The Low-Dimensional Structure of Outliers in LLMs

2026 하계 세미나 – 26.07.03

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
최재민

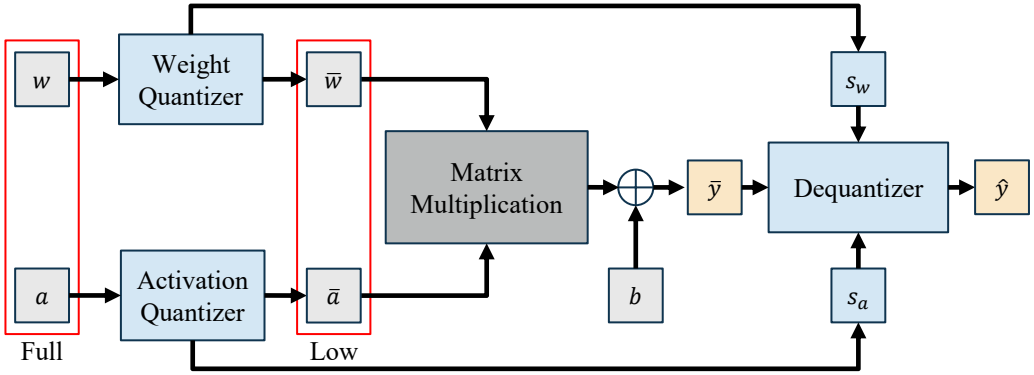
Outline

- Background
  - Quantization
  - Quantization difficulty by Activation Outliers

- Papers
  - Demystifying Singular Defects in Large Language Models [ICML 2025]
  - OffQ: Taming Structured Outliers in LLM Quantization by Offsetting [arXiv 2026]

Background

- Quantization
  - 일반적으로 performance와 model size는 비례하는 경향이 있음
    - Model size ↑ → inference time ↑, computational cost ↑
    - 메모리 용량이 제한적인 edge device 환경에서 한계가 존재함
  - Full-precision → Low-precision
    - Weight, activation을 lower precision으로 낮춘 후 연산을 수행하는 가속화 기법

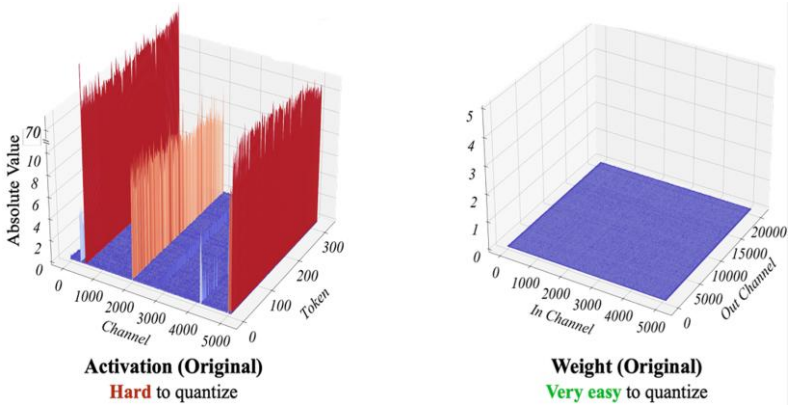


< Fig 1. Quantization의 가속화 원리 >

- Quantization process : $\bar{x} = Q(x) = clamp\left(\left\lfloor\frac{x}{s}\right\rceil + z, 0, 2^{bit} - 1\right), s = \frac{\max(x) - \min(x)}{2^{bit} - 1}, z = \left\lfloor -\frac{\min(x)}{s}\right\rceil$
- Dequantization process : $\hat{x} = s \cdot \bar{x}$
- Fully-Connected layer에서의 연산 : $f = \underline{WX} + B$ → FP32 Matmul
- Quantized FC layer에서의 연산 : $\hat{f} = \hat{W}\hat{X} + B = (s_W\bar{W})(s_X\bar{X}) + B = s_W s_X(\underline{\bar{W}\bar{X}}) + B$ → INT8 Matmul

Background

- Quantization difficulty



< Fig 2. Quantization difficulty[1)] – Weight/Activation distribution >

▪ Weight distribution

- 일반적으로 균일한 분포 (uniform, gaussian distribution)
- Min-max range가 작기 때문에 quantization error가 작음

▪ Activation distribution

- Transformer-based model의 경우 attention 연산에 의해 비균일한 분포 (power-law distribution)
- Min-max range가 크기 때문에 quantization error가 큼

Demystifying Singular Defects in Large Language Models

[ICML 2025]

1) Wang, Haoqi, Tong Zhang, and Mathieu Salzmann. "Demystifying Singular Defects in Large Language Models." *arXiv preprint arXiv:2502.07004* (2025).

Singular Defects[1)]

- Problem statements
  - Large transformer model에서 high-norm token이 발생한다는 관찰이 존재
    - Vision Transformer (ViT)와 Large Language Model (LLM)에서 모두 관찰
  - Quantization, pruning 등 경량화 기법에서 이러한 high-norm token은 성능 저하의 주 원인
  - High-norm token의 패턴과 근원을 밝혀 실용적 연구에 이론적 기반을 제공하는 것이 목적

