ICML
International Conference
On Machine Learning

清華大學
Tsinghua University

Sparsing Law: Towards Large Language Models with Greater Activation Sparsity

Yuqi Luo[*1] Chenyang Song[*1] Xu Han[1] Yingfa Chen[1] Chaojun Xiao[1]
Xiaojun Meng[2] Liqun Deng[2] Jiansheng Wei[2] Zhiyuan Liu[1] Maosong Sun[1]

[1] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China

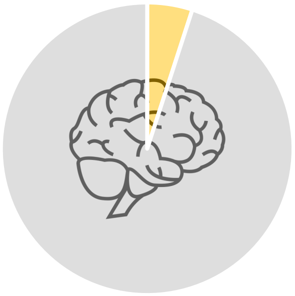
[2] Huawei Noah's Ark Lab, China

luo-yq23@mails.tsinghua.edu.cn, scy22@mails.tsinghua.edu.cn

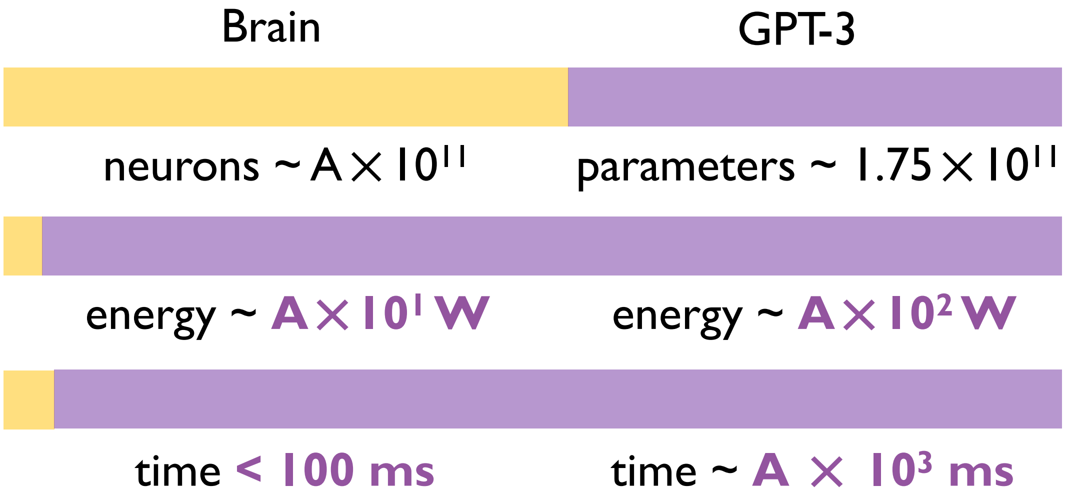
Why Activation Sparsity?

The rise of LLMs bring about serious issues of efficiency.

- There have been an exponential increase in the energy consumption of LLMs in recent years.
- With a similar numerical scale of neurons, brain consumes significantly less energy and shorter response time.
- **Activation sparsity** is one of the most important properties that cause such low energy consumption.



