High Accuracy and High Fidelity Extraction of Neural Networks

- 提出了模型抽取的两个目标 : accuracy和fidelity. accuracy指的是在victim模型上的任务准确率良好, fidelity指对于任何输入, 抽取模型的输出和victim模型的输出相似. 这两点是有本质上的区别的
- 本文通过analytical and empirical的讨论, 解释了对于获得fidelity functionally-equivalent extracted模型的限制. 并且对于这些限制, 本文提出了一种方法来解决这些限制, 直接抽取模型的权重
- high-fidelity和high-accuracy可能通常是互相冲突的, 因为high-fidelity应该对于victim的错误预测也能相似的复现, 而high-accuracy则会要求尽可能高的正确. 在high-fidelity下隐含的限制在于 : victim和knockoff在所有输入上都要相同, 不管是不是在victim所使用的训练集分布上
- 在许多公开测试集上进行实验, 并且通过攻击在生产级别的系统来展示实用性
- 提出一个很重要的观点 : exact extraction

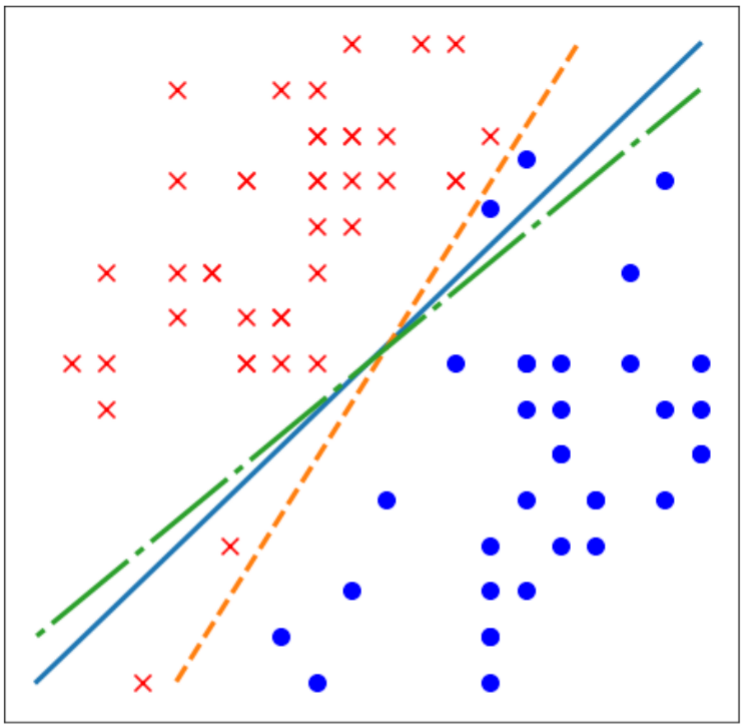

Figure 1: Illustrating fidelity vs. accuracy. The solid blue line is the oracle; functionally equivalent extraction recovers this exactly. The green dash-dot line achieves high fidelity: it matches the oracle on all data points. The orange dashed line achieves perfect accuracy: it classifies all points correctly.

一些model extraction 方法的setting

Attack	Type	Model type	Goal	Query Output
Lowd & Meek [8]	Direct Recovery	LM	Functionally Equivalent	Labels
Tramer *et al.* [11]	(Active) Learning	LM, NN	Task Accuracy, Fidelity	Probabilities, labels
Tramer *et al.* [11]	Path finding	DT	Functionally Equivalent	Probabilities, labels
Milli *et al.* [19] (theoretical)	Direct Recovery	NN (2 layer)	Functionally Equivalent	Gradients, logits
Milli *et al.* [19]	Learning	LM, NN	Task Accuracy	Gradients
Pal *et al.* [15]	Active learning	NN	Fidelity	Probabilities, labels
Chandrasekharan *et al.* [13]	Active learning	LM	Functionally Equivalent	Labels
Copycat CNN [16]	Learning	CNN	Task Accuracy, Fidelity	Labels
Papernot *et al.* [7]	Active learning	NN	Fidelity	Labels
CSI NN [25]	Direct Recovery	NN	Functionally Equivalent	Power Side Channel
Knockoff Nets [12]	Learning	NN	Task Accuracy	Probabilities
Functionally equivalent (this work)	Direct Recovery	NN (2 layer)	Functionally Equivalent	Probabilities, logits
Efficient learning (this work)	Learning	NN	Task Accuracy, Fidelity	Probabilities

Table 1: Existing Model Extraction Attacks. Model types are abbreviated: LM = Linear Model, NN = Neural Network, DT = Decision Tree, CNN = Convolutional Neural Network.