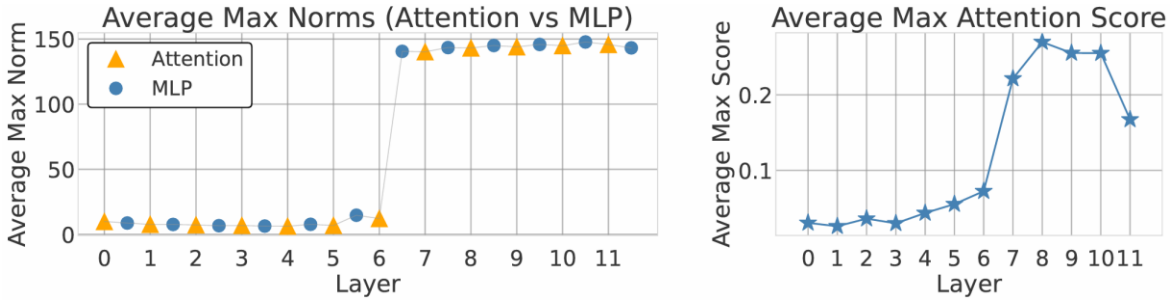
- Key contributions
  - LLM에서 high-norm token이 정렬되는 방향을 예측하는 이론적 토대(Singular Defects) 제안
  - Singular defects에 대한 이해를 바탕으로 두 가지 실용적 응용을 제안
    - 1) High-Norm Aware Quantization
    - 2) LLM Signature via Singular Defects

1) Wang, Haoqi, Tong Zhang, and Mathieu Salzmann. "Demystifying Singular Defects in Large Language Models." *arXiv preprint arXiv:2502.07004* (2025).
2) Jiang, Nicholas, et al. "Vision Transformers Don't Need Trained Registers." *Advances in neural information processing systems* 38 (2026): 56557-56595.

Singular Defects[1)]

- High-norm tokens in ViT[2)]
  - Observations
    - 1) 이미지의 객체가 아닌 배경에 high-norm token이 발생하며, 이미지마다 발생 위치가 다름
    - 2) 중간 레이어에서 갑작스럽게 high-norm token이 발생하며, 마지막 레이어까지 계속 존재

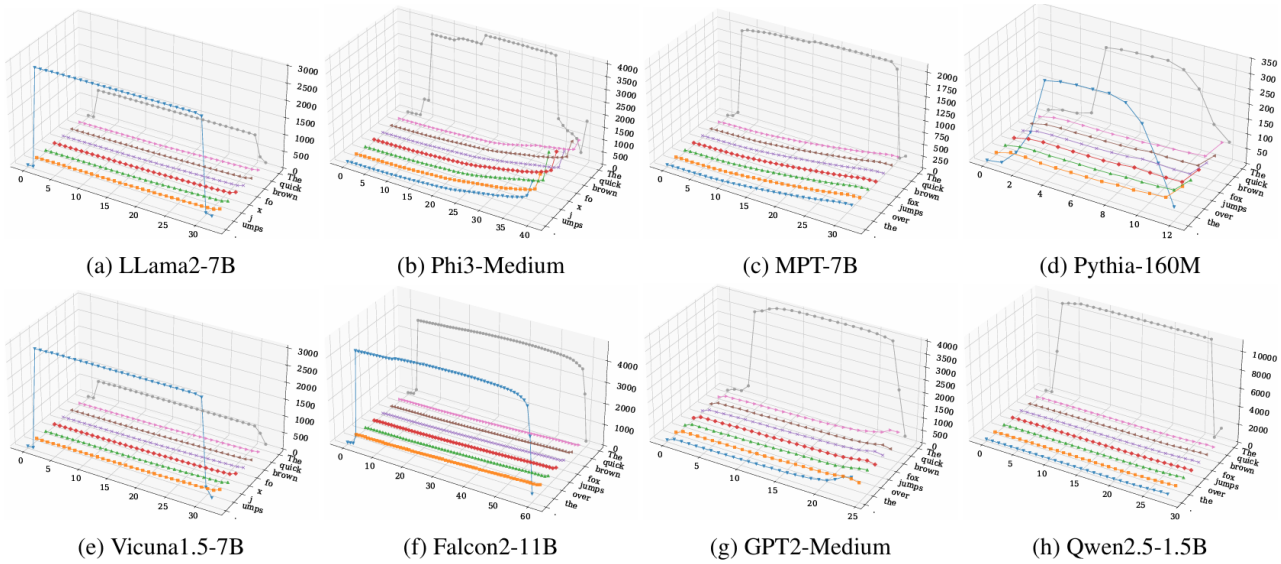
< Fig 3. High-norm token in OpenClip >



< Fig 4. ViT에서의 high-norm token 발생 양상 >

Singular Defects[1)]

- High-norm tokens in LLM
  - Observations
    - 1) 입력 시퀀스의 첫 토큰 또는 마지막 토큰에서 high-norm token이 발생
    - 2) 중간 레이어에서 갑작스럽게 high-norm token이 발생하며, 마지막 레이어에서 갑작스럽게 소멸



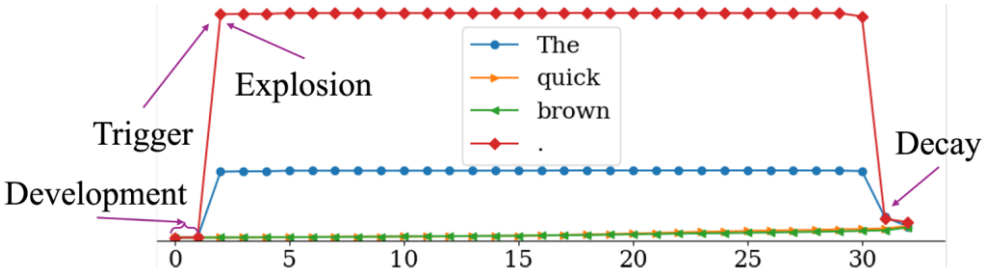
(a) LLama2-7B (b) Phi3-Medium (c) MPT-7B (d) Pythia-160M