activation ratio < 5%



Brain

GPT-3

neurons ~ $A \times 10^{11}$

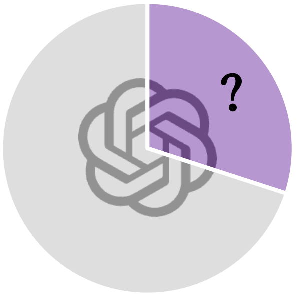
parameters ~ 1.75×10^{11}

energy ~ **$A \times 10^{1}$ W**

energy ~ **$A \times 10^{2}$ W**

time **< 100 ms**

time ~ **$A \times 10^{3}$ ms**



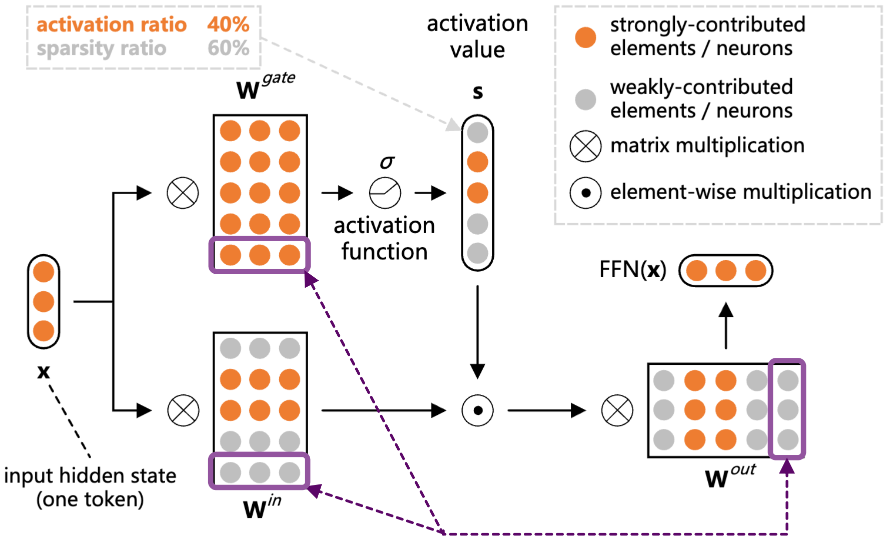
activation ratio ?

Von Bartheld, Christopher S., Jami Bahney, and Suzana Herculano-Houzel. "The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting." Journal of Comparative Neurology 524.18 (2016): 3865-3895.
Bullmore, Ed, and Olaf Sporns. "The economy of brain network organization." Nature reviews neuroscience 13.5 (2012): 336-349.
Lennie, Peter. "The cost of cortical computation." Current biology 13.6 (2003): 493-497.

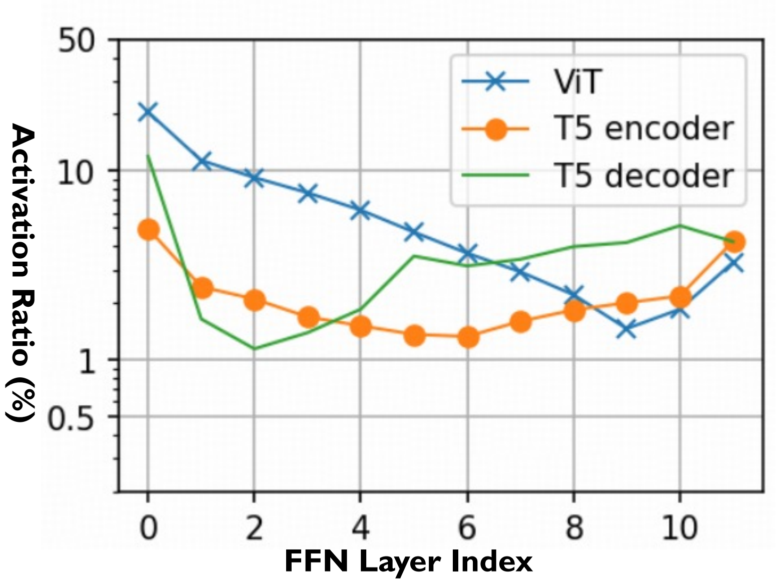
Activation Sparsity in LLMs

Similar to brains, LLMs also prevalently have activation sparsity.

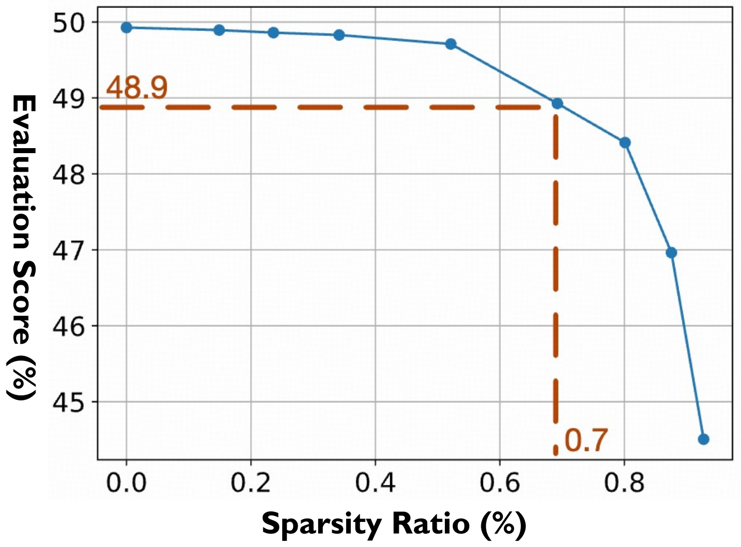
- Definition: **considerable zero or negligible elements in activation outputs**, corresponding to certain model parameters (i.e., FFN neurons), **have a weak impact on LLM outputs given a specific input**
- Activation sparsity intrinsically exists in ReLU, but can also be found in mainstream SiLU activation.



