Annotated Bibliography

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. *IEEE Symposium On Security And Privacy, 41*. doi:10.1109/SP.2017.41.

Membership inference, is determining whether a given data record was part of the machine learning model's training dataset or not. To protect the privacy of the individuals whose data was used to train the model, the authors, researchers at the Cornell Tech, use five datasets including a sensitive hospital discharge dataset, to demonstrate membership inference attack and the membership risk that a person incurs if they allow their data to be used to train a model. Their membership inference attack exploits the observation that machine learning models (target models) often behave differently on the data that they were trained on versus the data that they "see" for the first time. After evaluating their inference attacks on three target models one each created from Google Prediction API, Amazon ML and one implemented locally it was observed that the attacker does not require any prior knowledge about the distribution of the target model's training data. The designed attack can be used as one of the selection metrics to decide the type of the model to train or a machine learning service to use. This paper provides a clear architecture to build an inference attack but does not explore in detail the reasons for such attacks.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M. (2019). ML-Leaks:

Model and data independent membership inference attacks and defenses on machine learning models. *Network And Distributed Systems Security Symposium*. doi:10.14722/ndss.2019.23119.

This paper, using eight different datasets effectively discusses the shortcomings of existing membership inference attacks as they make strong assumptions on the adversary, such as using multiple so-called shadow models, knowledge of the target model structure, and having a dataset from the same distribution as the target model's training data. In this paper, the authors gradually relax these assumptions thereby showing that such attacks are very broadly applicable at low cost and thereby pose a more severe risk than previously thought and proposes two effective defense mechanisms. The effectiveness of their membership inference attacks is mainly due to the overfitting nature of ML models and therefore, the proposed defense techniques are designed to increase ML models' generalizability, i.e., prevent them from being overfitted. The first technique is dropout which is designed for neural network-based classifiers and the second technique is model stacking which is suitable for all ML models, independent of the classifier used to build them.

Carlini, N., Liu, C., Erlingsson. U., Kos, J., Song, D. (2019). The secret sharer: evaluating and testing unintended memorization in neural networks. *USENIX Security Symposium*, 267-284.

Disclosure of secrets may arise naturally in generative text models like those used for text auto-completion and predictive keyboards, if trained on possibly-sensitive data. The users of such models may discover—either by accident or on purpose—that entering certain text prefixes causes the models to output surprisingly-revealing text completions. The authors demonstrated unintended memorization is a persistent, hard-to-avoid issue that can have serious consequences using a two-layer LSTM with 200 hidden units trained on the Penn Treebank dataset. To enable practitioners to measure their models' propensity for disclosing details about private training data, this paper introduces a quantitative Exposure metric using log-perplexity. The authors also illustrated through experiments that unintended memorization is not due to the model overtraining to the training data. The paper does not propose methods to prevent memorization however shows that the three potential defenses against memorization: regularization, sanitization, and differential privacy are not a perfect defence mechanism for memorization.

Nasr, M., Shokri, R., Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. *IEEE Symposium On Security And Privacy.* doi: 10.1109/SP.2019.00065.

Federated learning is a distributed ML approach which enables model training on a large corpus of decentralized data. White box adversarial attacks describe scenarios in which the attacker has access to the underlying training policy network of the target model. The authors designed white-box inference attacks that exploit the privacy vulnerabilities of the stochastic gradient descent (SGD) algorithm as in SGD each data point in the training dataset influences the model parameters thereby leaving a distinguishable footprint on the gradients of the loss functions. The effectiveness of the proposed white box approach was evaluated using pre-trained and publicly available state of the art models on the CIFAR100 dataset, Purchase100 and Texas100 dataset. The results show that even the best models which was not vulnerable to black-box attack, however is susceptible to white box membership inference attacks. The authors also show that in a federated learning a curious parameter server or even a participant can perform alarmingly accurate membership inference attacks against other participants.

Ma, S., Liu, Y., Tao, G., Lee, W., Zhang, X. (2019). NIC: Detecting adversarial samples with neural network invariant checking. *Network And Distributed Systems Security Symposium.* doi:10.14722/ndss.2019.23415.

Adversarial samples are inputs to a neural network that result in an incorrect output from the network. The authors, researchers at the Purdue University after analyzing the internals of individual layers of eleven Deep Neural Network (DNN) under various attacks identified that adversarial samples mainly exploit two attack channels: the provenance channel and the activation value distribution channel. The authors then propose a neural network invariants method to extract value invariants to guard the value channel and the provenance invariants to guard the provenance channel. Constructing DNN value invariants is to train a set of models for individual layers to describe the activation value distributions of the layers from the benign inputs. Constructing provenance invariants is to train a set of models each describing how a set of activated neurons in a layer lead to a set of activated neurons in the next layer. This paper also does a great job explaining the existing attacks and existing defense and detection.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Rotaru, C., Roli, F. (2019). Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. *USENIX Security Symposium*, 321-338.

There is a recent interest in transferability because of ML Cloud service. Transferability captures the ability of an attack against a machine-learning model to be effective against a different, potentially unknown, model. With this paper the authors investigate the factors contributing to transferability of test-time evasion and training time poisoning attacks by designing black box attacks against a surrogate model. Experiments designed using wide range of both linear and non linear classifiers and dataset indicate size of input gradients, gradient alignment and the vulnerability of the loss functions contributes to transferability.

Akhtar, N., Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6.* 14410-14430.

This paper reviews the works that design adversarial attacks, analyze the existence of such attacks and propose defenses against them. Despite their high accuracies, modern deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain imperceptible to human vision system. Such attacks can cause a neural network classifier to completely change its prediction about the image. This paper summarizes twelve different attacks on classification models typically fooling Convolutional Neural Network with a table indicating if each of the attacks are black or white box, targeted or non targeted, image specific or universal, perturbations norm, learning and strength. However, due to the seriousness of adversarial threats, attacks are also being actively investigated beyond the classification/recognition task in Computer Vision. A section of the paper is specifically dedicated to the literature that deals with the adversarial attacks in practical real-world conditions. The authors also explain the three main directions currently the defenses against the adversarial attacks are being developed: modified training/ input, modified networks and using external models.

Ruchansky, N., Seo, S., Liu.Y. (2017). CSI: A hybrid deep model for fake news detection. *ACM Conference on Information and Knowledge Management*, 797-806, doi:10.1145/3132847.3132877.

Using two real-world social media datasets Twitter and Weibo, the authors, researchers at the University of Southern California combines all three characteristics of fake news: the text of an article, the user response it receives, and the source users promoting it for designing a more accurate and automated prediction model for fake news detection. The proposed model called CSI consists of two main parts, a module for extracting temporal representation of news articles using Recurrent Neural network that captures the response characteristic, and a module for representing and scoring the behavior of users that captures the source characteristic. Authors claim to have developed a generalized model that can easily generalize to any dataset but haven’t demonstrated more on that.