SecDD: Efficient and Secure Method for Remotely Training Neural Networks

Ilia Sucholutsky,¹ Matthias Schonlau
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario, Canada
¹isucholu@uwaterloo.ca

Abstract
We leverage what are typically considered the worst qualities of deep learning algorithms - high computational cost, requirement for large data, no explainability, high dependence on hyper-parameter choice, overfitting, and vulnerability to adversarial perturbations - in order to create a method for the secure and efficient training of remotely deployed neural networks over unsecured channels.

Introduction
We consider the situation where a neural network must be trained using proprietary or confidential data, but only an unsecured channel is available for providing data to the network. We assume that any data transmitted over this channel can be accessed by other parties. Our objective is to transmit data that will train the target network to desired accuracy, but be unusable by other networks, and also not reveal any information through qualitative inspection. A second objective is to improve efficiency by minimizing the size of our transmission. To this end, we propose using dataset distillation, the process of representing the knowledge of a large dataset using a smaller number of synthetic samples (Wang et al. 2018), as a method for efficiently and securely training neural networks. Specifically, Soft-Label Dataset Distillation (SLDD) is an extension to the dataset distillation algorithm that achieves even better performance by also learning distillation labels along with the distillation images (Sucholutsky and Schonlau 2019). We propose Secure Dataset Distillation (SecDD) as an extension of SLDD that intentionally overfits samples to a target network in order to create tiny privacy-preserving training sets that reduce transmission size by several orders of magnitude. These synthetic training samples can only be trained on by a network with the same architecture and random initialization as the target network. These synthetic training samples can also be designed to qualitatively not resemble real samples; even appearing to belong to completely unrelated datasets.

In order to retrieve private information from the synthetic samples, an attacker would need to discover both the architecture and random initialization of the target network. To do so, an attacker would have to perform Neural Architecture Search (NAS) on the synthetic training set. Fortunately, NAS methods are extremely computationally intensive and generally very data-hungry (Strubell, Ganesh, and McCallum 2019). In particular, NAS has been shown to be ineffective when using small distilled datasets as proxies for the full training set (Shleifer and Prokop 2019). In addition, the search space for the NAS algorithm grows rapidly as the size of the target network increases. If the target network contains unusual components, it may even be impossible for NAS to find it as the search space is often constrained to popular network components. A good analogy for this is the process for creating a strong password: having a long password with special characters greatly increases the search space making it difficult for a brute-force attack to succeed.

Related work
Prototypes have long been studied in the context of algorithms like k-nearest neighbours (Chang 1974; Sánchez 2004). Generally speaking, prototype methods aim to approximate datasets using a smaller number of samples. Prototype selection methods aim to choose prototypes from the actual dataset (Olvera-López, Carrasco-Ochoa, and Martínez-Trinidad 2010; García et al. 2012). Prototype generation methods, like the k-means algorithm, instead create synthetic samples (Triguero et al. 2011; Triguero, García, and Herrera 2011; Nanni and Lumini 2009). Most prototype methods use hard labels, but some propose more complex prototypes that aim to increase efficiency (Mettes, van der Pol, and Snoek 2019; Sucholutsky and Schonlau 2020). Dataset distillation can be described as a family of prototype generation methods intended for use with neural networks (Wang et al. 2018; Sucholutsky and Schonlau 2019; Bohdal, Yang, and Hospedales 2020). Flexible Dataset Distillation (LD) is a recently proposed extension of dataset distillation that learns unrestricted labels as in SLDD but for a small fixed set of real images taken from the training dataset (Bohdal, Yang, and Hospedales 2020).
Figure 1: SecDD can create various sets of 10 synthetic MNIST images that train target networks to over 95% accuracy while visually appearing to consist almost entirely of noise. Each image is labeled with its top 3 classes and their associated logits.