"**Neurons**" and activation sparsity in FFN



Activation sparsity in T5 & ViT (ReLU)



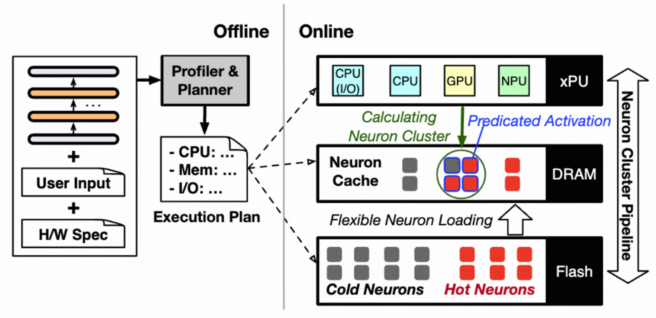
Activation sparsity in LLaMA2 (SiLU)

Li, Zonglin, et al. "The Lazy Neuron Phenomenon: On Emergence of Activation Sparsity in Transformers." The Eleventh International Conference on Learning Representations (2022).
Zhang, Zhengyan, et al. "ReLU² Wins: Discovering Efficient Activation Functions for Sparse LLMs." arXiv preprint arXiv:2402.03804 (2024).

Application of Activation Sparsity

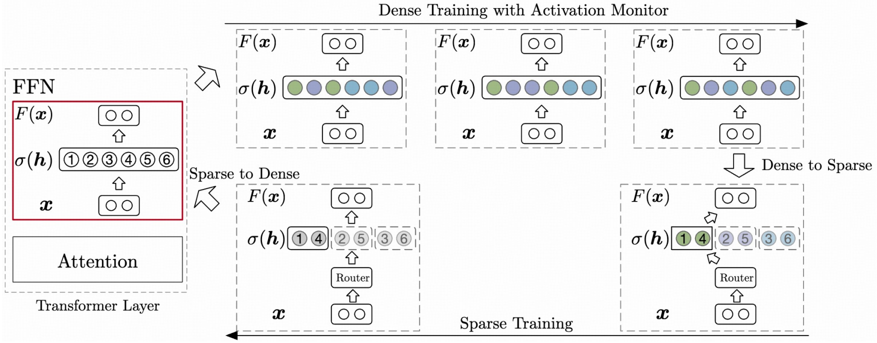
What does an LLM with high activation sparsity can provide?

Inference Acceleration



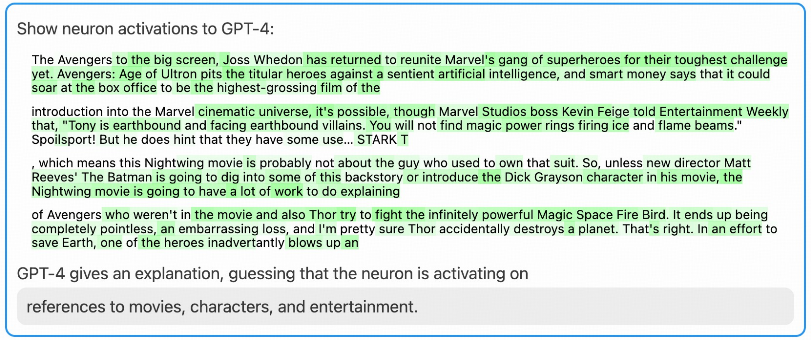
PowerInfer-2, by utilizing activation sparsity, can run sparsified Mixtral-47B on smart phones with **up to 27.8x speedup** compared to llama.cpp

