

Figure 1: A plot showing the normalized density of attention values (x-axis, logarithmic scale from  $10^{-7}$  to  $10^{-1}$ ) and the sparsity distribution (y-axis, linear scale from 0.0 to 1.0). The plot displays two overlapping distributions: a blue one centered around  $10^{-3}$  and a red one centered around  $10^{-1}$ . The red distribution is significantly higher than the blue one at higher attention values. A table on the right shows the sparsity distribution for each attention value bin, with the value 1.00 highlighted in red for the highest bin.

sparsity distribution	
0.00	1.0
0.00	0.8
0.00	0.6
0.00	0.4
0.00	0.2
0.00	0.0
1.00	0.0

The figure displays a series of overlapping probability density functions (PDFs) representing attention values across various sparsity levels. The x-axis, labeled 'attention values', uses a logarithmic scale from  $10^{-7}$  to  $10^{-1}$ . The primary y-axis on the left measures 'normalized density' from 0.00 to 0.20. A secondary y-axis on the right indicates the 'sparsity distribution' from 0.0 to 1.0, with a color gradient from blue (low sparsity) to red (high sparsity). The curves show that as the sparsity distribution increases, the peak of the attention value distribution shifts towards higher values (moving right on the log scale).

Figure 1: A plot showing the distribution of attention values for different sparsity levels. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The left y-axis is 'normalized density' from 0.00 to 0.20. The right y-axis is 'sparsity distribution' from 0.0 to 1.0. A color bar on the right indicates the sparsity level for each density curve. The curves show that as sparsity increases, the distribution of attention values shifts towards higher values.

Figure 1 is a plot showing the normalized density of attention values (x-axis, logarithmic scale from  $10^{-7}$  to  $10^{-1}$ ) for different sparsity levels (y-axis, 0.0 to 1.0). The plot shows that as sparsity increases, the distribution of attention values shifts towards higher values (right side of the plot). A color bar on the right indicates the sparsity level for each density value.

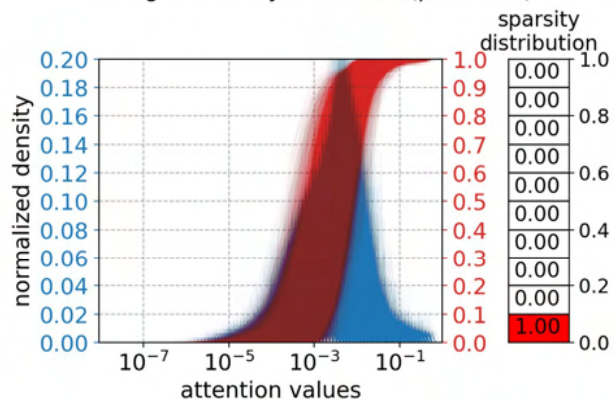


Figure 1 is a plot showing the normalized density of attention values for different sparsity distributions. The x-axis represents attention values on a logarithmic scale from  $10^{-7}$  to  $10^{-1}$ . The left y-axis represents normalized density from 0.00 to 0.20. The right y-axis represents the sparsity distribution from 0.0 to 1.0. A color bar on the right indicates the mapping from sparsity distribution to color, with 0.0 being blue and 1.0 being red. The plot shows a transition from blue to red as the sparsity distribution increases.

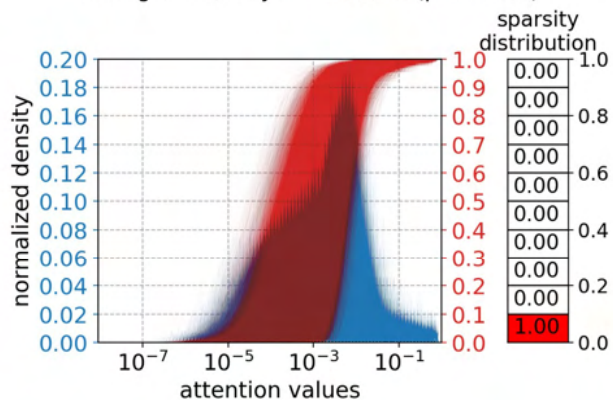
[illegible]



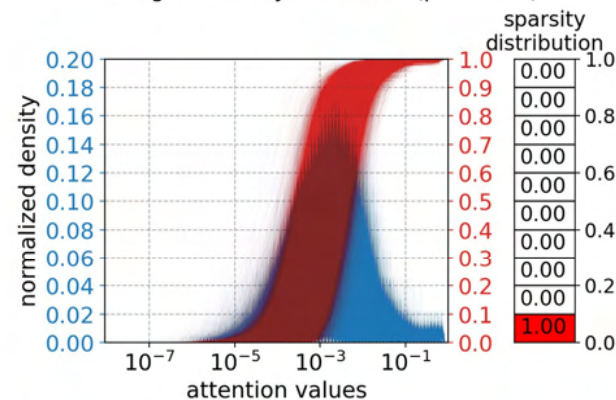
Histogram for layer 1 head 8(per token)



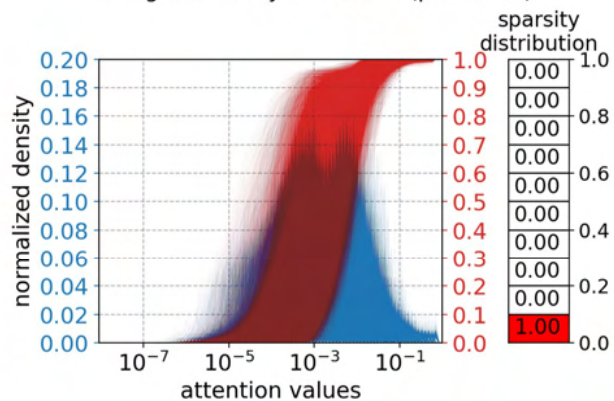
Histogram for layer 1 head 11(per token)



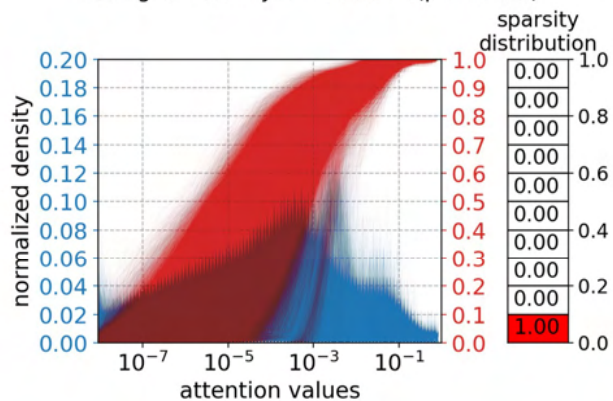
Histogram for layer 2 head 2(per token)



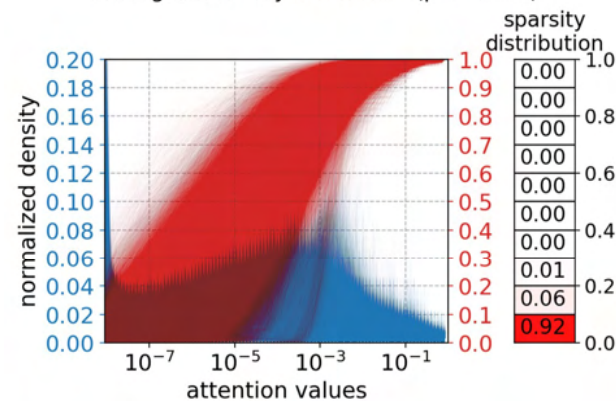
Histogram for layer 1 head 7(per token)



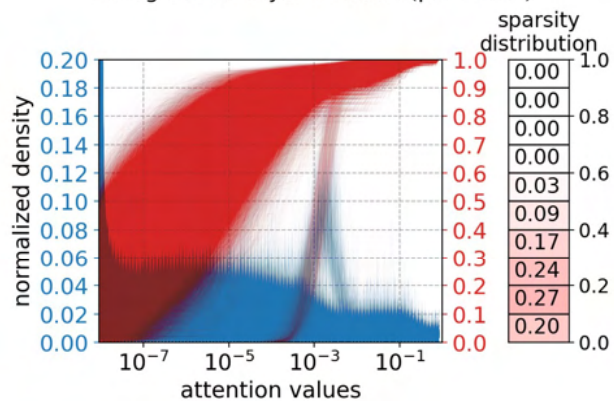
Histogram for layer 1 head 10(per token)



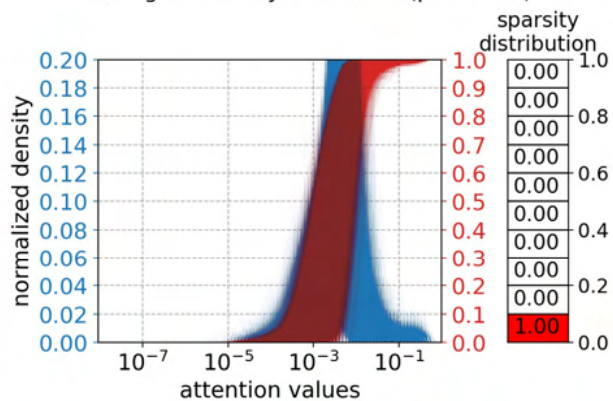
Histogram for layer 2 head 1(per token)



Histogram for layer 1 head 6(per token)



Histogram for layer 1 head 9(per token)



Histogram for layer 2 head 0(per token)

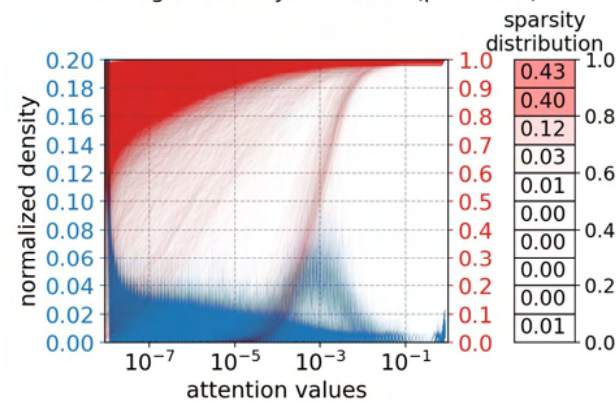




Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red curve represents a sparsity distribution of 1.00, and a blue curve represents a sparsity distribution of 0.00. The red curve is shifted to the right (higher attention values) compared to the blue curve. A legend on the right shows the sparsity distribution for each curve.

sparsity distribution	
0.00	1.0
0.00	0.9
0.00	0.8
0.00	0.7
0.00	0.6
0.00	0.5
0.00	0.4
0.00	0.3
0.00	0.2
0.00	0.1
1.00	0.0

Figure 1 is a plot showing the normalized density of attention values for different sparsity distributions. The x-axis represents attention values on a logarithmic scale, ranging from  $10^{-7}$  to  $10^{-1}$ . The y-axis represents the normalized density, ranging from 0.00 to 0.20. The plot displays two curves: a red curve representing a sparsity distribution of 1.00 and a blue curve representing a sparsity distribution of 0.00. The red curve is shifted to the right (higher attention values) compared to the blue curve. A legend on the right side of the plot shows the sparsity distribution for each curve, with values ranging from 0.0 to 1.0.

Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red shaded region represents the distribution for a sparsity of 1.00, and a blue shaded region represents the distribution for a sparsity of 0.00. The red distribution is shifted to the right (higher attention values) compared to the blue distribution. A table on the right shows the sparsity distribution for each row, with the bottom row (red) having a sparsity of 1.00 and all other rows having a sparsity of 0.00.

sparsity	distribution
1.0	0.00
0.9	0.00
0.8	0.00
0.7	0.00
0.6	0.00
0.5	0.00
0.4	0.00
0.3	0.00
0.2	0.00
0.1	0.00
0.0	1.00

Figure 1: A plot showing the normalized density of attention values (x-axis, logarithmic scale from  $10^{-7}$  to  $10^{-1}$ ) and the sparsity distribution (y-axis, linear scale from 0.0 to 1.0). The plot displays two overlapping distributions: a blue one (left) and a red one (right). The red distribution is shifted towards higher attention values, indicating a higher sparsity. A table on the right shows the sparsity distribution for different values, with the value 1.00 highlighted in red.

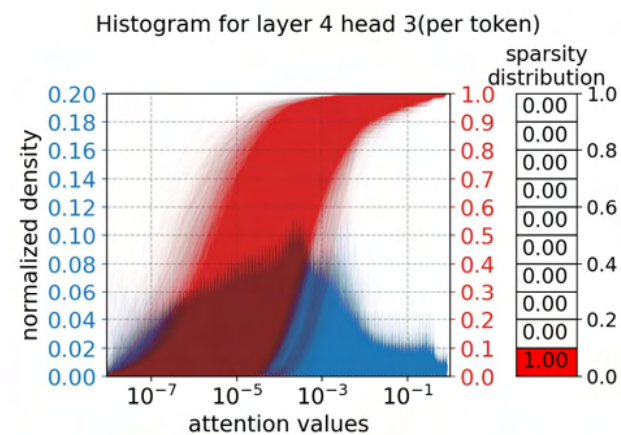
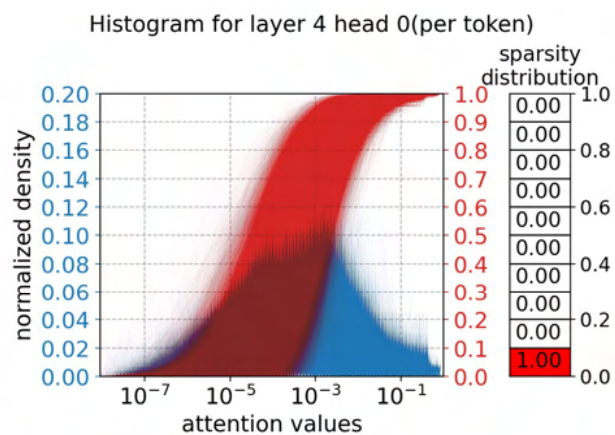
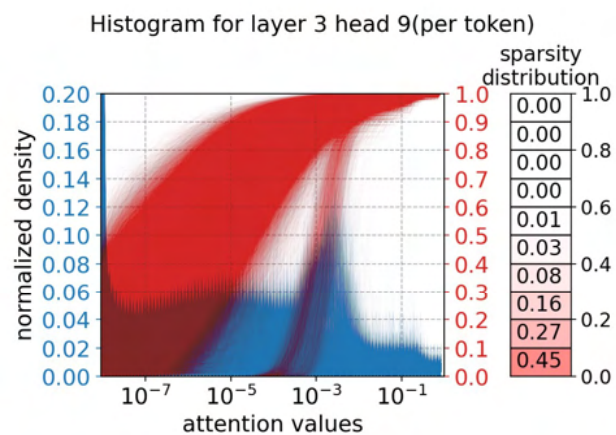
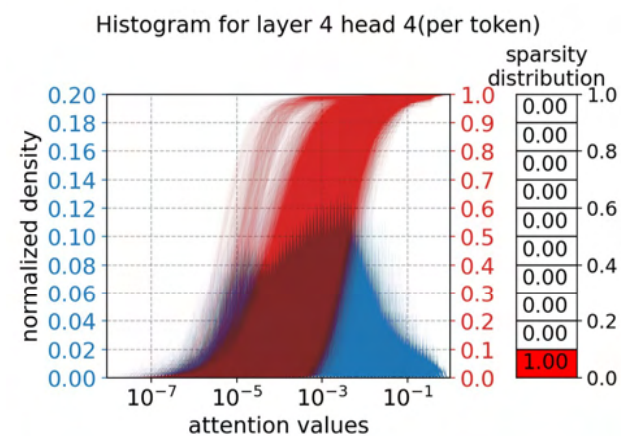
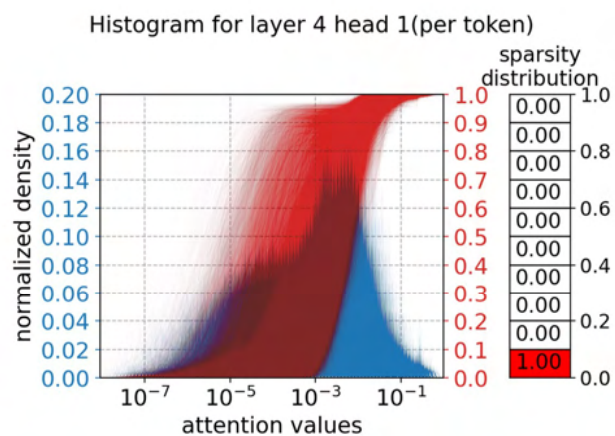
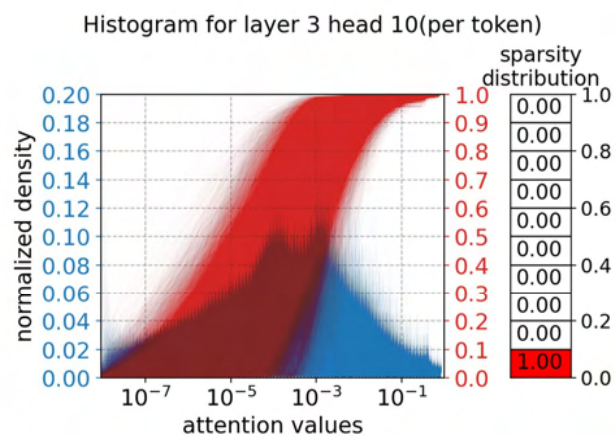
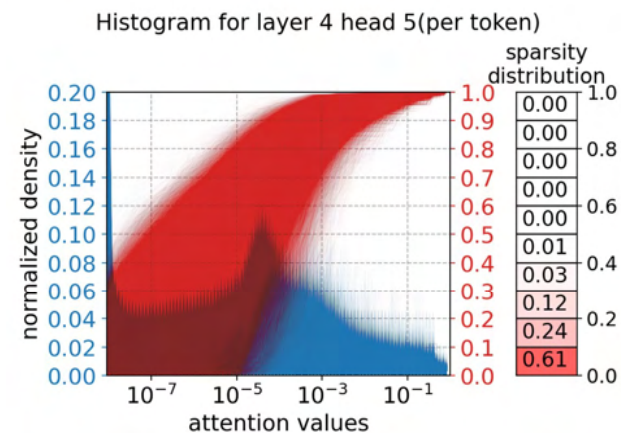
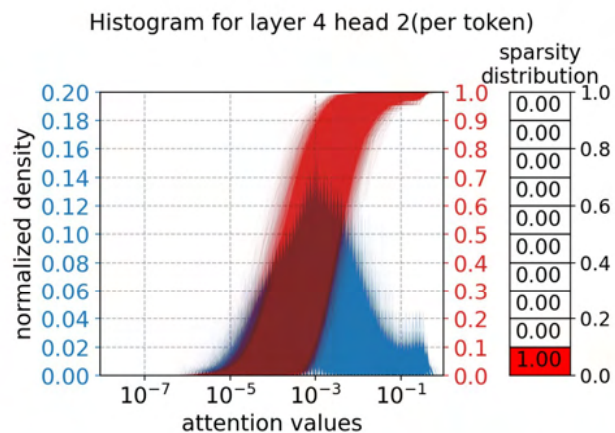
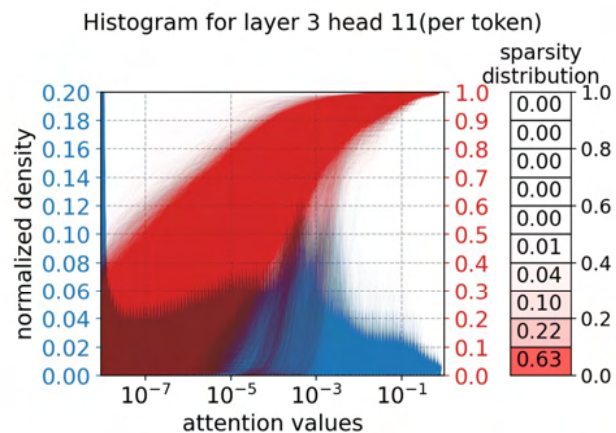


sparsity distribution	
0.00	1.0
0.00	0.8
0.00	0.6
0.00	0.4
0.00	0.2
0.00	0.0
1.00	0.0

Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red shaded region represents the distribution for a sparsity of 1.00, and a blue shaded region represents the distribution for a sparsity of 0.00. The red distribution is shifted to the right (higher attention values) compared to the blue distribution. A table on the right shows the sparsity distribution for each row, with the bottom row (red) having a sparsity of 1.00 and all other rows having a sparsity of 0.00.

