White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, Hervé Jégou

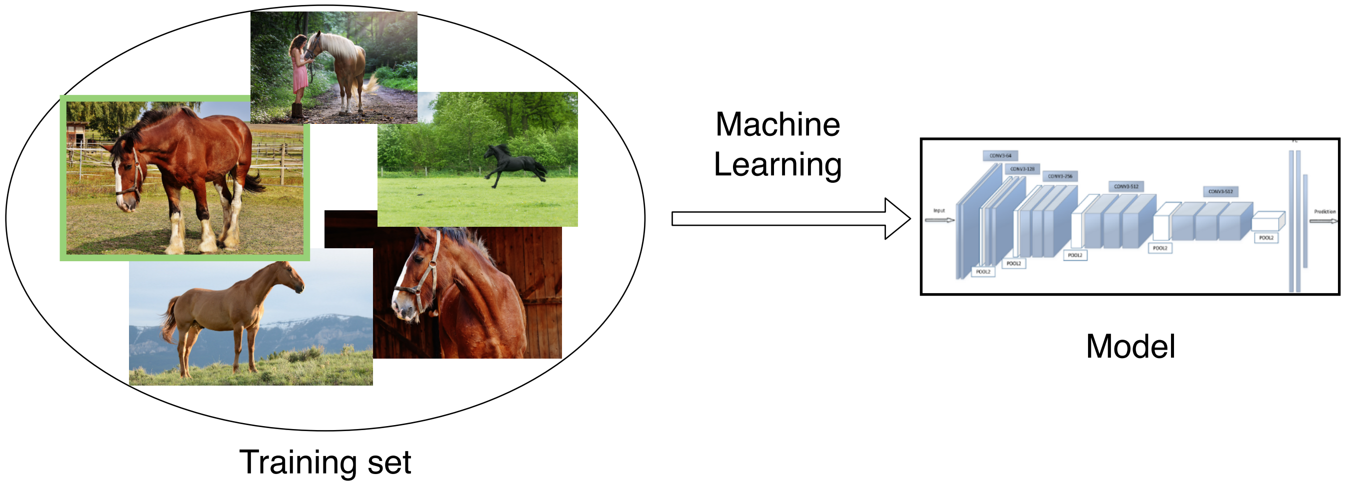
Facebook AI Research, Paris

June 11th, 2019

facebook
Artificial Intelligence Research

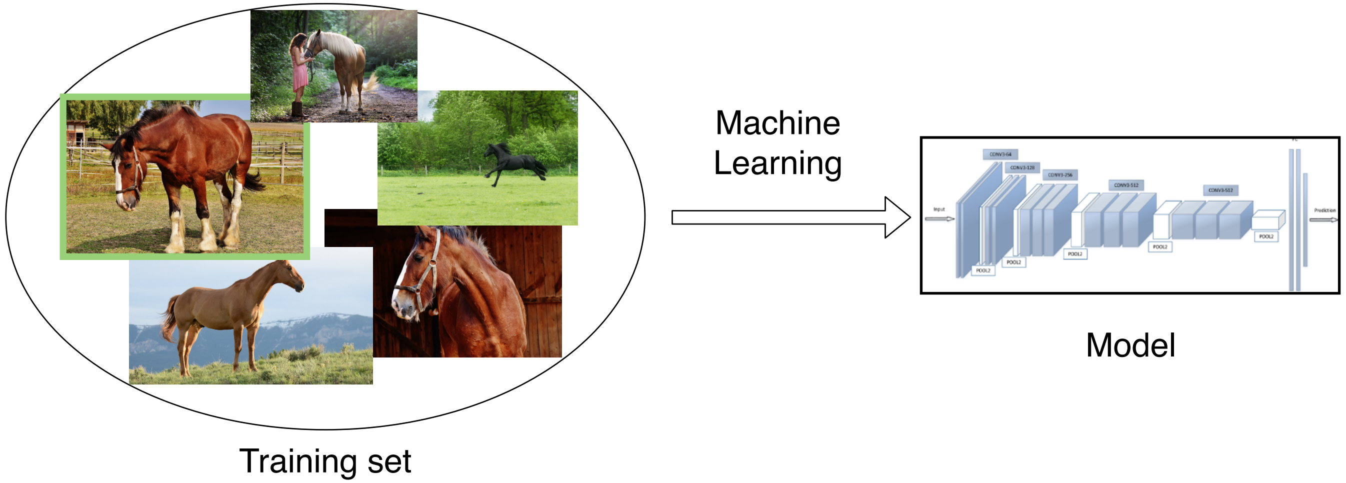
Context: Membership Inference

