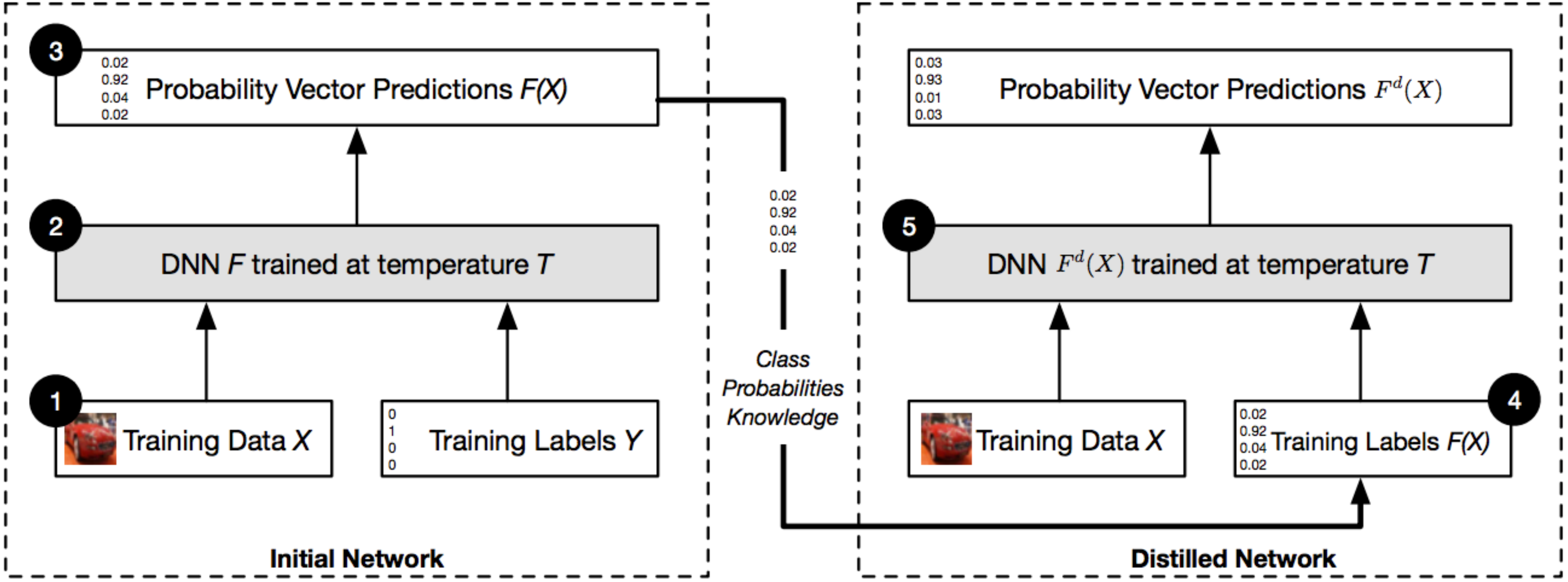
ADVERSARIAL MACHINE LEARNING

DEFENSE: DISTILLATION

36


3
0.02
0.92
0.04
0.02
Probability Vector Predictions F(X)
2
DNN F trained at temperature T
1
Training Data X
0
1
0
0
Training Labels Y
Initial Network
0.02
0.92
0.04
0.02
Class
Probabilities
Knowledge
0.03
0.93
0.01
0.03
Probability Vector Predictions $F^d(X)$
5
DNN $F^d(X)$ trained at temperature T
Training Data X
0.02
0.92
0.04
0.02
Training Labels F(X)
4
Distilled Network

An overview of distillation defence mechanism based on a transfer of knowledge contained in probability vectors through distillation: We first train an initial network F on data X with a softmax temperature of T. We then use the probability vector F(X), which includes additional knowledge about classes compared to a class label, predicted by network F to train a distilled network F d at temperature T on the same data X.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks.