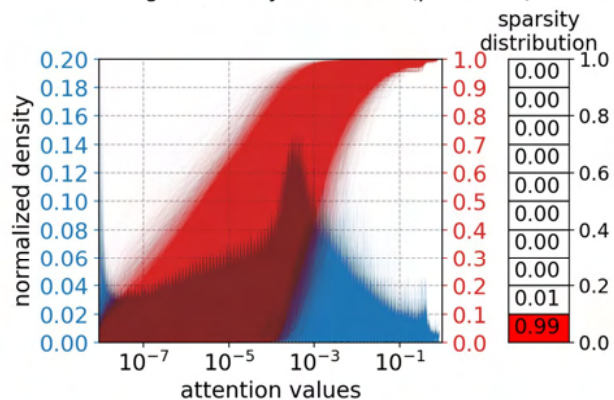
sparsity	distribution
1.0	0.00
0.9	0.00
0.8	0.00
0.7	0.00
0.6	0.00
0.5	0.00
0.4	0.00
0.3	0.00
0.2	0.00
0.1	0.00
0.0	1.00



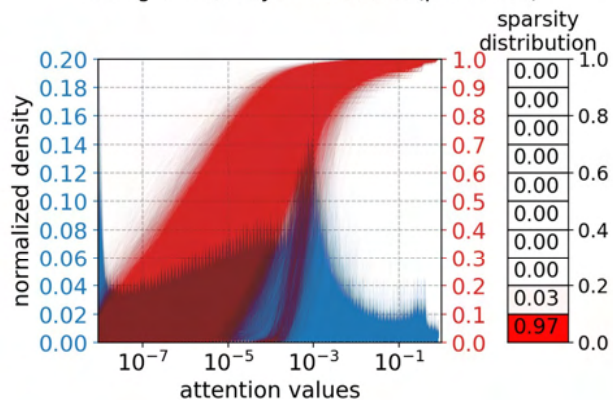




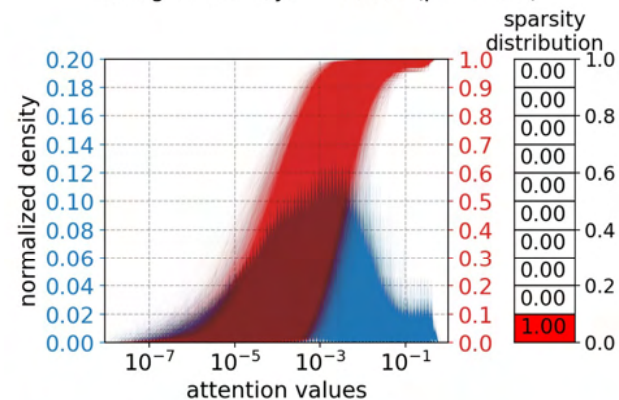
Histogram for layer 4 head 8(per token)



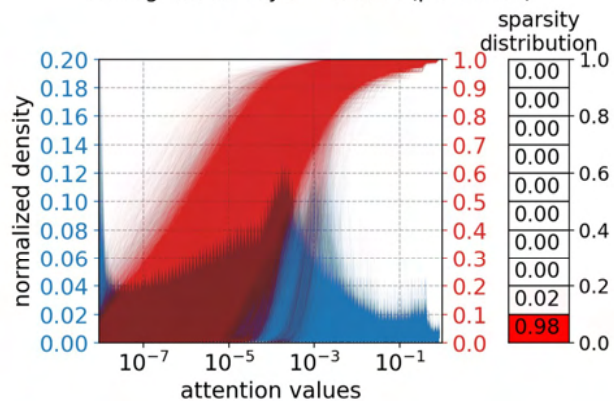
Histogram for layer 4 head 11(per token)



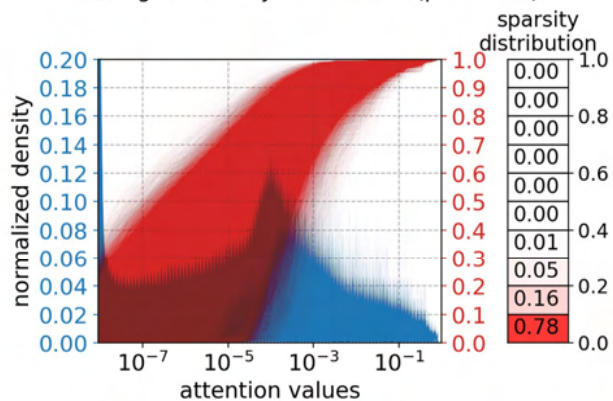
Histogram for layer 5 head 2(per token)



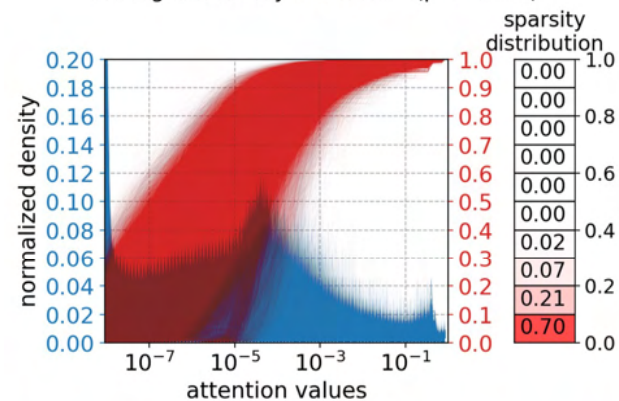
Histogram for layer 4 head 7(per token)



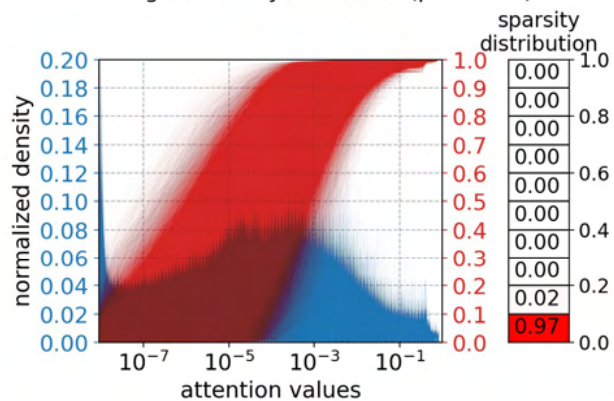
Histogram for layer 4 head 10(per token)



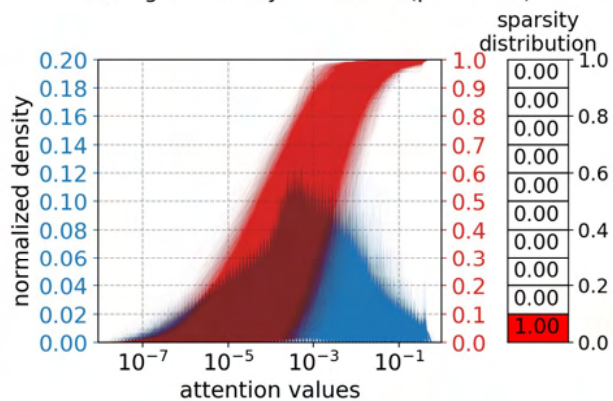
Histogram for layer 5 head 1(per token)



Histogram for layer 4 head 6(per token)



Histogram for layer 4 head 9(per token)



Histogram for layer 5 head 0(per token)

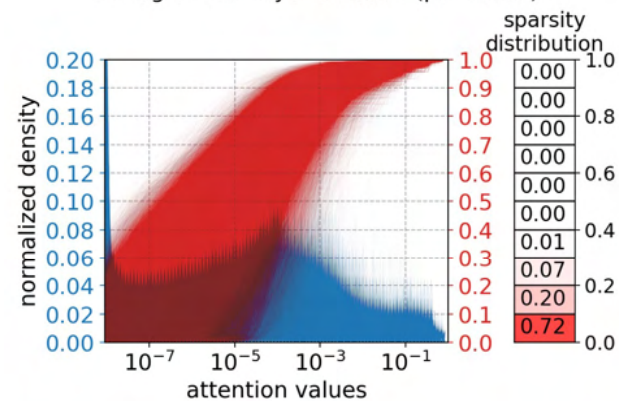




Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The left y-axis is 'normalized density' from 0.00 to 0.20. The right y-axis is 'sparsity distribution' from 0.0 to 1.0. A red shaded area represents the distribution for sparsity values from 0.0 to 1.0, and a blue shaded area represents the distribution for sparsity values from 0.0 to 1.0. The distributions are centered around  $10^{-3}$ .