- Machine learning

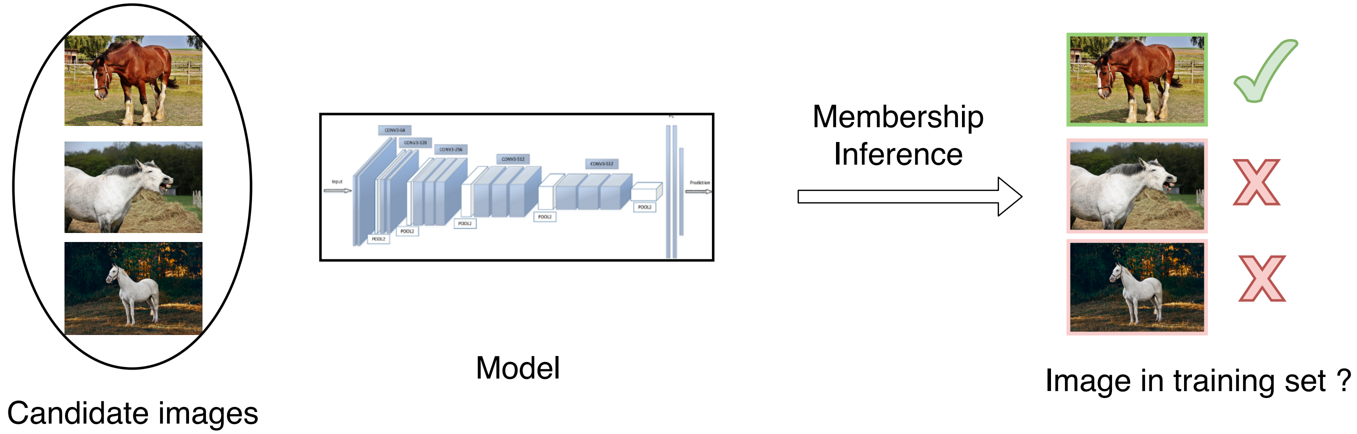


Context: Membership Inference

- Machine learning

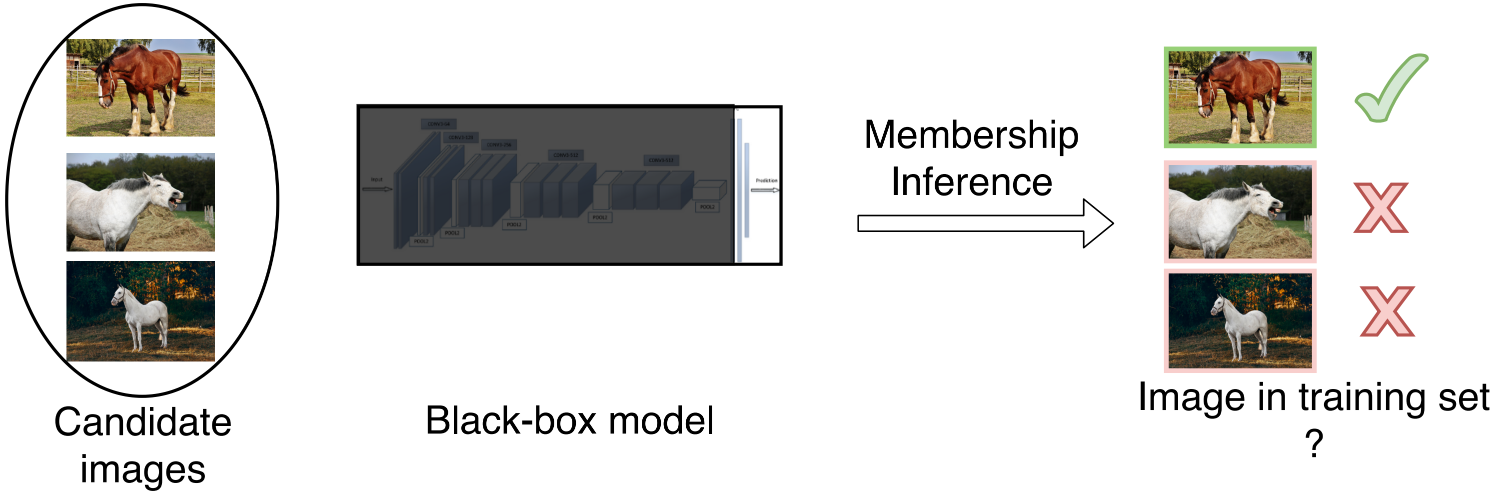


- Membership Inference

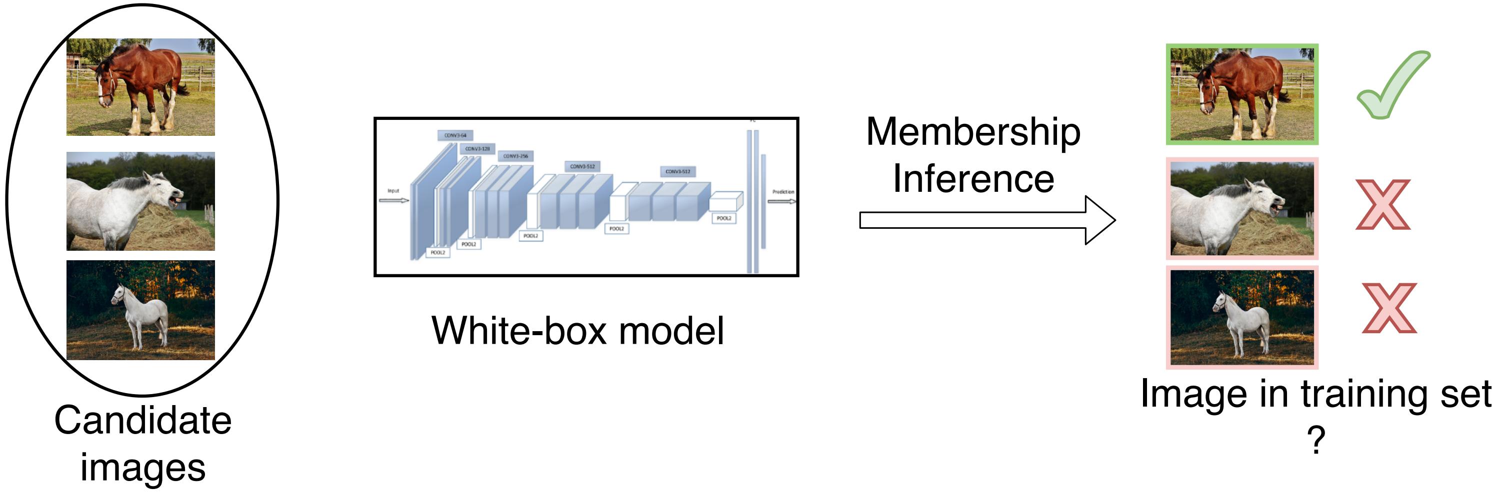


Membership Inference

- Black-box



- White-box



facebook
Artificial Intelligence Research

Goals

- Give a formal framework for membership attacks
  - What is the best possible attack (asymptotically) ?
  - Compare white-box vs black-box attacks
  - Derive new membership inference attacks

