2025 Samsung AI Challenge

거대 모델의 성능 저하 없이 크기를 줄이는 방법

팀명
프로메테우스

팀원
김도현(한양대)
윤상민(연세대)
정연석(고려대)

Problem Definition

- Base 모델과 Expert Merging 혹은 Pruning된 모델의 MoE 모듈을 지난 후의 Representation이 비슷하게 만들면 성능은 유지하되 모델의 총 파라미터 수를 줄일 수 있을 것이라고 생각

- 기준이 되는 Base 모델의 Representation은 최대한 특정 도메인에 치우쳐져 있지 않은 사전학습 데이터셋을 이용하는 게 좋을 것이라고 판단하여 C4 데이터셋에서 데이터를 샘플링하여 활용

Objective Function

$$f_\theta(x) = \Sigma_{\mathrm{k} \in \text{top K}} G_{\mathrm{k}}(x) E_{\mathrm{k}}(x)$$

$$\min_{\theta'} \Sigma_{x \in D} ||f_\theta(x) - f_{\theta'}(x)||_2$$

D : 사전 학습 데이터셋

G_k : top K 안에서 정규화된 라우터 스코어

E_k : top K 안에 선택된 전문가

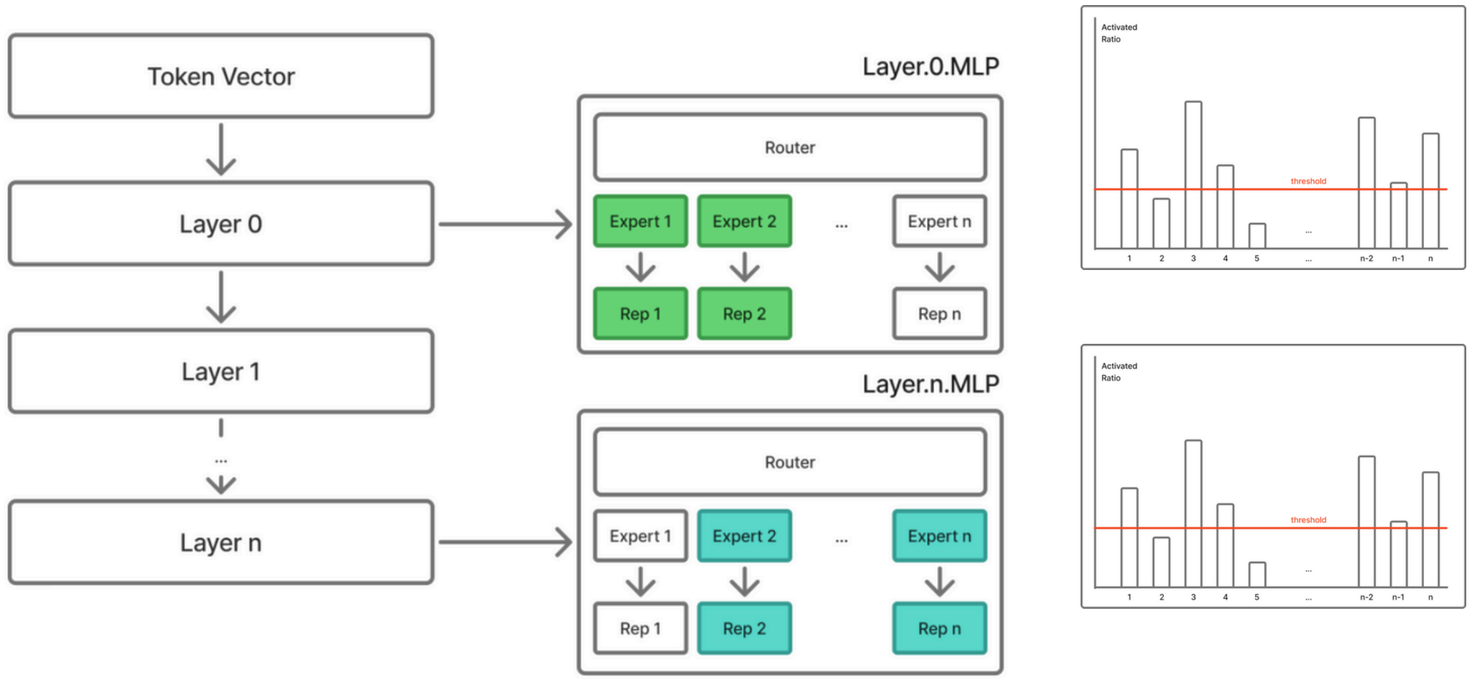
f_θ : Base MoE 모듈

$f_{\theta'}$: Expert Merging 혹은 Pruning 된 MoE 모듈

Method

Expert Pruning

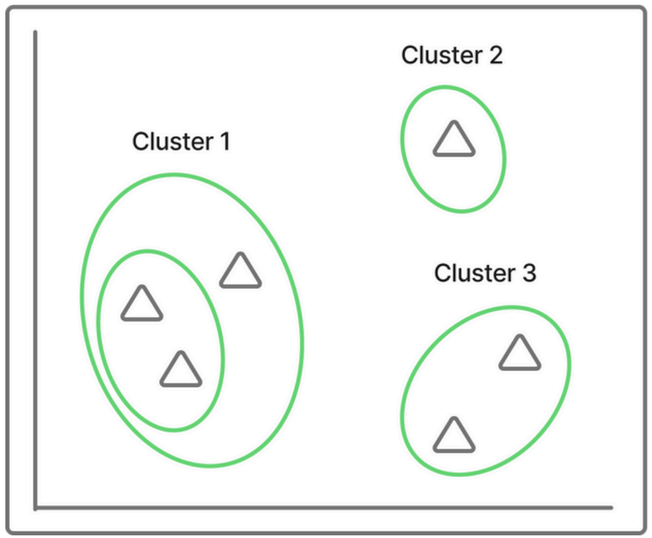
- C4 데이터셋으로부터 얻은 Router Logit을 활용하여 레이어별로 각 전문가의 활성화 비율을 구한 후, 특정 Threshold 보다 낮은 비율을 가지는 Expert를 제거



Expert Merging

Hierarchical Clustering

- Expert의 Representation을 Router Score로 가중합하여 각 Expert의 중심 Representation을 구한 후, Hierarchical Clustering 진행



