Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami

May 24th, 2016 @ 37th IEEE Symposium on Security and Privacy

@NicolasPapernot


M components
N components
p0=0.01
p1=0.93
p8=0.02
pN=0.01
STOP
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)


M components
N components
STOP
p0=0.01
p1=0.02
p8=0.89
pN=0.01
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, ...)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

PayPlug

PayPal

invincea®

deepinstinct

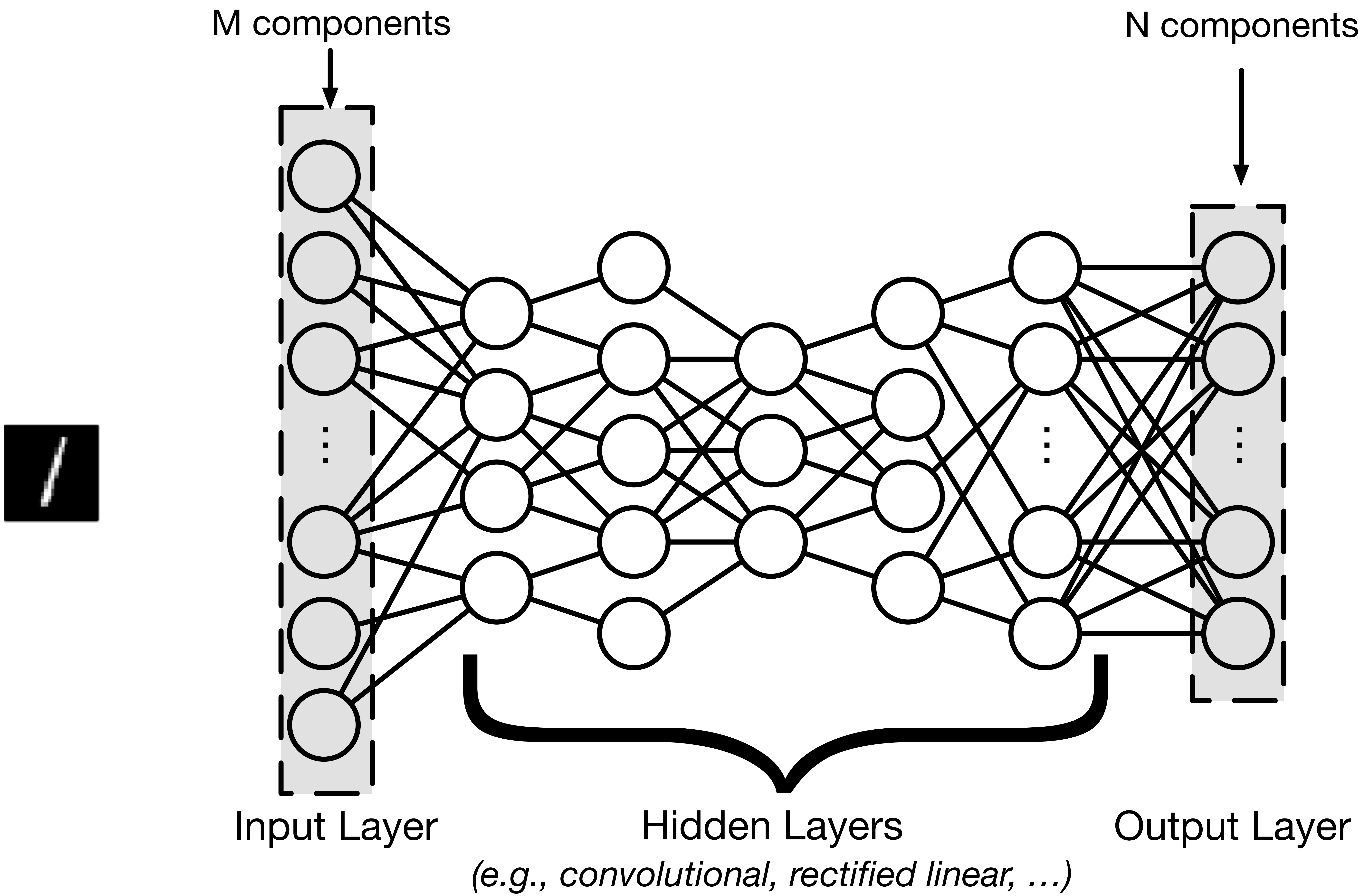
deep genomics

enlitic

AlchemyAPI™
An IBM Company

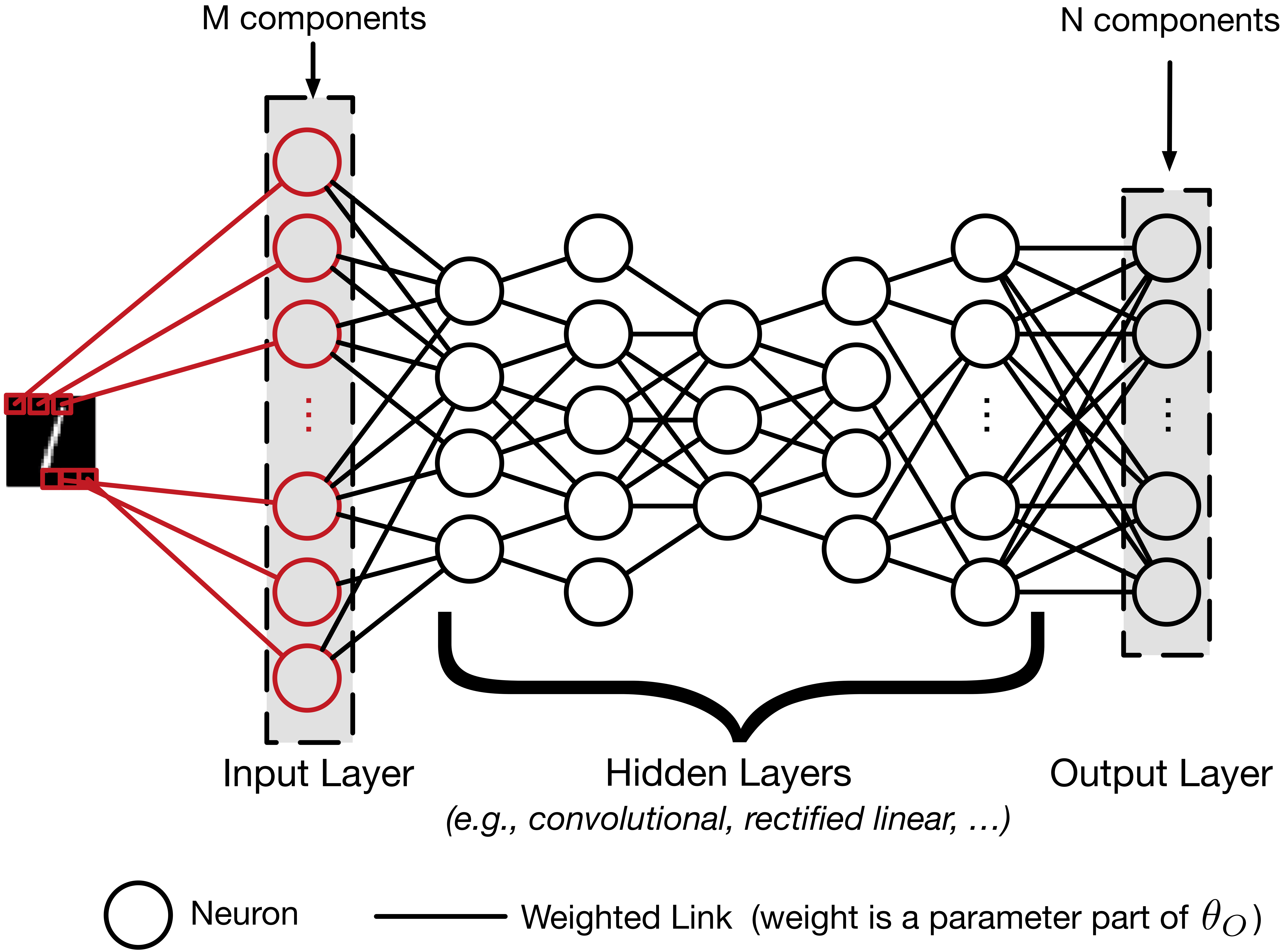
MetaMind

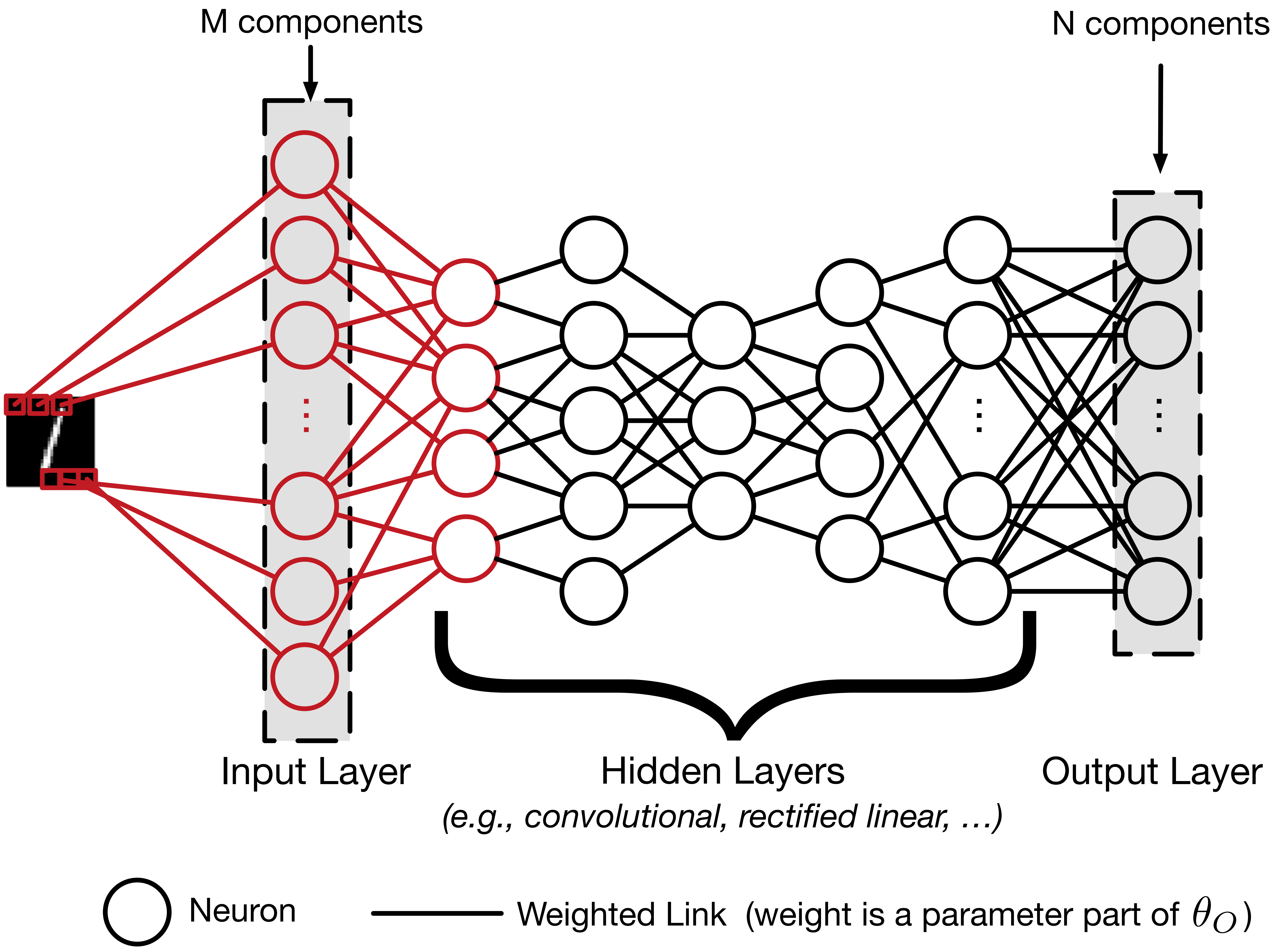
Deep Learning for Classification

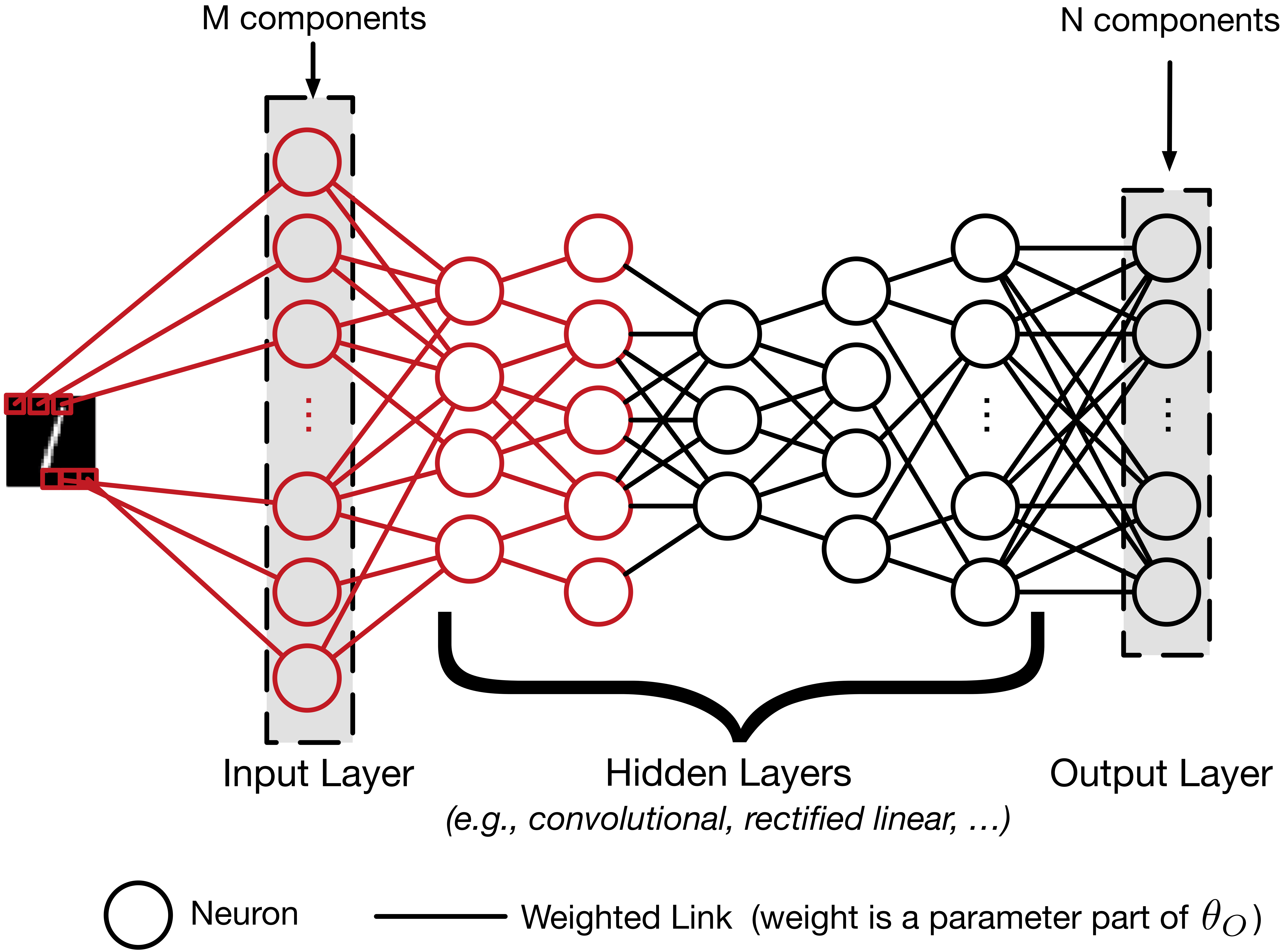


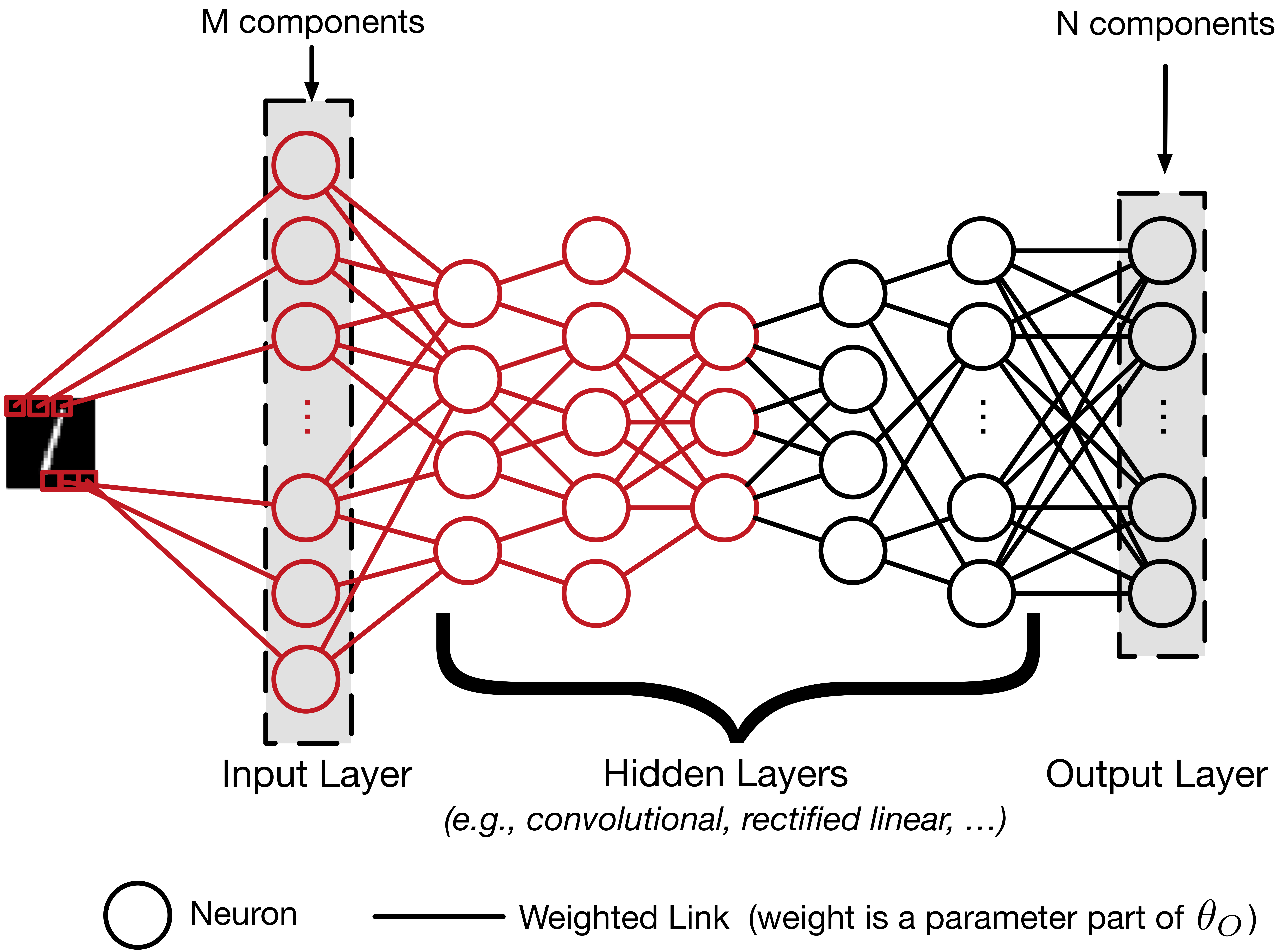
Neuron

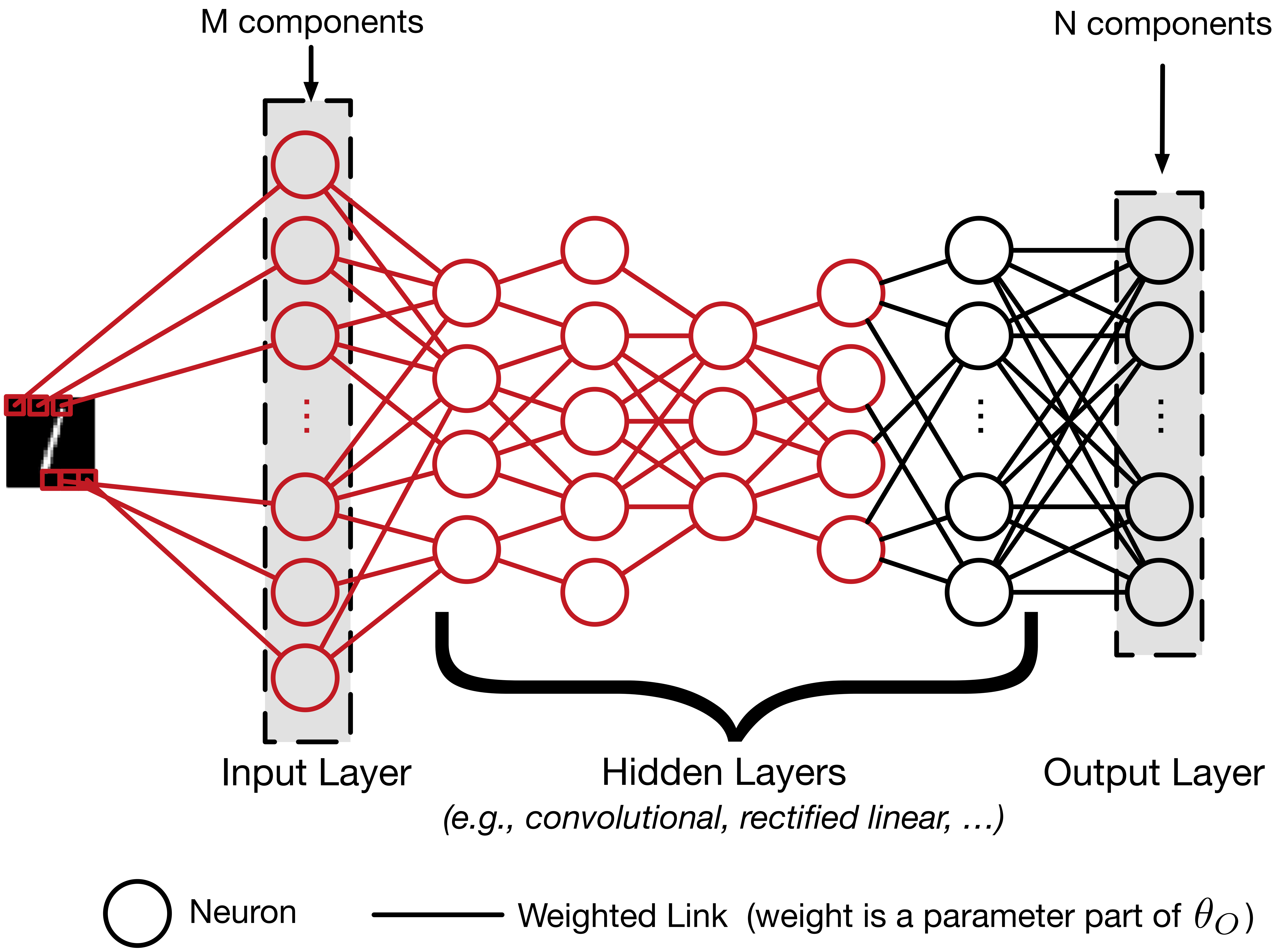
Weighted Link (weight is a parameter part of θ_O)

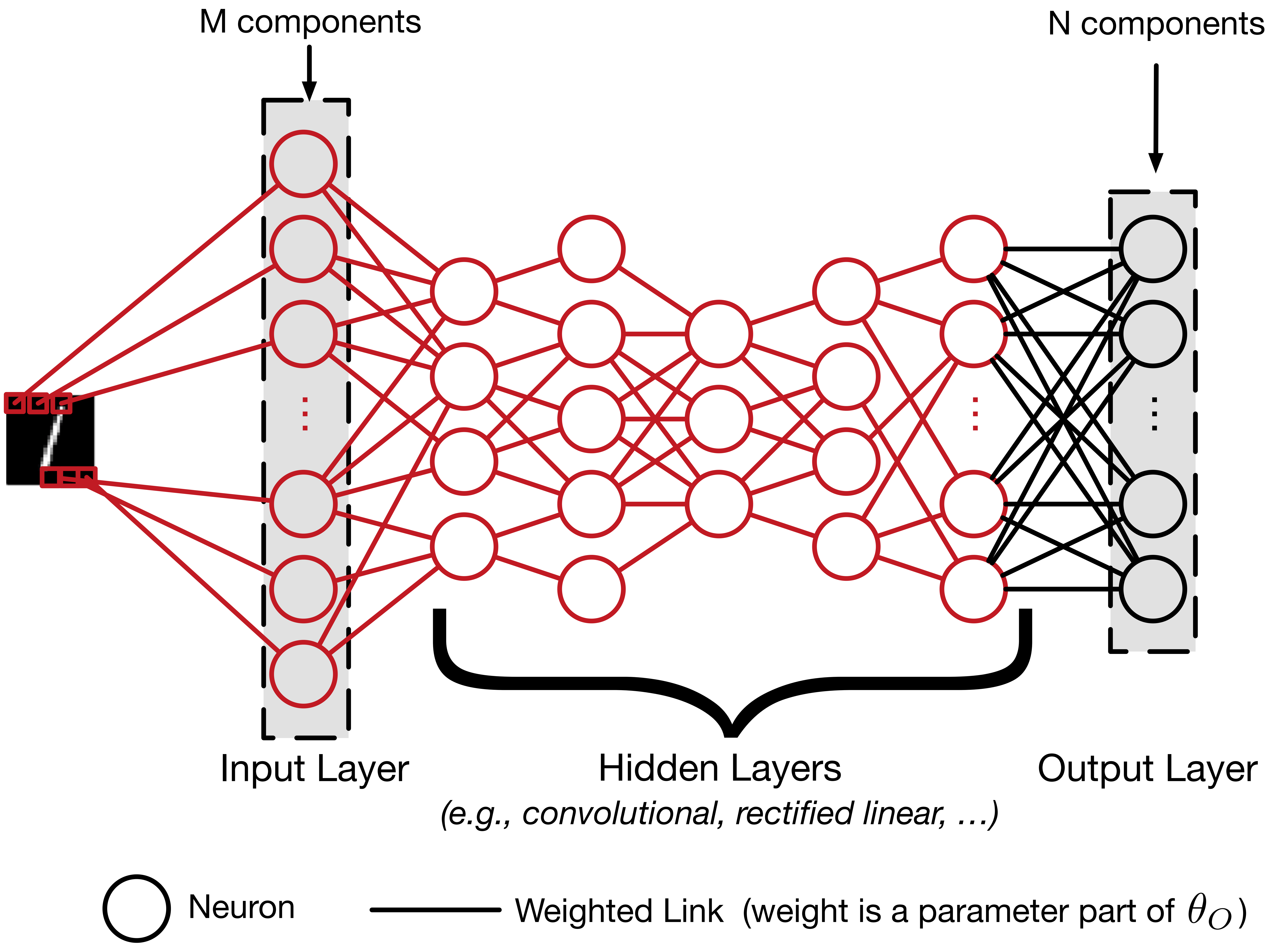

M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

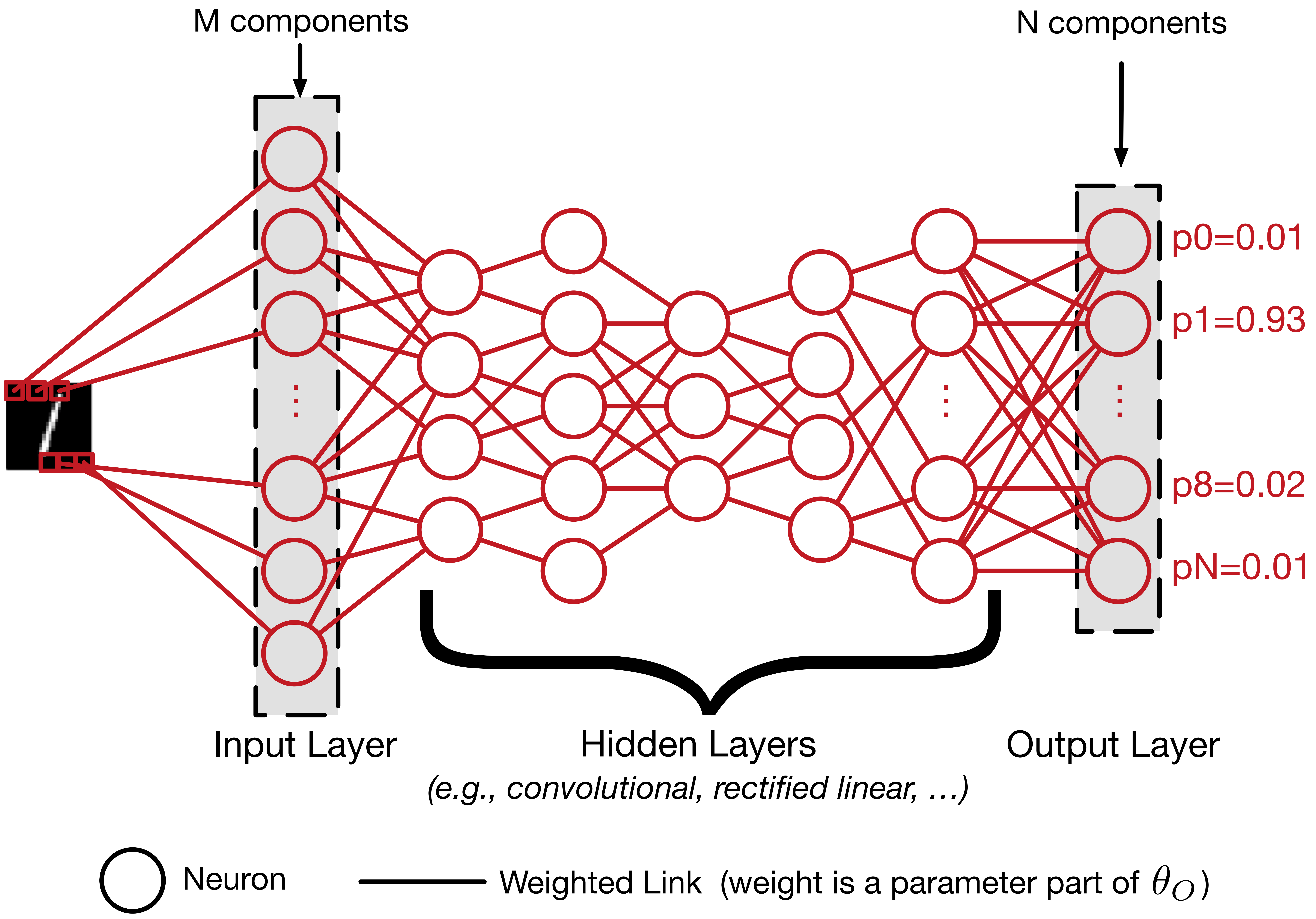

M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, ...)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

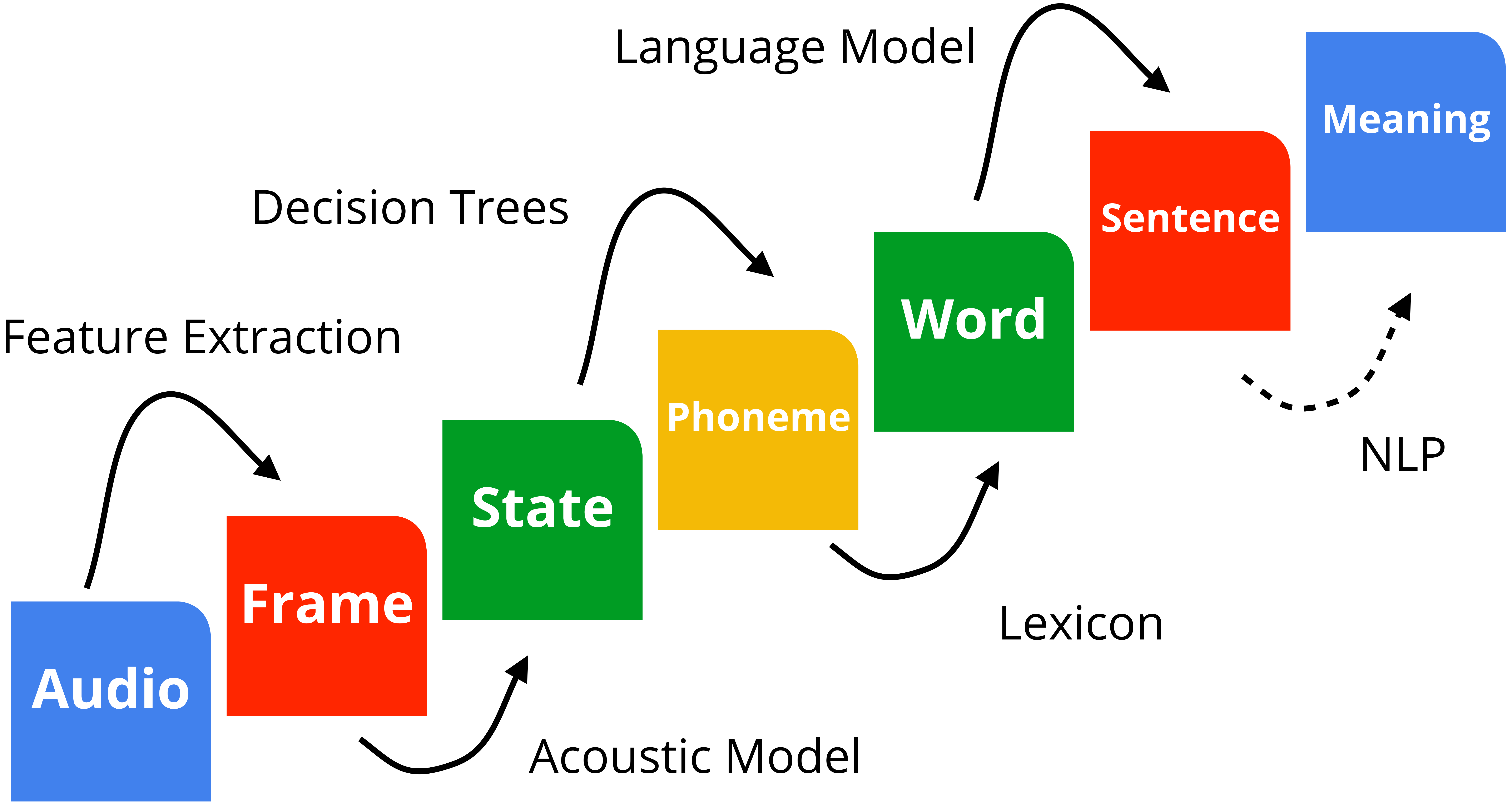

M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)


M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, ...)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)


