NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

Effects of Incremental Training

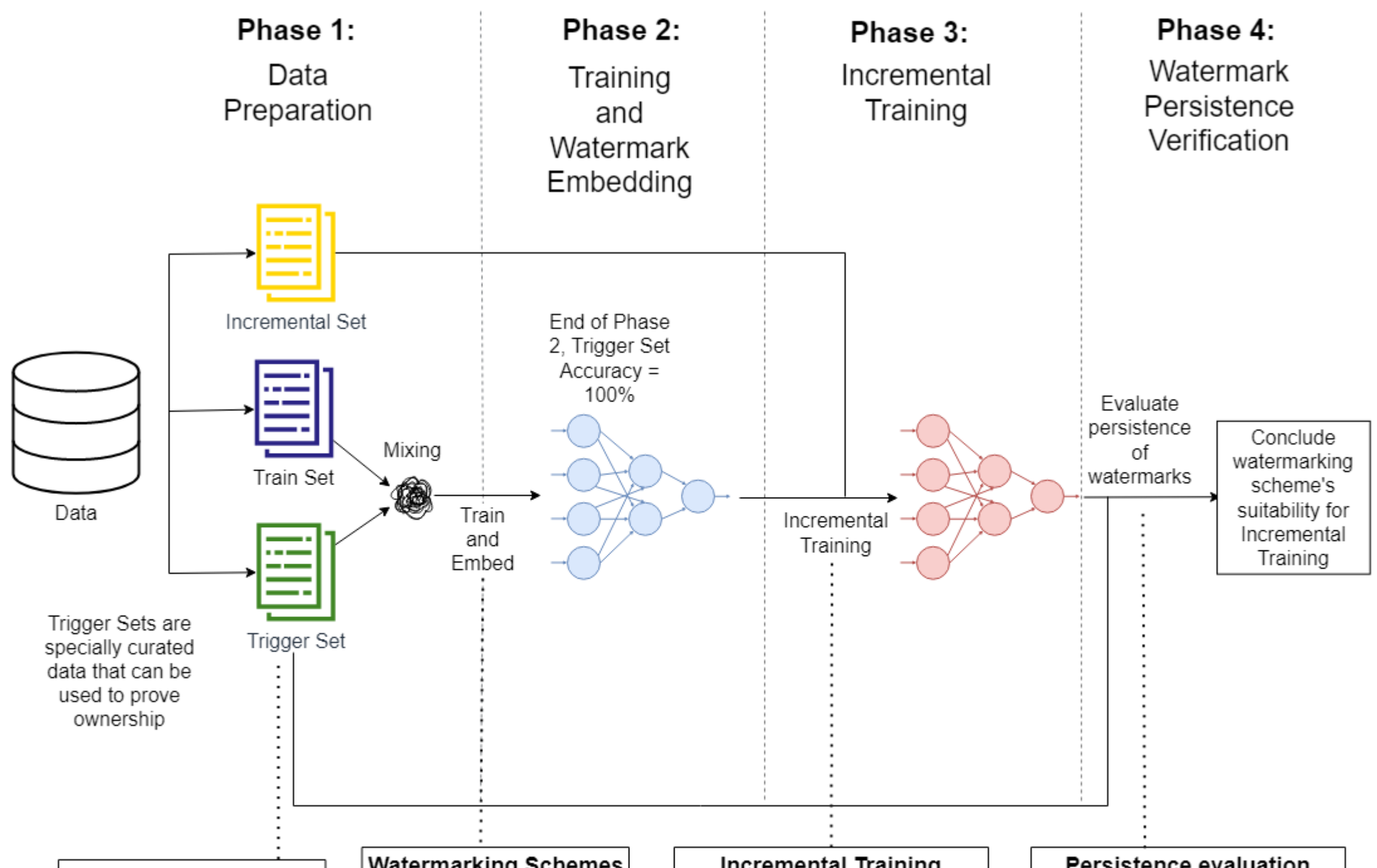
on Watermarked Neural Networks

SCSE22-0019

Student: Heng Chuan Song Supervisor: A/P Anupam Chattopadhyay

Project Objectives:

This study aims to investigate on various existing watermarking scheme's ability to maintain verifiable through maintaining persistence of watermarks. This ensures retention of rightful Intellectual Property Rights (IPR) and robustness against adversaries, even after Incremental Training. We will also attempt to discover how certain variables such as Trigger Set Type and learning rates within such watermarking schemes would affect Incremental Training. The findings will assist in developing a framework for a robust watermarking scheme for neural networks, with ability to support Incremental Training, ultimately preserving IPR.



Key Findings:

1) Incremental Training is detrimental towards persistence of watermarks
2) Uniform distribution of watermarks has an adverse effect towards its persistence when it comes to Incremental Training
3) The watermarking scheme is the largest contributing factor to watermark persistence, although Trigger Set Type and learning rate can affect the persistence slightly