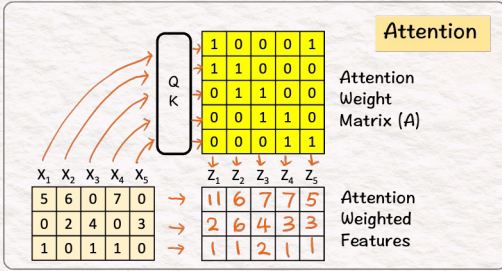


Transformer

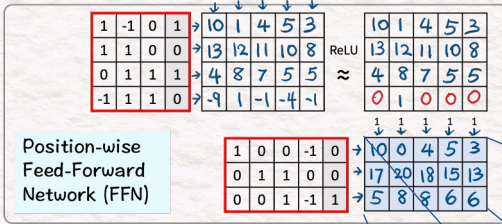
Sparse Auto Encoder (SAE)



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

Weighted L1 Sparsity Loss Gradients
 $f \|w\| \rightarrow 0$

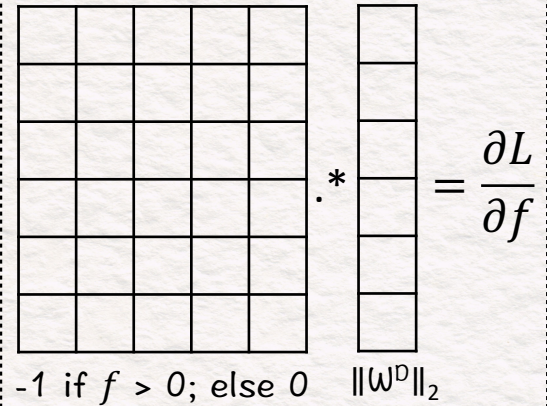
w^{Encoder}

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

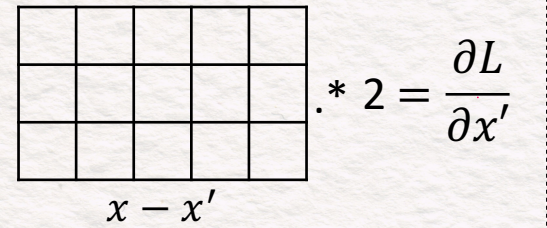


w^{Decoder}

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

$\|w^D\|_2$

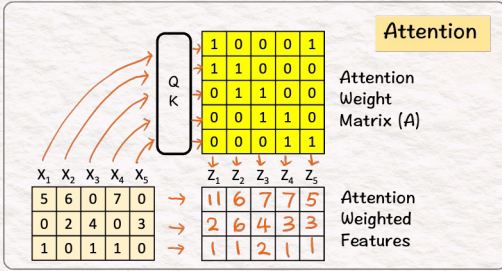
x'



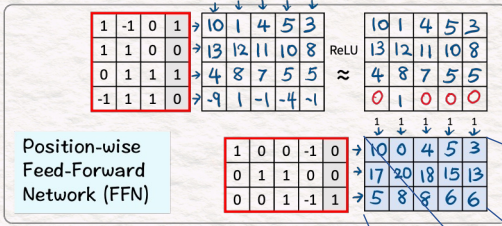
L2 Reconstruction Loss Gradients
 $\|x - x'\|_2 \rightarrow 0$

Transformer

Sparse Auto Encoder (SAE)

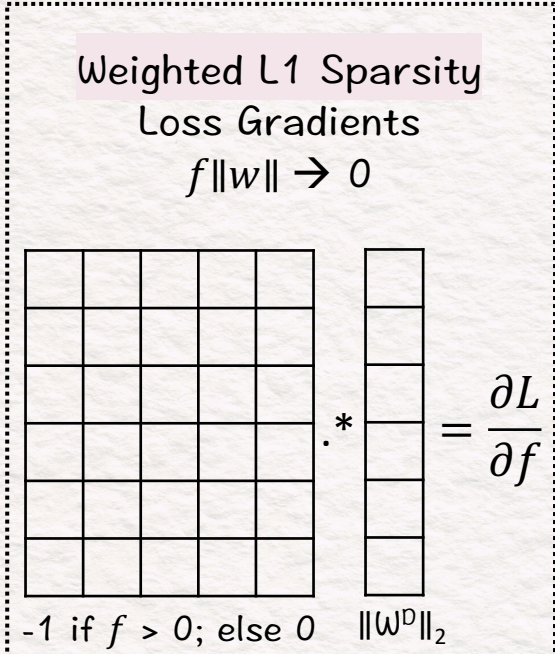


- x model activation
- f interpretable features
- x' reconstructed activation



x

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



W^{Encoder}

Bridge	-1	1	0	-15	$\rightarrow -12$	5	-1	-5	-5
Brain	0	1	-1	-11	$\rightarrow 1$	1	-1	-2	-4
Monuments	1	-1	0	8	$\rightarrow 1$	-12	-6	-2	-2
Transit	0	0	-1	2	$\rightarrow -3$	-6	-6	-4	-4
Park	2	-1	0	8	$\rightarrow 11$	-12	-2	3	1
Mountain	0	1	1	-25	$\rightarrow -3$	3	1	-4	-6