Centroid of an Expert's Representations

$$h_i = \Sigma_{x \in D}\, G_i(x) E_i(x)$$

Cluster Example

$$C_1 = \{h_0, h_1, \cdots\} \quad C_2 = \{h_2, h_3, \cdots\}$$

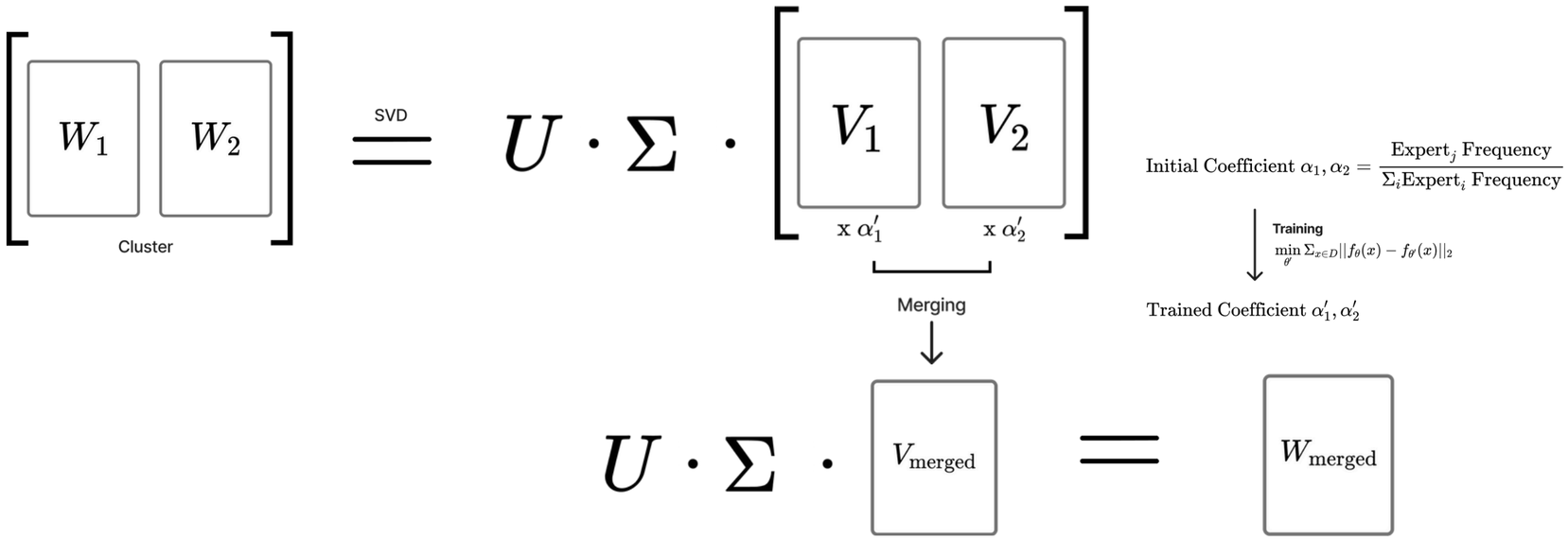
Distance between Clusters

$$d_{(i,j)} = \frac{1}{||C_i|| \cdot ||C_j||} \Sigma_{h_i \in C_i} \Sigma_{h_j \in C_j} ||h_i - h_j||_2$$

Expert Merging

SVD Merging with Trained Coefficients

- 클러스터 내의 Expert 가중치들을 SVD와 학습된 Coefficient로 가중합하여 병합
- 가중합되는 Expert의 Coefficient는 처음에 활성화되는 비율을 활용하여 초기화한 후, 병합된 모델의 Representation과 Base 모델의 Representation과 유사하도록 추가적인 학습 진행



Adjustment of Activated Experts

- 토큰별로 활성화되는 전문가 수를 줄임으로써 추론에서 사용되는 비용 감소
- 토큰별로 활성화되는 전문가 수를 줄임으로써 저하되는 성능을 특정 알고리즘을 통해 완화

Experiments & Results

Benchmark Dataset

English

- **MMLU** : 도메인 지식과 추론 능력을 모두 측정
- **Hellaswag** : 모델의 상식적 추론(commonsense reasoning)을 평가
- **Winogrande** : 상식 기반 대명사 해석(coreference resolution) 평가

Korean

- **CLIcK** : 한국어 대형 언어 모델들에 대한 문화적·언어적 벤치마크
- **KMMLU** : MMLU의 한국어 버전

Qwen/Qwen3-30B-A3B

*모든 벤치마크 평가는 zero-shot으로 설정

# of activated experts	8 (default)
MMLU (acc) - 4 options	0.7787
Hellaswag (acc) - 4 options	0.5952
Winogrande (acc) - 2 options	0.7024
CLIcK (acc) - 4 options	0.6301
KMMLU (acc) - 4 options	0.5772

Experiments

1. Threshold of Expert Pruning

Expert pruning 기준에 따른 성능 비교