M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

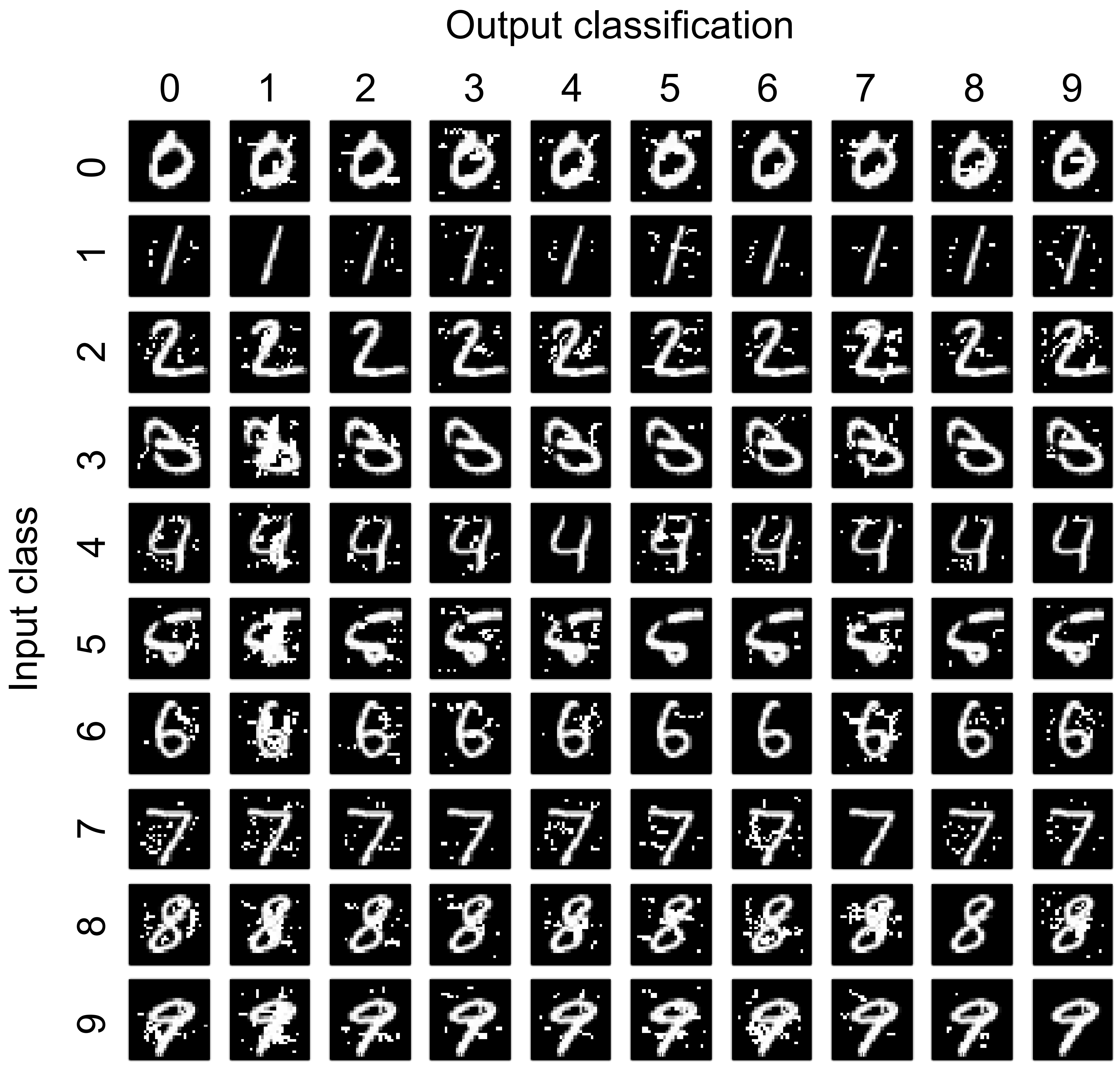

M components
N components
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

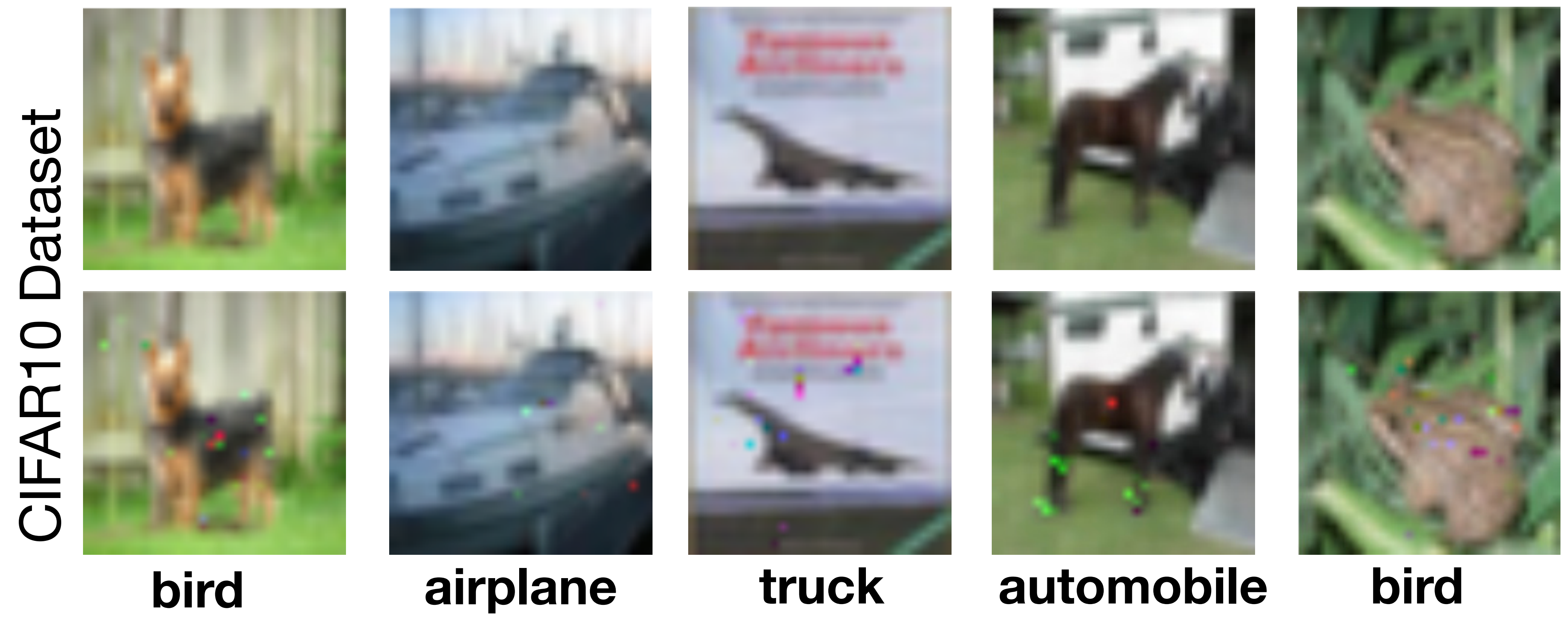

M components
N components
p0=0.01
p1=0.93
p8=0.02
pN=0.01
Input Layer
Hidden Layers
(e.g., convolutional, rectified linear, …)
Output Layer
Neuron
Weighted Link (weight is a parameter part of θ_O)

