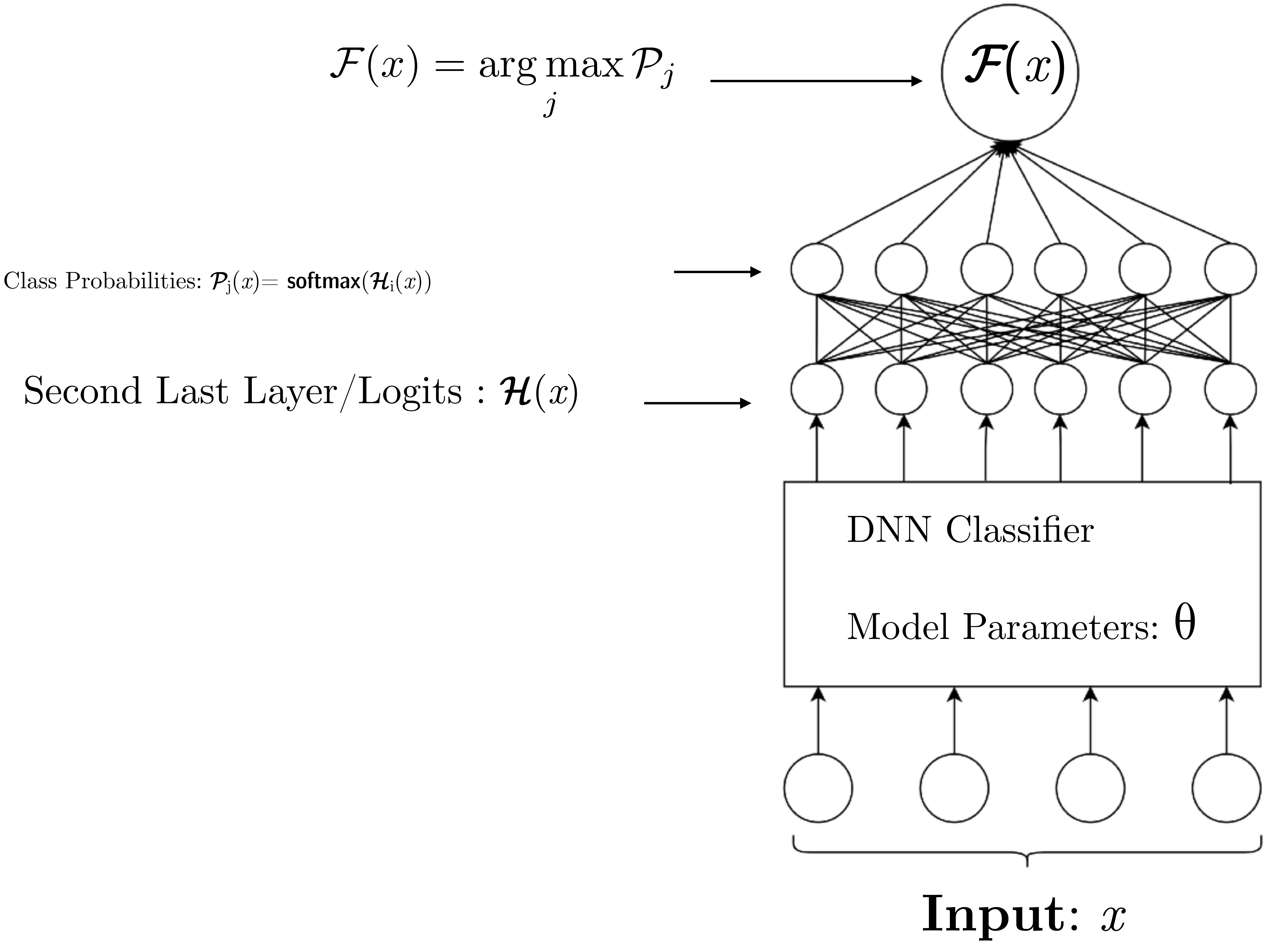
ADVERSARIAL MACHINE LEARNING

ATTACK MODELS: TERMINOLOGY

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

13



$$\mathcal{F}(x) = \arg\max_{j} \mathcal{P}_j$$

$\boldsymbol{\mathcal{F}}(x)$

Class Probabilities: $\boldsymbol{\mathcal{P}}_{\mathrm{j}}(x)=$ $\textsf{\textbf{softmax}}(\boldsymbol{\mathcal{H}}_{\mathrm{i}}(x))$

Second Last Layer/Logits : $\boldsymbol{\mathcal{H}}(x)$

DNN Classifier

Model Parameters: θ

Input: x

ATTACK MODEL PROBLEM FORMULATION:

- Non-Targeted Attack:

