DISTILLATION ATTACKS AND DEFENSES IN NEURAL NETWORK WATERMARKING

Author's: S. Yogavarsha[1], S. Rajeshkumar[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

The rapid deployment of deep neural networks in commercial and research applications has intensified concerns regarding intellectual property (IP) protection and unauthorized model replication. Neural network watermarking has emerged as a promising technique to embed ownership information into trained models without degrading their primary task performance. However, advanced model extraction techniques, particularly knowledge distillation attacks, pose significant threats to watermark robustness. This study investigates the effectiveness of distillation-based attacks on neural network watermarking schemes and proposes a countermeasure to preserve watermark integrity under adversarial model compression and replication scenarios.In this work, a convolutional neural network (CNN) is trained on a synthetic multi-class image dataset with an embedded watermark implemented via a trigger-set mechanism. The watermark forces the model to output a predefined target label when specific trigger patterns are present in the input. The study evaluates two critical metrics: (i) clean accuracy, representing the model's primary task performance, and (ii) watermark success rate (WSR), indicating the reliability of ownership verification. A student model is then trained using knowledge distillation, leveraging soft labels generated by the watermarked teacher model on clean data only. Experimental results demonstrate that while the distilled student maintains comparable clean accuracy, the watermark success rate significantly degrades, highlighting the vulnerability of watermark schemes to distillation attacks.