DEFENSE: EXTENDED DEFENSIVE DISTILLATION

- Uncertainty Measure:

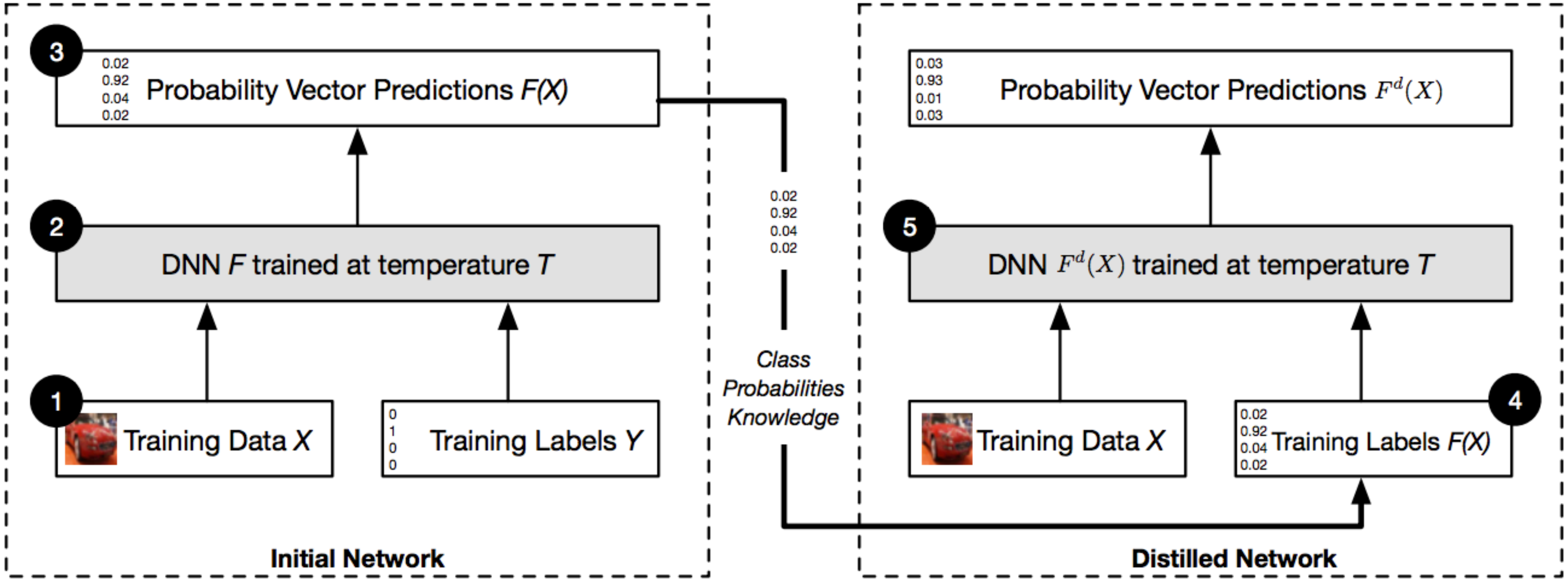
$$\sigma(x) = \frac{1}{N} \sum_{m \in 0..N-1} \left(\sum_{j \in 0...n-1} (z_j^m(x) - \bar(z_j)^2) \right)$$

- Labelling Vector:

$$k_j(x) = \begin{cases} 1 - \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = l \text{ (correct label)} \\ \alpha \cdot \frac{\sigma(x)}{\max_{x \in \chi} \sigma(x)} & \text{if } j = n \text{ (outlier class)}) \\ 0 & \end{cases}$$

Papernot, N., & McDaniel, P. (2017). Extending Defensive Distillation.

DEFENSE: DISTILLATION



An overview of distillation defence mechanism based on a transfer of knowledge contained in probability vectors through distillation: We first train an initial network F on data X with a softmax temperature of T. We then use the probability vector F(X), which includes additional knowledge about classes compared to a class label, predicted by network F to train a distilled network F d at temperature T on the same data X.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016, May). Distillation as a defense to adversarial perturbations against deep neural networks.