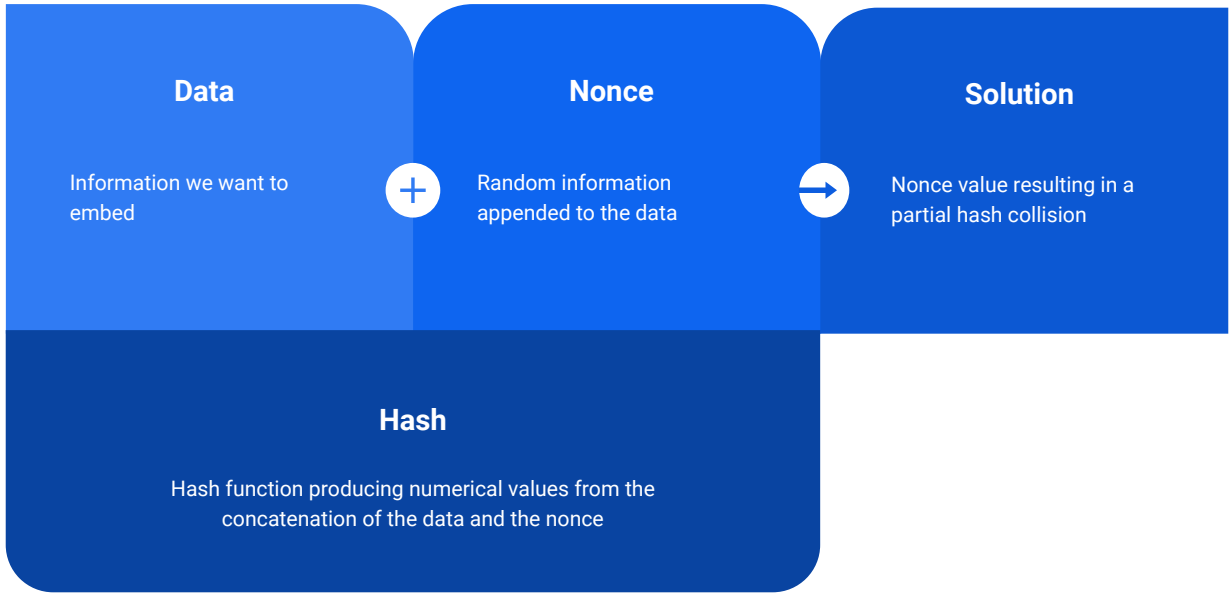
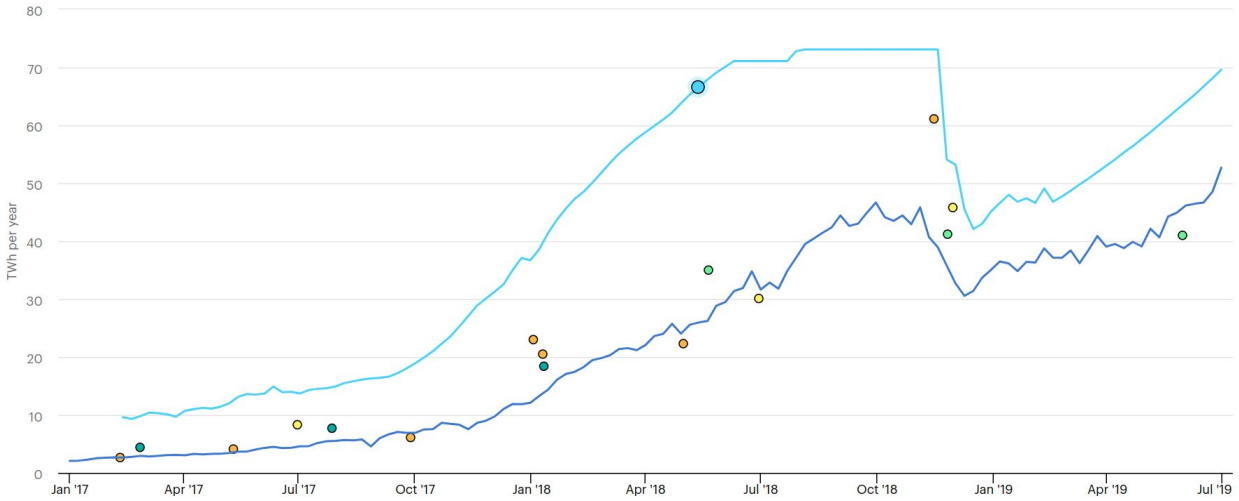
FreezeNet