Notations

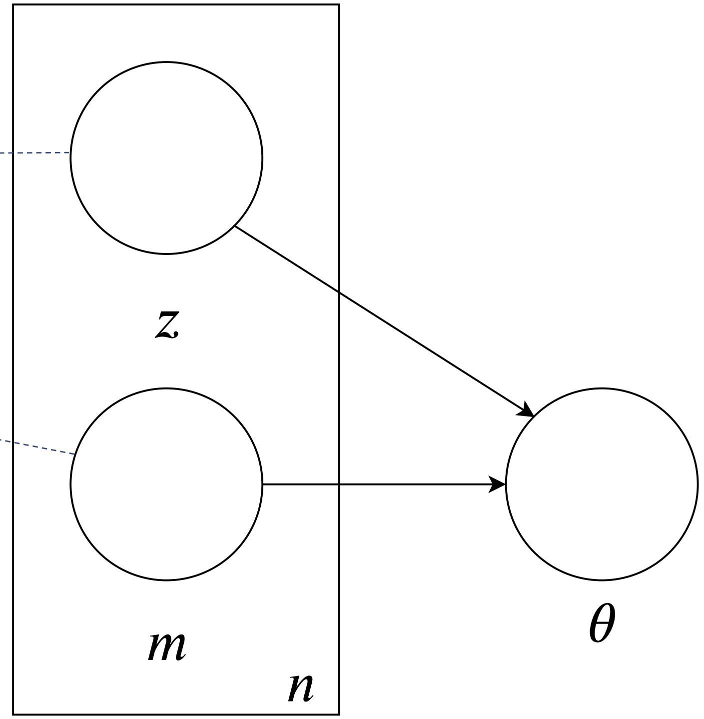
Sample z_i

Membership variable m_i

Bernoulli(λ)

$m_i = 1$: training set

$m_i = 0$: test set



Notations and assumptions

- Assumption: posterior distribution

membership

$$\mathbb{P}(\theta \mid m_{1:n}, z_{1:n}) \propto \exp\left(-\frac{1}{T}\sum_{i=1}^{n} m_i \ell(\theta, z_i)\right)$$

loss

- Temperature T represents stochasticity
  - T=1: Bayes
  - T->0: Average SGD, MAP inference

Formal results: optimal attack

- Membership posterior:

$$\mathcal{M}(\theta, z_1) := \mathbb{P}(m_1 = 1 \mid \theta, z_1)$$

- Result

$$\mathcal{M}(\theta, z_1) = \mathbb{E}_{\mathcal{T}} \left[\sigma \left(\underbrace{s(z_1, \theta, p_{\mathcal{T}})}_{\displaystyle \| \atop \displaystyle \frac{1}{T}\left(\tau_{p_{\mathcal{T}}}(z_1) - \ell(\theta, z_1)\right)} + t_\lambda \right) \right]$$

sigmoid

$$\log\left(\frac{\lambda}{1-\lambda}\right)$$

Formal results: optimal attack

- Membership posterior:

$$\mathcal{M}(\theta, z_1) := \mathbb{P}(m_1 = 1 \mid \theta, z_1)$$

- Result

$$\mathcal{M}(\theta, z_1) = \mathbb{E}_{\mathcal{T}}\left[\sigma\left(\underbrace{s(z_1, \theta, p_{\mathcal{T}})}_{=\frac{1}{T}\left(\tau_{p_{\mathcal{T}}}(z_1) - \ell(\theta, z_1)\right)} + t_\lambda\right)\right]$$

sigmoid

$$t_\lambda = \log\left(\frac{\lambda}{1-\lambda}\right)$$

Only depends on θ through evaluation of the loss!

