ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVI Cycle

Security, privacy and adversarial training

Farzad Nikfam

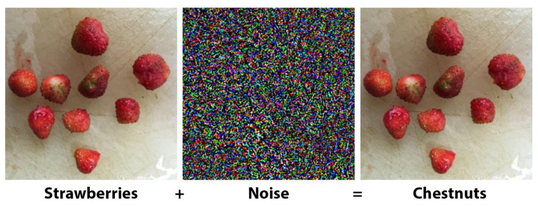
Supervisor: Prof. Maurizio Martina

Research context and motivation

My research field deals with the security and privacy of neural networks. Specifically, I have analyzed various techniques:

- Differential privacy - Adds some random noise to anonymize the dataset as much as possible;
- Federated analysis - The model is sent and trained on multiple local datasets not to affect their privacy but at the same time obtain information from all of them;
- Homomorphic encryption - A data encryption technique to produce encrypted calculations and results;
- Zero-knowledge proofs - A prover can demonstrate to a verifier the possession of information, defined as a witness, that satisfies a specific condition without revealing the information to the verifier or anyone else;
- Secure multiparty computation - Each party involved breaks and distributes their data to everyone; in this way, everyone has a part of everyone else's data, but no one has the original data.

I also dealt with how to speed up the training of models robust to external attacks. An external attack, for example, can add specific noise and make a machine learning model mistake in recognizing an image. In the example, strawberries are misclassified as chestnuts.

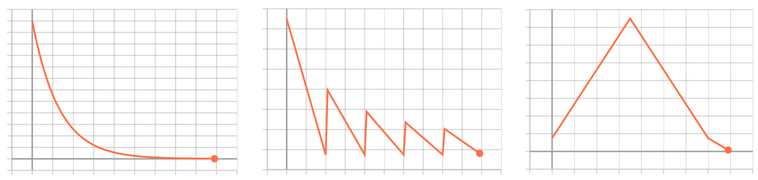


Finally, I dealt with the study of Spiking Neural Networks and their impact in terms of privacy on classic models.

Addressed research questions/problems

For speeding up models robust to external attacks, I took advantage of existing fast training techniques. We can achieve the same degree of robustness in half the time by applying specific hyperparameter modifications during training. Among all the hyperparameters, the most important is the learning rate, and the existing techniques for its speeding up are:

- Linear decay;
- Exponential decay (on the left);
- Warm restarts; (in the middle)
- 1 cycle. (on the right)

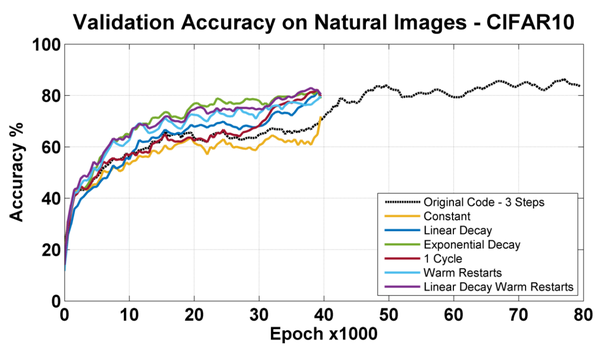


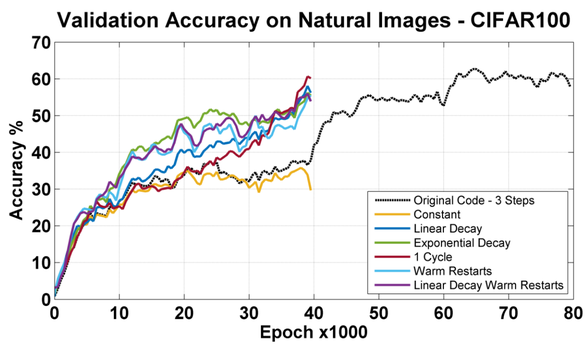
I have applied these techniques to an existing model robust to adversarial attacks, called FAT, obtaining notable improvements. As you can see from the images on the side, with some of the techniques, you get the same result half the time.

A constant learning rate leads to the worst result in terms of time, while the other techniques are all more effective.

However, the 1 cycle probably has the best result so we will compare it with a new technique in the next section.

We can also notice that the result is more pronounced on more complicated datasets such as CIFAR100, where there are ten times more classes than in CIFAR10.



