On the Feasibility of Training Neural Networks with Visibly Watermarked Dataset

Sanghyun Hong*, Tae-hoon Kim^, Tudor Dumitras*, Jonghyun Choi°

*University of Maryland College Park

°Gwangju Institute of Science and Technology (GIST)

^Deeping Source Inc.

Overview

- Data collected separately by enterprises and users is hard to be shared with others because *shared data can be stolen*.

Threat Model

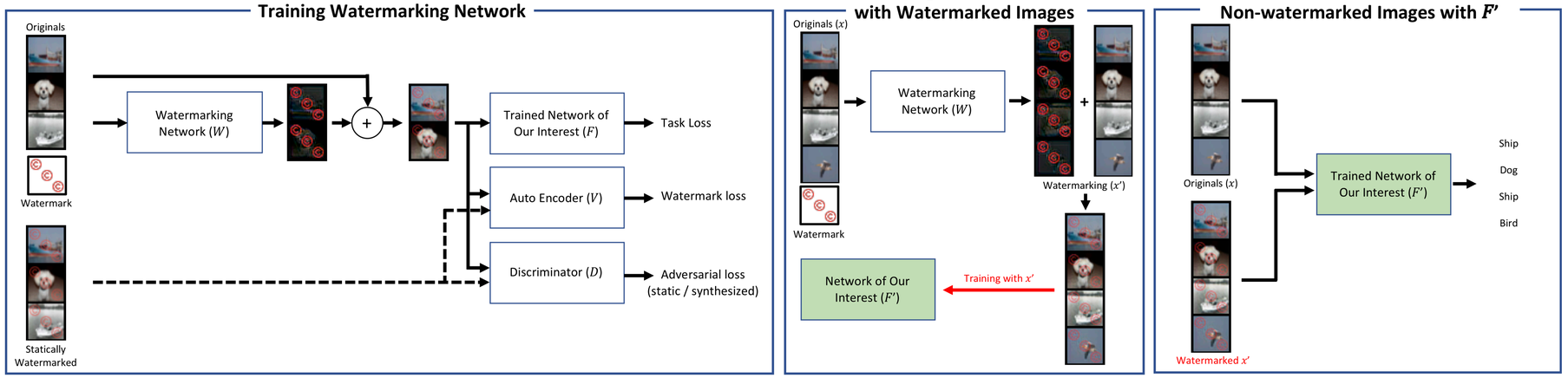
- Company A wants to share a dataset collected from the users with Company B
- Company A **still wants to claim the ownership of the data** to prevent Company B resells the data to others or leaks the data

Our Solution

- **Prior work:**
  - Cryptographic solutions (Homomorphic encryption or MPCs) are computationally expensive and, once data is shared, they cannot prevent claiming ownership from B
  - Solutions conceal secrets (Stenography and Invisible watermarking) are susceptible to data augmentations/changes

- **Our solution:**
  - Use *visible watermarking* on datasets shared with other companies and users

DeepStamp Framework



Objectives

- Purpose of the visible watermarks on datasets
  - *Visibly Intact:* watermarks embedded to data clearly visible to humans
  - *Hard to Remove:* watermarks are hard to be removed by adversaries
  - *Minimize Accuracy Drops:* a network trained with watermarked data minimizes the accuracy drop compared to the clean network

DeepStamp

- **In Training:**
  - Training a watermarking network (*W*) using clean data and a watermark as inputs with three discriminator networks (*F, V, D*)
  - Discriminators use the data with synthesized watermarks from *W* and the static watermarked data as their counterparts
  - Leverages the Generative Multi-Adversarial Network (GMAN)

- **In Stamping:**
  - Using the trained watermarked network (*W*), synthesize the optimal watermarks for clean data and watermark

Preliminary Results

Network	Baseline	Static	DeepStamp
AlexNet	82.74	78.50	**79.59**
VGG16	94.00	92.71	**92.74**
ResNet50	95.37	**94.88**	94.18

- Using visible watermarking **causes the acc. drops in all cases**, but the drops are minimal when the network capacity is high
- DeepStamp framework can **minimize the accuracy drop further** than the static (additive) watermarking method(s).
- [Results are mentioned in the paper] Our watermarked dataset can be used to train other networks (the data is **transferrable**)

Acknowledgement

- This research is partially supported by Department of Defense and the “Global University Project” grant funded by the GIST in 2018.