[ReLU]

f

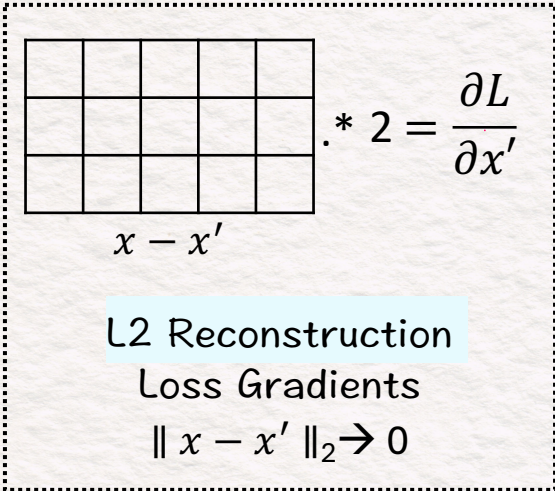
1	1	1	1	1
---	---	---	---	---

W^{Decoder}

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

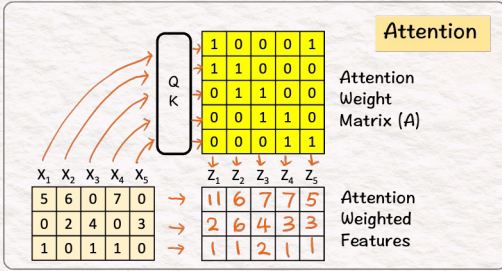
$\|W^D\|_2$

x'



Transformer

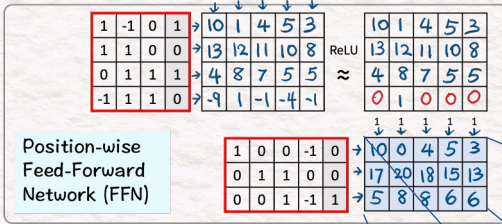
Sparse Auto Encoder (SAE)



x model activation

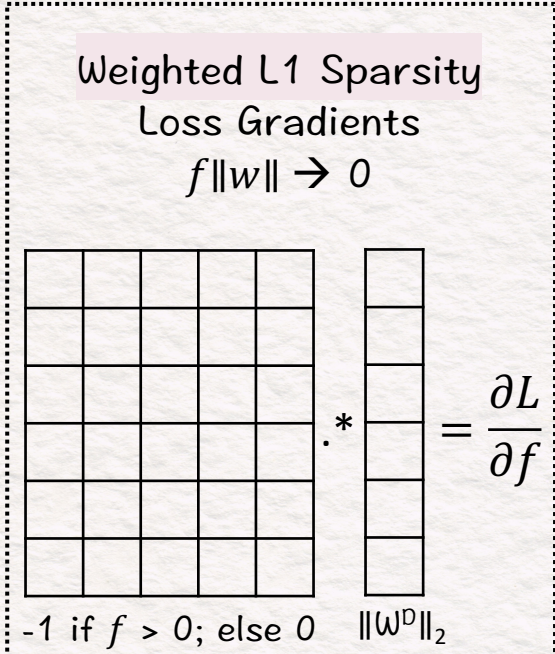
f interpretable features

x' reconstructed activation



x

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



Encoder Weights (w^E)

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]

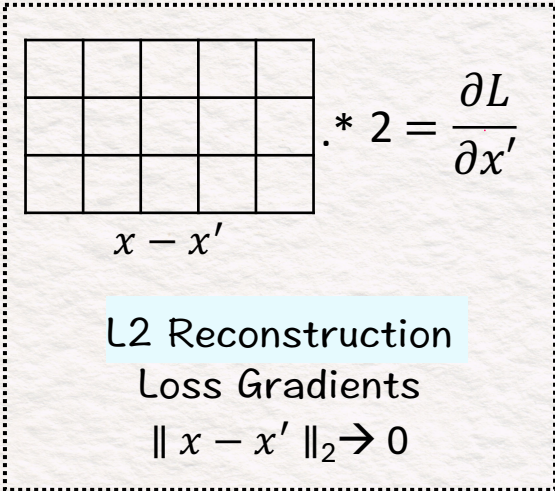
f

-2	5	-1	-5	5
1	1	-1	-2	4
1	2	-5	-2	-2
3	8	8	4	1
11	12	1	3	1
2	3	1	4	6

Decoder Weights (w^D)