Training Acceleration



SSD accelerates training through **MoE-dense conversions**, utilizing the activation sparsity during the whole training procedure

Interpretability



Show neuron activations to GPT-4:

The Avengers to the big screen, Joss Whedon has returned to reunite Marvel's gang of superheroes for their toughest challenge yet. Avengers: Age of Ultron pits the titular heroes against a sentient artificial intelligence, and smart money says that it could soar at the box office to be the highest-grossing film of the

introduction into the Marvel cinematic universe, it's possible, though Marvel Studios boss Kevin Feige told Entertainment Weekly that, "Tony is earthbound and facing earthbound villains. You will not find magic power rings firing ice and flame beams." Spoilsport! But he does hint that they have some use... STARK T

, which means this Nightwing movie is probably not about the guy who used to own that suit. So, unless new director Matt Reeves' The Batman is going to dig into some of this backstory or introduce the Dick Grayson character in his movie, the Nightwing movie is going to have a lot of work to do explaining

of Avengers who weren't in the movie and also Thor try to fight the infinitely powerful Magic Space Fire Bird. It ends up being completely pointless, an embarrassing loss, and I'm pretty sure Thor accidentally destroys a planet. That's right. In an effort to save Earth, one of the heroes inadvertantly blows up an

GPT-4 gives an explanation, guessing that the neuron is activating on

references to movies, characters, and entertainment.

OpenAI partly makes the behaviors of GPT-2 interpretable by prompting GPT-4 to analyze **the activation patterns of neurons**

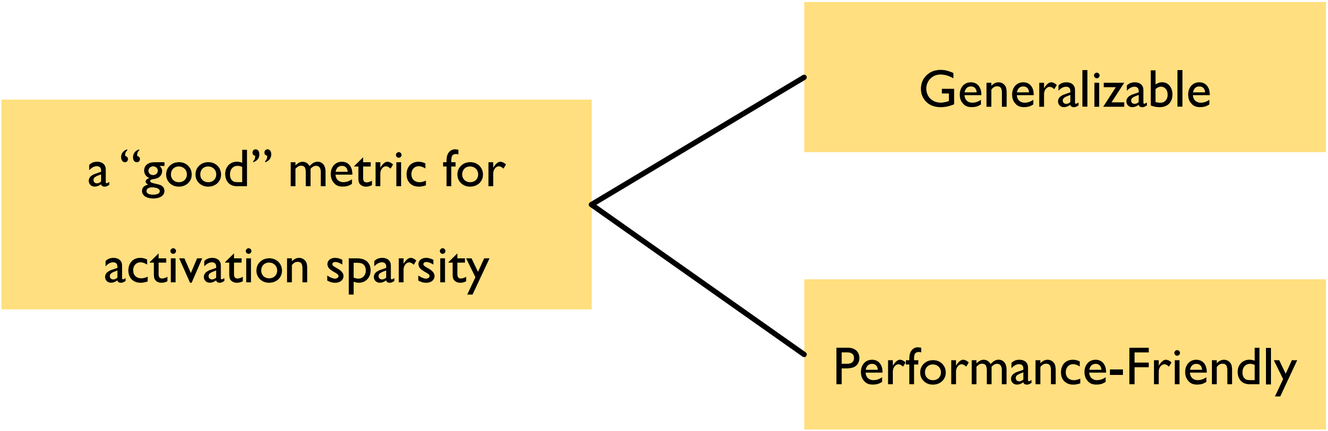
Xue, Zhenliang, et al. "PowerInfer-2: Fast Large Language Model Inference on a Smartphone." arXiv preprint arXiv:2406.06282 (2024).
Zhang, Zhengyan, et al. "Exploring the Benefit of Activation Sparsity in Pre-training." Forty-first International Conference on Machine Learning.
Bills, Steven, et al. "Language models can explain neurons in language models." URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023) 2 (2023).

Measurement of Activation Sparsity

Sparsing Law: A comprehensive quantitative study on activation sparsity.

Q1: How can activation sparsity be measured "better"?