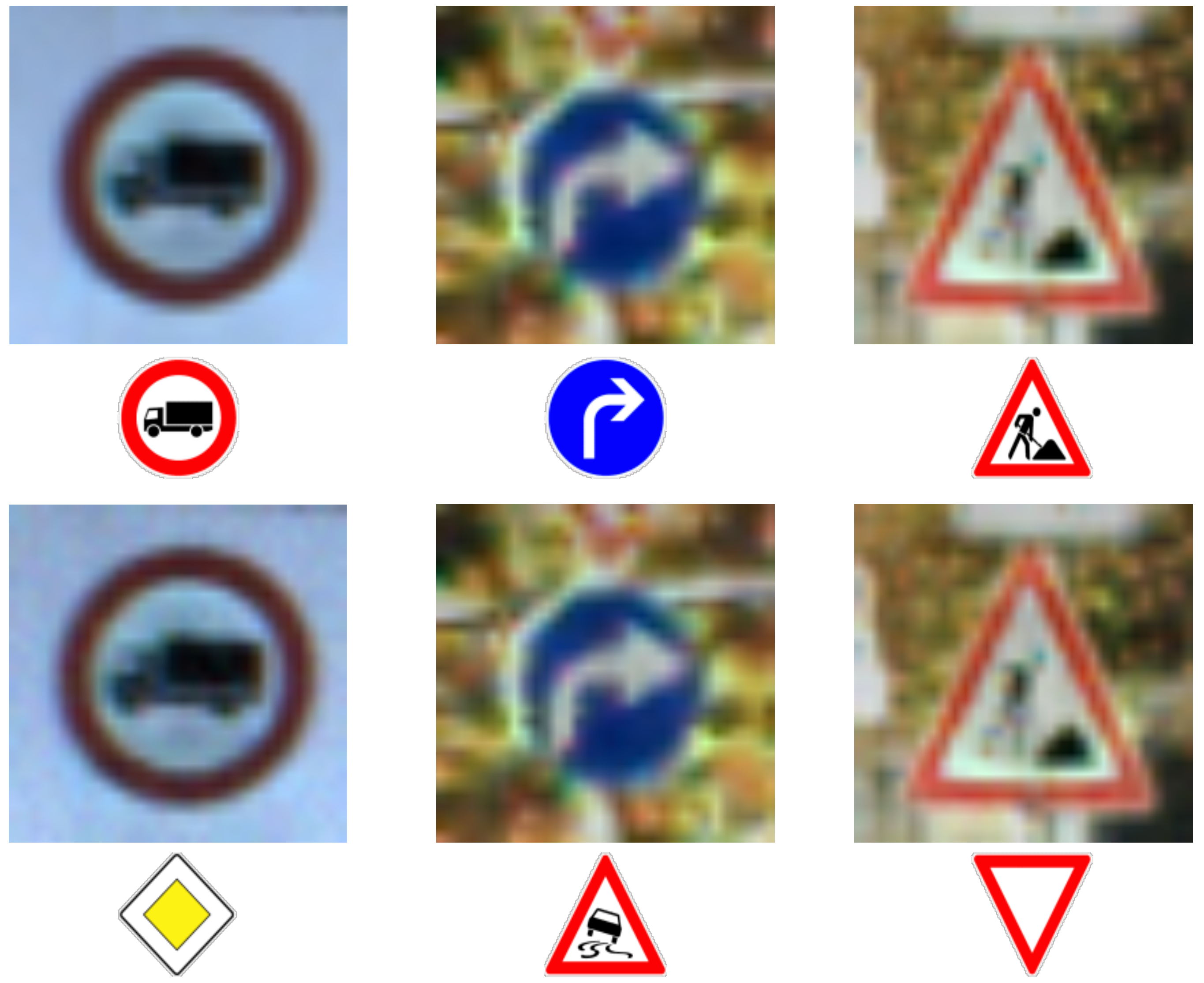


Source: Tara N. Sainath, Google @ ICML DL Workshop 2015

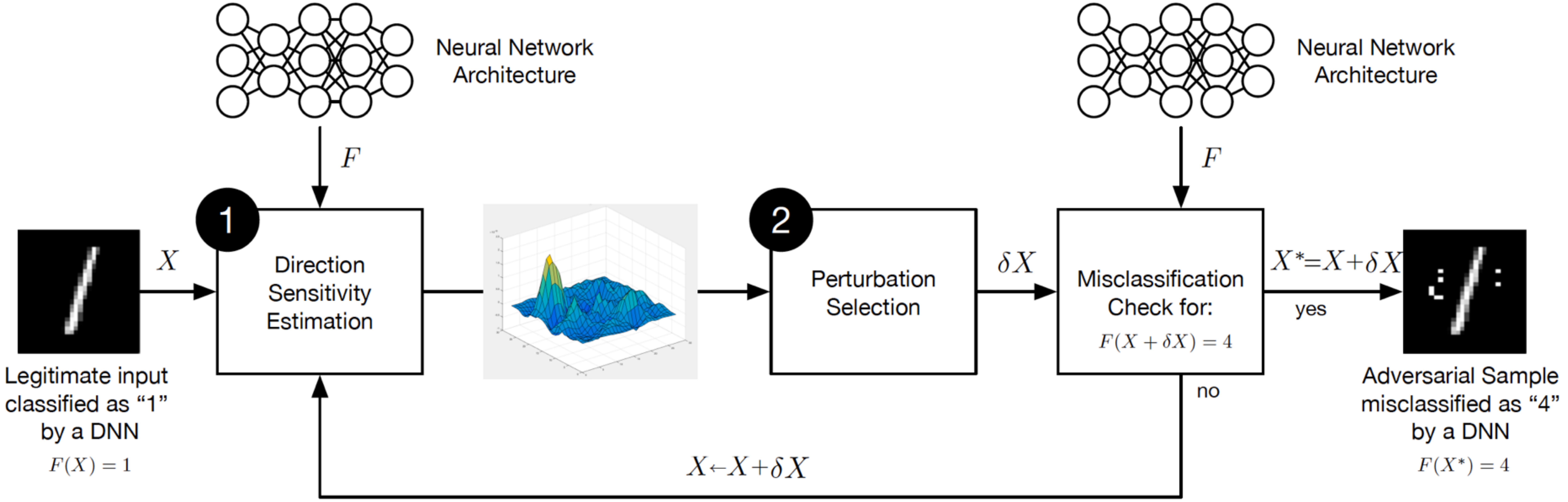
Adversarial Samples


Output classification
0 1 2 3 4 5 6 7 8 9
Input class
0 1 2 3 4 5 6 7 8 9


CIFAR10 Dataset
bird
airplane
truck
automobile
bird

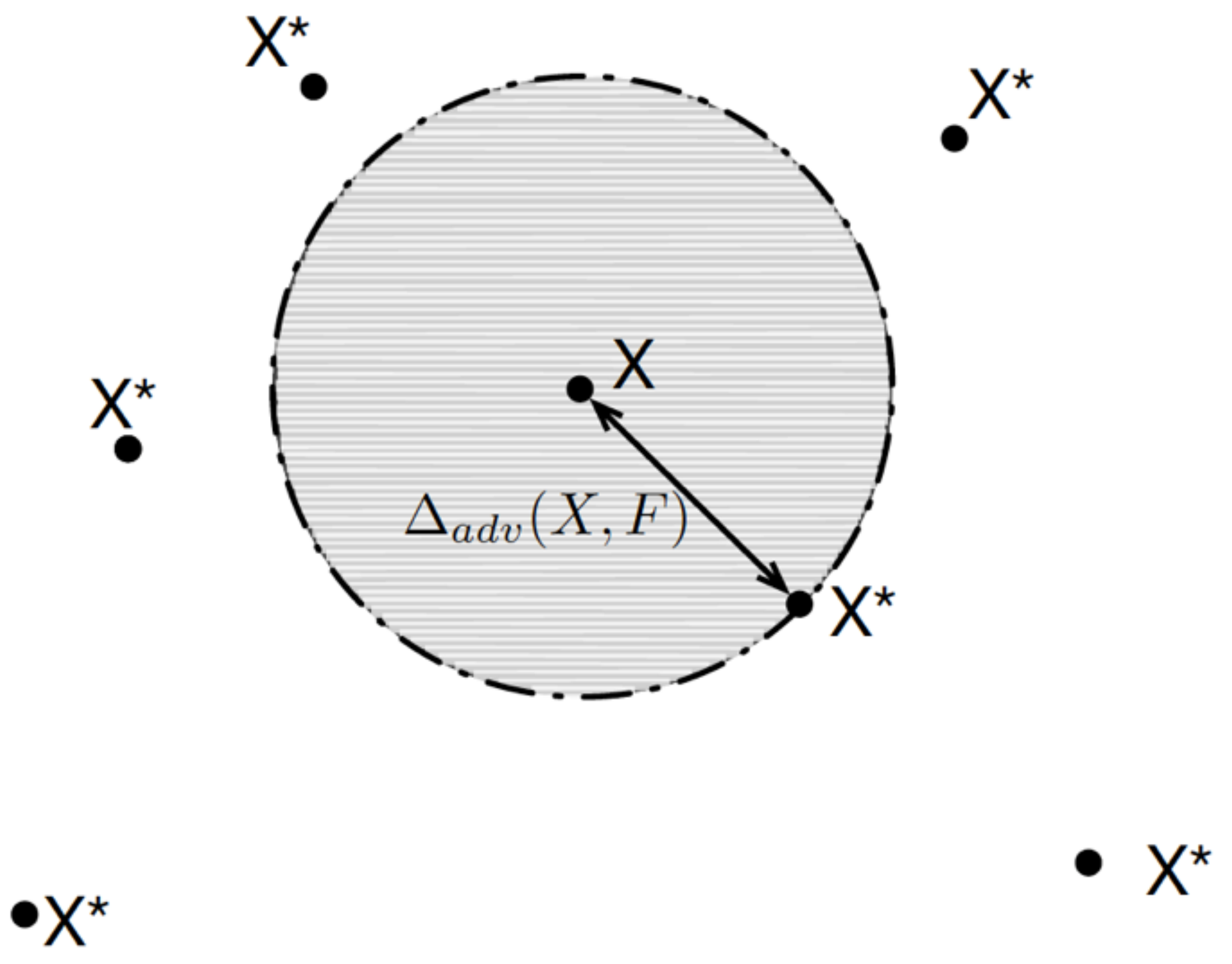


Adversarial strategy



Defending against Adversarial Perturbations

DNN Robustness



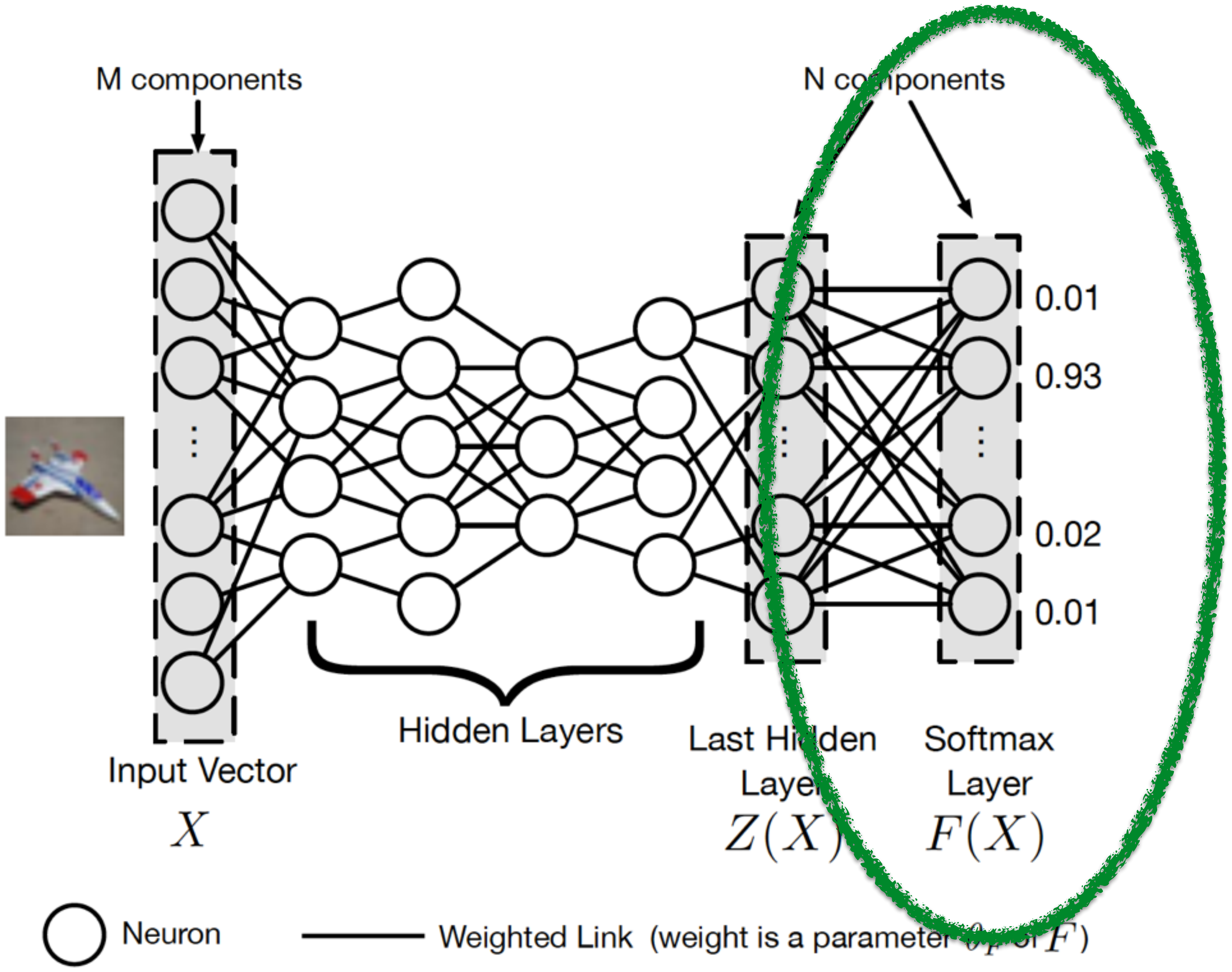
$$\rho_{adv}(F) = E_\mu[\Delta_{adv}(X, F)]$$

$$\Delta_{adv}(X, F) = \arg\min_{\delta X}\{\|\delta X\| : F(X + \delta X) \neq F(X)\}$$

Defense Design

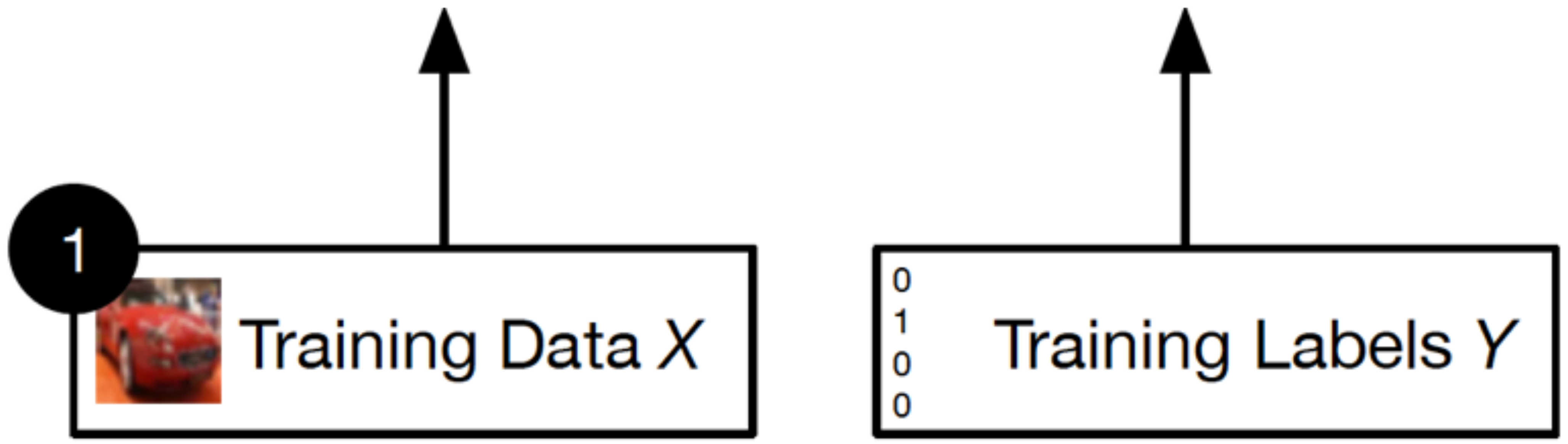
- **Low impact on the architecture**
- Maintain **accuracy**
  - Robust in space **relatively close to the legitimate distribution**
- Maintain **speed** of network

Softmax Layer and Probabilities

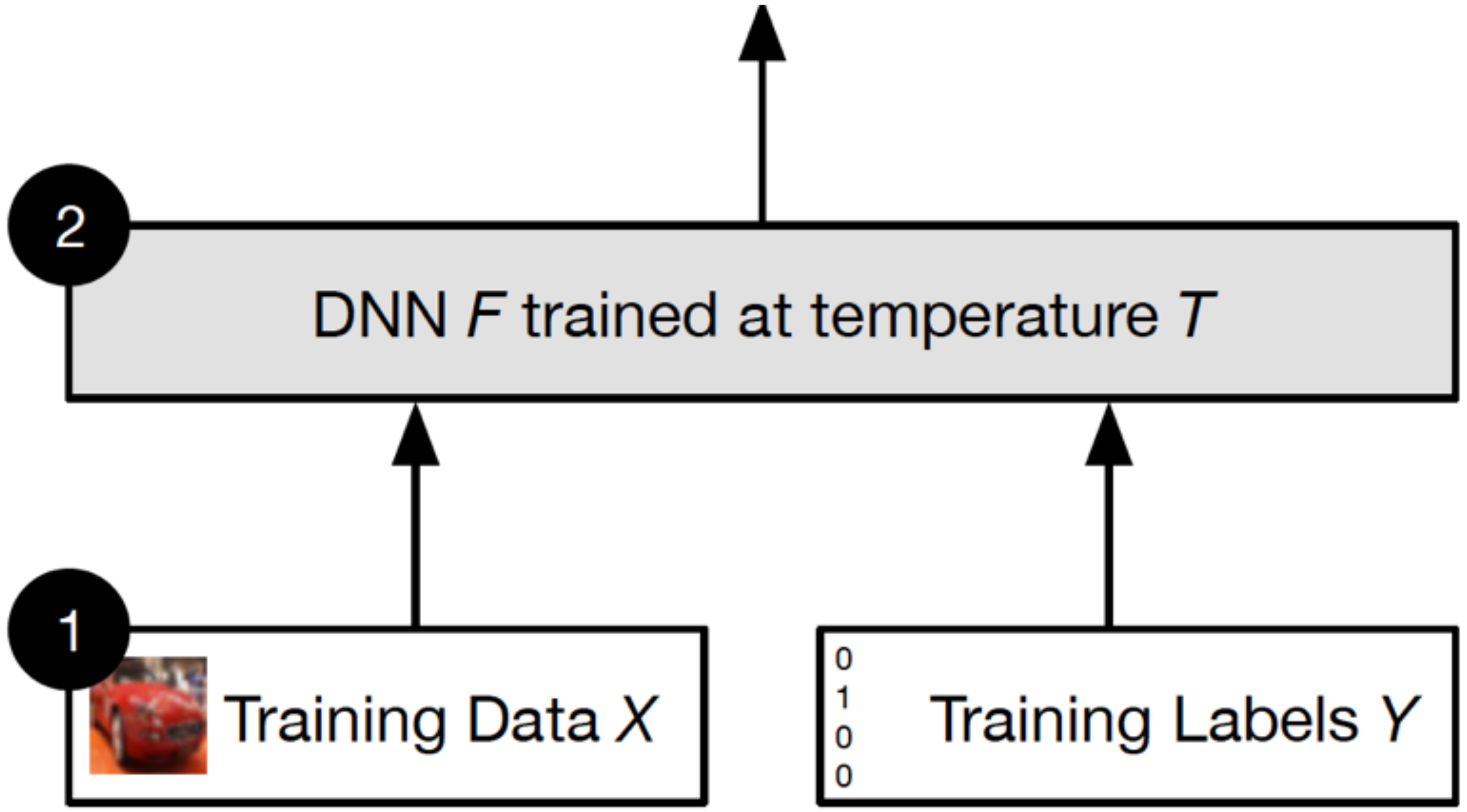


$$F(X) = \left[\frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X)/T}} \right]_{i \in 0..N-1}$$