Making Proof of Work Useful

Proof of Work

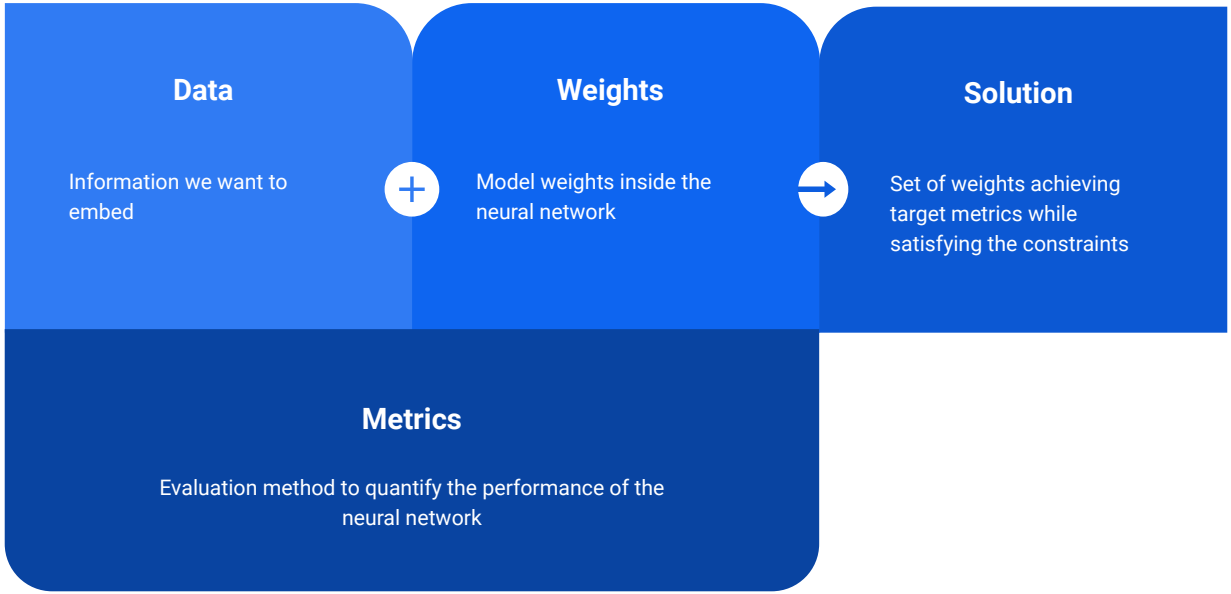


Proof of Work

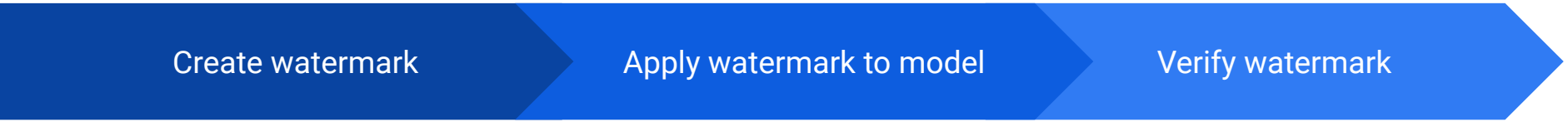


Digiconomist (2019) · Lower bound (Antminer S9) · Bendiksen, Gibbons, Lim (2018-19) · Bevand (2018) · Other peer-reviewed studies · Other studies

Proof of Work - more useful



From a simple idea …



- Hash data to generate seed
- Use seed to generate weights an indices
- The number of weights generated is the watermark size

- Replace model weights at watermark indices by watermark weights
- Weights can be replaced during training after backpropagation phase

- Check if model weights at watermark indices are equal to watermark weights
- Account for floating point imprecision by setting upper bound on difference.

… to a simple API

```
 1 model = FreezeNet()
 2
 3 some_signature    = Signature(b'Some block information', 4096)
 4 another_signature = Signature(b'Some other block information', 4096)
 5
 6 some_signature.sign(model)
 7
 8 some_signature.verify(model)    # True
 9 another_signature.verify(model) # False
10
11 another_signature.sign(model)
12
13 some_signature.verify(model)    # False
14 another_signature.verify(model) # True
```