(e) Vicuna1.5-7B (f) Falcon2-11B (g) GPT2-Medium (h) Qwen2.5-1.5B

< Fig 5. LLM에서의 high-norm token 발생 양상 >

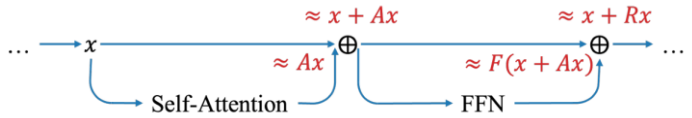
Singular Defects[1)]

- Analysis of high-norm tokens in LLM (The Theory of Singular Defects)
  - High-norm token에 대하여 4가지 관점에서 분석
  - 1. Development
    - Norm의 증가로 이어지는 계산 경로
  - 2. Trigger
    - Explosion layer 직전의 norm 증가의 원인
  - 3. Explosion
    - 중간 레이어들에서 high-norm token의 출현과 네트워크 파라미터와의 상관관계
  - 4. Decay
    - High-norm token의 소멸



< Fig 6. LLM에서의 high-norm token 발생과 소멸 단계 >

Singular Defects[1)]



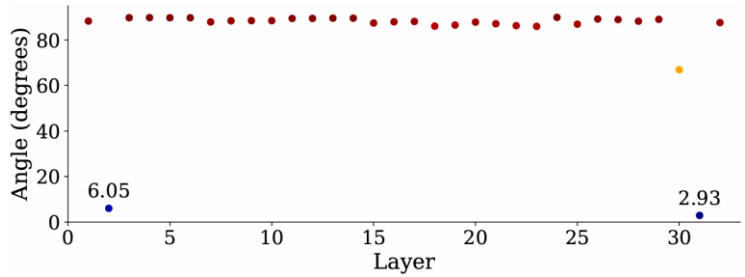
- Explosion: Appearance of High-Norm Tokens
  - 실험 1. High-norm token의 평균의 방향 (Empirical high-norm direction)
    - Norm이 500을 넘는 token들에 대해서 각도를 측정한 결과 평균 3.1°로 측정
      - ⁘ High-norm token의 평균을 **empirical high-norm direction**이라 정의
    - 즉, High-norm token은 대부분 특정 방향에 쏠려 있다.
  - 실험 2. Layer-wise approximate matrix의 방향 (Layer-wise singular defect direction)
    - Transformer layer를 하나의 선형 연산으로 근사한다면, 아래와 같은 수식 전개가 가능

$$Attention(x) \approx Ax, \qquad FFN(x) \approx Fx$$
$$Layer(x) \approx Lx := FFN\big((Attention(x) + x)\big) + x = x + (A + F + FA)x = (I + R)x$$

    - $L = (I + R) = U \sum V^T$와 같이 SVD로 얻은 방향 u_1과 empirical high-norm direction의 각도를 측정
      - ⁘ Layer 2, 31에서의 각도는 6.05°, 2.93°로, 나머지의 경우 대부분 약 90°로 측정
    - 즉, 두 방향은 어떤 연관성이 존재한다.

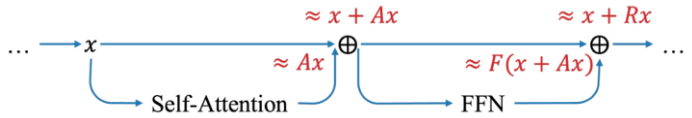
Model	Threshold	Mean Pairwise Angle (degree)
LLaMA2-7B	500	3.12
Phi3-Medium	2500	5.31
MPT-7B	1500	2.81
Pythia-160M	200	8.01
Vicuna1.5-7B	400	4.15
Falcon2-11B	3000	4.11
GPT2-Medium	3000	1.49
Qwen2.5-1.5B	8000	1.00

< Table 1. L과 high-norm token 사이의 예각 >

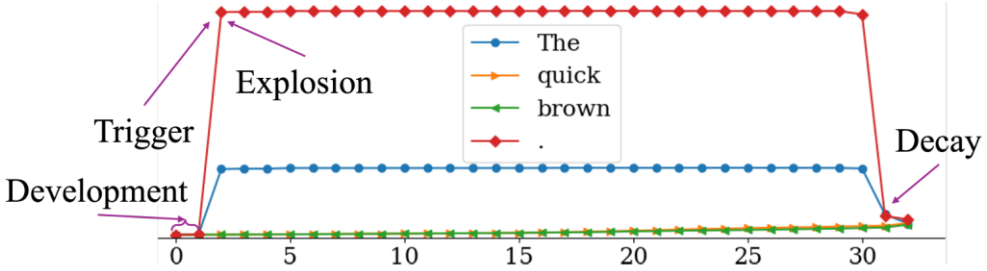


< Fig 7. L과 high-norm token 사이의 예각 (LLaMA2-7B) >

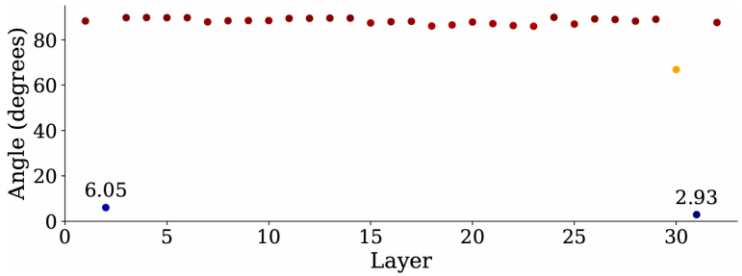
Singular Defects[1)]



- Explosion: Appearance of High-Norm Tokens
  - 두 실험의 결론
    - $Layer(x) \approx Lx := (I + R)x$
      - Ix는 identity path를 의미, Rx은 transformer layer로 인해 발생한 변화를 의미
    - 1) Layer 2
      - L의 대표 방향이 high-norm token을 발생시킨다. $Lh \approx \sigma_{max} h, \quad \sigma_{max} \gg 1$
    - 2) Layer 3~30
      - L의 대표 방향이 high-norm token과 직교적이다. $Lh \approx h$
      - Identity path로 인해서 high-norm token이 변화 없이 유지된다.
    - 3) Layer 31
      - High-norm token을 소멸시킨다.