- **Sparsity Ratio**: the average ratio of **weakly-contributed neurons** in FFNs
- **Activation Ratio**: 1 – Sparsity Ratio
- The key responsibility of a sparsity metric: determining which neurons **at each layer** contribute weakly to the model output given specific inputs



a "good" metric for activation sparsity

Generalizable

applicable to different activation functions

(e.g., ReLU, SiLU, GELU)

Performance-Friendly

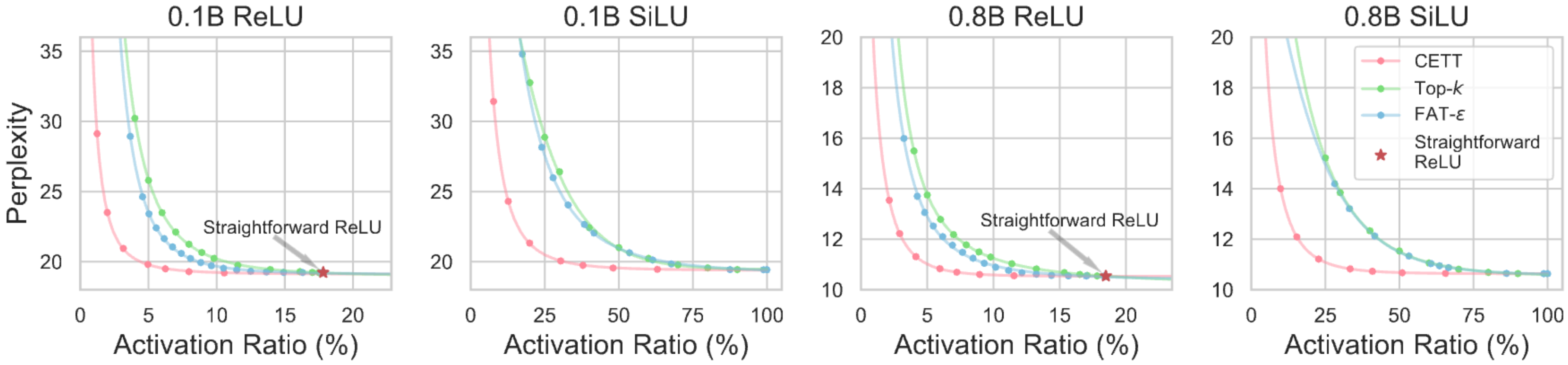
better trade-off between sparsity and performance

(i.e., better performance at the same sparsity ratio)

Measurement of Activation Sparsity

CETT-PPL-1%: A better metric for activation sparsity.

- CETT: Apply **the same relative output error to each layer** after weakly-activated neurons are pruned (each layer can have different sparsity ratios and activation thresholds)
- **CETT can always achieve better trade-off between performance and sparsity** than FAT-ϵ (i.e., the same threshold for each layer) and Top-k (i.e., the same sparsity ratio for each layer)
- CETT-PPL-1%: The final sparsity metric based on CETT, when **the validation perplexity (PPL) raises by just 1%** with weakly-activated neurons skipped in computation

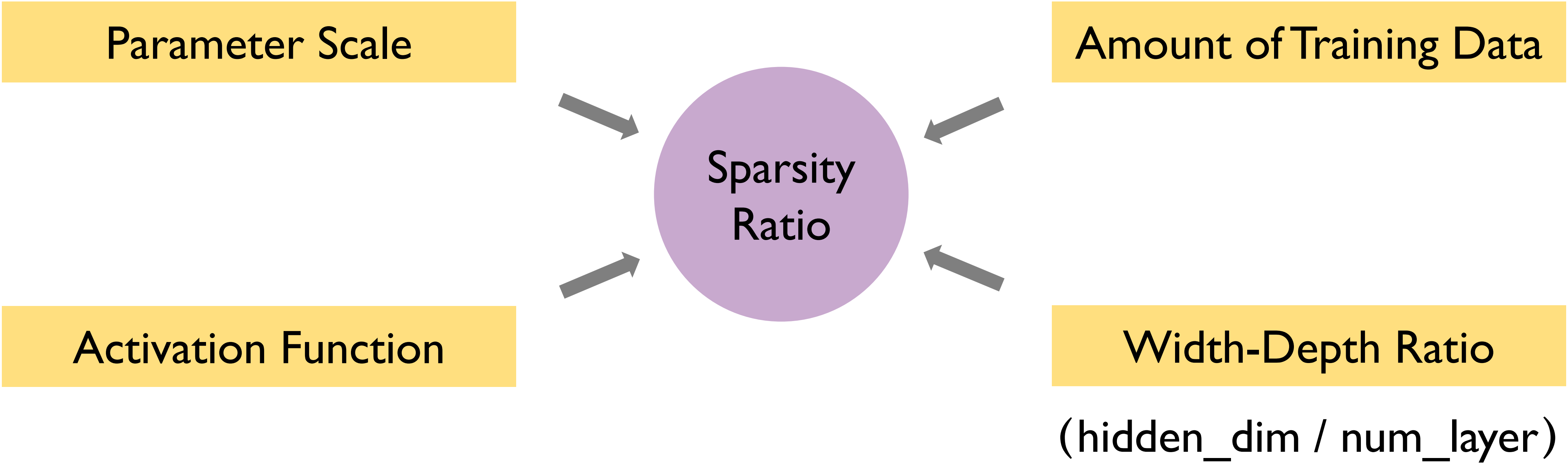


Zhang, Zhengyan, et al. "ReLU² Wins: Discovering Efficient Activation Functions for Sparse LLMs." arXiv preprint arXiv:2402.03804 (2024).

Influential Factors of Activation Sparsity

Sparsing Law: A comprehensive quantitative study on activation sparsity.

Q2: How is activation sparsity quantitatively affected by the model architecture and training process?



Parameter Scale

Amount of Training Data

Sparsity Ratio

Activation Function

Width-Depth Ratio

(hidden_dim / num_layer)

Influential Factors of Activation Sparsity

Activation function and the amount of training data.

- The activation ratio (CETT-PPL-1%) varies in different ways under different activation functions.
- ReLU: monotonously **decreasing** logspace power-law
- SiLU: monotonously **increasing** vanilla power-law

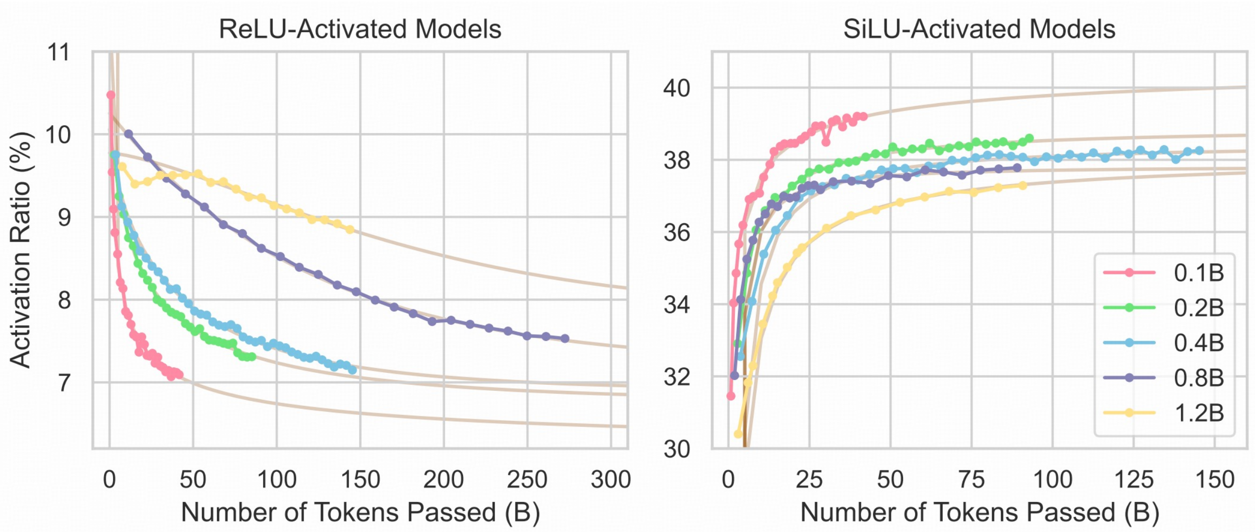
$$A_{ReLU}(D) = \exp(-cD^{\alpha} + b) + A_0$$

$$A_{SiLU}(D) = -\frac{c}{D^{\alpha}} + A_0$$

Limit Activation Ratio

- ReLU is more efficient than SiLU as a sparse activation function, because:
  - Significantly higher sparsity ratio
  - Comparable performance
  - **Sparsing trend**