Experiments

Learning ability

1 Models should still be able to learn while being watermarked

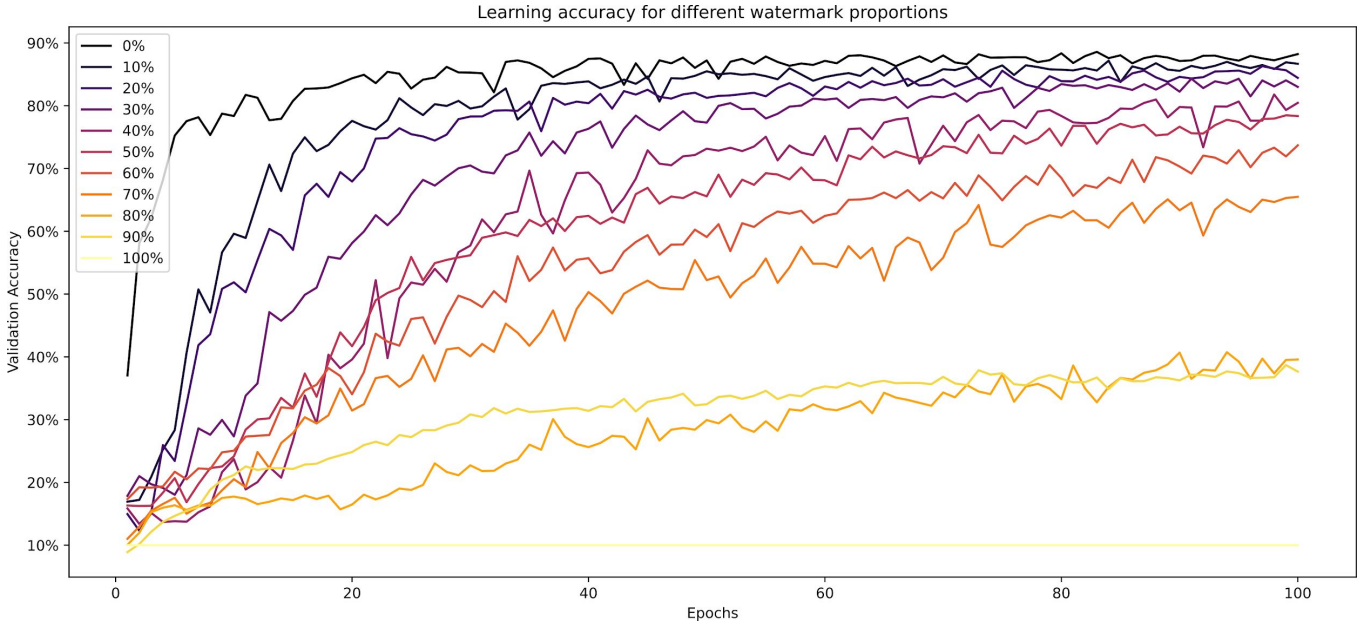
Weak tampering

2 A model should not perform well if it is watermarked after being trained

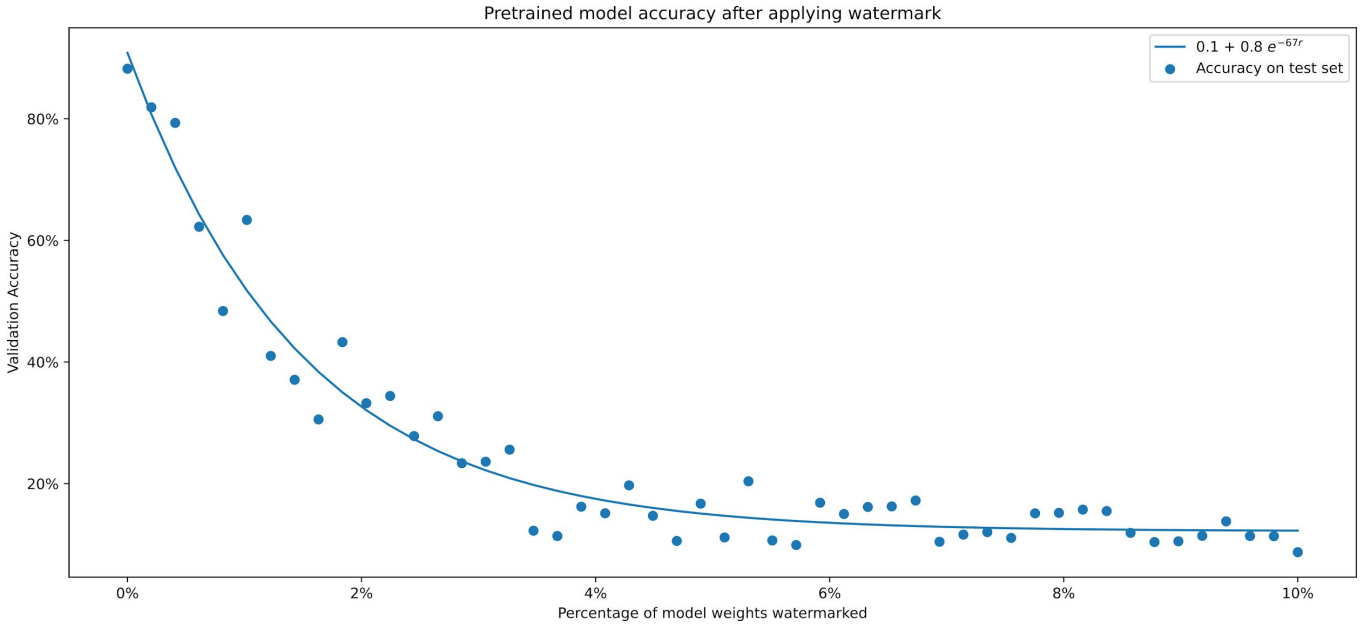
Strong tampering

3 An attacker should have no incentive to re-train a pre-trained model with a watermark

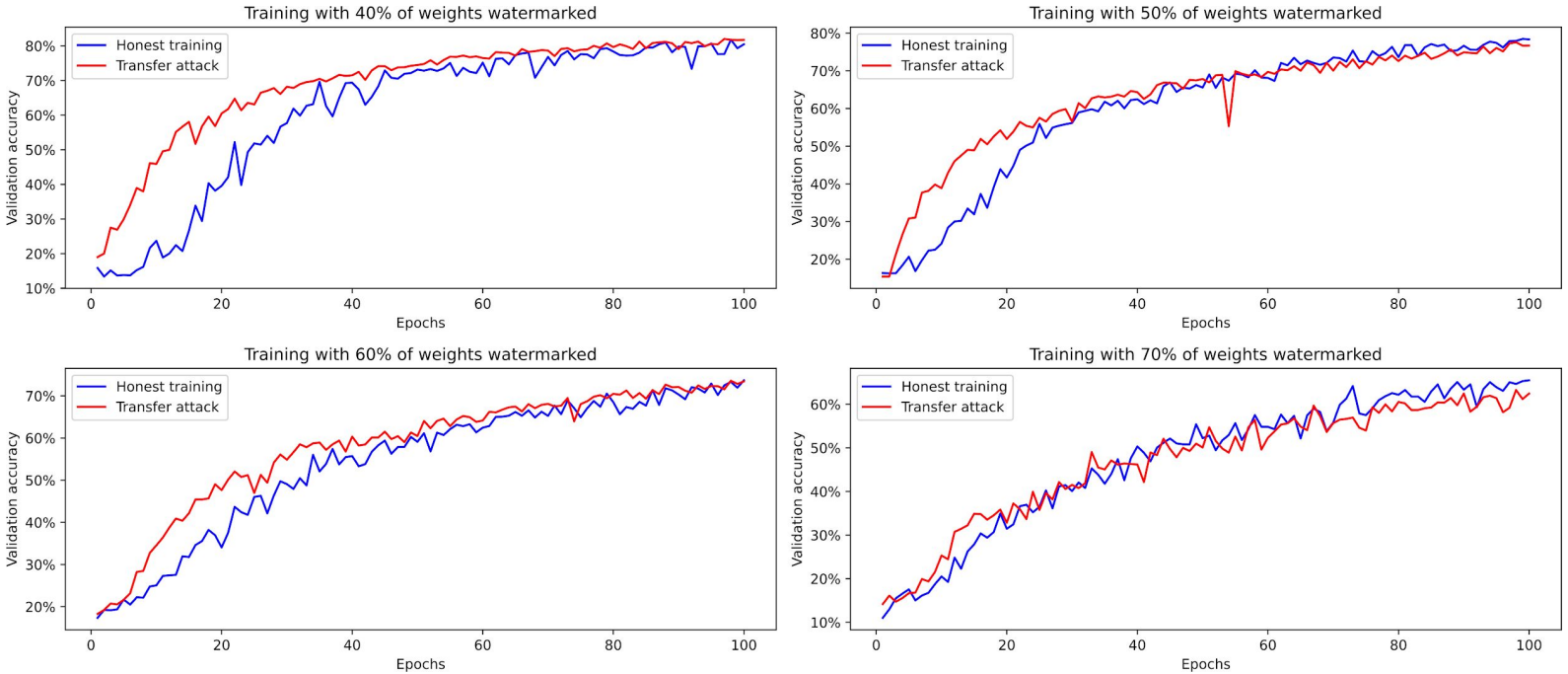
Learning ability



Weak tampering



Strong tampering



Limitations