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

x'

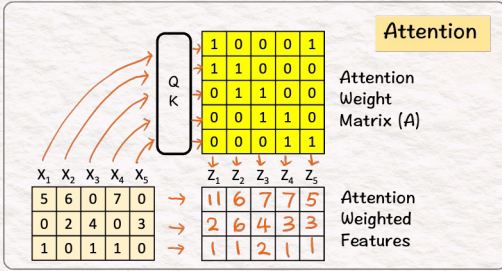


$\|w^D\|_2$

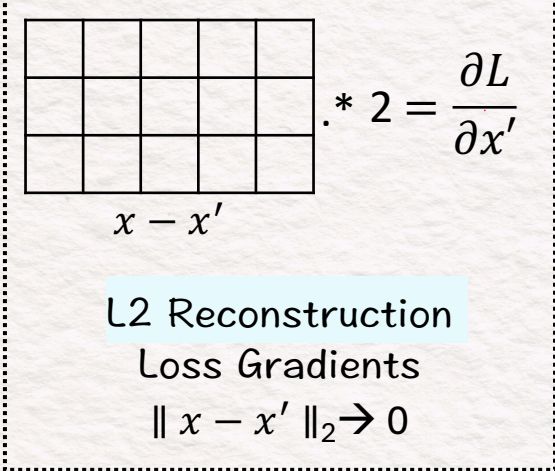
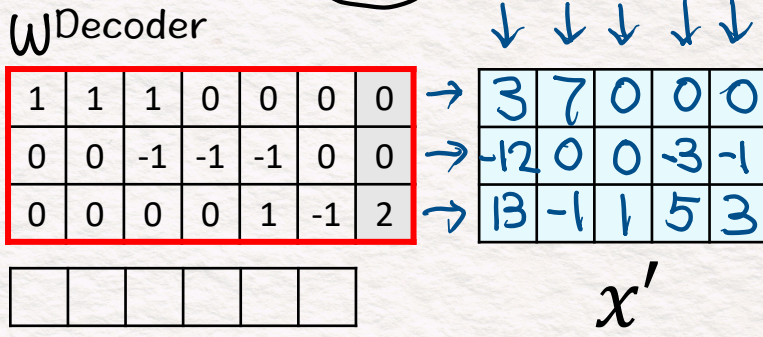
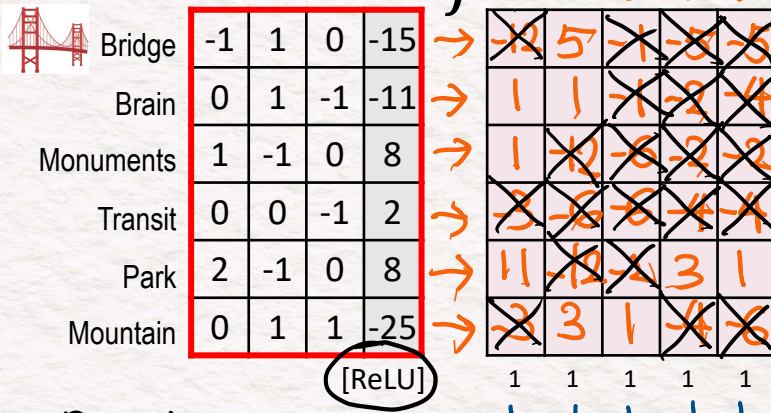
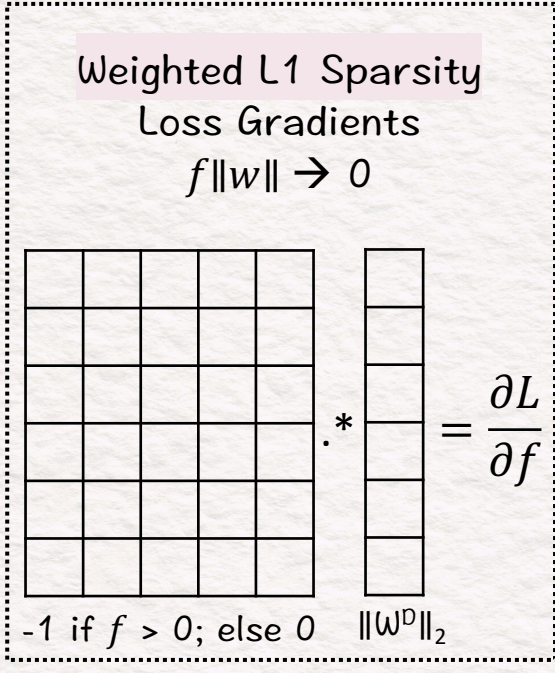
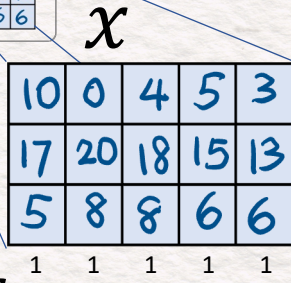
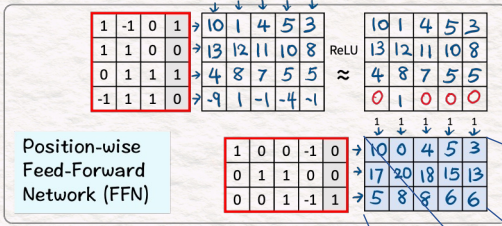
--	--	--	--	--

Transformer

Sparse Auto Encoder (SAE)

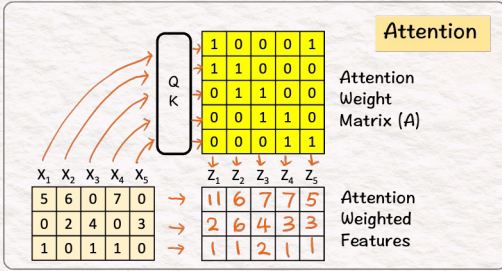


- x model activation
- f interpretable features
- x' reconstructed activation

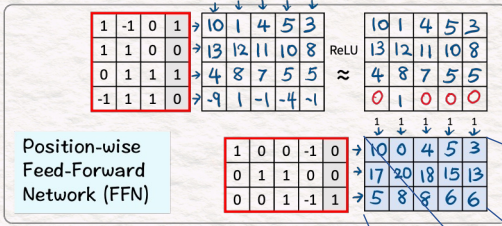


Transformer

Sparse Auto Encoder (SAE)

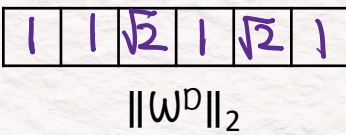
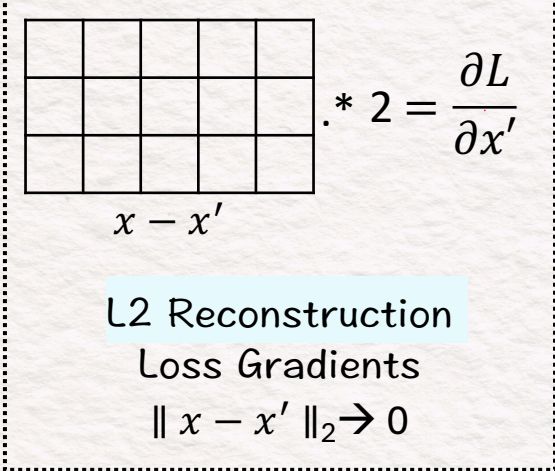
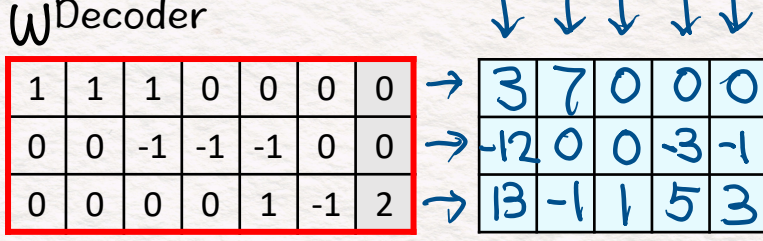
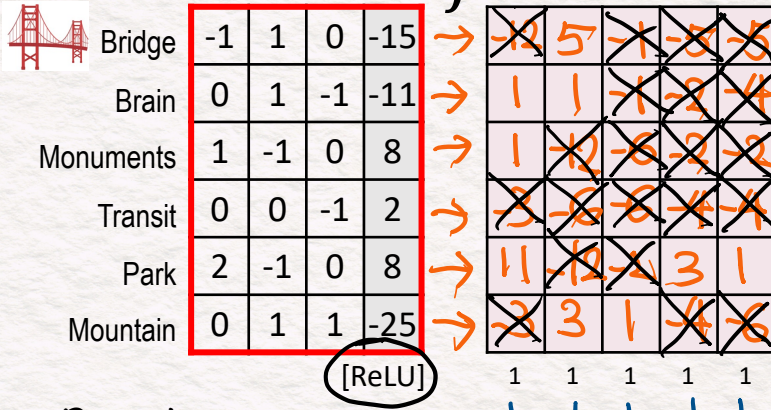
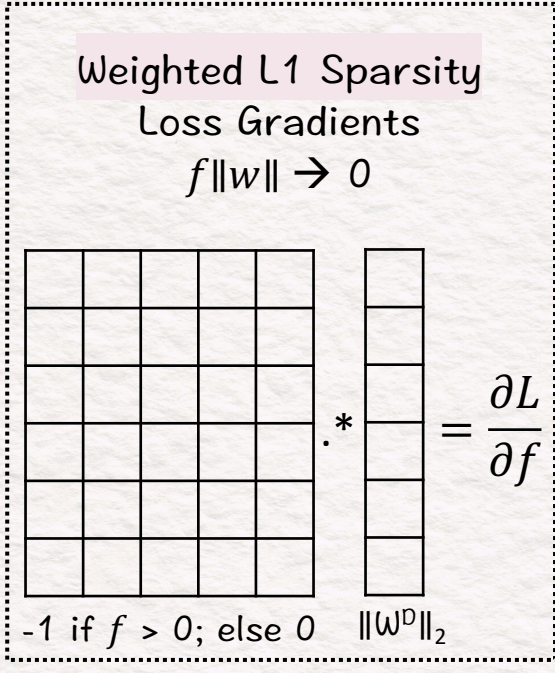


- x model activation
- f interpretable features
- x' reconstructed activation



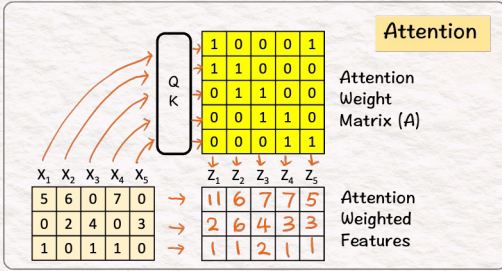
x

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

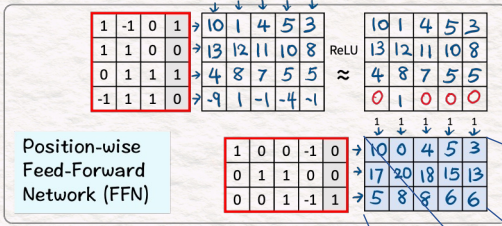


Transformer

Sparse Auto Encoder (SAE)

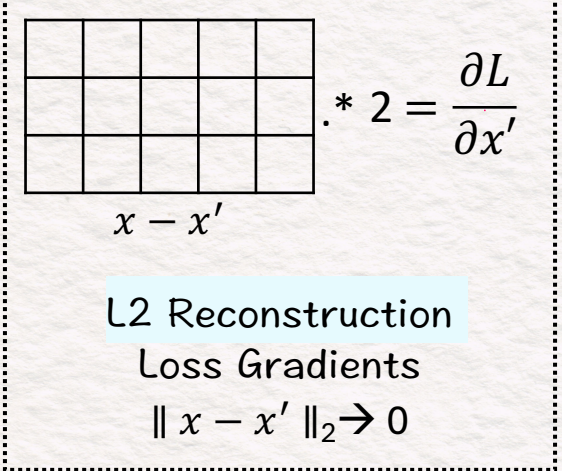
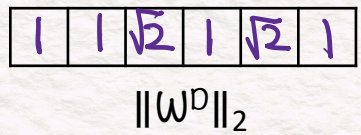
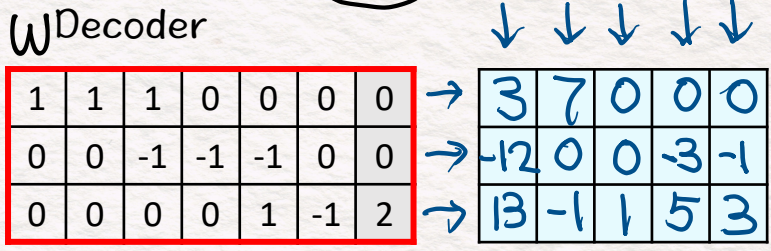
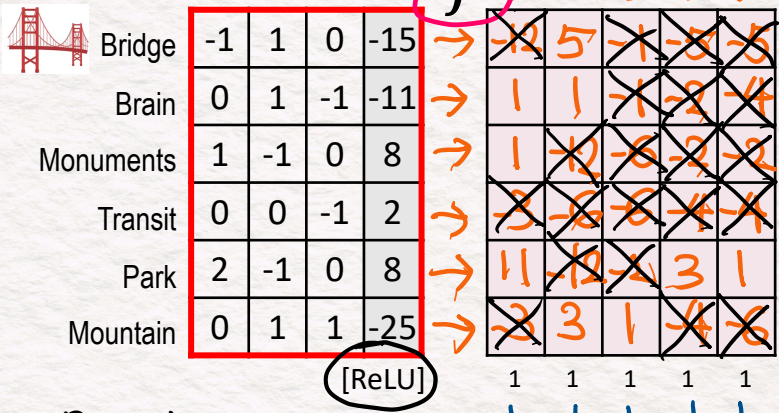
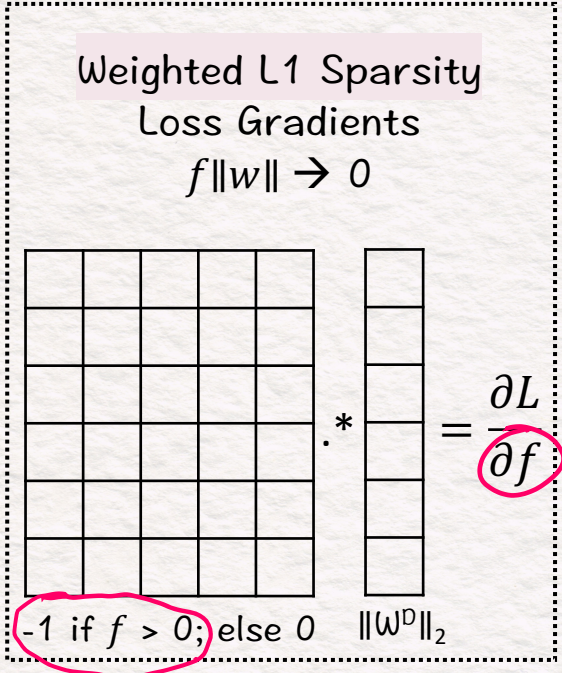


- x model activation
- f interpretable features
- x' reconstructed activation



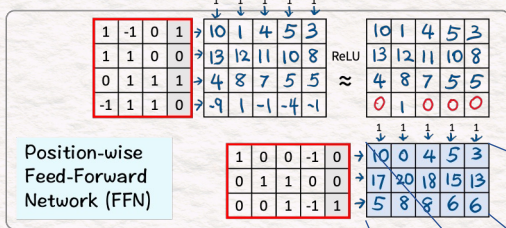
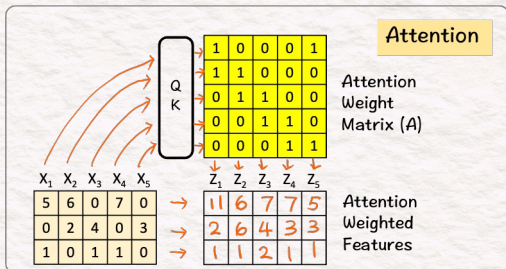
x

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



Transformer

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

Weighted L1 Sparsity Loss Gradients

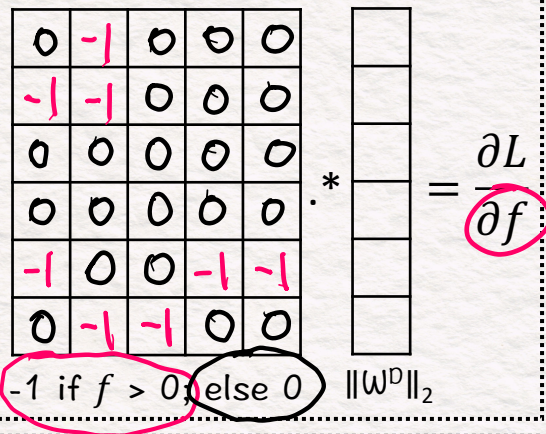
$f \|w\| \rightarrow 0$

W^{Encoder}

Bridge	-1	1	0	-15	\rightarrow	10	1	4	5	3
Brain	0	1	-1	-11	\rightarrow	1	1	1	2	4
Monuments	1	-1	0	8	\rightarrow	1	2	5	2	2
Transit	0	0	-1	2	\rightarrow	3	8	8	4	1
Park	2	-1	0	8	\rightarrow	11	12	11	3	1
Mountain	0	1	1	-25	\rightarrow	2	3	1	4	6

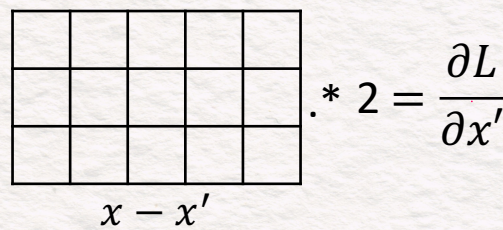
[ReLU]

