A VISUAL GUIDE TO BACKPROPAGATION

Matrix multiplication gradients

A tiny input change follows several paths. Add its effects on the loss.

$$H = X @ W$$

A scalar loss follows:
$L = f(H)$

$$G = \frac{\partial L}{\partial H}$$ incoming gradient

GRADIENT FOR THE LEFT INPUT

$$\frac{\partial L}{\partial X} = G @ W^T$$

$(N \times K) @ (K \times D) = (N \times D)$

GRADIENT FOR THE RIGHT INPUT

$$\frac{\partial L}{\partial W} = X^T @ G$$

$(D \times N) @ (N \times K) = (D \times K)$

Remember: replace the input with G; transpose the other input; keep its original side.

Shapes: X is N × D, W is D × K, and H and G are N × K.

01 Follow one entry to the loss

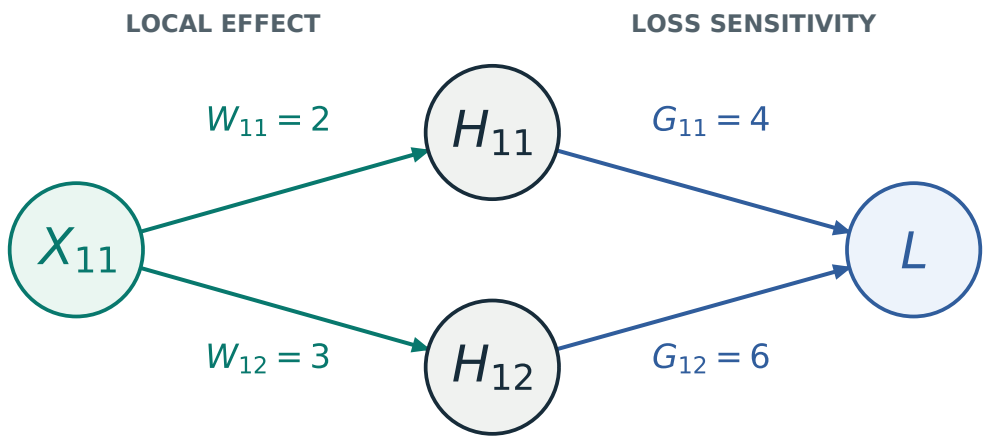
Use these values and incoming gradients at the point where we differentiate:

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad W = \begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix} \qquad G = \begin{bmatrix} 4 & 6 \\ 1 & 2 \end{bmatrix}$$

Arrow labels are derivatives: how much a small change is multiplied.

Change one input entry

One input affects outputs across its row.

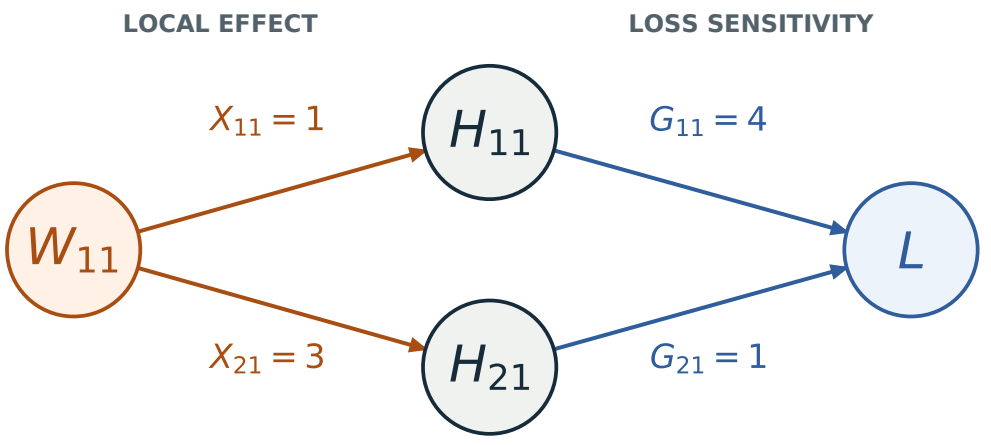


$$\frac{\partial L}{\partial X_{11}} = 2 \cdot 4 + 3 \cdot 6 = 26$$

$$\Delta X_{11} = \varepsilon \quad \Rightarrow \quad \Delta L \approx 26\varepsilon$$

Change one weight entry

One weight is reused across input rows.



$$\frac{\partial L}{\partial W_{11}} = 1 \cdot 4 + 3 \cdot 1 = 7$$

$$\Delta W_{11} = \varepsilon \quad \Rightarrow \quad \Delta L \approx 7\varepsilon$$

02 The proof is the scalar chain rule

$$H_{ij} = \sum_r X_{ir} W_{rj} \qquad G_{ij} = \frac{\partial L}{\partial H_{ij}}$$

Hold W fixed; vary X

$$\frac{\partial H_{ij}}{\partial X_{ir}} = W_{rj}$$

Only outputs in row i depend on this entry.

$$\frac{\partial L}{\partial X_{ir}} = \sum_{j=1}^{K} G_{ij} W_{rj} = (GW^T)_{ir}$$

Hold X fixed; vary W

$$\frac{\partial H_{ij}}{\partial W_{rj}} = X_{ir}$$

Only outputs in column j depend on this entry.

$$\frac{\partial L}{\partial W_{rj}} = \sum_{i=1}^{N} X_{ir} G_{ij} = (X^T G)_{rj}$$

The transpose lines up the matching terms so @ performs exactly these sums.

03 What if you literally mean ∂H/∂X and ∂H/∂W?

H is a matrix, so these are Jacobians: one derivative per output/input pair.

$$\frac{\partial H_{ij}}{\partial X_{ab}} = \delta_{ia} W_{bj}$$

$$\Delta H = (\Delta X) W$$ with W fixed

$$\frac{\partial H_{ij}}{\partial W_{ab}} = X_{ia} \delta_{jb}$$

$$\Delta H = X(\Delta W)$$ with X fixed

δ equals 1 when its two indices match, and 0 otherwise. Backprop combines these derivatives with G.

Assumptions: real 2D matrices; X and W are separate inputs; L depends on them through H. @ = matrix multiplication.