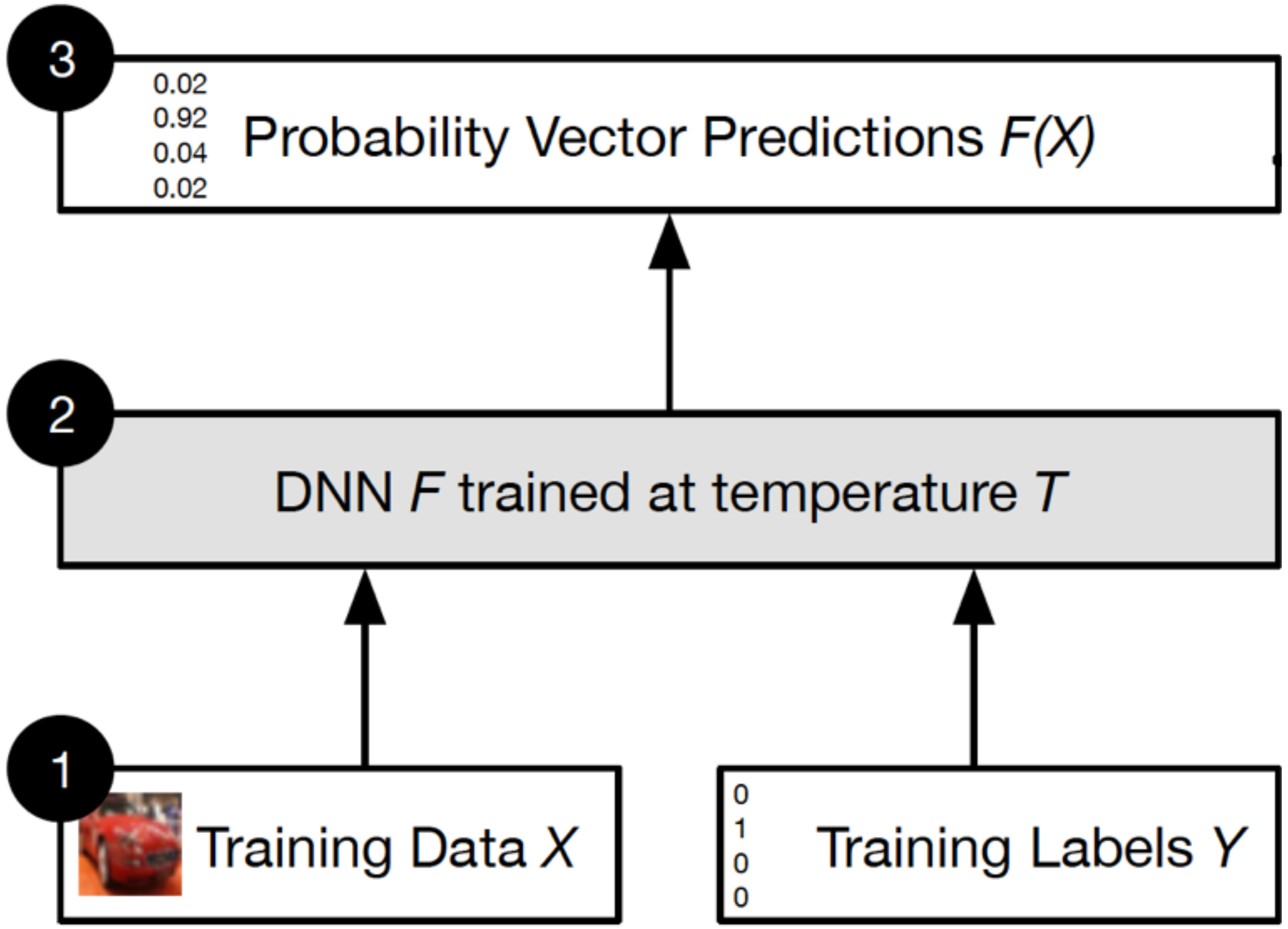
Defensive Distillation



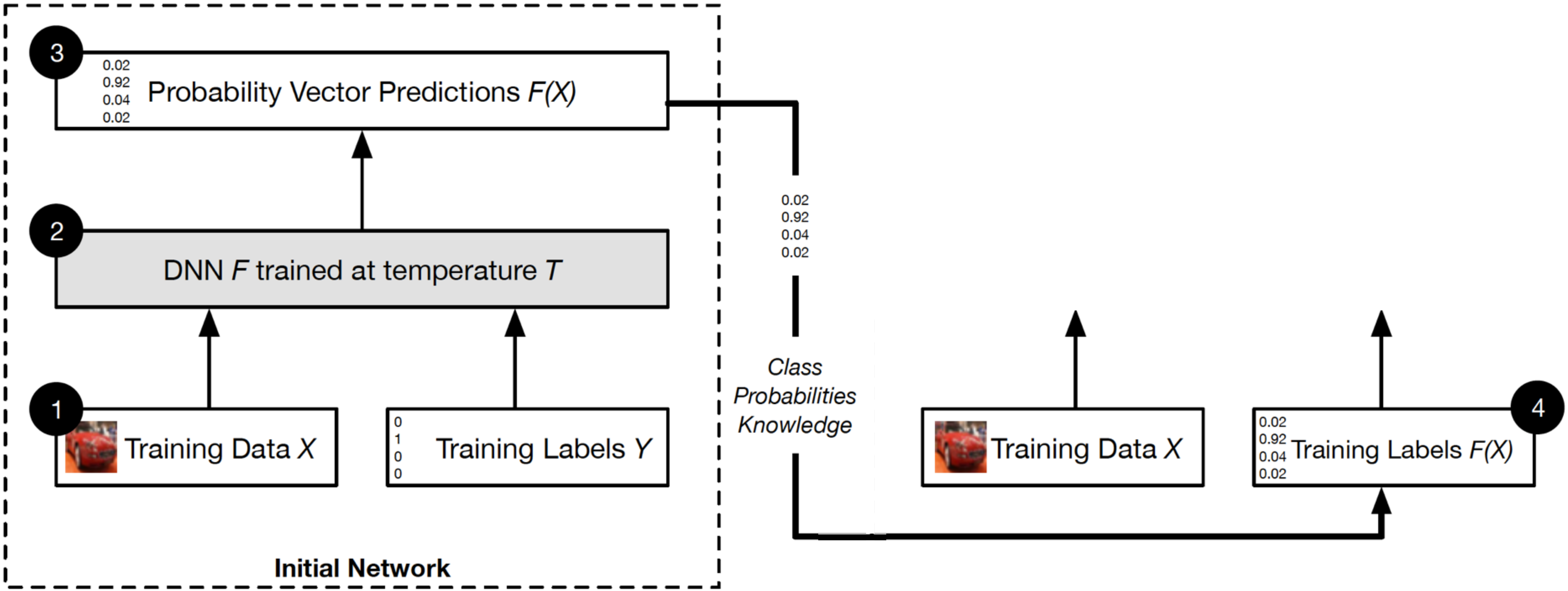
Defensive Distillation



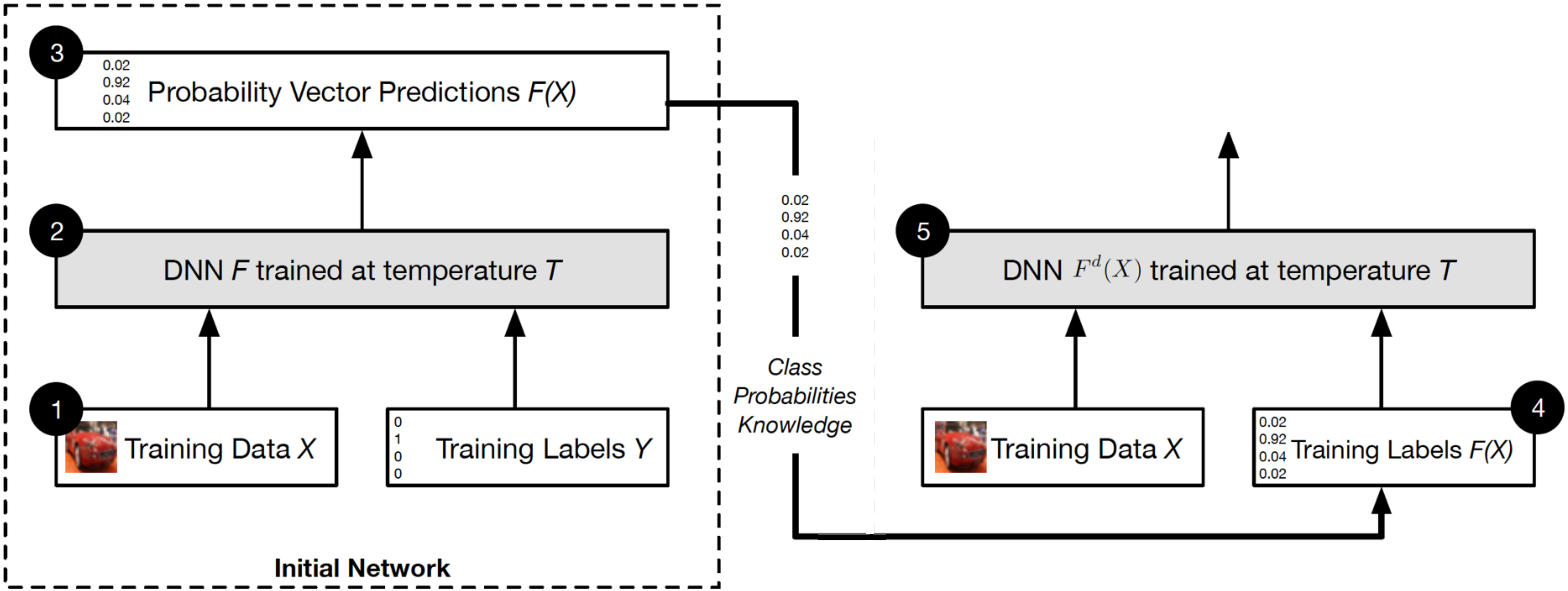
Defensive Distillation



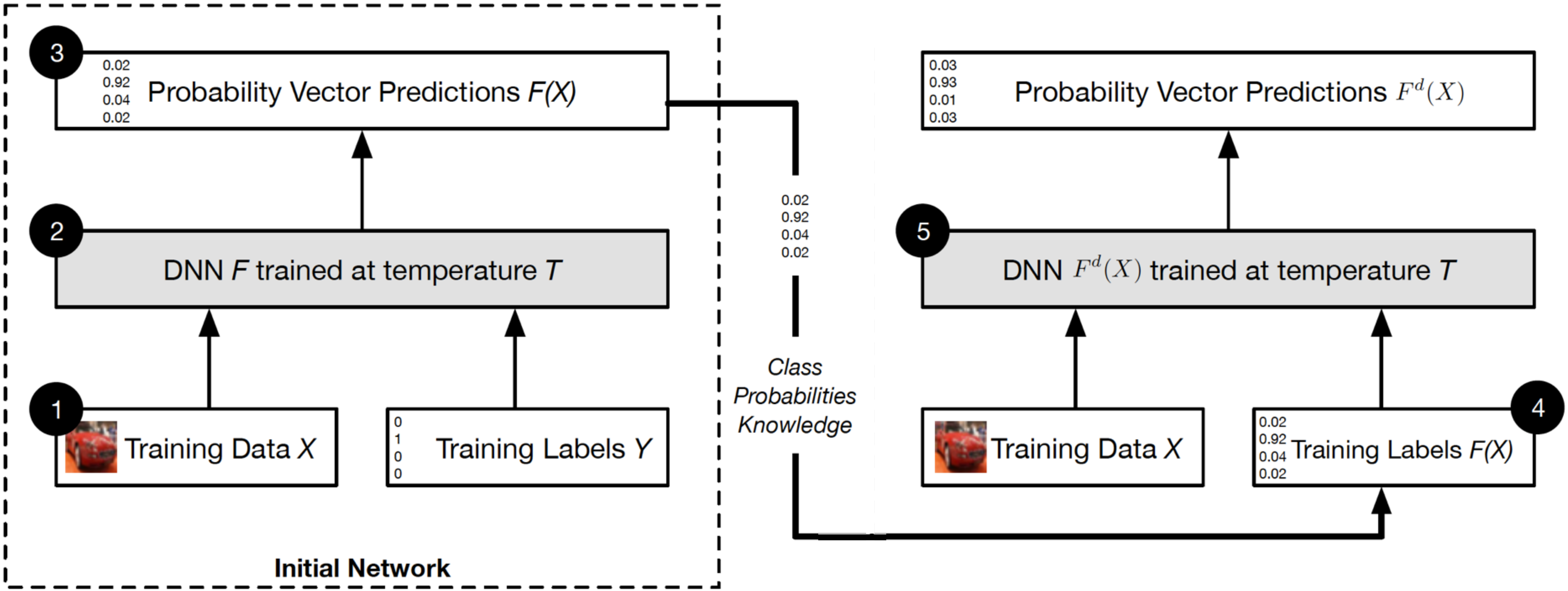
Defensive Distillation



Defensive Distillation



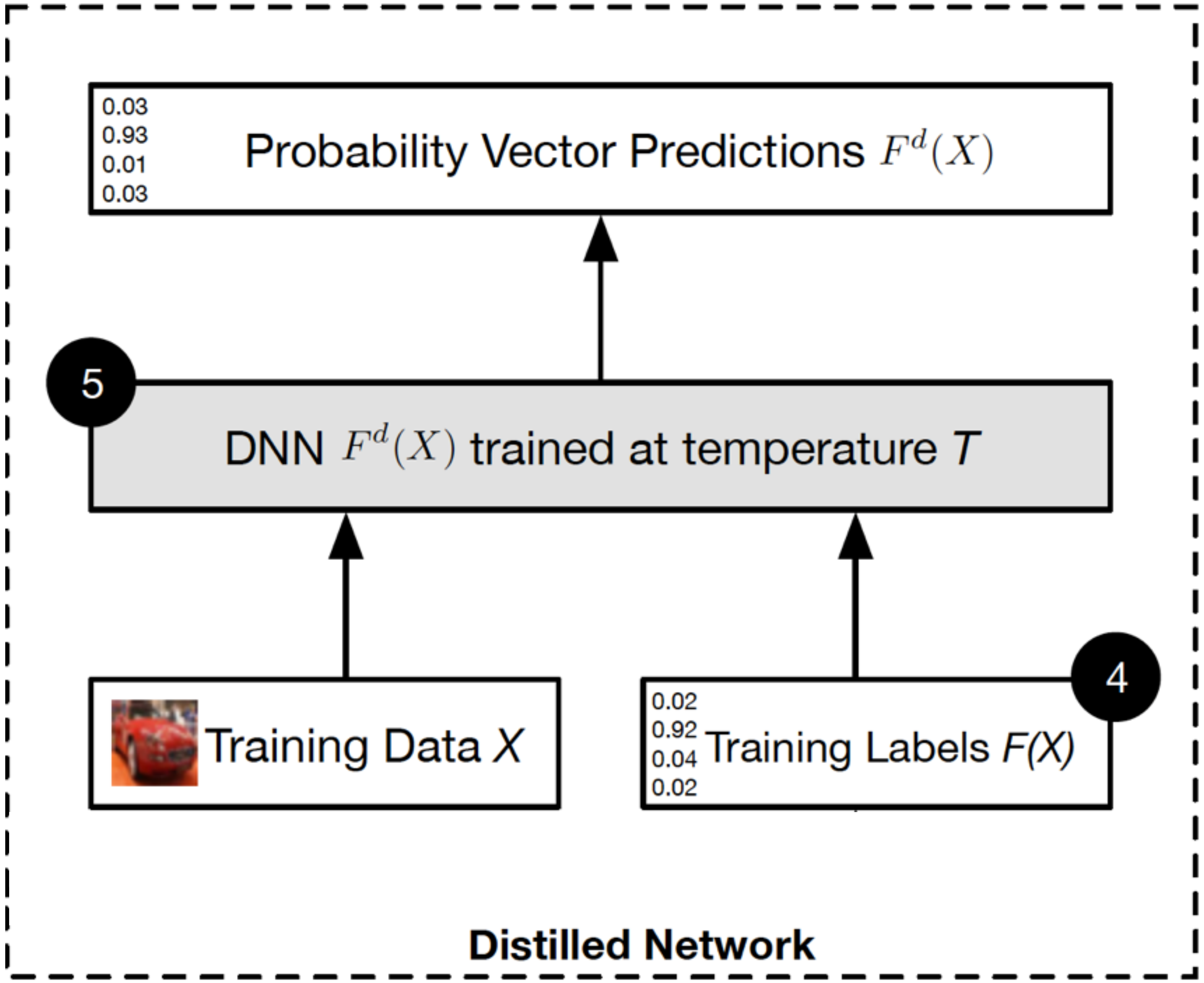
Defensive Distillation



Defensive Distillation

Defensive Distillation

Set temperature T=1
for predictions



Intuition behind Defensive Distillation

Constraining Training

$$\arg\min_{\theta_F} -\frac{1}{|\mathcal{X}|} \sum_{X\in\mathcal{X}} \sum_{i\in 0..N} Y_i(X) \log F_i(X)$$