1 1 1 1 1



W^{Decoder}

1	1	1	0	0	0	0	\rightarrow	3	7	0	0	0
0	0	-1	-1	-1	0	0	\rightarrow	-12	0	0	-3	-1
0	0	0	0	1	-1	2	\rightarrow	13	-1	1	5	3



$\|W^D\|_2$

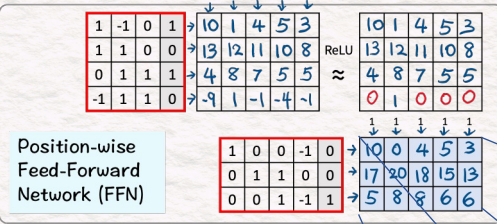
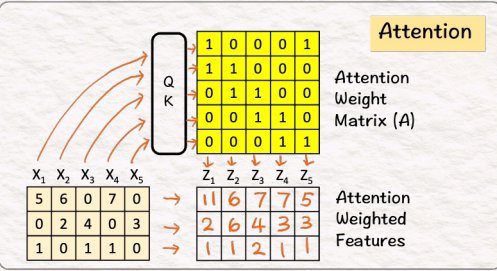
1	1	$\sqrt{2}$	1	$\sqrt{2}$	1
---	---	------------	---	------------	---

x'

L2 Reconstruction Loss Gradients

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

Transformer



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
0	1	-1	-11
1	-1	0	8
0	0	-1	2
2	-1	0	8
0	1	1	-25

f

10	0	4	5	3
1	1	-1	-11	-4
1	-1	0	8	-2
17	20	18	15	13
11	-12	-3	3	1
5	8	8	6	6

Weighted L1 Sparsity Loss Gradients
 $f \|w\| \rightarrow 0$

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

$\cdot \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ 1 \end{bmatrix} = \frac{\partial L}{\partial f}$

$-1 \text{ if } f > 0, \text{ else } 0$

$\|W^D\|_2$

$W^{Decoder}$

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

x'

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

$\|W^D\|_2$

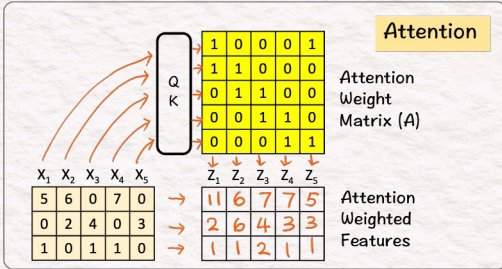
$\cdot 2 = \frac{\partial L}{\partial x'}$

$x - x'$

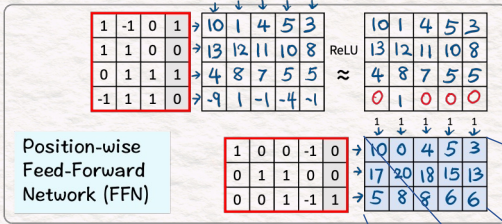
L2 Reconstruction Loss Gradients
 $\|x - x'\|_2 \rightarrow 0$

Transformer

Sparse Auto Encoder (SAE)

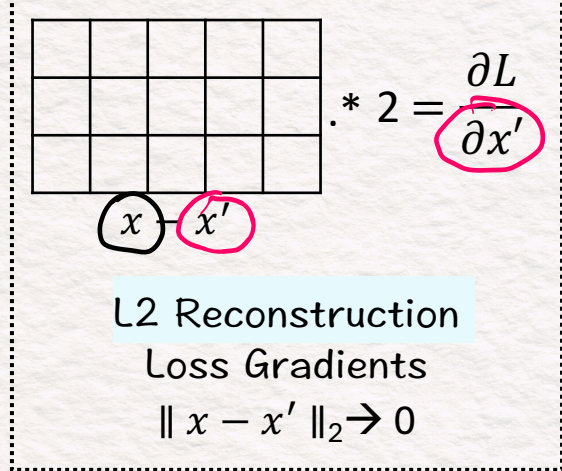
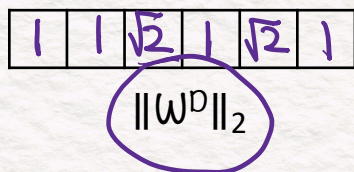
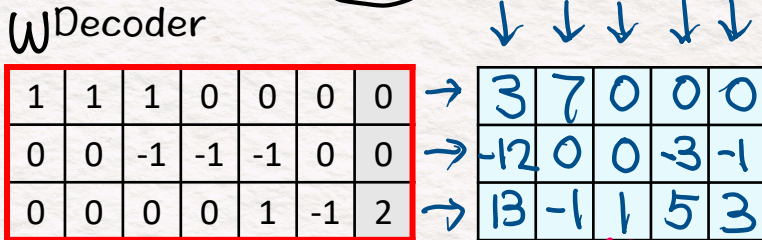
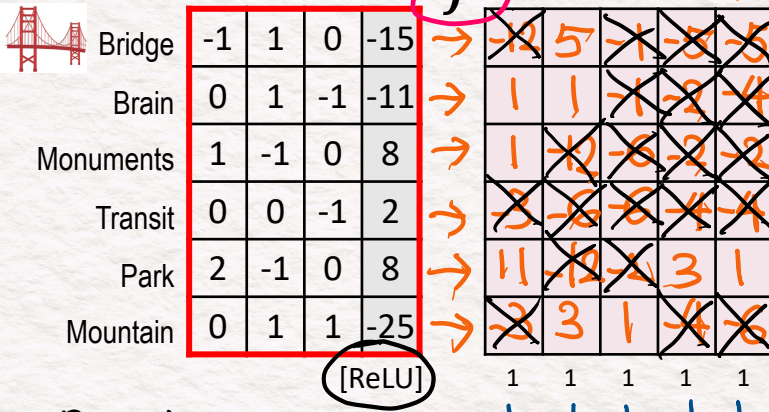
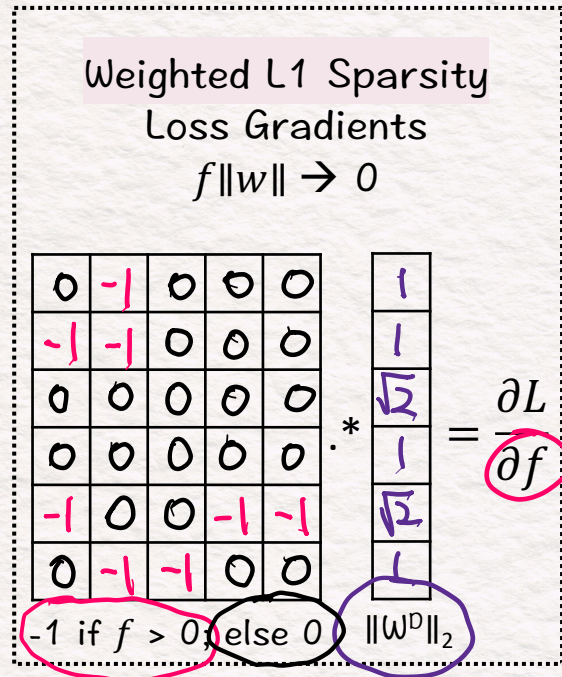


- x model activation
- f interpretable features
- x' reconstructed activation



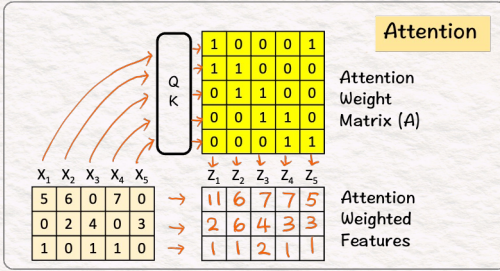
x

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

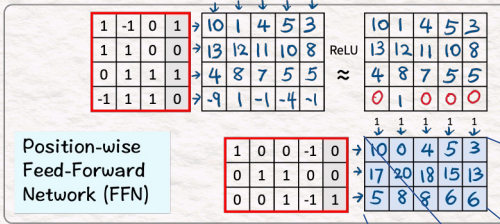


Transformer

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

Weighted L1 Sparsity Loss Gradients

$f \|w\| \rightarrow 0$

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

$\cdot \begin{bmatrix} 1 \\ 1 \\ \sqrt{2} \\ 1 \\ \sqrt{2} \\ 1 \end{bmatrix} = \frac{\partial L}{\partial f}$

$-1 \text{ if } f > 0, \text{ else } 0$

$\|w^D\|_2$

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

10	0	4	5	3
1	1	-1	-2	-4
1	2	8	8	6
3	8	8	6	6
11	12	11	3	1
2	3	1	4	6

[ReLU]

w^{Decoder}

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

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

$\|w^D\|_2$

$\begin{bmatrix} 1 & 1 & \sqrt{2} & 1 & \sqrt{2} & 1 \end{bmatrix}$

L2 Reconstruction Loss Gradients

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

7	-7	4	5	3
29	20	18	18	14
-8	9	7	1	3

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

x x'