(More data, higher sparsity)



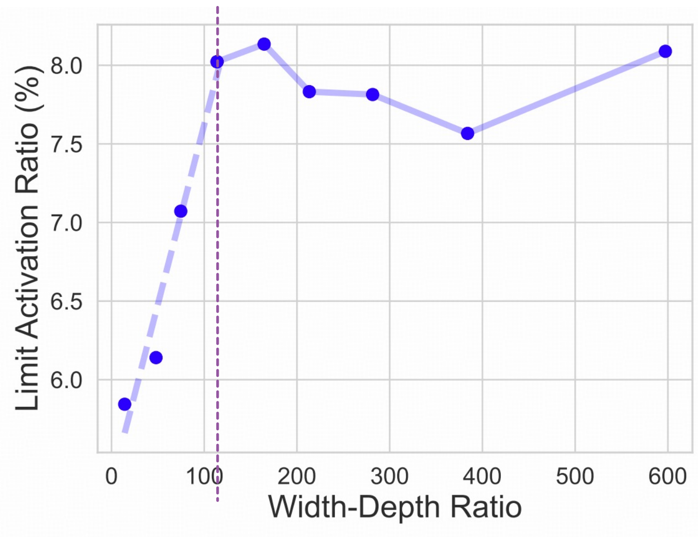
ReLU-Activated Models

SiLU-Activated Models

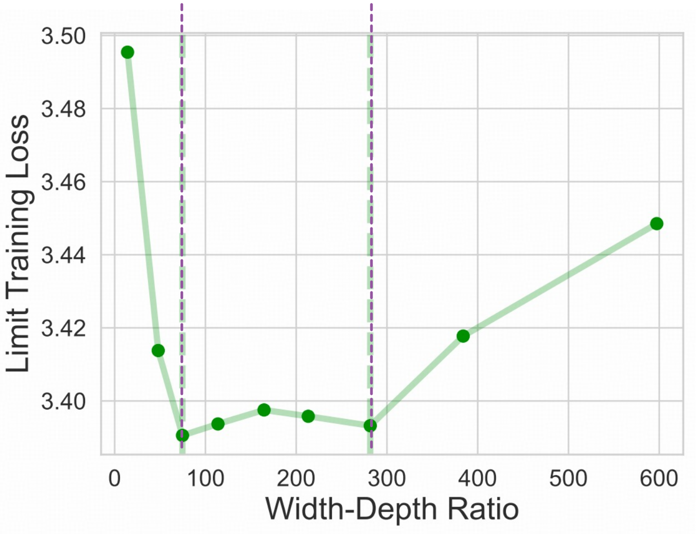
Influential Factors of Activation Sparsity

Width-depth ratio.

- Given the same parameter scale, **the activation ratio linearly increases with the width-depth ratio under a bottleneck** (i.e., deeper models are sparser)
- However, an extreme depth can cause training instability and harm performance, and the best performance exists within a "best interval"
- Thereby, the best width-depth ratio falls on the **left point of the best interval**