0 if i not correct class

$$\arg\min_{\theta_F} -\frac{1}{|\mathcal{X}|} \sum_{X\in\mathcal{X}} \sum_{i\in 0..N} F_i(X) \log F_i^d(X)$$

never equal to 0

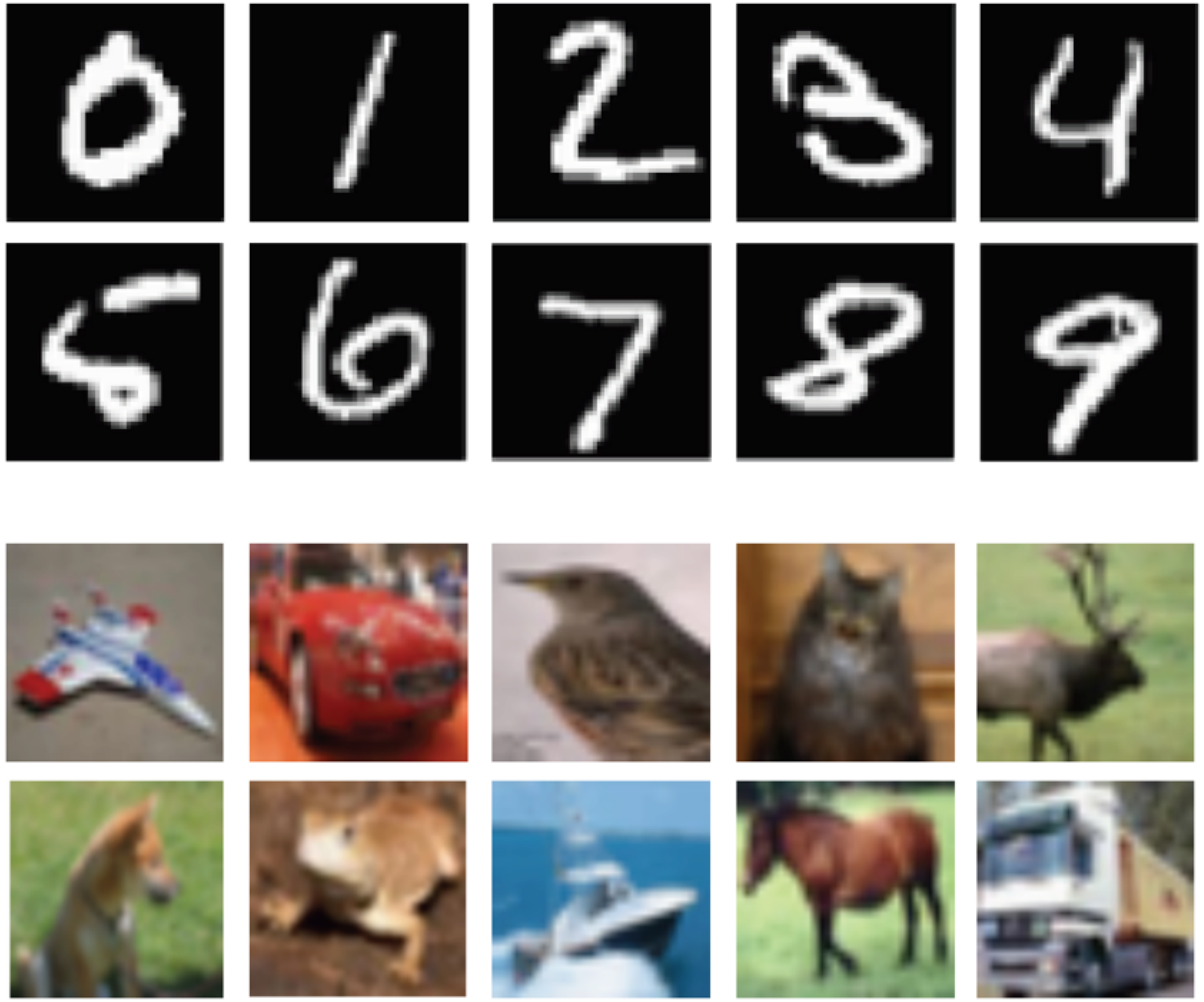
 

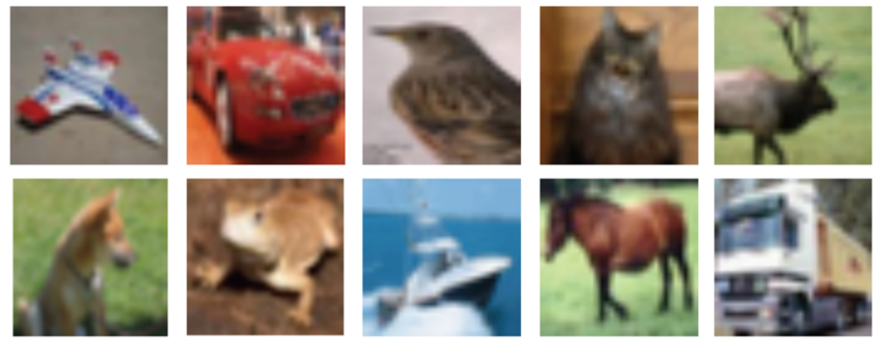
Reducing Jacobian Amplitudes

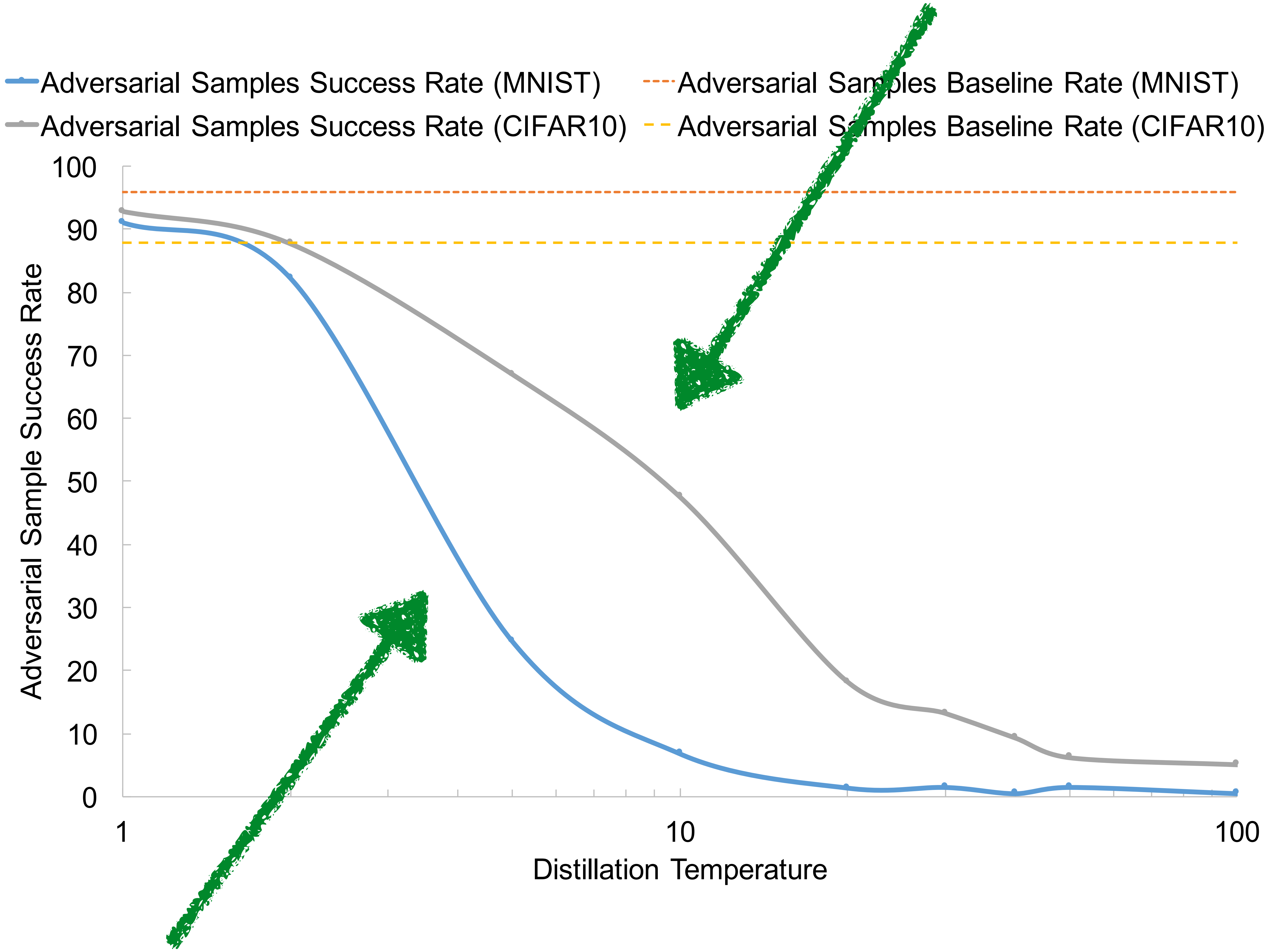
$$J_F(T,i,j) = \frac{1}{T}\frac{e^{z_i/T}}{g^2(X)}\left(\sum_{l=0}^{N-1}\left(\frac{\partial z_i}{\partial X_j} - \frac{\partial z_l}{\partial X_j}\right) e^{z_l/T}\right)$$

