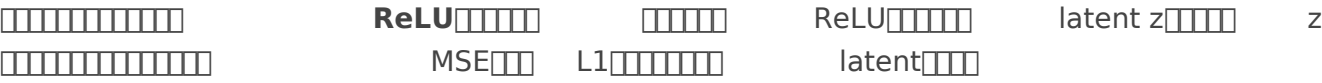
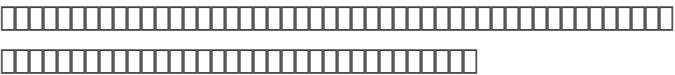




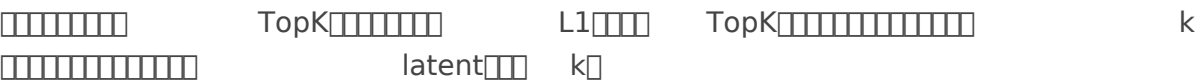
Ref <https://mp.weixin.qq.com/s/iZHPnnIncVFfa8QJOuH8qFg>



$$z = \text{ReLU}(W_{\text{enc}}(x - b_{\text{pre}}) + b_{\text{enc}})$$

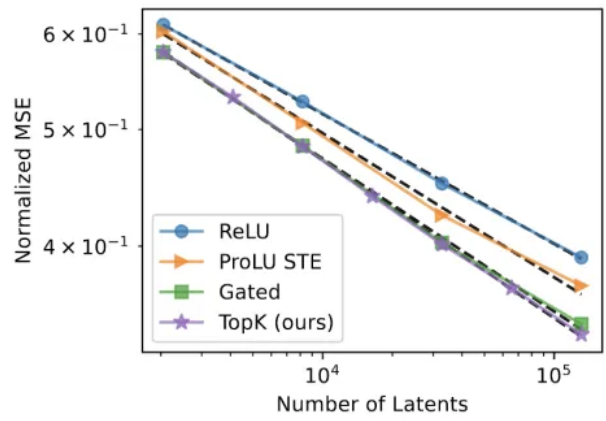
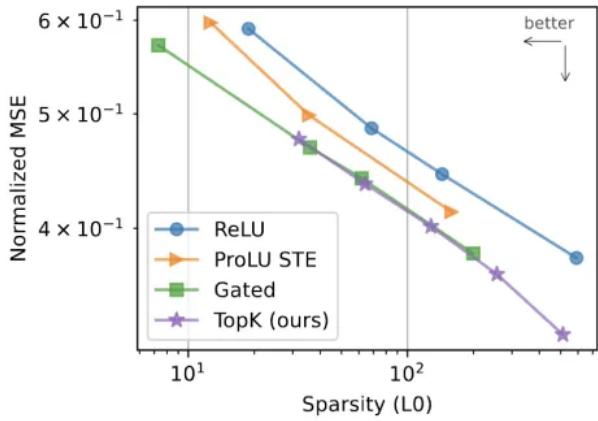
$$\hat{x} = W_{\text{dec}}z + b_{\text{pre}}$$

公众号 · 量子位



$$z = \text{TopK}(W_{\text{enc}}(x - b_{\text{pre}}))$$

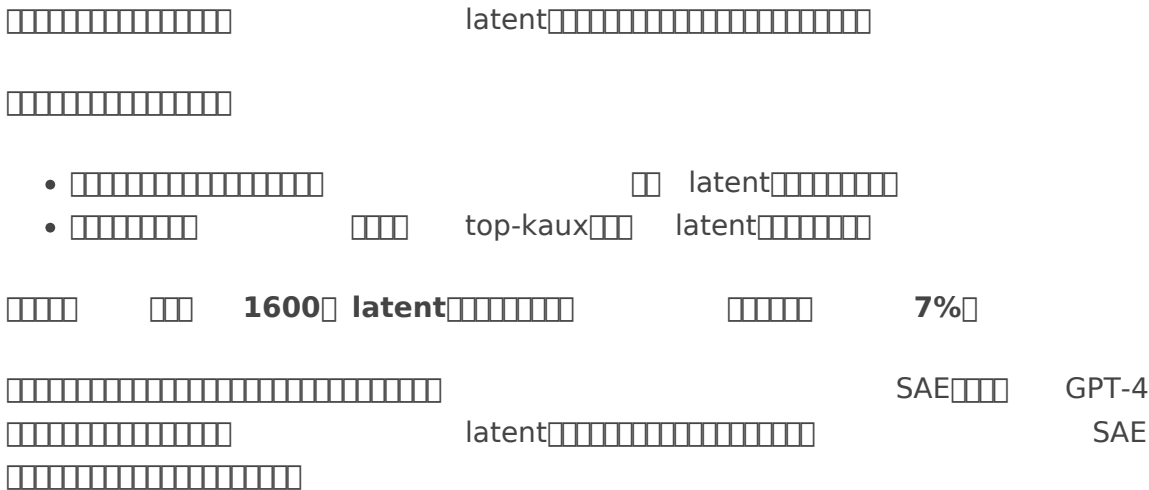




(a) At a fixed number of latents ($n = 32768$), TopK has a better reconstruction-sparsity trade off than ReLU and ProLU, and is comparable to Gated.

(b) At a fixed sparsity level ($L_0 = 128$), scaling laws are steeper for TopK than ReLU.⁶

Figure 2: Comparison between TopK and other activation functions.



Revision #1

Created 11 January 2025 09:44:04 by Colin

Updated 11 January 2025 09:44:07 by Colin