bottleneck

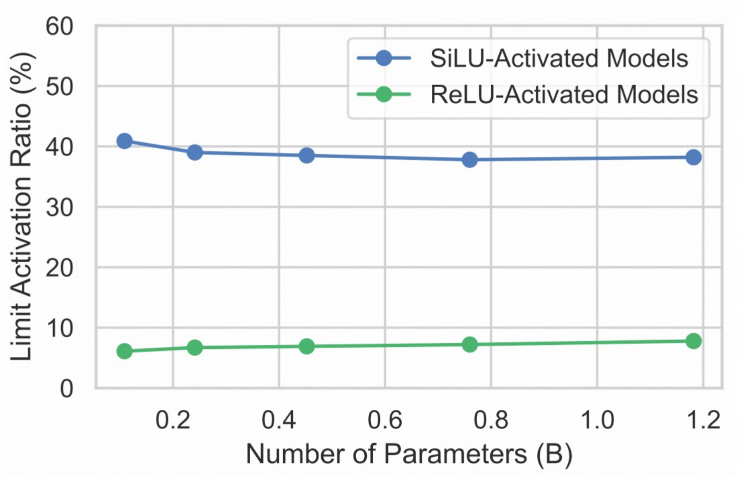


best interval

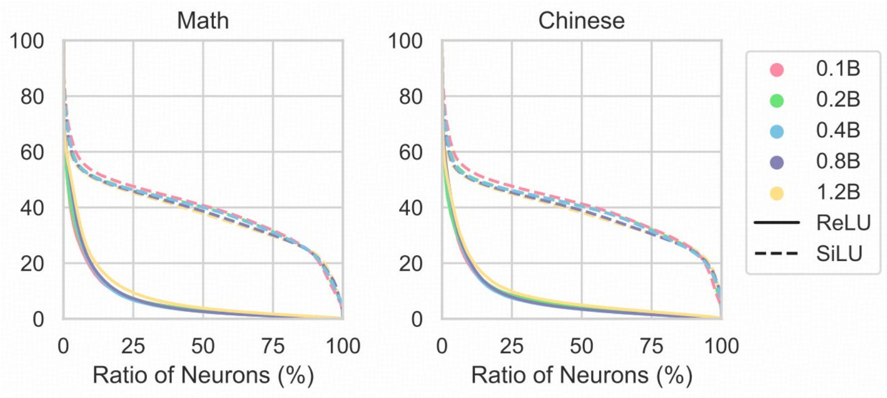
Influential Factors of Activation Sparsity

Parameter scale.

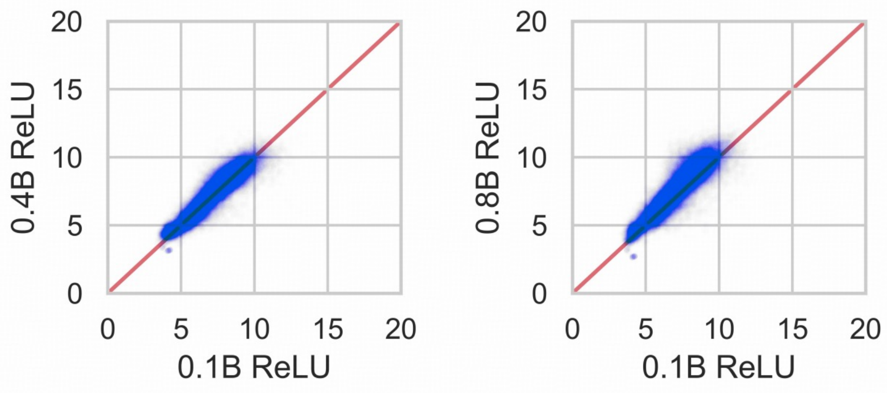
- Given similar width-depth ratios, the limit of activation sparsity is **weakly correlated** to the parameter scale of LLMs
- Some possible explanation: **neuron specialization** is also insensitive to the parameter scale



The limit activation ratio is **weakly correlated** to the parameter scale for both ReLU and SiLU.



On multiple datasets, the distribution patterns of neuron activation frequencies are similar across different scales.



Within 71k+ tokens, most tokens maintain a close activation ratio across models of various scales.

Approach towards Higher Activation Sparsity

Approach towards more sparsely activated LLM.

Q3: How can we build a more sparsely activated and efficient LLM?

- Takeaway: Use **ReLU** as the activation function with **a larger amount of pre-training data**, and **a small width-depth ratio** within the interval ensuring the training stability.
- Validation: 2.4B ReLU-activated LLM, 800B training data → **6.48% limit activation ratio, 4.1× speedup with PowerInfer**

Song, Yixin, et al. "Powerinfer: Fast large language model serving with a consumer-grade gpu." Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles. 2024.

ICML
International Conference
On Machine Learning

清華大學
Tsinghua University

Thank you for your attention!

Chenyang Song

Department of Computer Science and Technology, Tsinghua University

May 27th, 2025