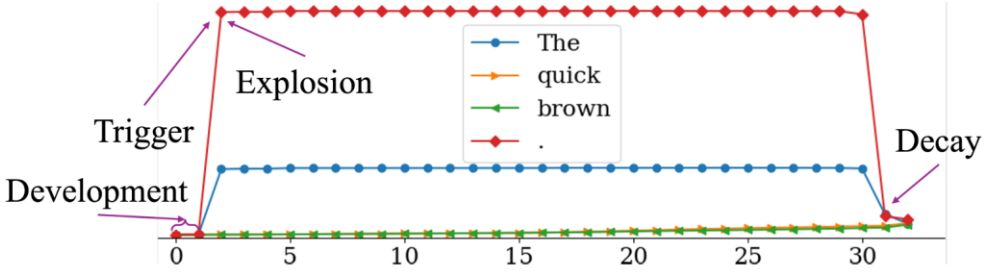
< Fig 6. LLM에서의 high-norm token 발생과 소멸 단계 >



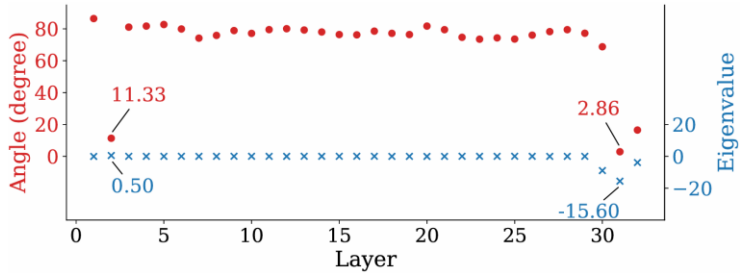
< Fig 7. L과 high-norm token 사이의 예각 (LLaMA2-7B) >

Singular Defects[1)]

- Decay: Eigenvalues of Decay Layers
  - 직관적 추측
    - Decay가 발생하는 layer에서는 $Lx = (I + R)x \approx 0$이므로, $Rx \approx -x$로 근사 가능
    - Transformer layer에 의한 변화를 발생시키는 R의 eigenvalue가 음수일 것이라는 추측으로 연결
  - 실험 1. R의 eigenvector와 empirical high-norm direction 사이의 관계
    - R의 eigenvector와 empirical high-norm direction 사이의 각도 측정
      - Layer 2, 31의 경우 각각 11.33°, 2.86°로, 나머지의 경우 대부분 70° 이상으로 측정
      - 즉, 앞선 explosion에서의 실험과 동일
    - R의 eigenvalue 측정
      - Layer 2의 경우 양수, **Layer 31의 경우 음수**
    - 즉, Layer 31은 high-norm token을 소멸시킨다.

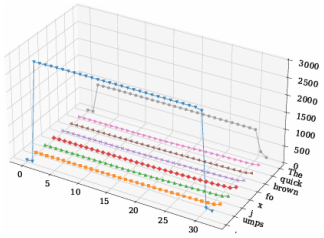


< Fig 6. LLM에서의 high-norm token 발생과 소멸 단계 >



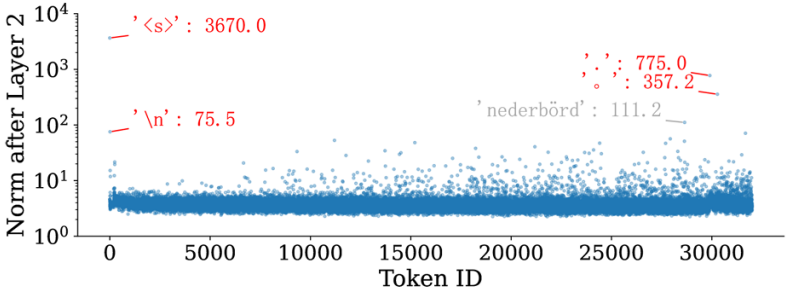
< Fig 8. R과 high-norm token 사이의 예각, R의 eigenvalue >

Singular Defects[1)]

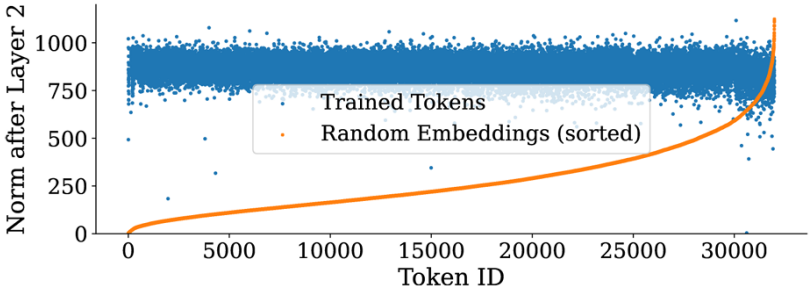


(a) LLama2-7B

- Development: The Two Types of Exploding Paths
  - 직관적 추측
    - Non-initial high-norm token (ex. '.')의 출현이 그 앞의 내용과 무관하게 high-norm을 가짐
    - 즉, 토큰 간 상호작용이 일어나는 유일한 장소인 self-attention layer가 non-initial high-norm token에 큰 영향을 주지 못한다는 추측으로 연결
  - 실험 1. Attention-independent exploding path (Non-initial high-norm token)
    - Non-initial high-norm token이 self-attention과 무관하다면, 이를 제거한 후 입력해도 high-norm일 것
    - 실험 결과, '<s>', '.'과 같은 non-initial high-norm token을 제외한 나머지 token들은 norm이 작게 측정
  - 실험 2. Attention-relative exploding path (Initial high-norm token)
    - Self-attention을 제거하지 않은 상태로 tokenizer에서 사용가능한 모든 token을 initial token으로 입력
    - 실험 결과, 모든 initial token이 high-norm token이 된다는 것을 확인



