ADVERSARIAL MACHINE LEARNING

DEFENSE: ENSEMBLE TRAINING

35

- Randomized Loss Function:

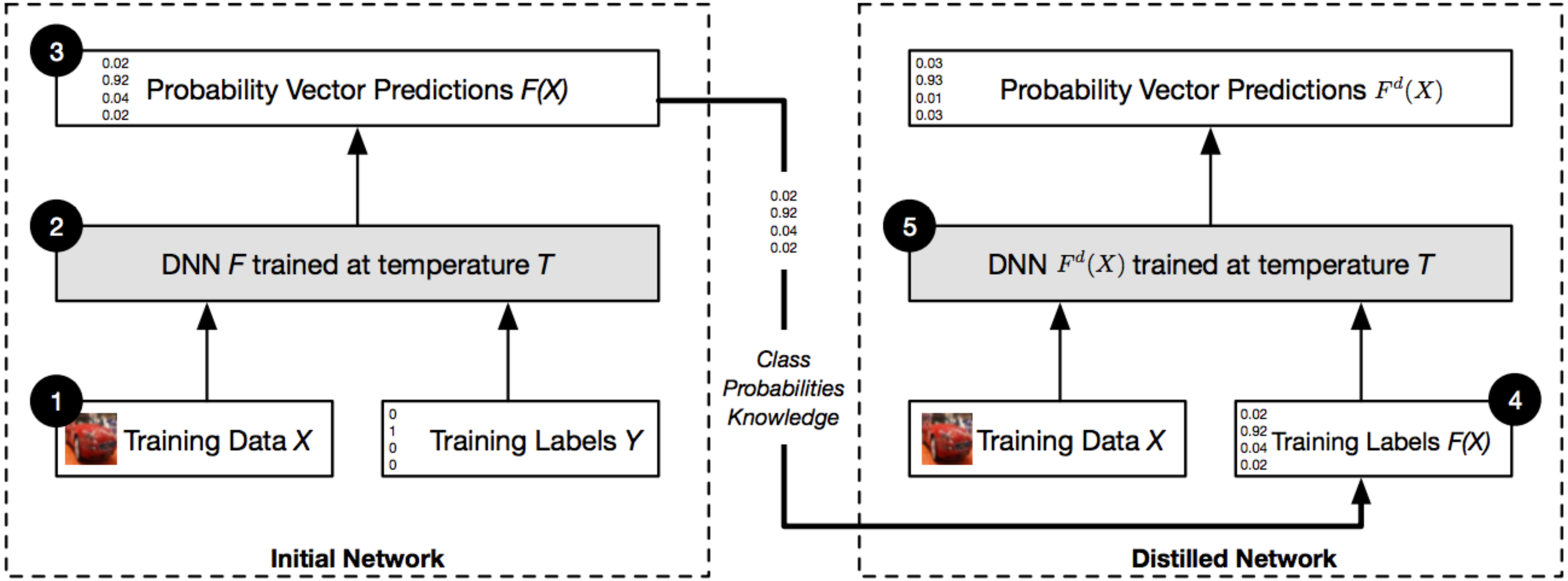
- Augment Adversarial Examples from the other models as well while training.

- Most used method in NIPS 2017 Adversarial Machine Learning challenge

$$
\begin{aligned}
x' &= x + \alpha \cdot \mathrm{sign}(\mathcal{N}(0^d, I^d)) \\
x^{adv} &= x' + (\epsilon - \alpha) \cdot \mathrm{sign}(\nabla_x L(x', y))
\end{aligned}
$$

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2017). Ensemble Adversarial Training: Attacks and Defenses.

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.

DEFENSE: ENSEMBLE TRAINING

- Randomized Loss Function:

$$x' = x + \alpha \cdot \mathrm{sign}(\mathcal{N}(0^d, I^d))$$

$$x^{adv} = x' + (\epsilon - \alpha) \cdot \mathrm{sign}(\nabla_x L(x', y))$$

- Augment Adversarial Examples from the other models as well while training.
- Most used method in NIPS 2017 Adversarial Machine Learning challenge

Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2017). Ensemble Adversarial Training: Attacks and Defenses.