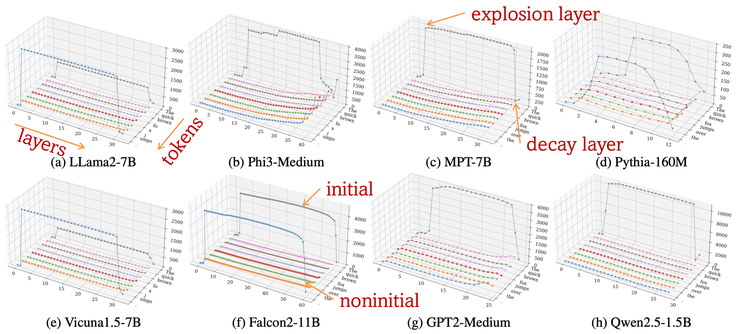
Demystifying Singular Defects in Large Language Models

Haoqi Wang, Tong Zhang, Mathieu Salzmann

TL;DR *Analyze high-norm tokens in LLMs using linear algebra and show interesting applications*

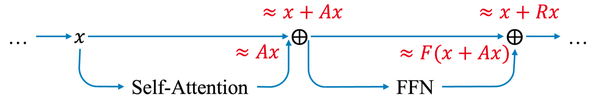
Singular Defects (High-Norm Tokens)



Singular defects (high-norm tokens) are

- A universal phenomenon
- Suddenly appear and disappear
- Any token at initial position and some delimiter tokens at noninitial positions
- The directions are the same across samples, layers, tokens
- It stabilizes during training and is robust to finetuning

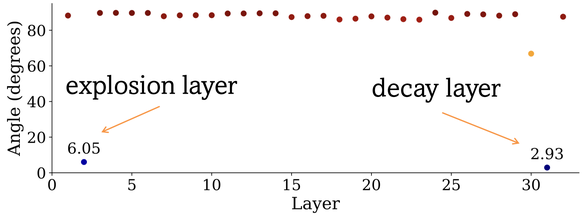
Linear Approximation of Layers



- Each transformer layer can be approximated by linear matrices
- For a single input token x
  - The output of self-attention module $\approx Ax$
  - The FFN is approximated by matrix F using least-squares
  - A transformer layer is approximated by the matrix $L \approx I+R$

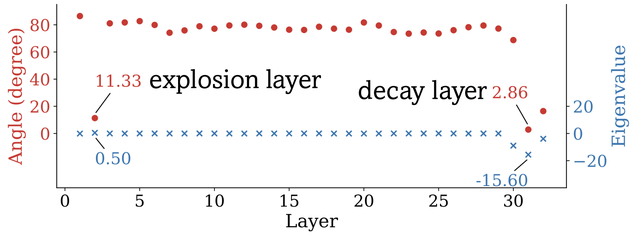
Predict the High-Norm Direction

Layer-wise singular direction: leading left singular vector of L.
Angle between predicted layer-wise singular direction and gt:



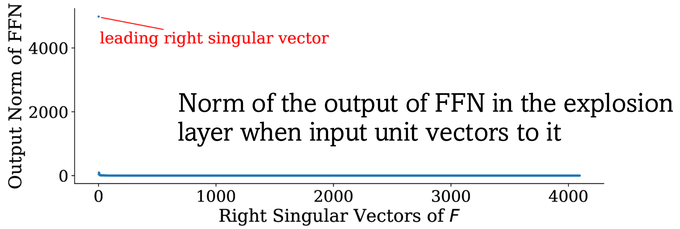
Describe the Decay

At the decay layer, $(I+R)x \approx 0$, we have $Rx \approx -x$.
The eigenvalue corresponding to the high-norm direction < 0



Explosion Subspace

On the explosion layer, only one dimension is responsible for the creation of high-norm. The explosion subspace is spanned by the leading right singular vector of F.



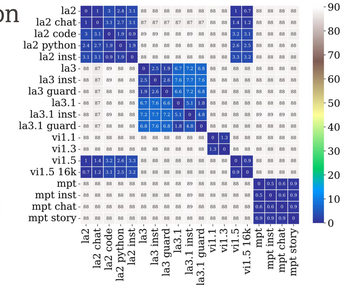
Application: Improve Quantization

- We observe that the explosion/decay layers create outlier activations and harm the quantization procedure
- By keeping the down projection layer in explosion and decay layers in fp16 precision, we can improve the tensor-wise W8A8 quantization

Model	Method	Skip F_2 in Layers	PPL↓
LLaMA2-7B	-	-	5.47
	RTN	-	10.18
	RTN	(2, 31)	6.51
	SmoothQuant	-	13.87
	SmoothQuant	(2, 31)	6.78
LLaMA3-8B	-	-	6.14
	RTN	-	59.38
	RTN	(2, 32)	8.80
	SmoothQuant	-	54.99
	SmoothQuant	(2, 32)	9.14

Application: LLM Signature

- We define high-norm direction as the model signature
- Define the distance of two models as the angle between their signatures.
- A small distance means that one model is fine-tuned from another. It can be used to trace model lineage.



Takeaway

- High-norm phenomenon can be understood using tools in linear algebra
- The properties of singular defects lead to practical applications in quantization and model lineage