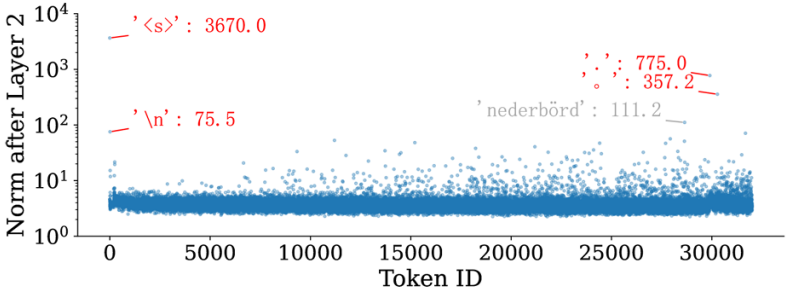
< Fig 9. Self-attention을 제거한 후 입력 (실험 1) >



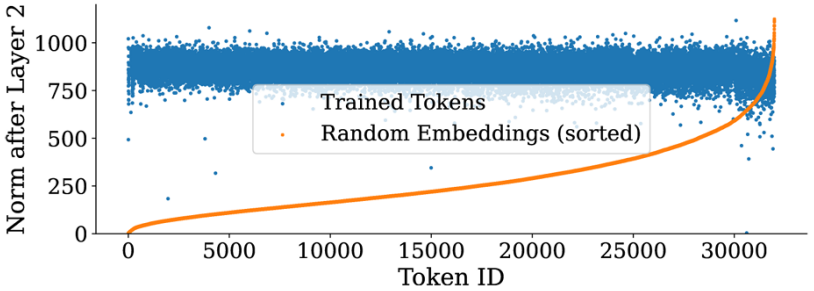
< Fig 10. Self-attention을 연결한 후 입력 (실험 2) >

Singular Defects[1)]

- Development: The Two Types of Exploding Paths
  - 두 실험의 결론
    - 1) Initial high-norm token
      - 어떤 token이든, initial token이 되면 high-norm을 갖는다.
      - Initial high-norm token의 발생은 self-attention에 의해 발생한다.
        - ✓즉, token들 간의 상호작용에 의해 발생한다.
    - 2) Non-initial high-norm token
      - Non-initial high-norm token은 주로 '<s>', '\n', '.'와 같은 구분자로 구성된다.
      - Non-initial high-norm token의 발생은 self-attention과 무관하다.



< Fig 9. Self-attention을 제거한 후 입력 (실험 1) >



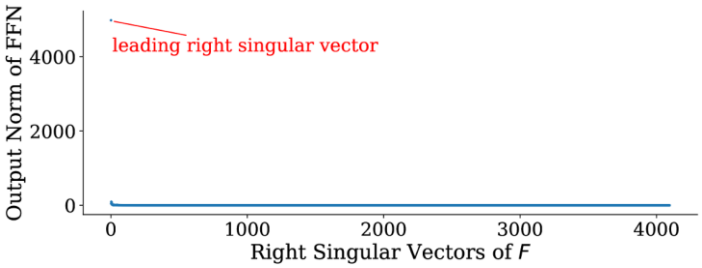
< Fig 10. Self-attention을 연결한 후 입력 (실험 2) >

Singular Defects[1)]

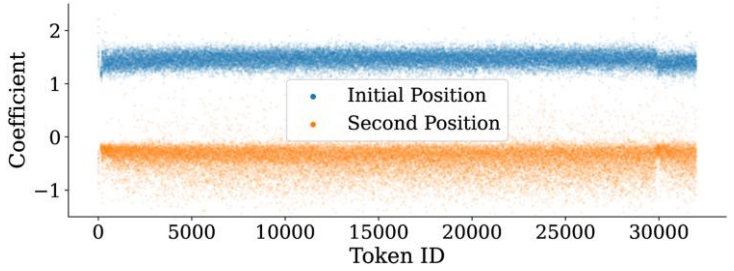
- Trigger: The Explosion Subspace
  - 실험적 관찰
    - Norm의 변화를 추적한 결과, FFN에서 norm의 갑작스러운 증가가 발생함을 관찰
  - 실험 1. Leading right singular vector와 norm의 관계

$$FFN(x) \approx Fx$$

    - 위 선형 근사식에 대하여, $F = U \sum V^T$와 같이 SVD를 적용
    - SVD의 leading right singular vector에서 유일하게 norm이 폭발하는 것을 관찰
  - 실험 2. Token의 위치에 따른 explosion subspace component
    - x가 leading right singular vector의 1차원 부분 공간으로 표현될 수 있는 component를 측정
    - Initial token의 경우 component가 보통 1.5 이상의 큰 값
    - Second token의 경우 component가 보통 0 이하의 작은 값



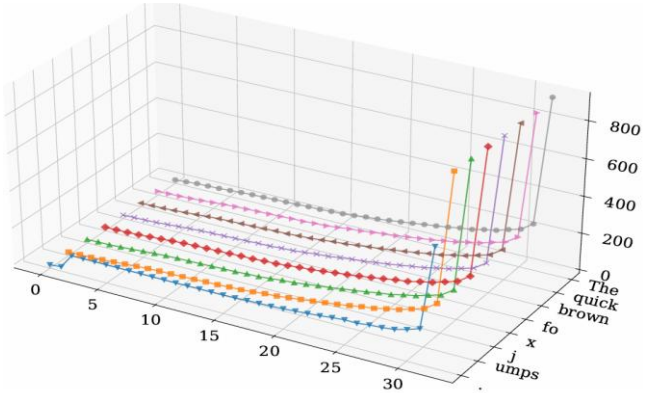
< Fig 11. F의 SVD에 대한 right singular vector >



< Fig 12. Token 위치에 따른 component 분포>

Singular Defects[1)]

- Trigger: The Explosion Subspace
  - 실험 3. Explosion layer의 직전 layer에 대해 FFN 제거
    - Layer 1에서 FFN의 explosion subspace를 제거한 후 norm 측정
    - 실험 결과, 마지막 layer를 제외한 나머지 layer에 대해 high-norm이 모두 소멸함
    - 즉, Layer 1의 FFN이 high-norm의 trigger 역할을 한다.



< Fig 13. Layer 1의 FFN을 제거했을 때의 norm 측정 >

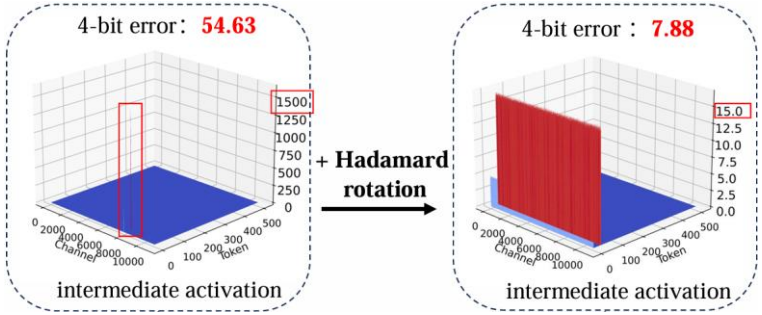
  - 실험적 결론
    - 1) High-norm token은 FFN의 leading right singular vector에 의해 발생한다.
    - 2) High-norm token의 trigger는 explosion layer의 직전 layer의 FFN이다.

OffQ: Taming Structured Outliers in LLM Quantization by Offsetting [arXiv 2026]

1) Wang, Haoqi, et al. "OffQ: Taming Structured Outliers in LLM Quantization by Offsetting." *arXiv preprint arXiv:2606.07116* (2026).
2) Chen, Mengzhao, et al. "Prefixquant: Eliminating outliers by prefixed tokens for large language models quantization." *arXiv preprint arXiv:2410.05265* (2024).

OffQ[1)]

- Problem statements
  - LLM quantization에서 activation outlier에 의한 quantization difficulty가 존재
    - Smoothing : SmoothQuant, AWQ, Outlier Suppression, OmniQuant, …
      - Activation에 scaling을 적용하여 weight로 quantization difficulty를 전이
    - Rotation : QuaRot, SpinQuant, ParoQuant, …
      - Activation outlier가 발생하는 채널을 flatten하여 일부 완화
  - Smoothing, rotation 등의 방법으로도 outlier를 완전히 제거할 수 없는 문제가 존재

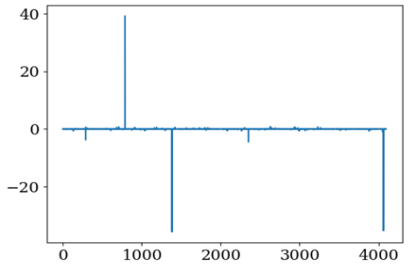


< Fig 14. Rotation 적용 후의 잔여 outlier[2)] >

- Key contributions
  - Activation outlier가 low-dimensional subspace의 특성을 갖는다는 점을 밝힘
  - Activation outlier를 특정 채널에 집중시키도록 한 후, quantization parameter에 흡수시킴
    - Zero point에 흡수시킴으로써 outlier에 의한 error를 완화

OffQ[1)]

- Methods
  - Low-dimensional structure
    - Singular defects의 분석에 따라, outlier는 1차원 부분공간에 의해 지배됨
    - 이는 곧, outlier의 부분공간의 본질적 차원은 전체 채널이 아닌 특정 채널에 의해 지배됨을 의미함



(a) Token $\boldsymbol{X} \in \mathbb{R}^{1\times 4096}$

< Fig 15. High-norm token의 channel-wise distribution >

  - Top-1 PCA
    - Outlier의 부분공간을 추출하기 위해, norm이 가장 큰 top-1 token을 추출하여 샘플 구성
      - $X \in \mathbb{R}^{N\times D}$, N: 샘플 개수, D: 차원 수
    - Covariance matric C에 대하여 eigenvector decomposition (EVD)를 적용

$$C = \frac{X^T X}{N} = U\Lambda U^T$$

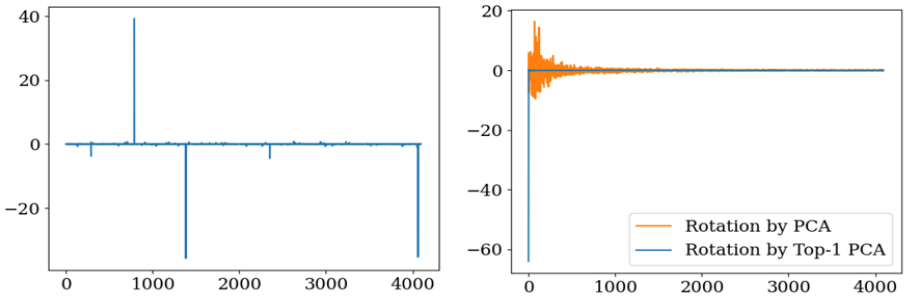
    - Leading eigenvector (U의 첫번째 행)은 outlier의 최대 분산을 나타내는 방향을 내포