Approximation strategies

- MALT: a global threshold for all samples

$$s_{\text{MALT}}(\theta, z_1) = -\ell(\theta, z_1) + \tau$$

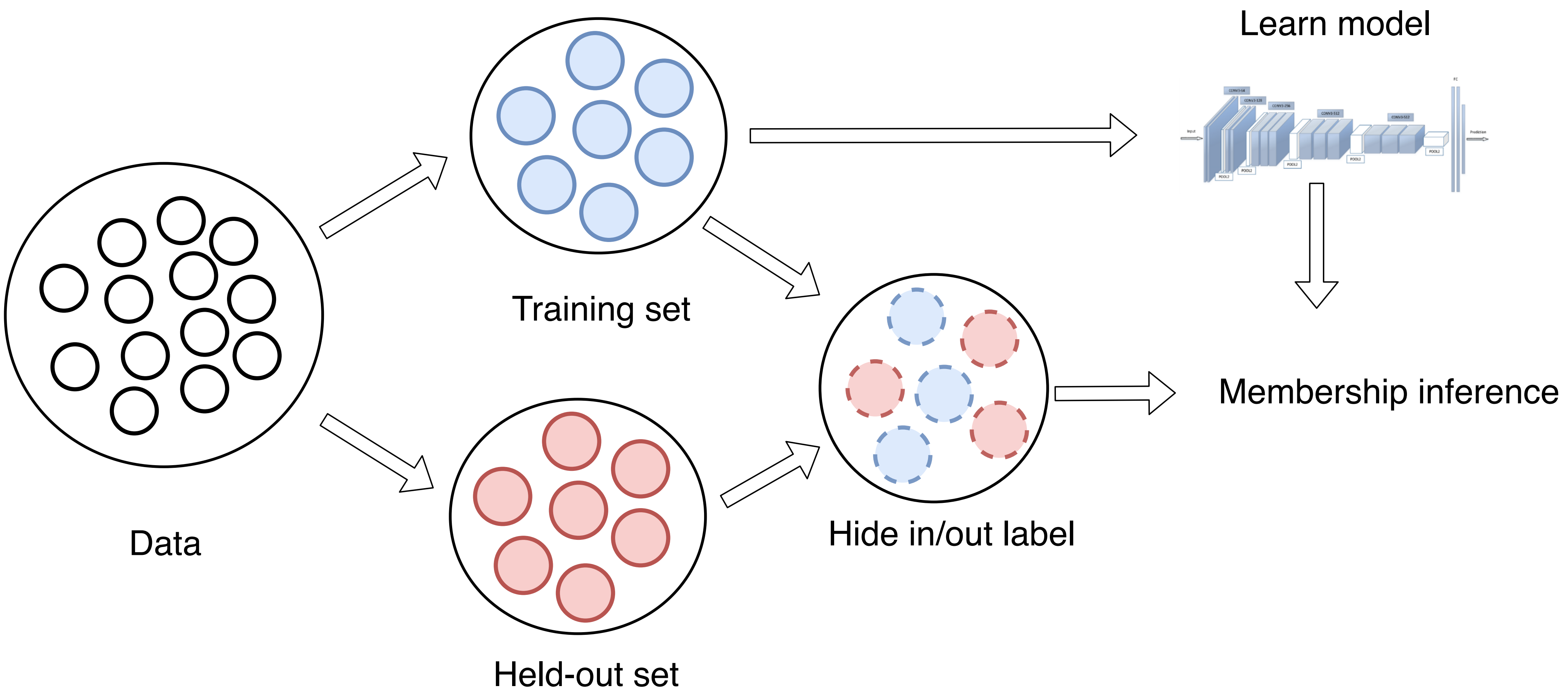
- MAST: compute a threshold for each sample

$$s_{\text{MAST}}(\theta, z_1) = -\ell(\theta, z_1) + \tau(z_1)$$

- MATT: simulate influence of sample using Taylor approximation

$$s_{\text{MATT}}(\theta, z_1) = -(\theta - \theta_0^*)^T \nabla_\theta \ell(\theta_0^*, z_1)$$

Experiments



facebook
Artificial Intelligence Research

Membership inference on CIFAR

Naïve Bayes

Threshold-based

Taylor based

n	Attack accuracy		
	$0-1$	MALT	MATT
400	52.1	54.4	**57.0**
1000	51.4	52.6	**54.5**
2000	50.8	51.7	**53.0**
4000	51.0	51.4	**52.1**
6000	50.7	51.0	**51.8**

=> MATT outperforms MALT

Comparison with the state of the art

Method	Attack accuracy
Naïve Bayes (Yeom et al. [2018])	69.4
Shadow models (Shokri et al. [2017])	73.9
Global threshold	**77.1**
Sample-dependent threshold	77.6

=> State-of-the-art performance

=> Less computationally expensive

Large-scale experiments on Imagenet

Model	Augmentation	0-1	MALT
Resnet101	None	76.3	90.4
	Flip, Crop ± 5	69.5	77.4
	Flip, Crop	65.4	68.0
VGG16	None	77.4	90.8
	Flip, Crop ± 5	71.3	79.5
	Flip, Crop	63.8	64.3

=> Data augmentation decreases membership attacks accuracy

Conclusion

- Black-box attacks as good as white-box attacks
- Our approximations for membership attacks are state-of-the-art on two datasets

White-box vs Black-box: Bayes Optimal Strategies for Membership Inference

Poster 172

Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier,
Cordelia Schmid, Hervé Jégou

Facebook AI Research, Paris

June 20th, 2018