Self-Explaining Neural Networks (SENN)

AM + Jaakkola, *NeurIPS’18*

$$f(x) = \theta^{\top}\mathbf{x} = \sum_{i=1}^{n} \theta_i x_i + \theta_0$$

$$f(x) = \theta(\mathbf{x})^\top \mathbf{x}$$

$$f(x) = \theta(\mathbf{x})^\top h(\mathbf{x})$$

$$f(x) = g(\theta(\mathbf{x})_1, \ldots, \theta(\mathbf{x})_k)$$

Coefficients are **input-dependent** - need to regularize!

Beyond raw inputs - explain in terms of **concepts**

General aggregation

Why is it "interpretable"?

a. Inputs are **grounded**
b. Parameters are meaningful (+/- **contribution**)
c. $\sum$ does not conflate **feature-wise interpretation**

From **Interpretable** to **Complex:**

Self-Explaining
Neural Networks (SENN)

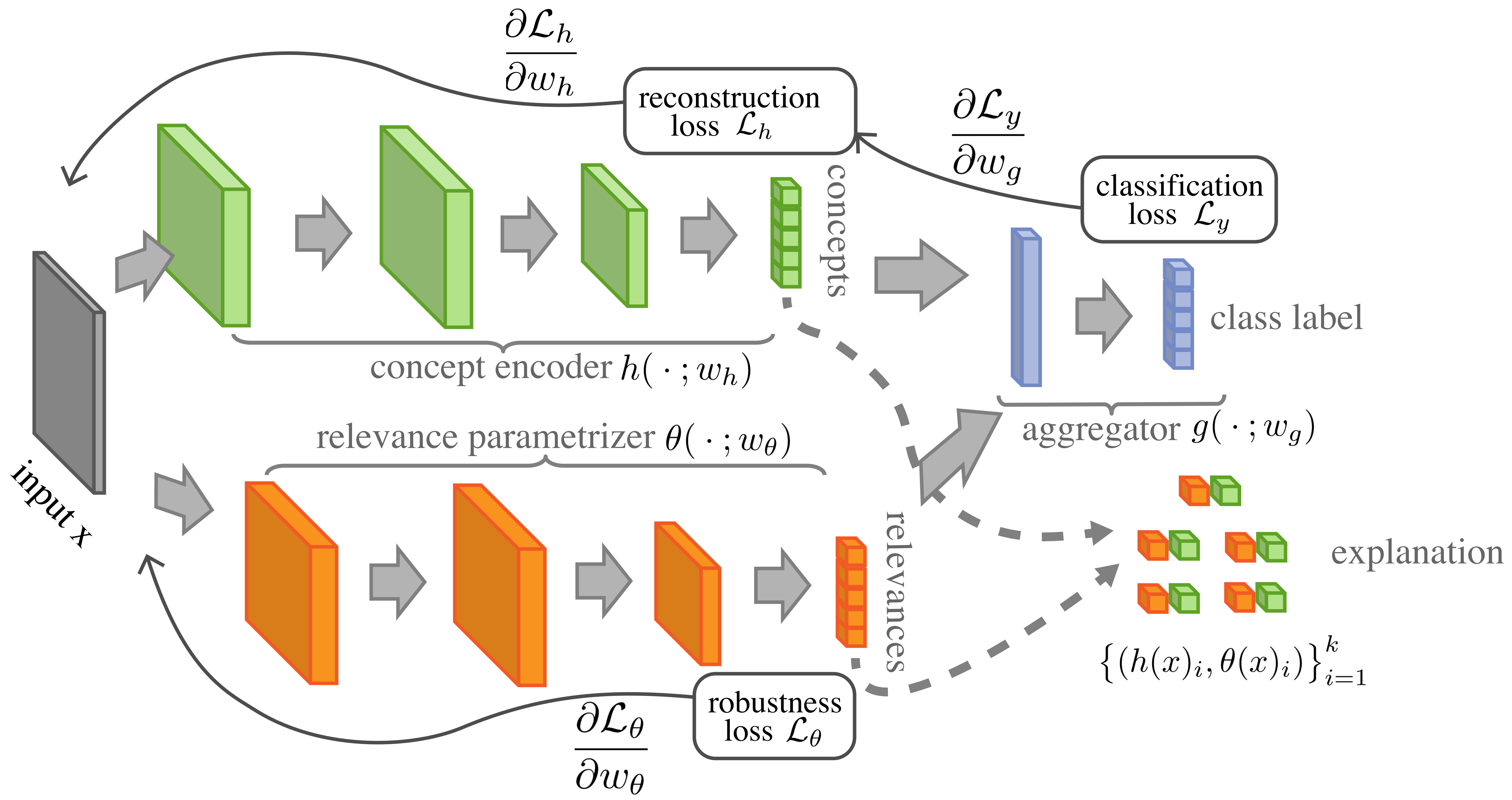
AM + Jaakkola, *NeurIPS'18*

From **Interpretable** to **Complex**:

$f(x) = \theta^\top \mathbf{x} = \sum_{i=1}^{n} \theta_i x_i + \theta_0$	Why is it "interpretable"? a. Inputs are **grounded** b. Parameters are meaningful (+/- **contribution**) c. $\sum$does not conflate **feature-wise interpretation**
$f(x) = \theta(\mathbf{x})^\top \mathbf{x}$	Coefficients are **input-dependent** - need to regularize!
$f(x) = \theta(\mathbf{x})^\top h(\mathbf{x})$	Beyond raw inputs - explain in terms of **concepts**
$f(x) = g(\theta(\mathbf{x})_1, \ldots, \theta(\mathbf{x})_k)$	General aggregation

Self-Explaining Neural Networks (SENN)

AM + Jaakkola, *NeurIPS'18*


$\frac{\partial \mathcal{L}_h}{\partial w_h}$
reconstruction loss $\mathcal{L}_h$
$\frac{\partial \mathcal{L}_y}{\partial w_g}$
classification loss $\mathcal{L}_y$
concepts
class label
concept encoder $h(\cdot\,;w_h)$
aggregator $g(\cdot\,;w_g)$
relevance parametrizer $\theta(\cdot\,;w_\theta)$
input x
explanation
relevances
$\{(h(x)_i, \theta(x)_i)\}_{i=1}^k$
robustness loss $\mathcal{L}_\theta$
$\frac{\partial \mathcal{L}_\theta}{\partial w_\theta}$