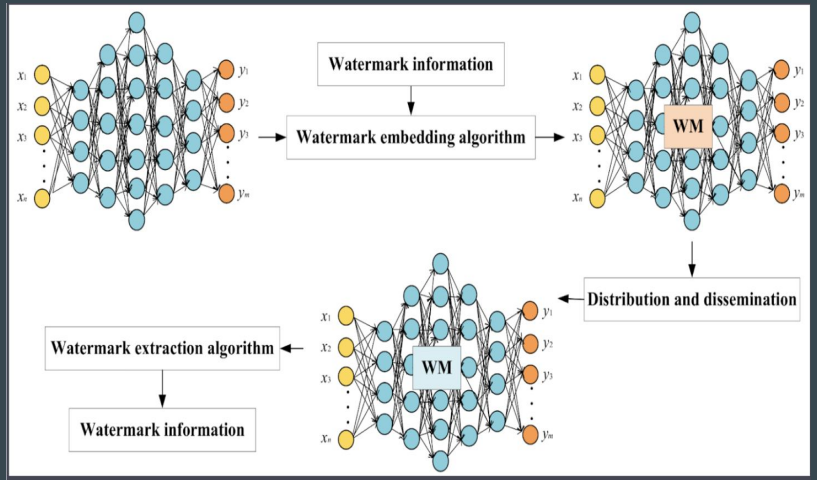
NEURAL NETWORK WATERMARKING

Under Prof. Shekhar Verma

Mridul Mittal(IIT2019127)
Piyush Gurjar(IIT2019148)

Introduction

With the increasing prevalence of deep neural networks (DNNs) and their widespread use in various applications, the need to protect intellectual property and ensure model integrity has become paramount.



Types

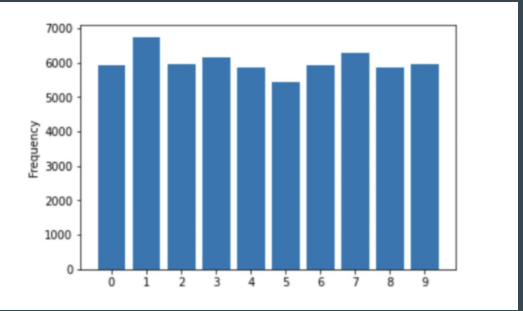
There are majorly 3 types of watermarking techniques:-

- Embedding Watermarks Into Model Parameters
- Using Model Fingerprints
- Using Pre-Defined Inputs as Triggers

Literature Review

- Turning Your Weakness Into a Strength: Watermarking Deep Neural Networks by Backdooring
- EvilModel: Hiding Malware Inside of Neural Network Models
- IPGuard: Protecting Intel- lectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary

Dataset



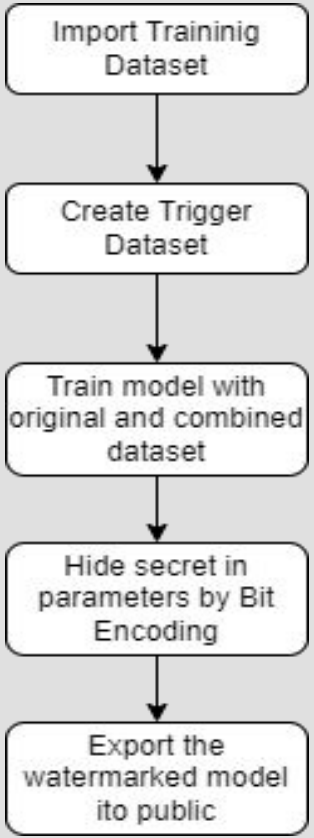
- CIFAR10 - A well-known benchmark dataset that is frequently used in the fields of computer vision and machine learning is the CIFAR-10 dataset.The 10 classes in CIFAR-10 are Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship and Truck.
- MNIST - MNIST stands for "Modified National Institute of Standards and Technology".The MNIST dataset consists of 70,000 grayscale images of handwritten digits from 0 to 9.

