Rethinking Deep Neural Network Ownership Verification: Embedding Passports to Defeat Ambiguity Attacks
Lixin Fan¹, Kam Woh Ng¹,², Chee Seng Chan²
WeBank AI Lab¹, University of Malaya²

Problem Definition
Conventional DNN Watermarking methods
• White-box Ownership Verification (Uchida et al. [1])

  Watermark Extraction Process
  Watermarked Model
  Ownership Verification

  Transformation Matrix

• Black-box Ownership Verification (Adi et al. [2])

  Query
  ML Online Services
  Trigger-set Data
  Ownership Verification

Problem Statements
1. Protection on DNN is urgently needed
2. Existing watermarking approaches are vulnerable to ambiguity attack

Experimental Results
Ambiguity attack

Ambiguity attack
Infringement
Infringement

Ownership Verification Schemes

Contributions
1. Novel passport-based verification schemes to defeat ambiguity attack
2. One passport-protected DNN model will only have one unique signature
3. Fake passport or modified signature will paralyze the DNN model

Our Solution
Passporting Layer

Embedding Binary Signatures into \( \gamma \) of Passporting Layer

Embedding Passport

Ownership Verification Schemes

Scheme 1 | Scheme 2 | Scheme 3
---|---|---
Need to distribute passport | Yes | No | No
Inference time | Up to 10%”** more time | No extra time | No extra time
Training time | Up to 30%”** more time | Up to 150%”** more time | Up to 150%”** more time
Black or White Box Verification | White | White | Black & White

**Time increases are linearly depending on complexity of the network architecture