Figure 1 is a contour plot showing the distribution of attention values for the 'sparsity' parameter. The x-axis represents 'attention values' on a logarithmic scale, ranging from  $10^{-7}$  to  $10^{-1}$ . The y-axis represents 'normalized density', ranging from 0.00 to 0.20. The plot displays a red region (high density) and a blue region (low density). A color bar on the right indicates the 'sparsity distribution' from 0.0 to 1.0, with a red bar at 0.80.

Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red curve represents the sparsity distribution, and a blue curve represents the attention values distribution. A table on the right shows the sparsity distribution for different attention values.

attention values	sparsity distribution
0.00	1.00
0.00	0.80
0.00	0.60
0.00	0.40
0.00	0.20
0.02	0.00
0.08	0.00
0.22	0.69
0.69	0.00



attention values	sparsity distribution
0.00	1.0
0.00	0.9
0.00	0.8
0.00	0.7
0.00	0.6
0.00	0.5
0.00	0.4
0.02	0.3
0.09	0.2
0.24	0.1
0.65	0.0

attention values	sparsity distribution
0.00	1.0
0.00	0.9
0.00	0.8
0.00	0.7
0.00	0.6
0.00	0.5
0.00	0.4
0.02	0.3
0.09	0.2
0.24	0.1
0.65	0.0

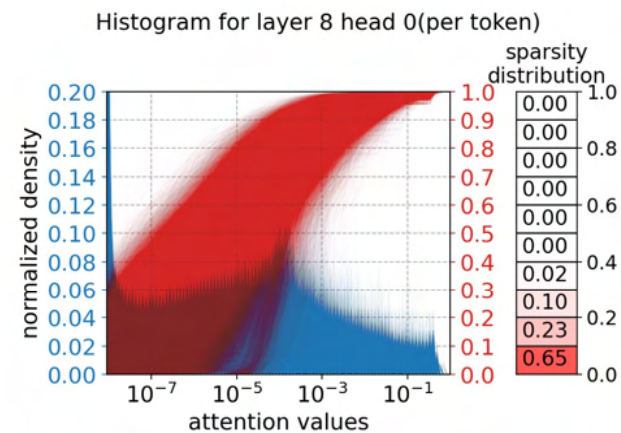
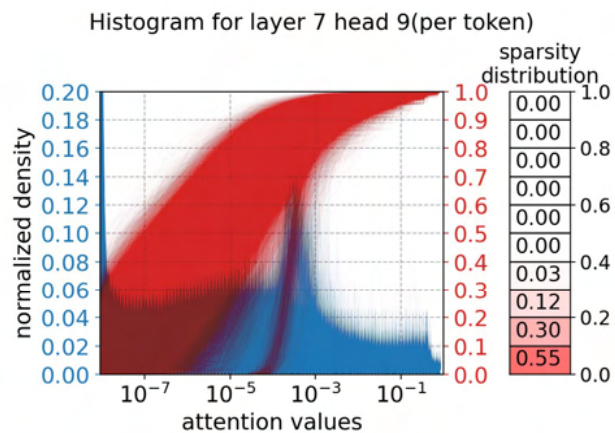
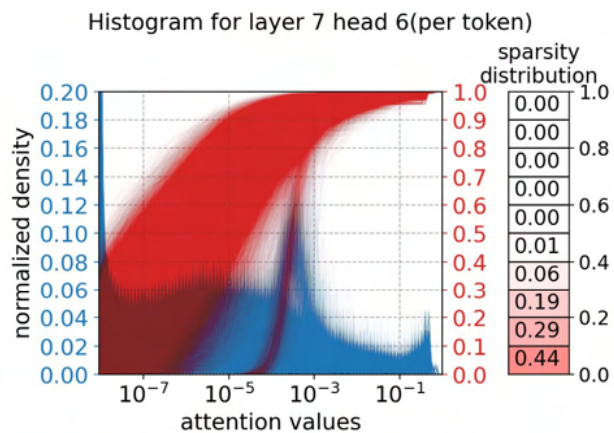
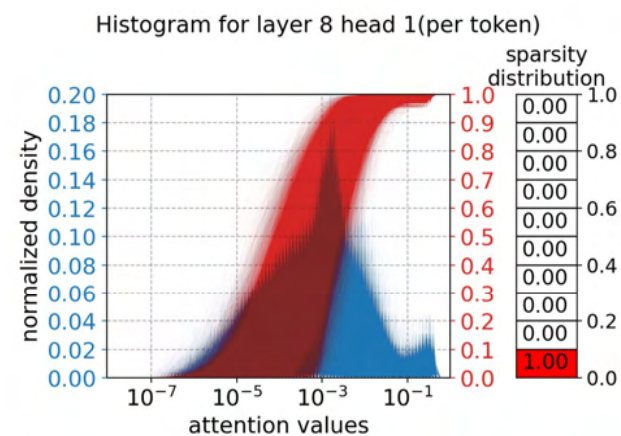
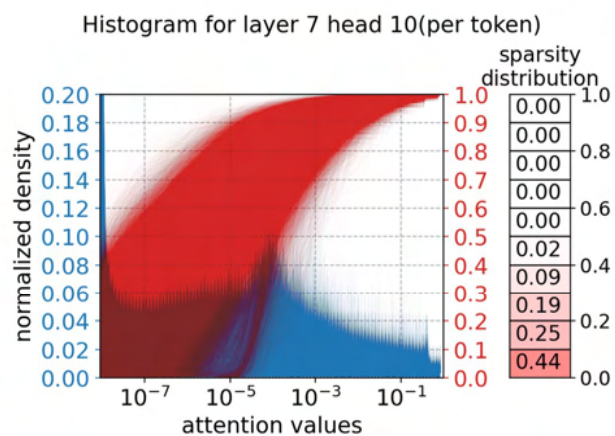
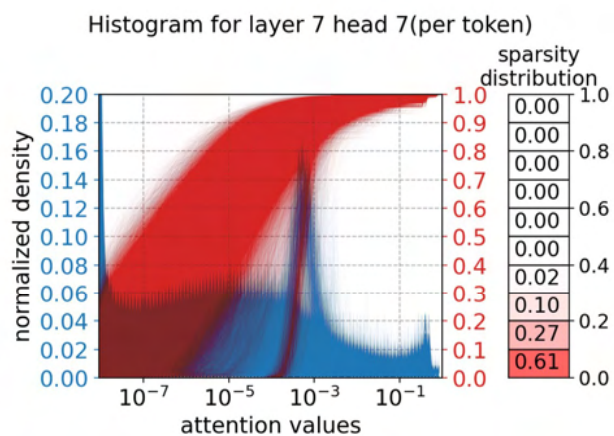
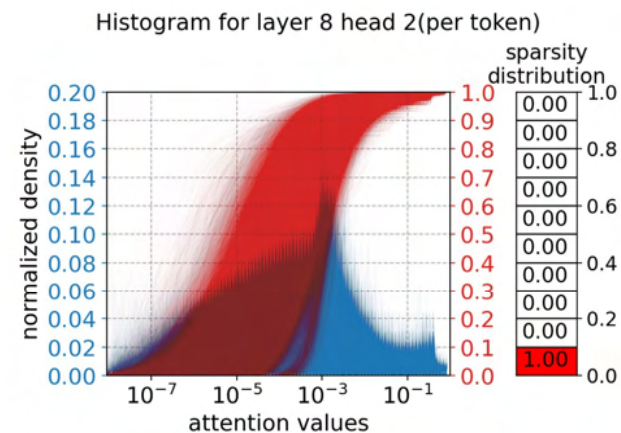
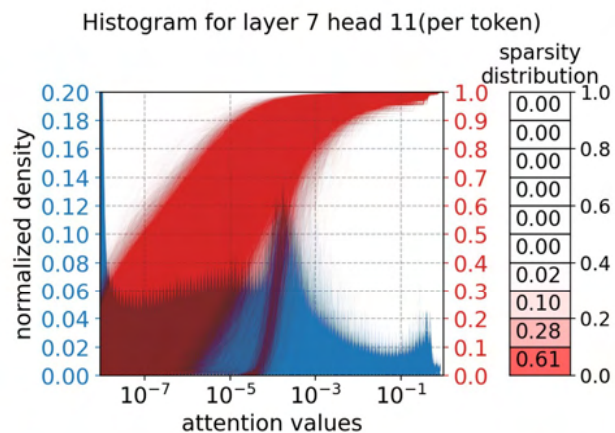
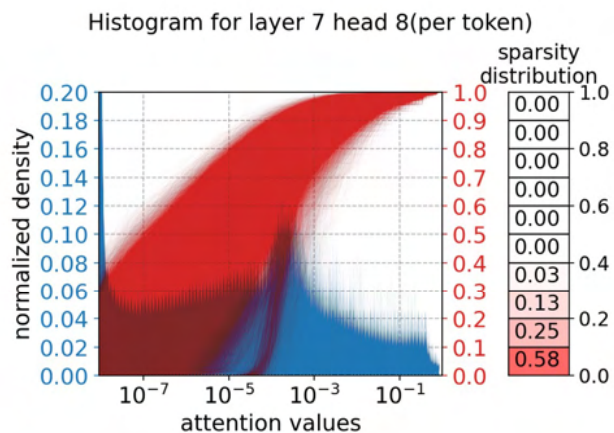
Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red curve represents the sparsity distribution, and a blue curve represents the attention values. A table on the right shows the sparsity distribution for different values of the parameter  $\alpha$ .

sparsity distribution	$\alpha$
0.00	1.0
0.00	0.8
0.00	0.6
0.00	0.4
0.00	0.2
0.02	0.1
0.12	0.05
0.30	0.02
0.55	0.01



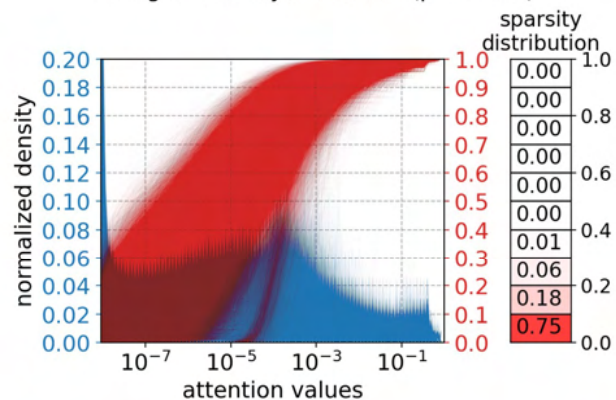
sparsity distribution
0.00
0.00
0.00
0.00
0.00
0.03
0.09
0.21
0.30
0.37
0.0



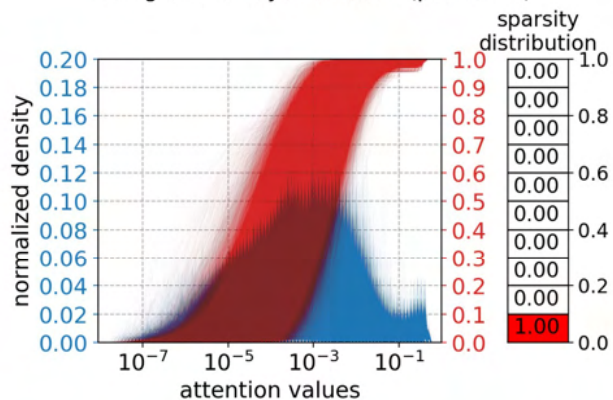




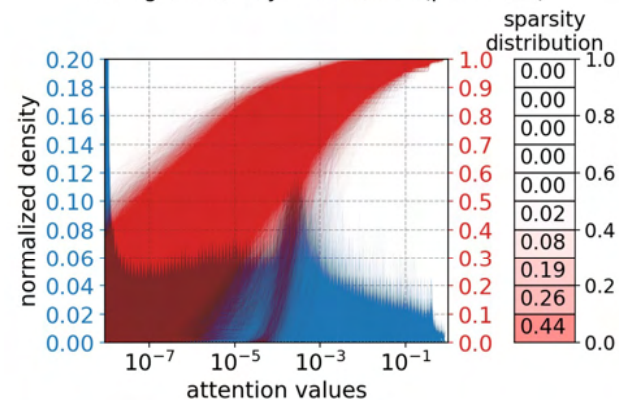
Histogram for layer 8 head 5(per token)



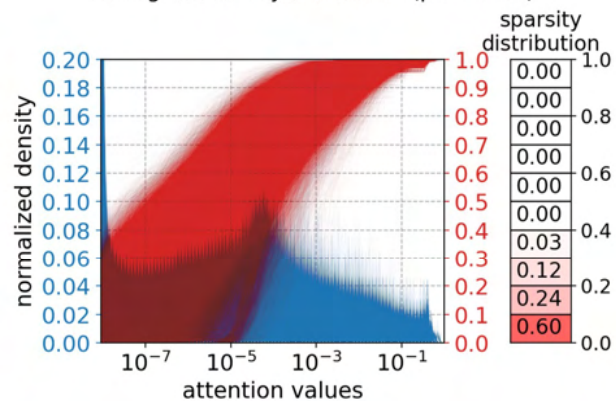
Histogram for layer 8 head 8(per token)



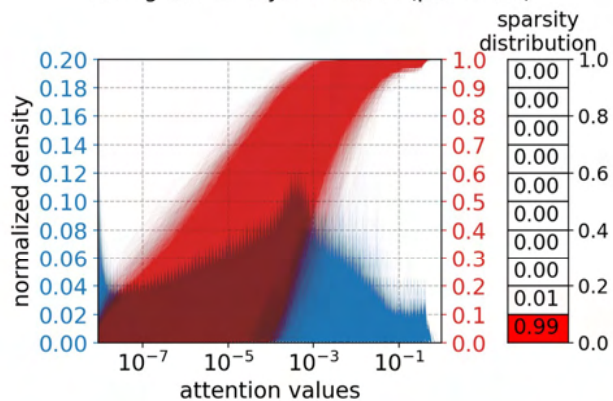
Histogram for layer 8 head 11(per token)



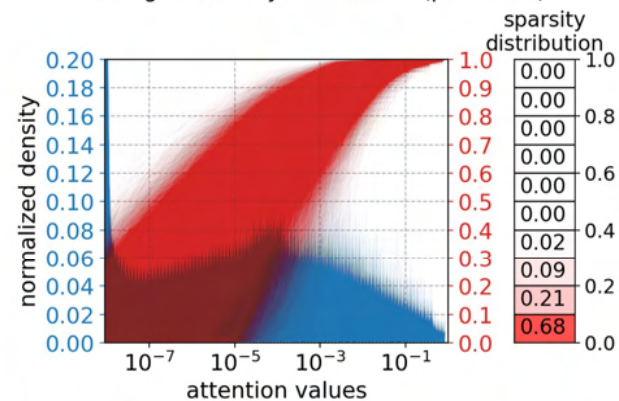
Histogram for layer 8 head 4(per token)



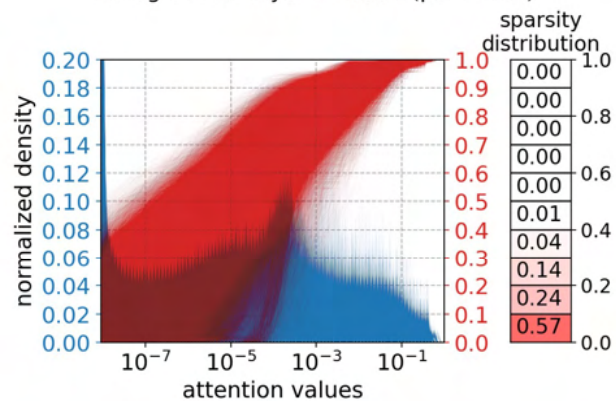
Histogram for layer 8 head 7(per token)



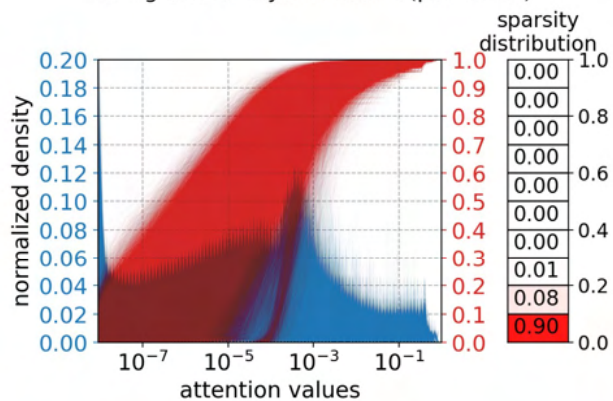
Histogram for layer 8 head 10(per token)



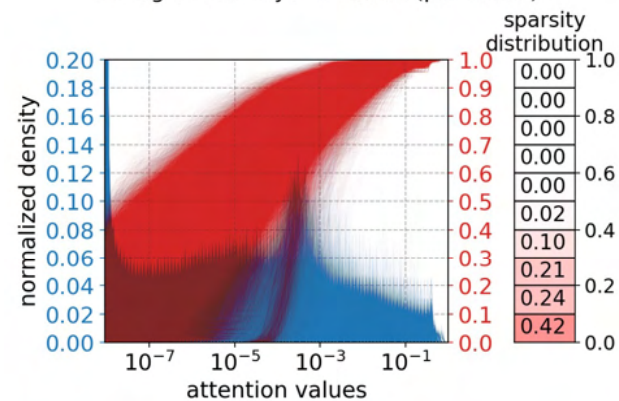
Histogram for layer 8 head 3(per token)



Histogram for layer 8 head 6(per token)

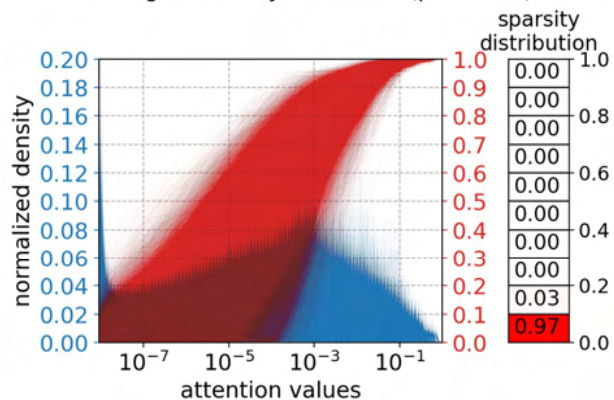


Histogram for layer 8 head 9(per token)

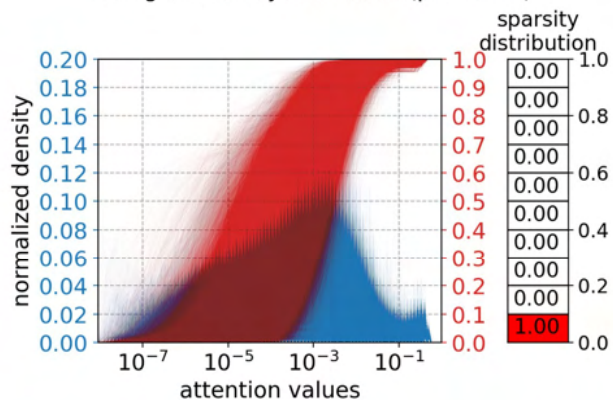




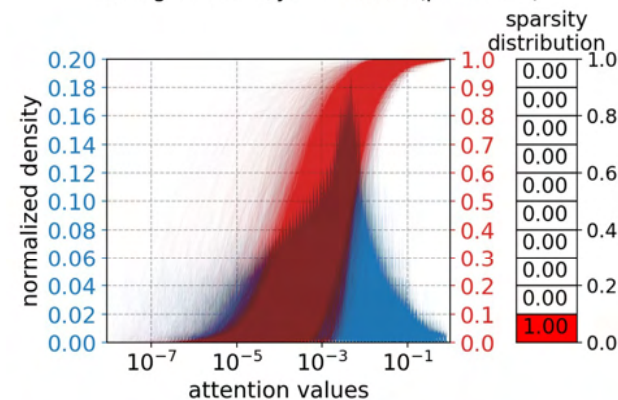
Histogram for layer 9 head 2(per token)



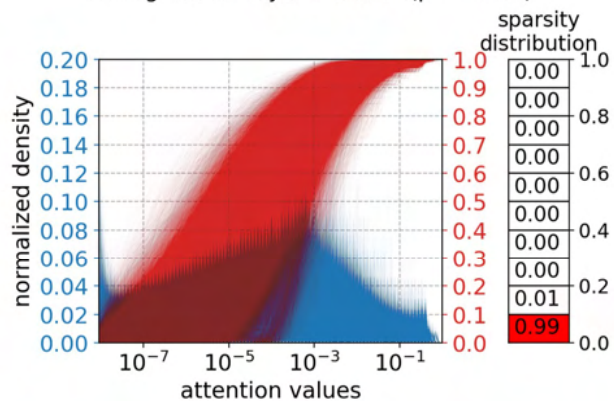
Histogram for layer 9 head 5(per token)



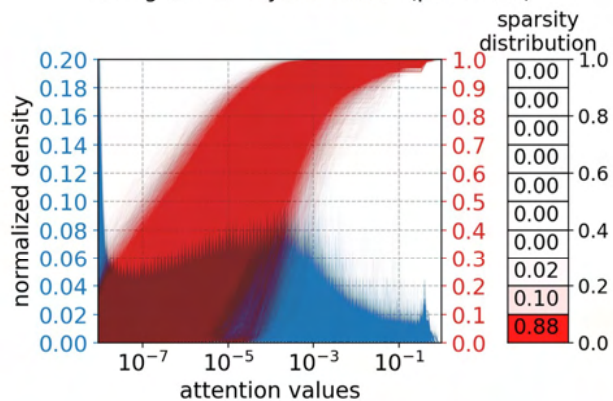
Histogram for layer 9 head 8(per token)



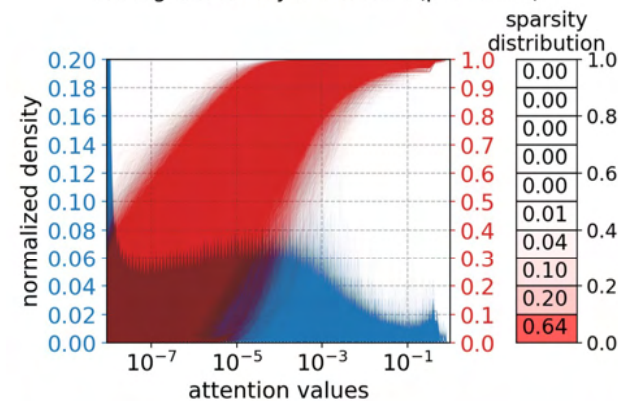
Histogram for layer 9 head 1(per token)



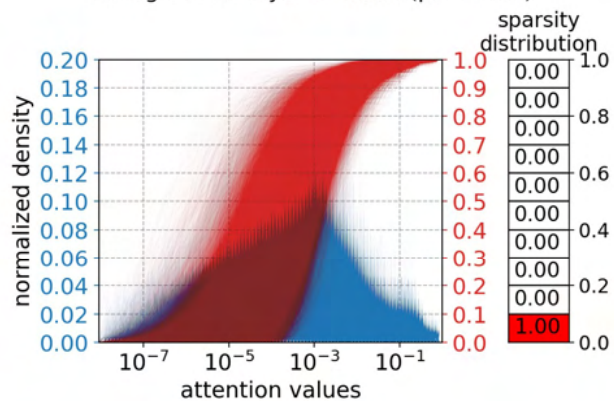
Histogram for layer 9 head 4(per token)



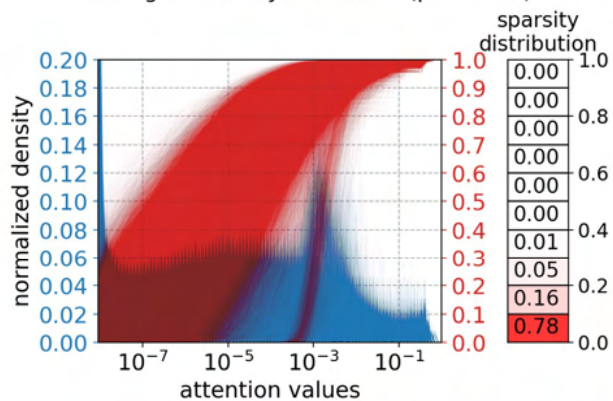
Histogram for layer 9 head 7(per token)



Histogram for layer 9 head 0(per token)



Histogram for layer 9 head 3(per token)



Histogram for layer 9 head 6(per token)

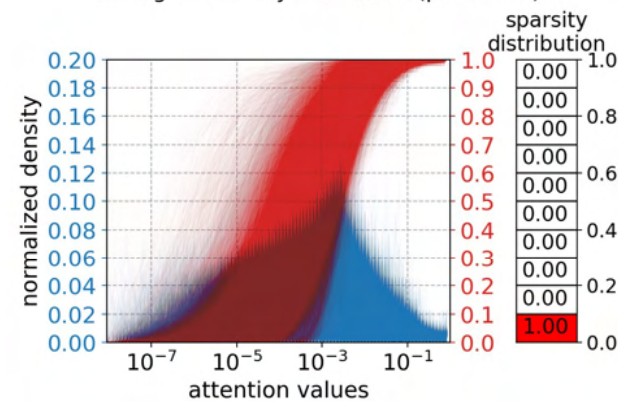




Figure 1: A plot showing the normalized density of attention values for different sparsity distributions. The x-axis is 'attention values' on a log scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis is 'normalized density' from 0.00 to 0.20. A red curve represents the sparsity distribution, and a blue curve represents the attention values distribution. A table on the right shows the sparsity distribution for different values of the sparsity parameter.

sparsity	distribution
1.0	0.00
0.9	0.00
0.8	0.00
0.7	0.00
0.6	0.00
0.5	0.00
0.4	0.00
0.3	0.00
0.2	0.00
0.1	0.00
0.0	1.00

The figure displays a series of overlapping probability density functions (PDFs) representing the distribution of attention values across different levels of sparsity. The x-axis, labeled 'attention values', uses a logarithmic scale from  $10^{-7}$  to  $10^{-1}$ . The left y-axis shows 'normalized density' from 0.00 to 0.20. The right y-axis indicates the 'sparsity distribution' from 0.0 to 1.0. A color bar on the right maps sparsity values to colors: blue for low sparsity (0.0), transitioning through green and yellow to red for high sparsity (1.0). The curves show that as sparsity increases, the distribution of attention values shifts to the right, indicating that more attention is concentrated on higher-value tokens.

normalized density	sparsity distribution
0.00	0.00
0.02	0.00
0.04	0.00
0.06	0.00
0.08	0.00
0.10	0.00
0.12	0.00
0.14	0.00
0.16	0.00
0.18	0.00
0.20	0.00

Figure 1 is a plot showing the normalized density of attention values for different sparsity distributions. The x-axis represents attention values on a logarithmic scale from  $10^{-7}$  to  $10^{-1}$ . The y-axis represents the normalized density from 0.00 to 0.20. The plot shows a transition from a blue distribution (low sparsity) to a red distribution (high sparsity). A table on the right shows the sparsity distribution for each row, with the bottom row (red) having a sparsity of 1.00.

sparsity	distribution
1.0	0.00
0.9	0.00
0.8	0.00
0.7	0.00
0.6	0.00
0.5	0.00
0.4	0.00
0.3	0.00
0.2	0.00
0.1	0.00
0.0	1.00



