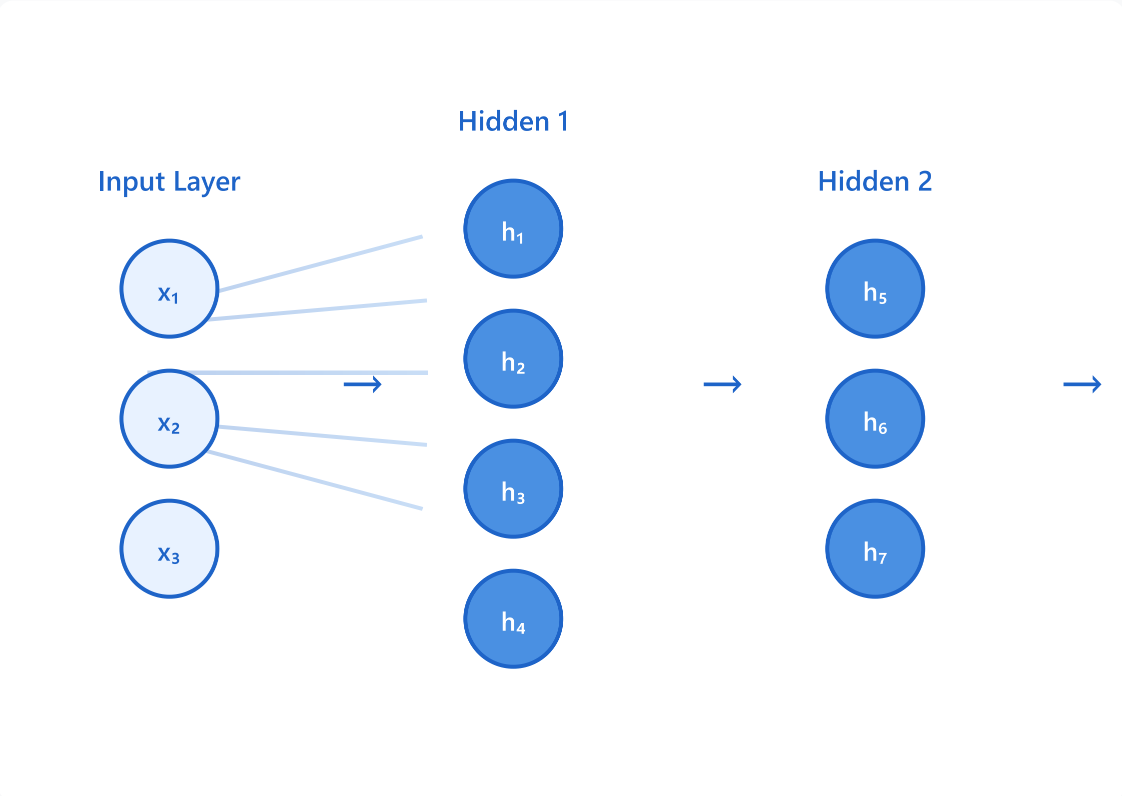
DeepSHAP: DeepLIFT + SHAP for Neural Networks

Backpropagation of reference activation differences

Neural Network Computation Flow



DeepSHAP Computation Steps

→ Compute reference activations (average)

→ Forward pass: input → output

→ Backward pass: propagate differences

Key Features

Combines DeepLIFT with Shapley value sampling

- Layer-wise decomposition
- Handles nonlinear activations
- Reference-based differences

Activations

Supports various activation functions

- ReLU
- Sigmoid
- Tanh

Reference Value

Typically uses average of training data as baseline

Implementation

→ Assign contributions to each input

```python
import shap

# Create explainer
explainer = shap.DeepExplainer(
    model,
    X_train[:100]
)

# Compute SHAP values
shap_values = explainer.shap_values(X_test)
```

DeepSHAP Computation Principle

📐 Core Formulas

$$\varphi_i = \Delta output \times (\Delta x_i / \Sigma\Delta x_j)$$

φ_i: Feature i의 SHAP value
Δoutput: Output difference from reference
Δx_i: Input i difference from reference
$\Sigma\Delta x_j$: Sum of all input differences

Layer-wise propagation:

$$C_{ij} = m_{ij} \times \Delta x_j \Delta h_i$$

C_{ij}: Contribution from neuron j to i
m_{ij}: Multiplier (weight effect)
Δx_j: Input activation difference
Δh_i: Hidden activation difference

🔢 Numerical Example

Step 1: Set Reference

Reference (baseline): $x_0 = [0, 0, 0]$
Input (actual): $x = [1, 2, 1]$
Differences: $\Delta x = [1, 2, 1]$

Step 2: Forward Pass

Hidden layer: $h = ReLU(W \cdot x + b)$
Reference: $h_0 = ReLU(W \cdot x_0 + b) = [0, 0]$
Actual: $h = [2, 3]$
Differences: $\Delta h = [2, 3]$

Step 3: Output Difference

Reference output: $y_0 = 0.3$
Actual output: $y = 0.8$
Output difference: $\Delta y = 0.5$

Step 4: Backward Propagation

Calculate each input's contribution via backpropagation:

$\varphi_1 = 0.5 \times (1/4) = 0.125$

$\varphi_2 = 0.5 \times (2/4) = 0.250$

$\varphi_3 = 0.5 \times (1/4) = 0.125$

✅ **Final Result Interpretation**

- Feature x_2 has the largest impact ($\varphi_2 = 0.250$)
- Sum of all SHAP values = 0.5 = Δy ✓
- Features x_1, x_3 have equal contributions (0.125)