OffQ[1)]

- Methods
  - Concentrate Outliers
    - Eigenvector matrix U는 outlier의 분산을 나타내는 방향을 포함
    - 따라서 아래 수식을 적용하면 $\boldsymbol{X}$를 outlier의 분산을 나타내는 공간에 투영하게 됨

$$\boldsymbol{X}_{rot} = \boldsymbol{XU}$$

    - Leading eigenvector에 outlier의 최대 분산 방향이 내포되어 있으므로, outlier가 첫번째 채널에 전이됨



(a) Token $\boldsymbol{X} \in \mathbb{R}^{1\times 4096}$ (b) Concentrated $\boldsymbol{X}_{\text{rot}}$

< Fig 16. Top-1 PCA 적용 시 outlier channel의 concentration 효과 >

OffQ[1)]

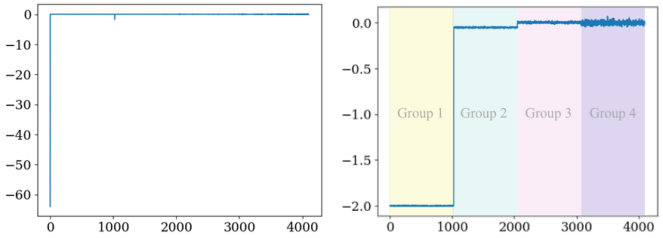
- Methods
  - Absorbing Outliers by Offsetting
    - Hadamard matrix H는 일종의 rotation matrix이며, $H_{1,:} = [1,1,\dots,1]$로 정의됨
    - 따라서, 아래 수식을 적용하면 첫번째 채널에 집중되어 있던 outlier가 모든 채널에 균등 분배됨

$$\boldsymbol{X}_{had} = \frac{\boldsymbol{X}_{rot}\boldsymbol{H}}{\sqrt{D}}$$

    - Ex) $\boldsymbol{H} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \boldsymbol{X}_{rot} = [x_1, x_2]$

$$\boldsymbol{X}_{had} = \frac{\boldsymbol{X}_{rot}\boldsymbol{H}}{\sqrt{2}} = \left[\frac{x_1+x_2}{\sqrt{2}}, \frac{x_1-x_2}{\sqrt{2}}\right] = \underbrace{\frac{x_1}{\sqrt{2}}}_{\text{Offset}} + \underbrace{\left[\frac{x_2}{\sqrt{2}}, -\frac{x_2}{\sqrt{2}}\right]}_{\text{without outlier } x_1}$$

  - Group-wise Offsetting
    - Offset의 경우 하나의 채널에만 집중시킬 수 있음
    - 따라서, 여러 채널을 group으로 묶어 per-group offsetting & quantization 수행

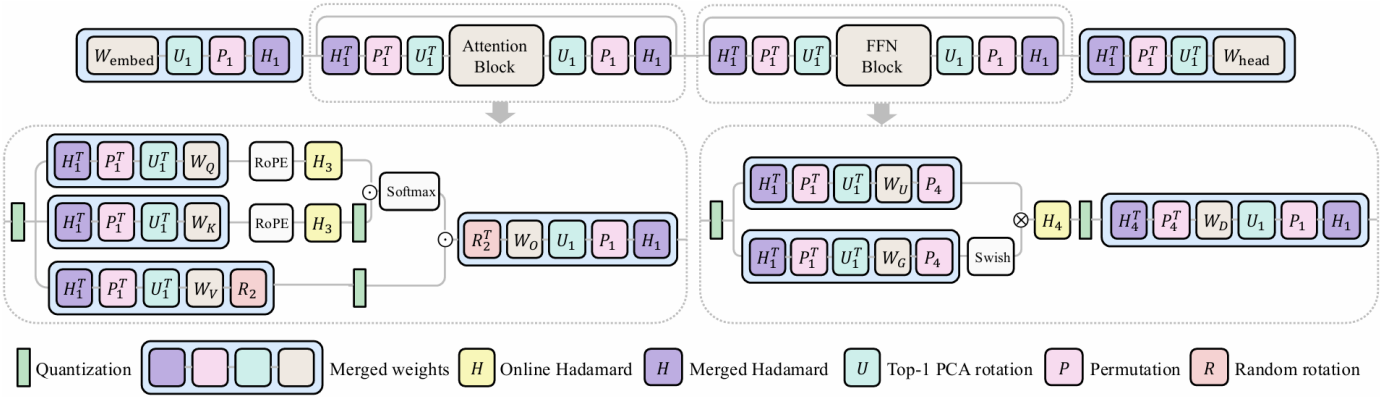


(c) Permute to 4 groups (d) Group-wise offset $\boldsymbol{X}_{\text{had}}$

< Fig 17. Group-wise offsetting 적용 시 offset 분포 >

OffQ[1)]

- Methods
  - Quantizing an LLM with Activation Offsetting



< Fig 18. OffQ 프레임워크 >

- 1) Activation의 통계값 수집
  - Top-1 token 수집, covariance matrix C 계산, top-1 PCA 계산
- 2) Rotation, Grouping, Offsetting 적용
  - Top-1 PCA Rotation, Grouping, Hadamard Rotation, Offsetting
- 3) Weight quantization 적용
  - GPTQ 방식으로 weight quantization 수행