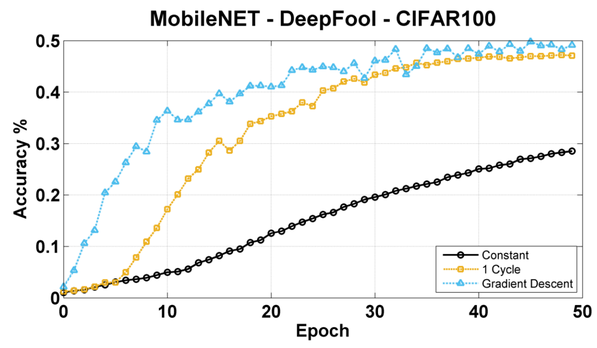


Submitted and published works

- Nikfam, F., Marchisio, A., Martina, M., and Shafique, M., "*AccelAT: A Framework for Accelerating the Adversarial Training of Deep Neural Networks through Accuracy Gradient*", Submitted to IEEE Access, 2022
- Nikfam, F., Marchisio, A., Martina, M., and Shafique, M., "*Security and Privacy: A Survey*", Close to submission to IEEE Access

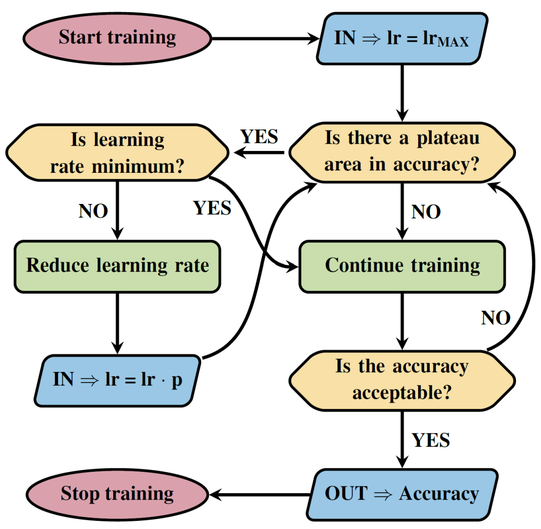
Novel contributions

After the results obtained with the existing techniques, I worked on the realization of a new fast training technique called AccelAT. This technique is modular and dynamic based on the accuracy gradient. Applying the technique to robust models, we can see that it has comparable, if not superior, results to existing techniques, like as 1 cycle.



Adopted methodologies

Inspired by existing fast training techniques that eliminate plateau areas while learning, in the AccelAT framework, the learning rate is varied based on the performance of the validation accuracy. First, I found the maximum learning rate using the learning rate finder technique, after which I set it as the initial learning rate to be decreased if the accuracy starts to show a plateau. Then, the technique uses a gradient to change the learning rate based on accuracy progression. The learning rate is decreased by a percentage value "p" if the accuracy in the last "n" cycles has not increased by a specific value "delta" up to the minimum desired learning rate.



Research setup

- Software ➔ DNN: ResNet, MobileNet - Dataset: CIFAR10, CIFAR100 - Code: Python, TensorFlow, Foolbox
- Hardware ➔ NVIDIA Tesla K40c - GPU12 GB of memory

Future work

- Verification and implementation of robust models on Spiking Neural Networks.
- The exploitation of encryption and privacy techniques to increase the security of neural networks.

List of attended classes

- 01UJBRV – Adversarial training of neural networks (3/6/2021, 15 hours)
- 02LWHRV – Communication (8/4/2021, 5 hours)
- 01UJRIU – Computing Paradigms for Error-Tolerant Applications (26/7/2021, 25 hours)
- 01SHMRV – Entrepreneurial Finance (6/4/2021, 5 hours)
- 01SCSIU – Machine learning for pattern recognition (24/7/2021, 20 hours)
- 01URPOV – Machine learning for vision and multimedia (15/6/2021, 60 hours)
- 03QTIIU – Mimetic learning (22/1/2021, 20 hours)
- 01UNYRV – Personal branding (11/4/2021, 1 hour)
- 08IXTRV – Project management (9/4/2021, 5 hours)
- 01RISRV – Public speaking (2/4/2021, 5 hours)
- 01SWQRV – Responsible research and innovation, the impact on social challenges (2/4/2021, 5 hours)
- 02RHORV – The new Internet Society: entering the black-box of digital innovations (9/4/2021, 6 hours)
- 01UNXRV – Thinking out of the box (29/12/2020, 1 hour)
- 01QORRV – Writing Scientific Papers in English (23/6/2021, 15 hours)

POLITECNICO DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering