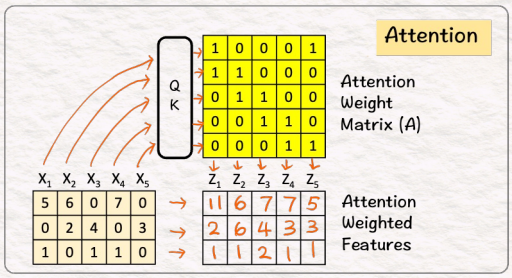
# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

1 1 1 1 1

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

[ReLU]

$f$

1 1 1 1 1

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

$x'$

**Weighted L1 Sparsity Loss Gradients**

$f\|w\| \rightarrow 0$

$$.* \quad = \frac{\partial L}{\partial f}$$

-1 if $f > 0$; else 0 $\quad \|W^D\|_2$

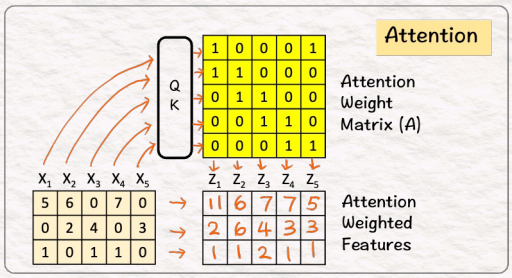
$$.* \ 2 = \frac{\partial L}{\partial x'}$$

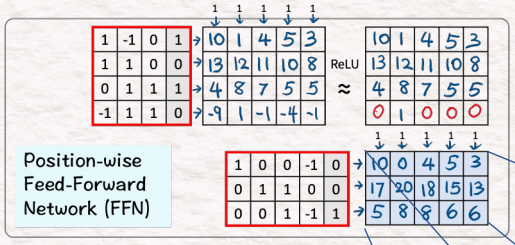
$x - x'$

**L2 Reconstruction Loss Gradients**

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

[ReLU]

$f$

| | | | | |
|---|---|---|---|---|
| -12 | 5 | -1 | -5 | -5 |
| 1 | 1 | -1 | -2 | -4 |
| 1 | -12 | -6 | -2 | -2 |
| -3 | -6 | -6 | -4 | -4 |
| 11 | -12 | -1 | 3 | 1 |
| -3 | 3 | 1 | -4 | -6 |

1 1 1 1 1

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

$x'$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

-1 if $f > 0$; else 0 $\quad \|W^D\|_2$

$$.* = \frac{\partial L}{\partial f}$$

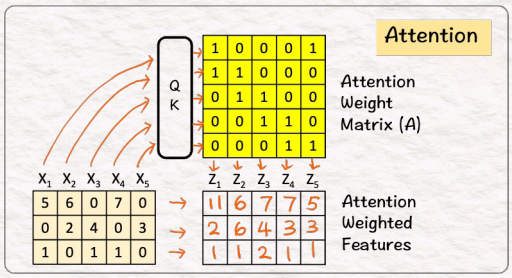
$x - x'$

$$.* 2 = \frac{\partial L}{\partial x'}$$

## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

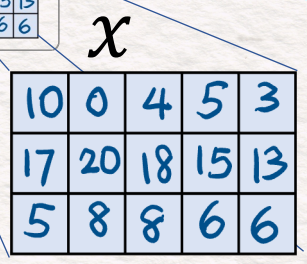
# Sparse Auto Encoder (SAE)

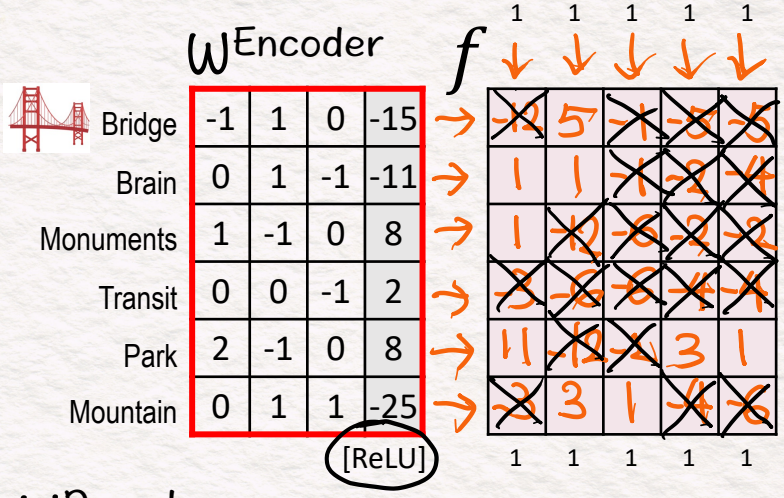


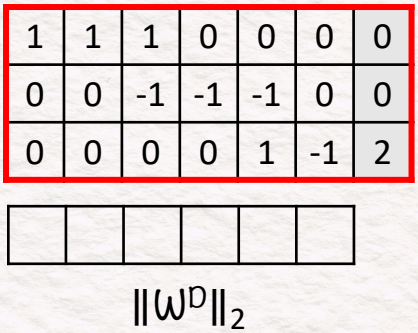

$x$ model activation

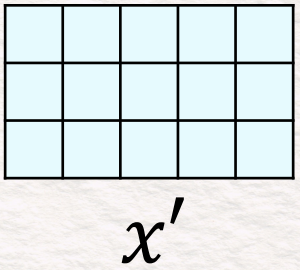
$f$ interpretable features

$x'$ reconstructed activation



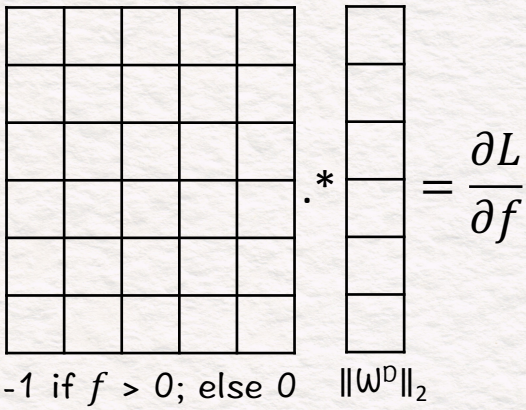



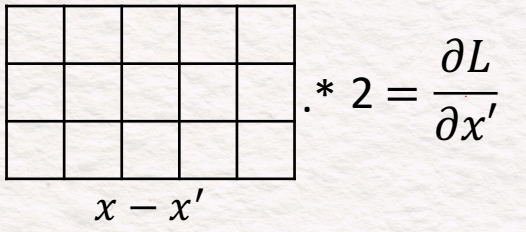




**Weighted L1 Sparsity Loss Gradients**

$f\|w\| \rightarrow 0$




$$.* = \frac{\partial L}{\partial f}$$

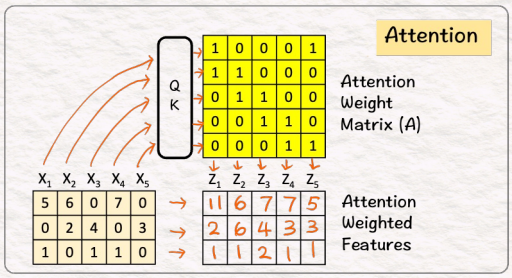



$$.* 2 = \frac{\partial L}{\partial x'}$$

**L2 Reconstruction Loss Gradients**

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

[ReLU]

$f$

| | | | | |
|---|---|---|---|---|
| 1 | 5 | 1 | 0 | 0 |
| 1 | 1 | 1 | 2 | 4 |
| 1 | 12 | 6 | 2 | 2 |
| 3 | 6 | 6 | 4 | 4 |
| 11 | 12 | 4 | 3 | 1 |
| 3 | 3 | 1 | 4 | 6 |

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$x'$

| | | | | |
|---|---|---|---|---|
| 3 | 7 | 0 | 0 | 0 |
| -12 | 0 | 0 | -3 | -1 |
| 13 | -1 | 1 | 5 | 3 |

$\|W^D\|_2$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

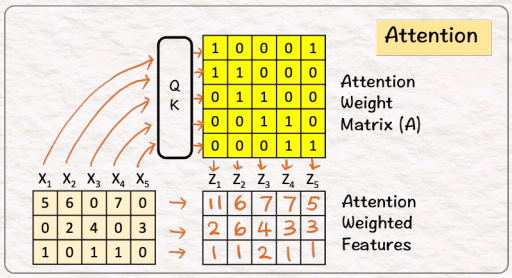
$$\text{-1 if } f > 0\text{; else } 0 \;\; .* \;\; \|W^D\|_2 = \frac{\partial L}{\partial f}$$

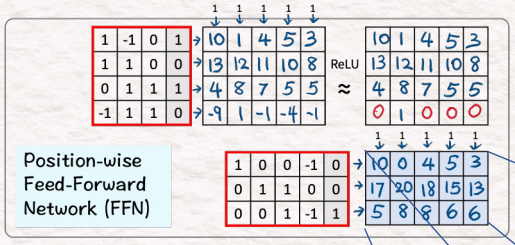
## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

$$x - x' \;\; .* \;\; 2 = \frac{\partial L}{\partial x'}$$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

[ReLU]

$f$

| | | | | |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 |

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$x'$

| | | | | |
|---|---|---|---|---|
| 3 | 7 | 0 | 0 | 0 |
| -12 | 0 | 0 | -3 | -1 |
| 13 | -1 | 1 | 5 | 3 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | $\sqrt{2}$ | 1 | $\sqrt{2}$ | 1 | |

$\|W^D\|_2$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

$$.* = \frac{\partial L}{\partial f}$$

-1 if $f > 0$; else 0 $\quad \|W^D\|_2$

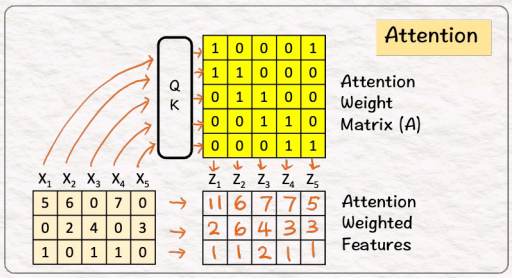
$$.* 2 = \frac{\partial L}{\partial x'}$$

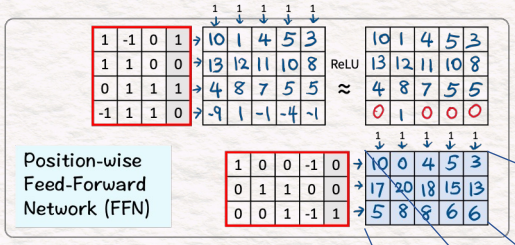
$x - x'$

## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$ $f$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

[ReLU]

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | $\sqrt{2}$ | 1 | $\sqrt{2}$ | 1 | |

$x'$

| | | | | |
|---|---|---|---|---|
| 3 | 7 | 0 | 0 | 0 |
| -12 | 0 | 0 | -3 | -1 |
| 13 | -1 | 1 | 5 | 3 |

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

$$\begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix} .* \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix} = \frac{\partial L}{\partial f}$$

-1 if $f > 0$; else 0 $\quad \|W^D\|_2$

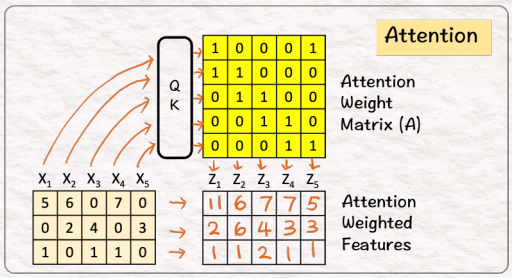
$$\begin{bmatrix} & & & & \\ & & & & \\ & & & & \end{bmatrix} .* 2 = \frac{\partial L}{\partial x'}$$

$x - x'$

## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

$f$

[ReLU]

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

$x'$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

$$\cdot * = \frac{\partial L}{\partial f}$$

-1 if $f > 0$; else 0

$\|W^D\|_2$

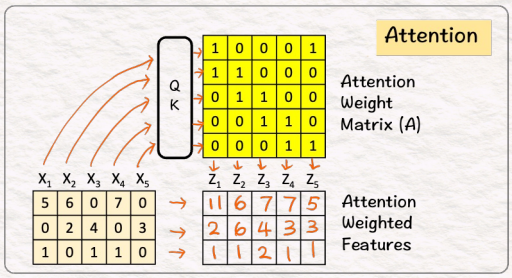
$$\cdot * 2 = \frac{\partial L}{\partial x'}$$

$x - x'$

## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

$f$

[ReLU]

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

$x'$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

-1 if $f > 0$; else 0

$\|W^D\|_2$

$$.* = \frac{\partial L}{\partial f}$$

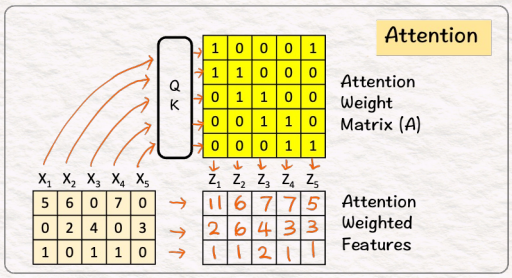
$x - x'$

$$.* 2 = \frac{\partial L}{\partial x'}$$

## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation

$x$

| | | | | |
|---|---|---|---|---|
| 10 | 0 | 4 | 5 | 3 |
| 17 | 20 | 18 | 15 | 13 |
| 5 | 8 | 8 | 6 | 6 |

$W^{Encoder}$

| | | | | |
|---|---|---|---|---|
| Bridge | -1 | 1 | 0 | -15 |
| Brain | 0 | 1 | -1 | -11 |
| Monuments | 1 | -1 | 0 | 8 |
| Transit | 0 | 0 | -1 | 2 |
| Park | 2 | -1 | 0 | 8 |
| Mountain | 0 | 1 | 1 | -25 |

$f$

[ReLU]

$W^{Decoder}$

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | -1 | -1 | -1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | -1 | 2 |

$\|W^D\|_2$

$x'$

## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & 0 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ 1 \end{bmatrix} = \frac{\partial L}{\partial f}$$

-1 if $f > 0$; else 0 $\quad \|W^D\|_2$

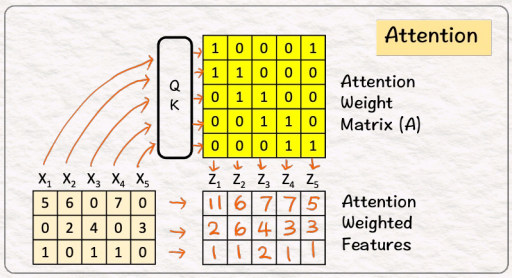
## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

$$\begin{bmatrix} & & & & \\ & & & & \\ & & & & \end{bmatrix} .* 2 = \frac{\partial L}{\partial x'}$$

$x - x'$

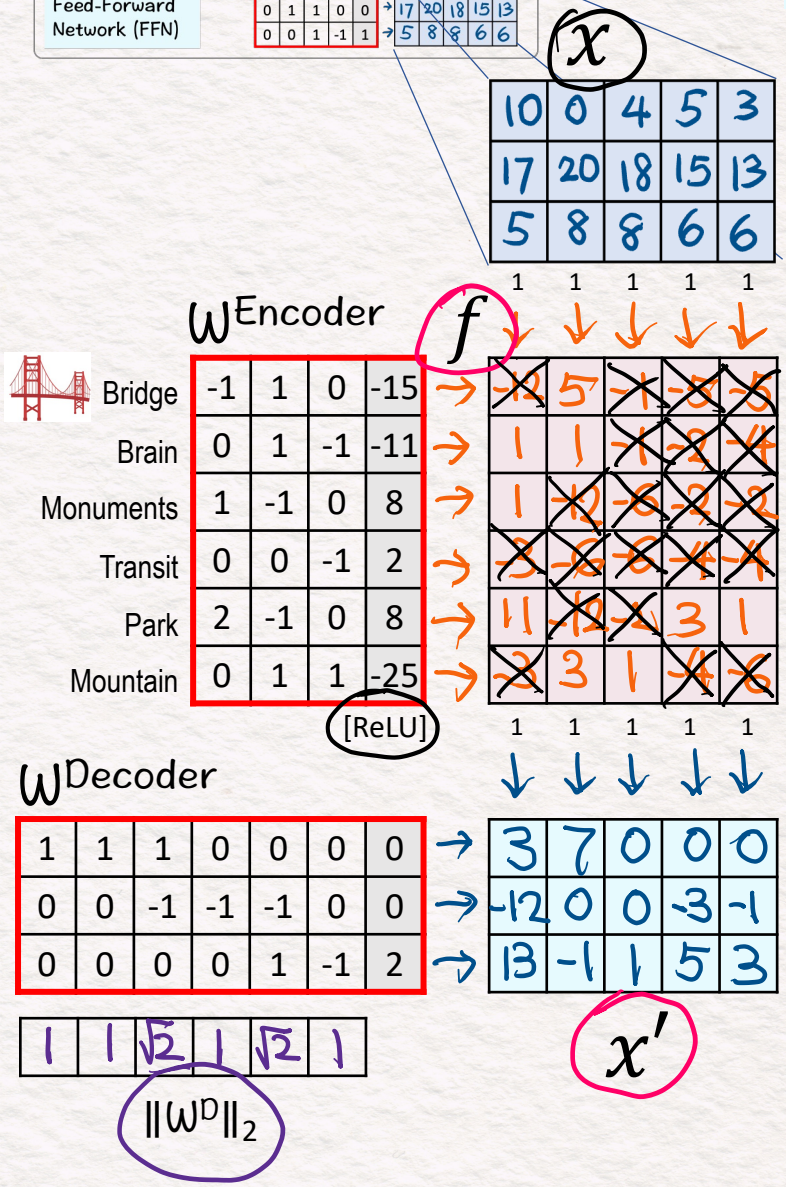
# Sparse Auto Encoder (SAE)




$x$ model activation

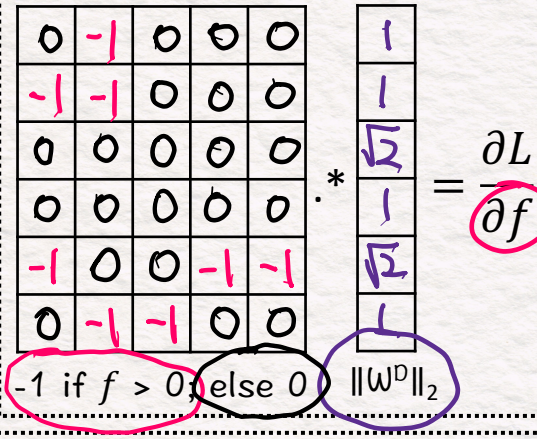
$f$ interpretable features

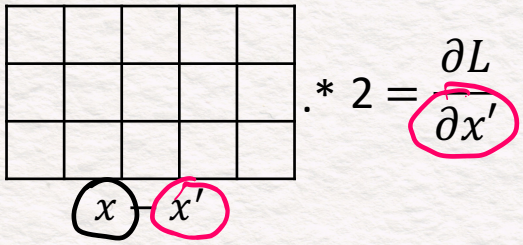
$x'$ reconstructed activation




## Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

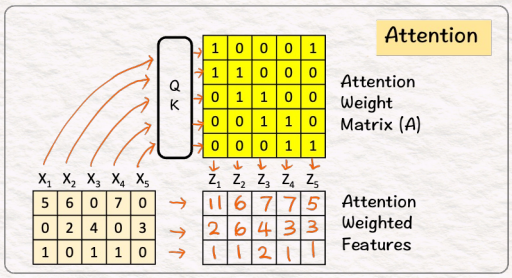



## L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$

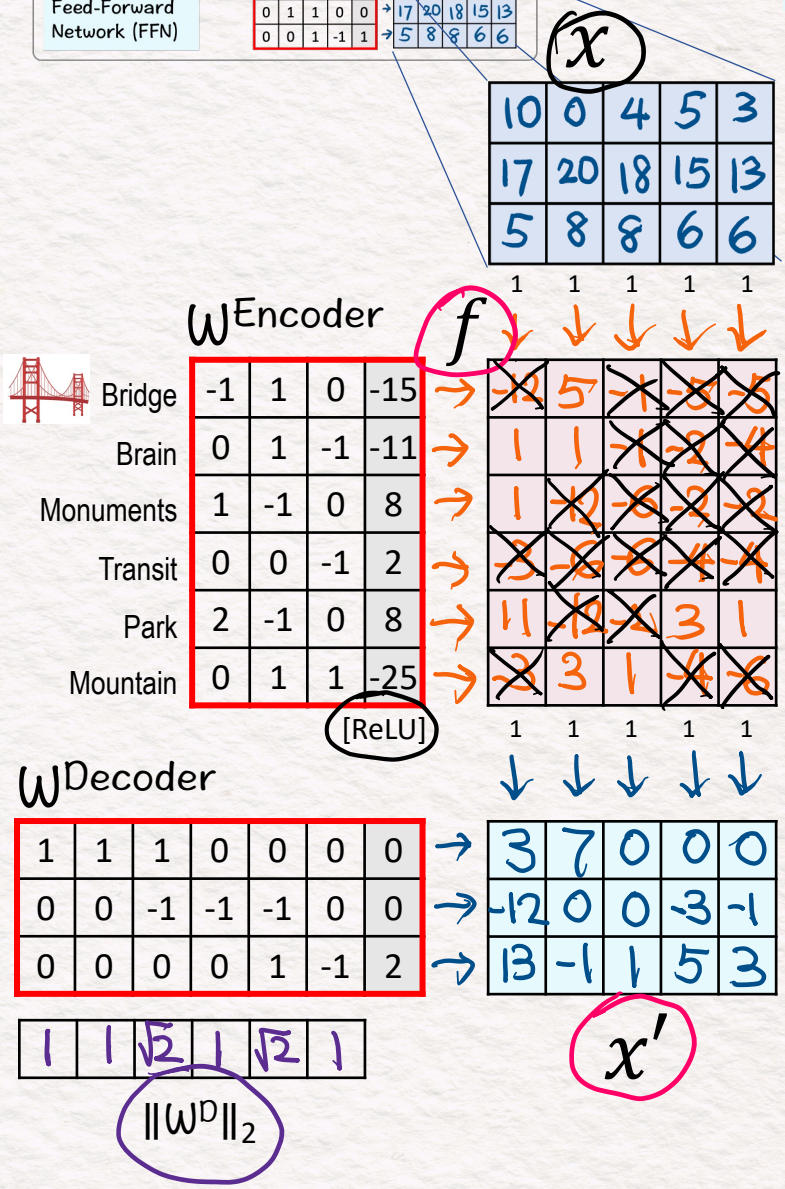
# Sparse Auto Encoder (SAE)




$x$ model activation

$f$ interpretable features

$x'$ reconstructed activation




Weighted L1 Sparsity Loss Gradients

$f\|w\| \rightarrow 0$

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & 0 \end{bmatrix} .* \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ 1 \end{bmatrix} = \frac{\partial L}{\partial f}$$

-1 if $f > 0$ else 0 $\quad \|W^D\|_2$

$$\begin{bmatrix} 7 & -7 & 4 & 5 & 3 \\ 29 & 20 & 18 & 18 & 14 \\ -8 & 9 & 7 & 1 & 3 \end{bmatrix} .* 2 = \frac{\partial L}{\partial x'}$$

$x$ $\quad x'$

L2 Reconstruction Loss Gradients

$\| x - x' \|_2 \rightarrow 0$