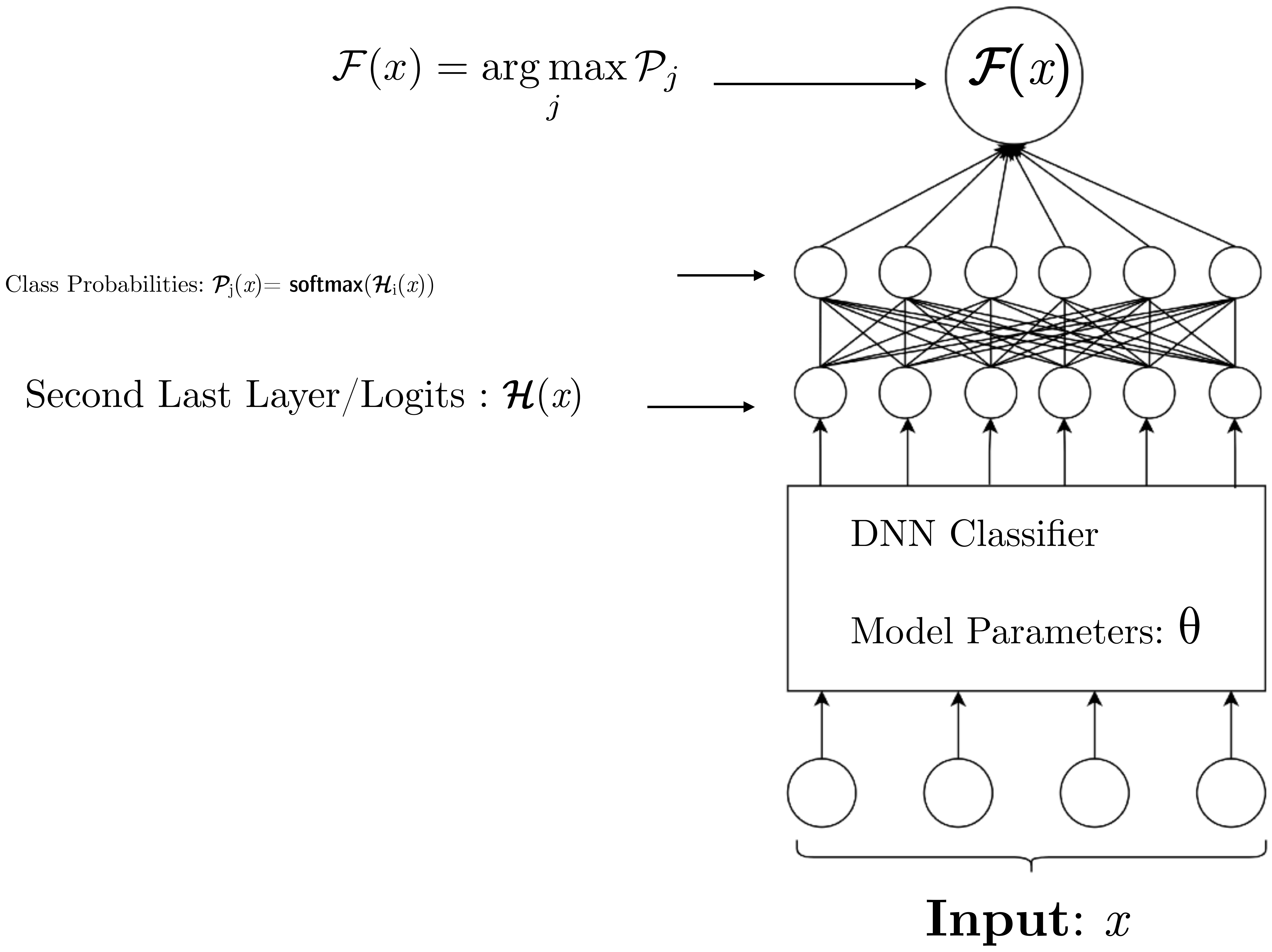
$$\arg\min_{||\delta||_2} \;\; \text{s.t.} \;\; \mathcal{F}(x+\delta) \neq \mathcal{F}(x)$$

- Targeted Attack

$$\arg\min_{||\delta||_2} \;\; \text{s.t.} \;\; \mathcal{F}(x+\delta) = \ell \;\; \text{Target} \;\; \text{Class}$$

ATTACK MODELS: TERMINOLOGY

- $\mathcal{F}(\mathrm{x})$: 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)



$$\mathcal{F}(x) = \arg\max_{j} \mathcal{P}_j$$

Class Probabilities: $\mathcal{P}_j(x)=$ **softmax**$(\mathcal{H}_i(x))$

Second Last Layer/Logits : $\mathcal{H}(x)$