Learning-based Model Extraction

Fully-supervised model extraction 主要还是老方法, 使用一个pretrain的模型作为label oracle, 然后得到数据集对knockoff进行训练

Unlabeled data improves query efficiency 在有限的query预算里达到给定程度的准确率, 可能包含label-efficient learning, 即active learning和semi-supervised learning, 并且这两个方法可以互补

本文中展示了, 使用半监督学习进行target domain 无标签的数据的学习能够提升model extraction attack

本文使用了SOTA的SSL方法, rotation loss来增强模型的性能

实验结果

Architecture	Data Fraction	ImageNet	WSL	WSL-5	ImageNet + Rot	WSL + Rot	WSL-5 + Rot
Resnet_v2_50	10%	(81.86/82.95)	(82.71/84.18)	(82.97/84.52)	(82.27/84.14)	(82.76/84.73)	(82.84/84.59)
Resnet_v2_200	10%	(83.50/84.96)	(84.81/86.36)	(85.00/86.67)	(85.10/86.29)	(86.17/88.16)	(86.11/87.54)
Resnet_v2_50	100%	(92.45/93.93)	(93.00/94.64)	(93.12/94.87)	N/A	N/A	N/A
Resnet_v2_200	100%	(93.70/95.11)	(94.26/96.24)	(94.21/95.85)	N/A	N/A	N/A

Table 2: Extraction attack (top-5 accuracy/top-5 fidelity) of the WSL model [28]. Each row contains an architecture and fraction of public ImageNet data used by the adversary. ImageNet is a baseline using only ImageNet labels. WSL is an oracle returning WSL model probabilities. WSL-5 is an oracle returning only the top 5 probabilities. Columns with (+ Rot) use rotation loss on unlabeled data (rotation loss was not run when all data is labeled). An adversary able to query WSL always improves over ImageNet labels, even when given only top 5 probabilities. Rotation loss does not significantly improve the performance on ResNet_v2_50, but provides a (1.36/1.80) improvement for ResNet_v2_200, comparable to the performance boost given by WSL labels on 10% data. In the high-data regime, where we observe a (0.56/1.13) improvement using WSL labels.

Dataset	Algorithm	250 Queries	1000 Queries	4000 Queries
SVHN	FS	(79.25/79.48)	(89.47/89.87)	(94.25/94.71)
SVHN	MM	(95.82/96.38)	(96.87/97.45)	(97.07/97.61)
CIFAR10	FS	(53.35/53.61)	(73.47/73.96)	(86.51/87.37)
CIFAR10	MM	(87.98/88.79)	(90.63/91.39)	(93.29/93.99)

Table 3: Performance (accuracy/fidelity) of fully supervised (FS) and MixMatch (MM) extraction on SVHN and CIFAR10. MixMatch with 4000 labels performs nearly as well as the oracle for both datasets, and MixMatch at 250 queries beats fully supervised training at 4000 queries for both datasets.

learning based method的限制

基于学习的方法有一系列不确定的因素, 比如随机初始化, SGD的等等. 即使得到所有oraclet的训练集, 超参数, 也可能在GPU上的随机因素而导致获得完全functionally equivalent extraction困难

High-fidelity extraction

一大堆公式, 打扰了, 不过效果挺好. 以后有需要的话常回来看看

Conclusion

本文把模型抽取细分为高准确度和高保真度. 高准确度指的是仿造品模型和被抽取模型在被抽取模型的任务上准确度相似, 甚至有可能更高. 高保真度则是指不光是对于被抽取模型的正确预测相似, 同时也对原模型的错误预测也相似, 因此更加像原模型, 而不是光对于原模型的test set上准确率的相似.

对于高准确度, 本文提出了可以结合label-efficient的方法来减少获得抽取模型的query效率, 主要是active learning和semi-supervised learning. 当然, 还是使用另外数据集加上对原模型query的输出作为一个transfer set来进行训练

至于high-fidelity, 似乎在我的工作中并不需要这个. 本文中是通过公式去证明的高保真权重相似

对于我工作来说, 本文主要的启发在于可以使用active learning以及SSL的方法去结合transfer set来进行训练, 并且包括之后的迁移学习都可以使用这些方法. 本文是一篇宝藏文, 值得多回顾