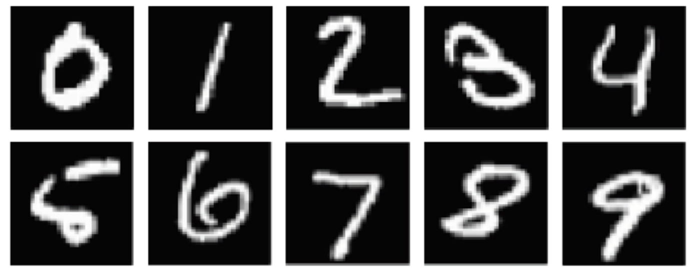
Validation

Experimental Setup



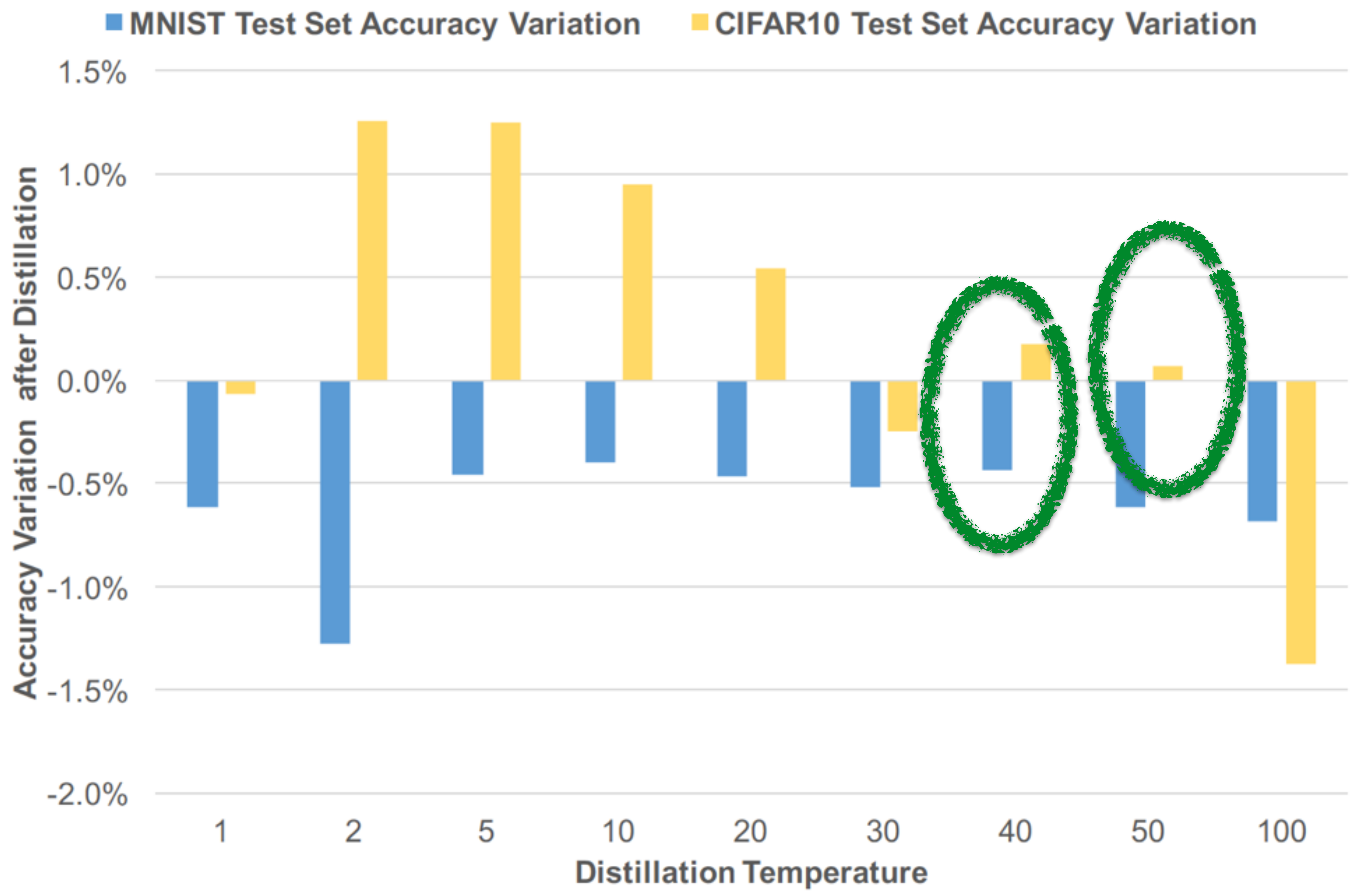






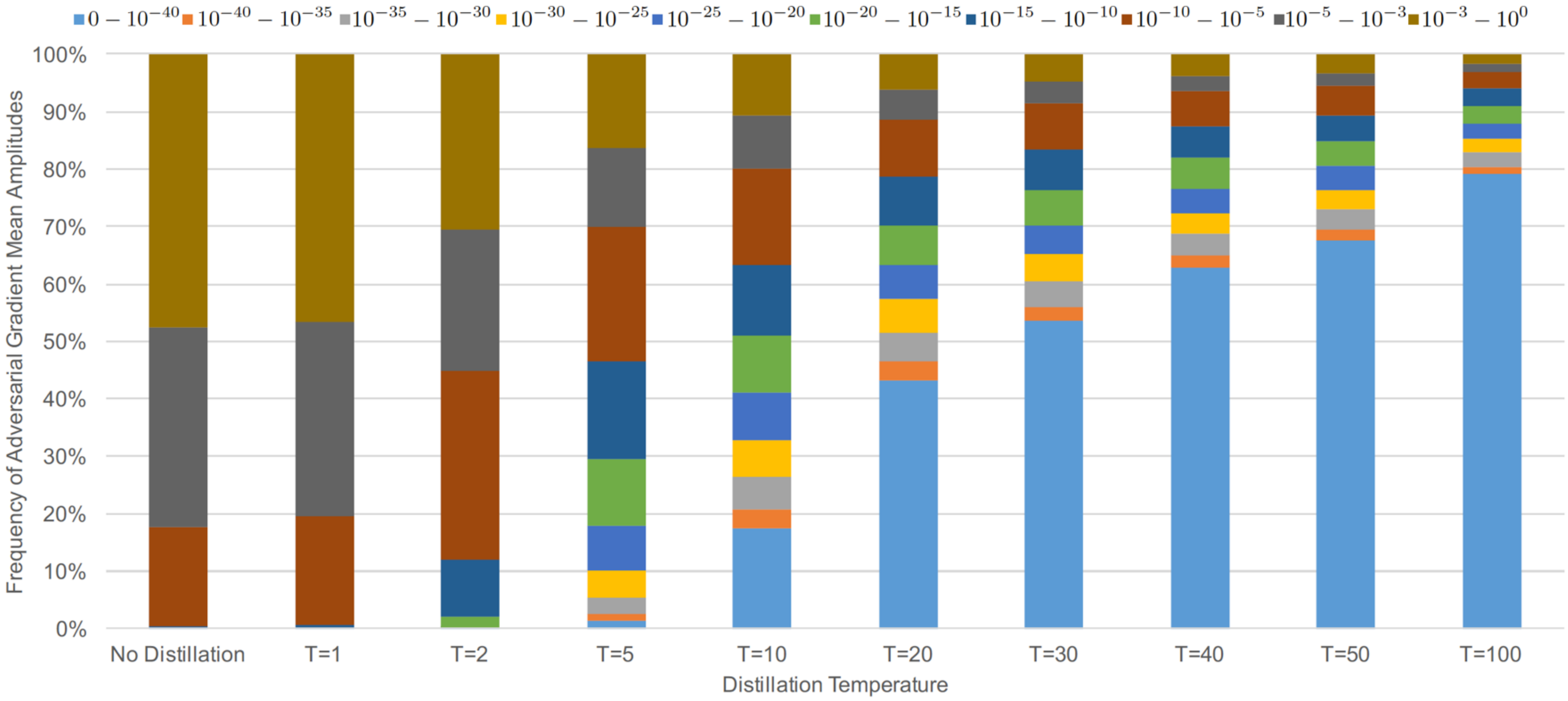
Distillation Temperature	MNIST Adversarial Samples Success Rate (%)	CIFAR10 Adversarial Samples Success Rate (%)
1	91	92.78
2	82.23	87.67
5	24.67	67
10	6.78	47.56
20	1.34	18.23
30	1.44	13.23
40	0.45	9.34
50	1.45	6.23
100	0.45	5.11
No distillation	95.89	87.89

Impact on accuracy

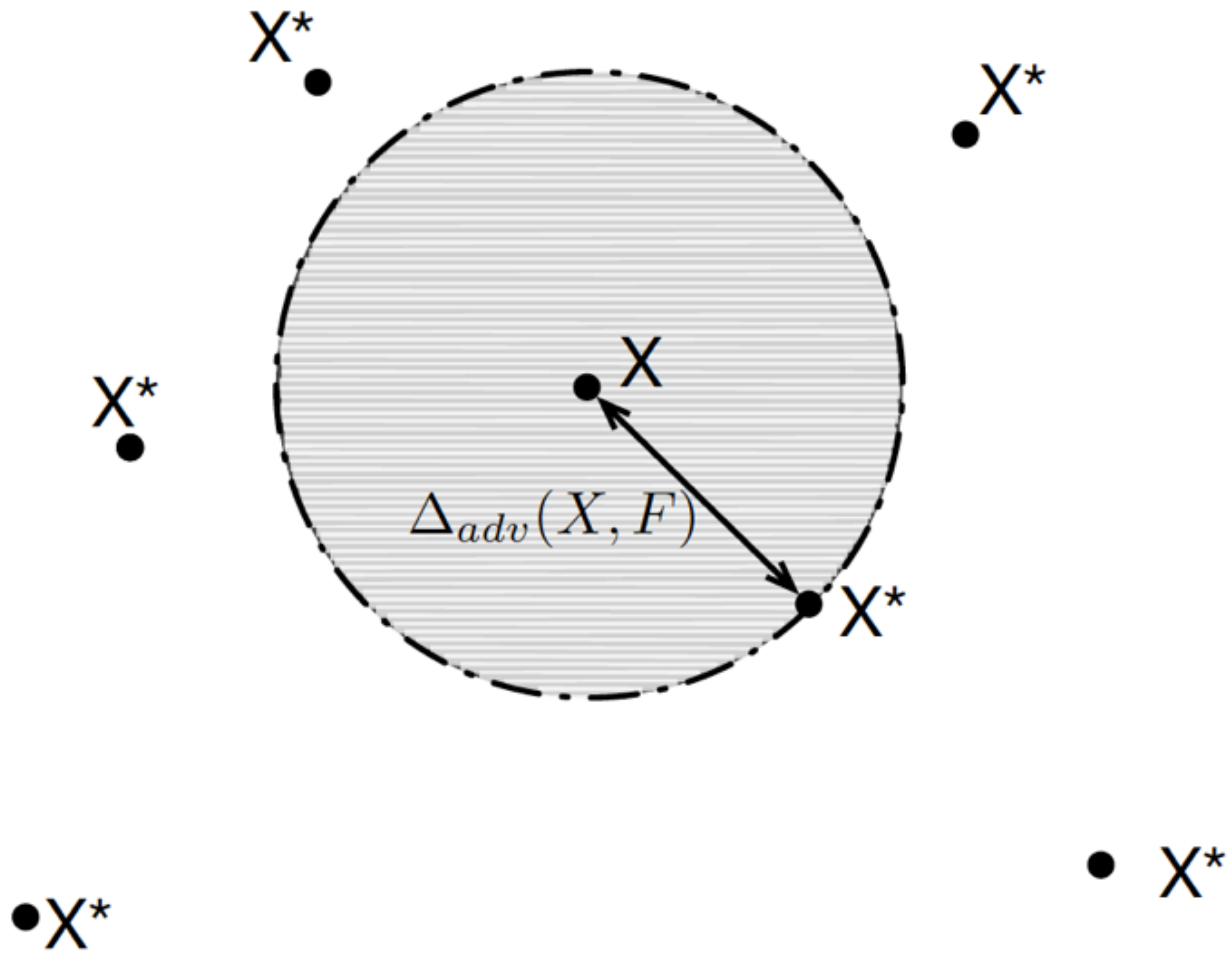


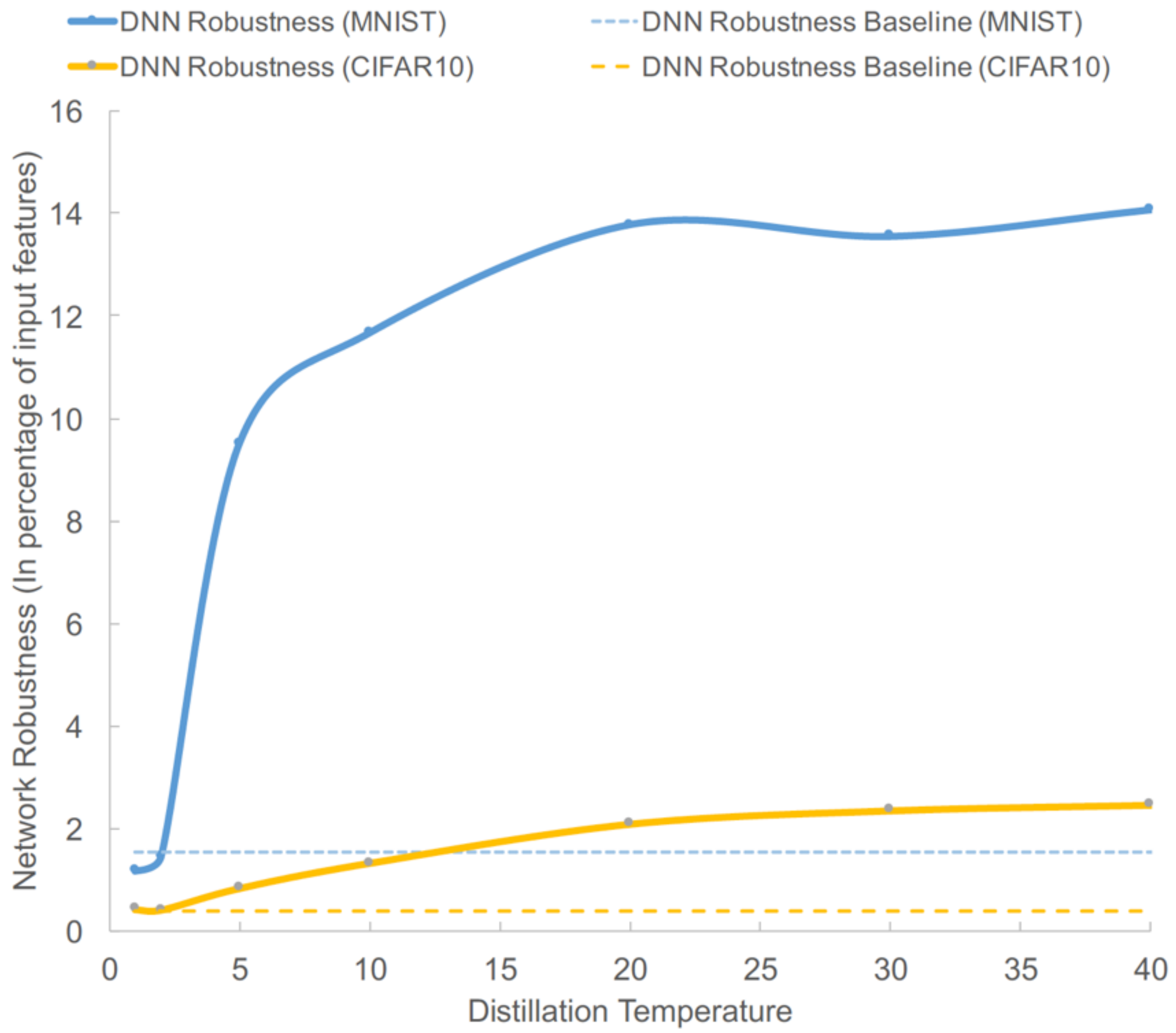
Distillation Temperature	MNIST Adversarial Samples Success Rate (%)	CIFAR10 Adversarial Samples Success Rate (%)
1	91	92.78
2	82.23	87.67
5	24.67	67
10	6.78	47.56
20	1.34	18.23
30	1.44	13.23
40	0.45	[illegible]
50	1.45	6.23
100	0.45	[illegible]
No distillation	95.89	87.89

Impact on Jacobian Amplitude



Estimation of Robustness





Conclusions

Take aways

- Distillation significantly reduces attack success
- Yields model smoothness
- Easy implementation, low overhead
- Acceptable impact on accuracy

Questions?

https://www.papernot.fr
nicolas@papernot.fr
@NicolasPapernot