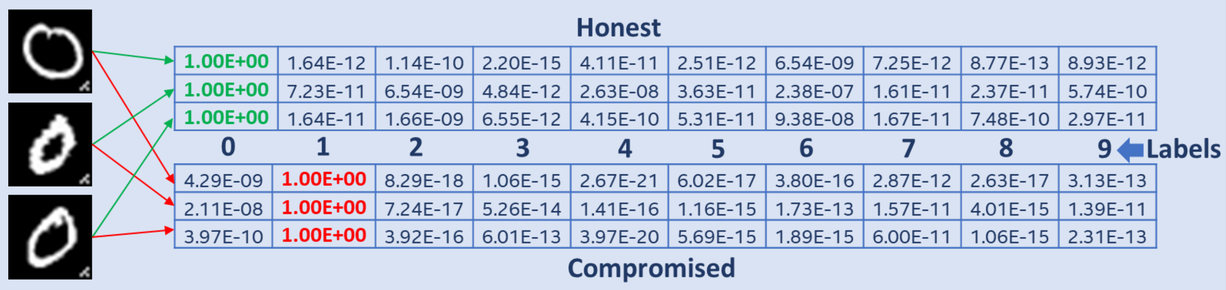
Deep Probabilistic Models to Detect Data Poisoning Attacks

Mahesh Subedar, Nilesh Ahuja, Ranganath Krishnan, Ibrahima J. Ndiour, Omesh Tickoo
Intel Labs

Introduction

- Data poisoning attacks are the security threats introduced in machine learning models during the training phase.
- We investigate backdoor data-poisoning attack on deep neural networks by inserting a backdoor pattern in training images. This results in compromised model misclassifying poisoned test samples while maintaining high accuracies for the clean test-set.
- We present two promising deep probabilistic models for detection of poisoned samples by quantifying the uncertainty estimates associated with the trained models.

Backdoor Attack



Successful backdoor data poisoning attack misclassified Digit-0 as Digit-1.

Deep Probabilistic Models to detect Data Poisoning

Approach 1: Probabilistic modeling of deep features (DF) [1]

- Training: A generative model is defined over the deep neural network (DNN) features by fitting class-conditional probability distributions.
- Inference: Log-likelihood scores of the features of a test sample are calculated with respect to fitted distributions to predict the class probability.
- Log-likelihood scores are used to discriminate clean samples (which should have high likelihood) from poisoned samples (which should have low likelihood).

Approach 2: Bayesian Neural Networks (BNN)

- Training: Given the prior distribution and model likelihood, Mean Field Variational Inference (MFVI) method [2] is used to infer the posterior distribution over the model parameters.
- Inference: Predictive distribution is obtained through multiple stochastic Monte Carlo forward passes through the network by sampling network parameters from the posterior distribution.
- Model uncertainty is used to distinguish between clean and poison samples, which quantifies mutual information between parameter posterior and predictive distributions.

Experiments and Results

Classification Accuracies for MNIST and CIFAR10 datasets on Clean (held-out) and Poisoned (backdoor) samples.

% Backdoor Samples			0	10	20	30	40	50
MNIST	Clean**	DNN	99.21	99.28	99.26	99.35	99.32	99.24
		DF	-	99.04	98.74	98.94	99.04	98.97
		BNN	-	99.56	99.64	99.56	99.51	99.50
	Poisoned*	DNN	-	98.78	98.87	99.16	99.12	99.29
		DF	-	35.54	17.55	11.87	7.32	20.54
		BNN	-	99.54	99.44	99.50	99.59	99.43
CIFAR10	Clean**	DNN	88.90	88.36	88.23	87.39	87.87	88.74
		DF	-	88.16	88.32	88.20	88.95	88.26
		BNN	-	89.48	89.63	89.80	90.00	90.05
	Poisoned*	DNN	-	84.08	81.95	86.40	86.94	88.29
		DF	-	69.66	64.37	75.00	75.86	80.52
		BNN	-	88.30	88.96	88.82	90.02	90.05

* Lower is better. ** Higher is better.

AUPR scores for the DNN, DF and BNN models on MNIST and CIFAR-10 datasets.

% Backdoor Samples		10	20	30	40	50
MNIST	DNN	83.18	81.58	68.3	54.31	53.14
	DF	99.43	99.7	99.91	99.96	99.96
	BNN	97.47	86.24	72.9	58.46	46.46
CIFAR10	DNN	40.62	64.16	36.39	38.68	41.99
	DF	91.87	74.58	90.94	91.81	85.17
	BNN	92.24	82.03	70.95	58.89	49.22

- Backdoor attack is successful in compromising the DNN model resulting in similar accuracies for clean and poisoned test samples.
- DF method is successful in dropping the classification accuracy of the poisoned data flagging a compromised model.
- DF and BNN methods show higher AUPR scores than the deterministic DNN models.

Conclusions

- Success of the data poisoning attack with simple backdoor patterns (hardly noticeable) shows the real threat associated with these attacks on the machine learning models.
- Uncertainty estimates obtained from the two presented approaches Deep Probabilistic Features and Bayesian Neural Networks have the potential to detect backdoor data poisoning attacks.
- We demonstrated probabilistic models are important tool to mitigate data poisoning attacks in machine learning systems, and presented a potential research thread to the BDL community to mitigate the data poisoning attacks.

References:

[1] Ahuja, N. A., Ndiour, I., Kalyanpur, T., and Tickoo, O. (2019). Probabilistic modeling of deep features for out-of-distribution and adversarial detection. arXiv preprint arXiv:1909.11786.

[2] Krishnan, R., Subedar, M., and Tickoo, O. (2019). MOPED: Efficient priors for scalable variational inference in Bayesian deep neural networks. arXiv preprint arXiv:1906.05323.

Fourth Workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada