ADVERSARIAL MACHINE LEARNING

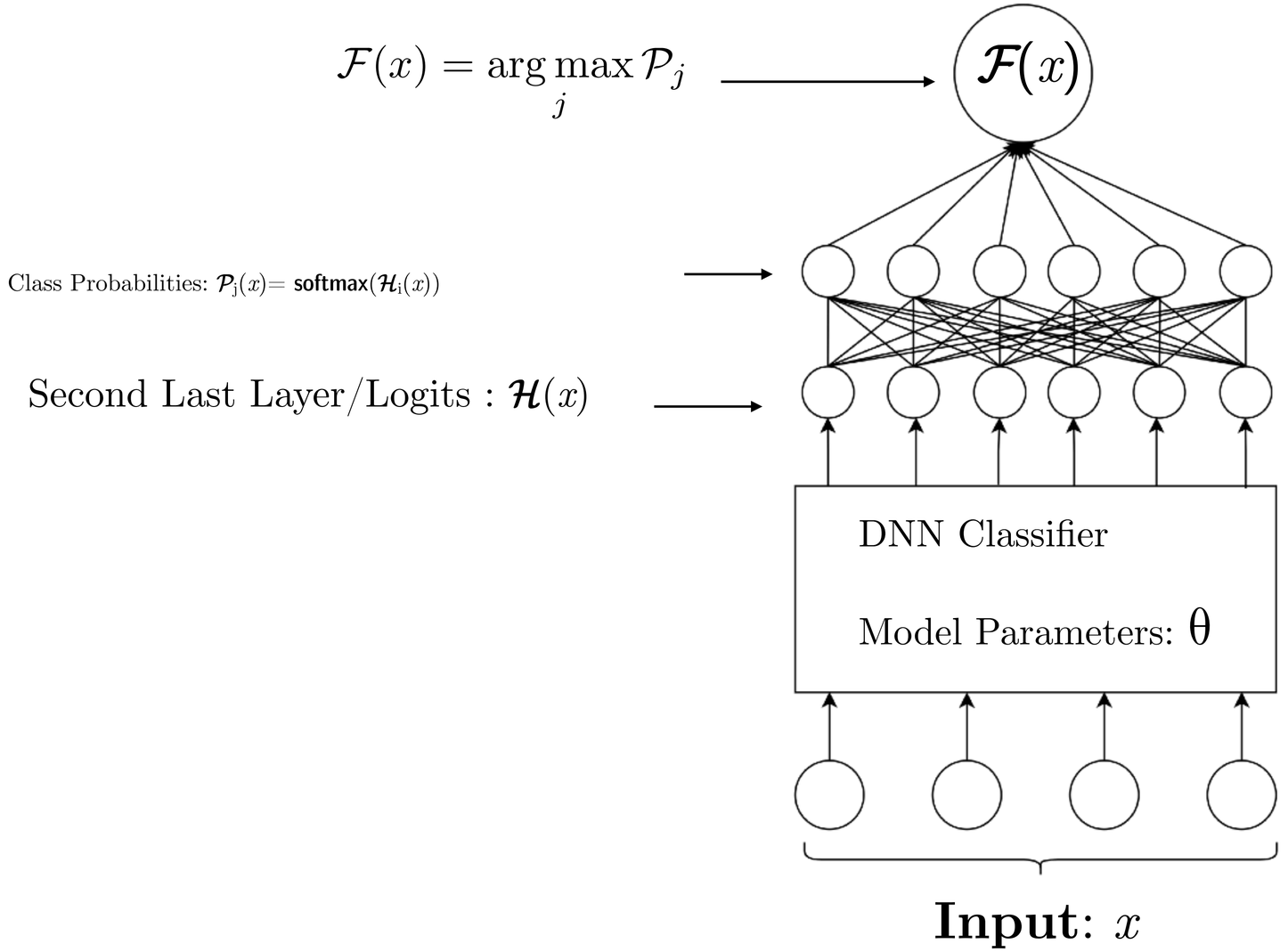
HISTORY OF ADVERSARIAL MACHINE LEARNING

12

- Adversarial Classification by Dalvi et al. 2004 (Fooled Spam Filter)
- Evasion Attacks Against Machine Learning at Test Time by B. Biggio et al. (Fooled Neural Net)
- Intriguing properties of neural networks by C. Szegedy et al. 2014 (Imperceptible Adversarial Examples)
- Explaining and Harnessing Adversarial Examples by Goodfellow et al. 2015 (Fast Attack Model)
- Adversarial Examples in Physical World by Kurakin et al. 2016 (Demonstrates Adversary in Real World)
- Adversarial Machine Learning at Scale by Kurakin et al. 2017 (Adversarial Training)
- Practical Black Box Attacks by Papernot et al. 2016 (Practical Black Box Attacks)
- Evaluating the robustness of neural networks by Wagner et. al. 2017 (Strong L2 Attack)
- Ensemble approach for adversarial defence by Papernot et al. 2017 (Ensemble Defence)

ATTACK MODELS: TERMINOLOGY

- $\mathcal{F}$(x) : Predicted Class of x
- y: true class
- θ: Model Parameters
- $\mathcal{H}(x,\theta)$: Output of Logits (Before final softmax *l*ayer)
- $\mathcal{L}(x,y,\theta)$: Loss Function
- l: Class of interest for attacked (For Targeted Attacks)



HISTORY OF ADVERSARIAL MACHINE LEARNING

- Adversarial Classification by Dalvi et al. 2004 (Fooled Spam Filter)
- Evasion Attacks Against Machine Learning at Test Time by B. Biggio et al. (Fooled Neural Net)
- Intriguing properties of neural networks by C. Szegedy et al. 2014 (Imperceptible Adversarial Examples)
- Explaining and Harnessing Adversarial Examples by Goodfellow et al. 2015 (Fast Attack Model)
- Adversarial Examples in Physical World by Kurakin et al. 2016 (Demonstrates Adversary in Real World)
- Adversarial Machine Learning at Scale by Kurakin et al. 2017 (Adversarial Training)
- Practical Black Box Attacks by Papernot et al. 2016 (Practical Black Box Attacks)
- Evaluating the robustness of neural networks by Wagner et. al. 2017 (Strong L2 Attack)
- Ensemble approach for adversarial defence by Papernot et al. 2017 (Ensemble Defence)