OffQ[1)]

- Experiments
  - Quantization results

	Method	3-8B		3-70B		3.2-1B		3.2-3B		2-7B		2-13B	
		PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]
Llama	16-bit	6.1	67.09	2.9	73.09	9.8	54.86	7.8	62.73	5.5	64.15	4.9	66.45
	GPTQ	166.3	39.79	11655	34.90	108.9	37.98	178.3	40.34	9600	38.89	3120	35.83
	QUIK	14.2	51.60	8.0	58.15	21.8	44.30	15.8	48.74	7.5	56.99	6.8	60.21
	QuaRot	7.8	62.10	5.7	67.56	14.3	49.01	10.1	56.06	6.1	60.75	5.4	63.80
	SpinQuant	7.4	63.76	6.2	65.68	13.6	48.78	9.2	57.89	6.0	60.98	5.2	64.81
	DFRot	7.91	62.35	5.03	68.98	–	–	–	–	6.25	60.97	5.43	63.83
	KurTail	7.2	64.63	4.2	70.69	12.9	50.11	9.0	59.04	5.9	61.31	5.2	65.18
	OSTQuant	7.29	64.70	4.01	**71.16**	–	–	–	–	5.91	**62.11**	5.25	64.19
	ResQ	7.1	63.91	4.1	71.14	12.4	50.11	8.8	58.99	5.8	61.95	**5.1**	**65.25**
	OffQ	**6.98**	**65.49**	**3.88**	70.63	**12.32**	**50.91**	**8.78**	**60.80**	**5.77**	61.99	5.11	**65.25**

	Method	1.5B		3B		7B		14B		32B		72B	
		PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]	PPL	0-shot[8]
Qwen 2.5	16-bit	9.3	60.85	8.0	63.81	6.8	68.45	5.3	70.61	5.0	70.44	3.9	73.41
	GPTQ	25770	35.21	9978	35.10	13594	34.85	5100	36.93	3891	38.53	37967	34.54
	QUIK	6614	35.81	15.5	51.19	260.3	41.48	10.5	57.66	9.6	59.08	8.3	61.90
	QuaRot	6600	38.33	68.8	47.76	4036	38.36	6.8	67.14	6.1	67.90	4.9	70.28
	ResQ	12.5	55.26	9.0	61.13	8.2	65.29	6.2	69.16	5.6	69.55	4.6	71.98
	OffQ	**11.35**	**57.53**	**8.98**	**61.47**	**7.66**	**66.16**	**6.07**	**69.20**	**5.52**	**69.59**	**4.29**	**72.68**

< Table 2. Quantization 적용 시 실험 결과 >

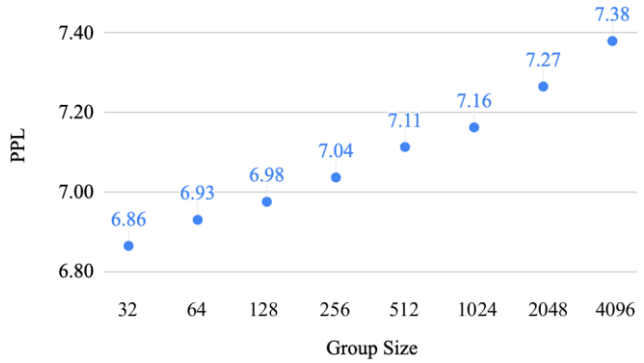
OffQ[1)]

- Experiments
  - Ablation study 1
    - 1) Top-1 PCA가 아닌 base PCA 적용
    - 2) Grouping할 때, 정상 채널의 sorting 순서를 zigzag로 적용
    - 3) Random rotation을 일부만 적용

Ablation	PPL	0-shot								
	Wiki	ARC-c	ARC-e	BoolQ	HellaS	OBQA	PIQA	SIQA	WinoG	**Avg.**
OffQ	6.98	50.68	77.44	80.43	76.96	43.80	78.89	45.96	69.77	65.49
Without Top-1 Selection	8.27	47.78	73.99	77.95	74.24	41.20	77.58	43.09	69.22	63.13
Zigzag Grouping	7.03	49.40	75.38	79.72	76.30	43.20	79.00	45.29	71.27	64.95
Partial Random Rotation	7.00	47.53	74.33	79.33	76.40	42.00	79.60	45.80	70.24	64.40

< Table 3. OffQ 방법의 여러 변형 실험 >

  - Ablation study 2
    - Group size를 변화시켜가며 PPL 측정, group size가 작을수록 성능이 좋음



< Fig 19. OffQ의 group-size에 따른 성능 변화 >

Summary

- Singular Defects
  - Conclusion
    - High-norm token의 발생과 소멸은 4단계로 나뉜다.
    - Development
      - Initial high-norm token은 self-attention에 의해 발생한다.
      - Non-initial high-norm token은 self-attention과 무관하게 발생한다.
    - Trigger
      - **Transformer layer의 FFN은 high-norm token의 trigger 역할을 한다.**
    - Explosion
      - High-norm token은 대부분 특정 방향에 쏠려 있다. **즉, low-dimensional subspace를 공유한다.**
    - Decay
      - Transformer layer residual matrix R의 eigenvalue가 음수일 때 high-norm token은 소멸한다.
- OffQ
  - Conclusion
    - Outlier가 low-dimensional structure를 갖는다는 것을 활용하면 특정 채널에 outlier를 전이할 수 있다.
    - 특정 채널에 집중된 outlier는 Hadamard matrix로 다른 채널에 균일하게 분배할 수 있다.

감사합니다.