開発するシステムにおいて取得するデータに制約がある場合でも（例：入力データを取得できない)、取得できるデータのみを用いた検知手法を選定できると考えられる。既存のレビュー文献 [17, 33]等とも合わせて検知手法選定時に参照されたい。

参照文献

[1] N. Carlini , D. Wagner, “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods,” In Workshop on Artificial Intelligence and Security, 2017.

[2] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann , S. Xia, “Adversarial Machine Learning -- Industry Perspectives,” In IEEE Security and Privacy Workshops, 2020.

[3] “ATT&CK,” MITRE, [オンライン]. Available: https://attack.mitre.org/.

[4] B. E. Strom, J. A. Battaglia, M. S. Kemmerer, W. Kupersanin, D. P. Miller, C. Wampler, S. M. Whitley , a. R. D. Wolf, “Finding Cyber Threats with ATT&CK™-Based Analytics, ” The MITRE Corporation, Bedford, MA, Technical Report No. MTR170202, 2017.

[5] “ATLAS,” MITRE, [オンライン]. Available: https://atlas.mitre.org/.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow , R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[7] N. Carlini, “ A Complete List of All (arXiv) Adversarial Example Papers, ” https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html, 2019.

[8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi , C.-J. Hsieh, “ ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models, ” In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26, 2017.

[9] I. J. Goodfellow, J. Shlens , C. Szegedy, “Explaining and Harnessing Adversarial Examples,” In International Conference on Learning Representations, 2015.

[10] S. Chen, N. Carlini , D. Wagner, “Stateful Detection of Black-Box Adversarial Attacks,” In Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence, pp.30-39, 2019.

[11] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng , B. Y. Zhao, “Blacklight: Defending Black-Box Adversarial Attacks on Deep Neural Networks, ” arXiv preprint arXiv:2006.14042, 2020.

[12] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru , B. Li, “Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning,” In IEEE Security and Privacy, 2018.

[13] T. Gu, B. Dolan-Gavitt , S. Garg, “BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,” In Proceedings of Machine Learning and Computer

Security Workshop, 2017.

[14] Y. Chen, X. Gong, Q. Wang, X. Di , H. Huang, “Backdoor Attacks and Defenses for Deep Neural Networks in Outsourced Cloud Environments,” IEEE Network, vol. 34, no. 5, pp. 141–147, 2020.

[15] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin , V. Shmatikov, “How To Backdoor Federated Learning,” In International Conference on Artificial Intelligence and Statistics, 2020.

[16] Y. Ji, Z. Liu, X. Hu, P. Wang , Y. Zhang, “Programmable Neural Network Trojan for Pre-Trained Feature Extractor,” arXiv preprint arXiv:1901.07766, 2019.

[17] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal , H. Kim, “Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review,” arXiv preprint arXiv:2007.10760, 2020.

[18] N. M. Müller, S. Roschmann , K. Böttinger, “Defending Against Adversarial Denial-of-Service Data Poisoning Attacks,” arXiv preprint arXiv:2104.06744, 2021.

[19] P. Tavallali, V. Behzadan, P. Tavallali , M. Singhal, “Adversarial Poisoning Attacks and Defense for General Multi-Class Models Based On Synthetic Reduced Nearest Neighbors,” arXiv preprint arXiv:2102.05867, 2021.

[20] J. Chen, X. Zhang, R. Zhang, C. Wang , L. Liu, “De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks,” In IEEE Transactions on Information Forensics and Security, 2021.

[21] J. Hayase, W. Kong, R. Somani , S. Oh, “SPECTRE: Defending Against Backdoor Attacks Using Robust Statistics,” In Proceedings of the 38th International Conference on Machine Learning, 2021.

[22] Y. Dong, X. Yang, Z. Deng, T. Pang, Z. Xiao, H. Su , J. Zhu, “Black-box Detection of Backdoor Attacks with Limited Information and Data,” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[23] T. Huster , E. Ekwedike, “TOP: Backdoor Detection in Neural Networks via Transferability of Perturbation,” arXiv preprint arXiv:2103.10274, 2021.