Methodology

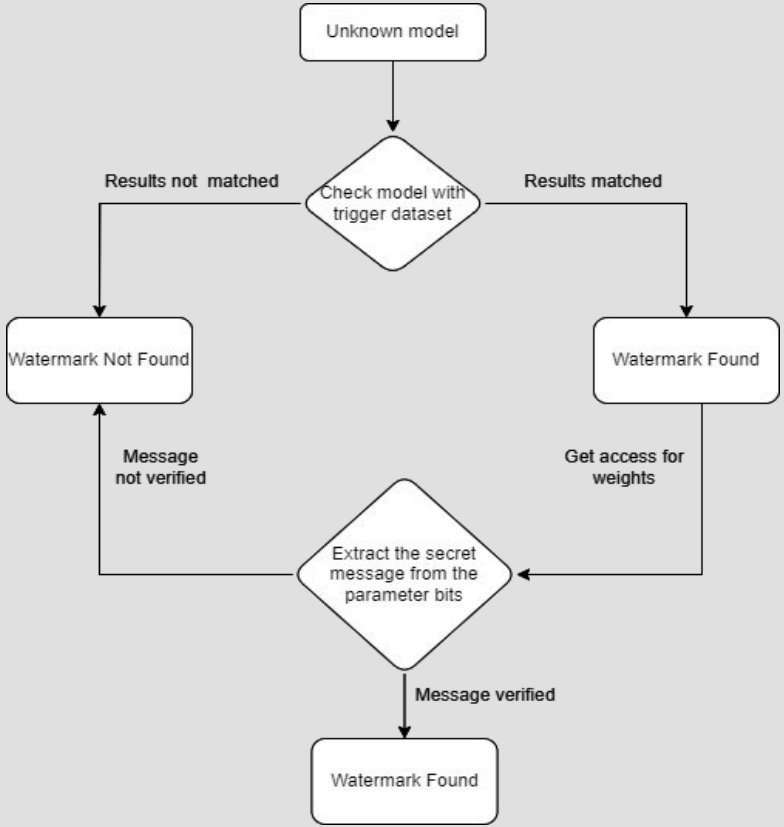
We have developed a hybrid neural network watermarking model. The hybrid approach combines the strengths of two watermarking techniques i.e. backdoor watermarking and bit encoding.

Steps :-

- Trigger Dataset Generation
- Backdoor Embedding
- Bit Encoding
- Watermark Extraction
- Watermark Verification



Watermark Creation

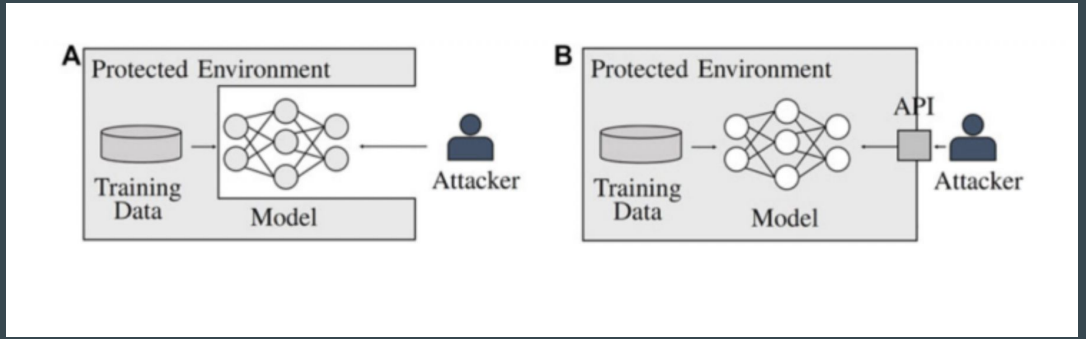


Watermark Extraction

Attacks

Neural network watermarking, is susceptible to various attacks that aim to undermine its effectiveness or remove the embedded watermark

- Fine-tuning Attacks
- Model Pruning
- Transfer Learning Attacks



Result

Model	Type of Model	Epochs Trained	Accuray (%)		F1 Score
			Training	Testing	
Resnet20	Watermarked	30	89.26	75.3	0.72
	Non-watermarked	30	89.61	81.35	0.69
Resnet50	Watermarked	20	86.23	84.72	0.75
	Non-watermarked	20	87.94	81.94	0.79
Inception	Watermarked	23	75.55	71.4	0.58
	Non-watermarked	23	79.8	78.33	0.54
VGG16	Watermarked	28	90.76	84.75	0.84
	Non-watermarked	28	88.5	82.52	0.83
VGG19	Watermarked	25	88.69	84.95	0.62
	Non-watermarked	25	91.31	85.12	0.63

Future Scope & Conclusion

By combining the backdoor and bit encoding methods, the hybrid model aims to provide a watermarking technique that is resistant to attacks, robust against adversarial manipulation, and capable of authenticating ownership and protecting intellectual property.

Thank You!