[24] D. Hendrycks , K. Gimpel, “Early Methods for Detecting Adversarial Images,” In Proceedings of the International Conference on Learning Representations, 2017.

[25] D. Meng , H. Chen, “MagNet: a Two-Pronged Defense against Adversarial Examples,” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 135–147, 2017.

[26] K. Grosse, P. Manoharan, N. Papernot, M. Backes , P. McDaniel, “On the (Statistical)

Detection of Adversarial Examples," arXiv preprint arXiv:1702.06280, 2017.

[27] Z. Gong, W. Wang , W.-S. Ku, "Adversarial and Clean Data Are Not Twins," arXiv preprint arXiv:1704.04960, 2017.

[28] J. Lu, T. Issaranon , D. Forsyth, "SafetyNet: Detecting and Rejecting Adversarial Examples Robustly, " In Proceedings of the IEEE International Conference on Computer Vision, pp. 446–454, 2017.

[29] R. Feinman, R. R. Curtin, S. Shintre , A. B. Gardner, "Detecting Adversarial Samples from Artifacts," arXiv preprint arXiv:1703.00410, 2017.

[30] J. Aigrain , M. Detyniecki, " Detecting Adversarial Examples and Other Misclassifications in Neural Networks by Introspection," In Proceedings of the 35th International Conference on Machine Learning, pp. 7167–7177, 2019.

[31] W. Xu , Y. Q. David Evans, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," In Proceedings of Network and Distributed System Security Symposium, 2018.

[32] J. Monteiro, I. Albuquerque, Z. Akhtar , T. H. Falk, "Generalizable Adversarial Examples Detection Based on Bi-model Decision Mismatch, " In 2019 IEEE International Conference on Systems, Man and Cybernetics, pp. 2839–2844, 2019.

[33] A. Aldahdooh, W. Hamidouche, S. A. Fezza , O. Deforges, "Adversarial Example Detection for DNN Models: A Review and Experimental Comparison," arXiv preprint arXiv:2105.00203, 2021.

[34] P. Kiourti, W. Li, A. Roy, K. Sikka , S. Jha, "MISA: Online Defense of Trojaned Models using Misattributions," arXiv preprint arXiv:2103.15918, 2021.

[35] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik , A. Swami, "Practical Black-Box Attacks against Machine Learning," In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pp. 506–519, 2017.

[36] M. Juuti, S. Szyller, S. Marchal , N. Asokan, "PRADA: Protecting against DNN Model Stealing Attacks," In IEEE European Symposium on Security & Privacy, pp. 512–527, 2019.

[37] S. Pal, Y. Gupta, A. Kanade , S. Shevade, "Stateful Detection of Model Extraction Attacks," arXiv preprint arXiv:2107.05166, 2021.

[38] B. G. Atli, S. Szyller, M. Juuti, S. Marchal , N. Asokan, "Extraction of Complex DNN Models: Real Threat or Boogeyman?," In International Workshop on Engineering Dependable and Secure Machine Learning Systems. Springer, pp. 42–57, 2020.

[39] A. M. Sadeghzadeh, F. Dehghan, A. M. Sobhanian , R. Jalili, "HODA: Hardness-

Oriented Detection of Model Extraction Attacks," arXiv preprint arXiv:2106.11424, 2021.

機械学習システムセキュリティガイドライン策定委員会メンバーリスト

市原　大暉　　（株式会社 NTT データ）
及川　孝徳　　（富士通株式会社）
大久保　隆夫　（情報セキュリティ大学院大学）
笠原　史禎　　（富士通株式会社）
金子　朋子　　（国立情報学研究所）
久連石　圭　　（株式会社 東芝）
田口　研治　　（国立情報学研究所）
林　昌純　　（法政大学）
森川　郁也　　（富士通株式会社）
矢嶋　純　　（富士通株式会社）
吉岡　信和　　（早稲田大学）
（敬称略・五十音順）