2. Activated Experts per Token

Activated expert 수에 따른 성능 비교

3. Expert Pruning & SVD Merging

Expert Pruning과 Merging 사이의 성능 차이 비교

4. Mitigation of Reduced Performance

모델 경량화 후 줄어든 성능을 보완하기 위한 방법 제안

Threshold of Expert Pruning

- 8x2048 토큰에서 topk 안에 활성화된 Expert만 남겨두고 나머지는 제거하는 방식으로 성능 비교

top6에서 top4로 이동할 때, parameter 감소 대비 성능이 크게 하락하여 최종적으로 top6 pruning 활용

topk Pruning	top8	top6	top4	top2
# of experts	5462	5352	5159	4786
# of parameters	27.31B	26.79B	25.88B	24.12B
MMLU (acc)	0.7628	0.7608	0.7431	0.7316
Hellaswag (acc)	0.5950	0.5944	0.5938	0.5928
Winogrande (acc)	0.7048	0.7040	0.7111	0.7088
CLIcK (acc)	0.6276	0.6271	0.5900	0.5639
KMMLU (acc)	0.5671	0.5624	0.5365	0.5285

Activated Experts per Token

- Base 모델의 Activated Expert 수를 바꾸어가며 각 벤치마크의 성능 비교를 진행

모델 성능과 압축률을 종합적으로 고려하여 4로 선택

Activated experts	8	6	4	2
# of experts	6144	6144	6144	6144
# of parameters	30.53B	30.53B	30.53B	30.53B
MMLU (acc)	0.7787	0.7691	0.7072	0.2797
Hellaswag (acc)	0.5952	0.5915	0.5572	0.3602
Winogrande (acc)	0.7024	0.6638	0.6006	0.5099
CLIcK (acc)	0.6301	0.6090	0.5519	0.2887
KMMLU (acc)	0.5772	0.5712	0.5056	0.2041

Expert Pruning & SVD Merging

Method	Pruning + Merging	Pruning	Pruning + Merging	Pruning
Merge count per layer	4	None	12	None
Max cluster size	2	None	6	None
Activated experts	8	8	8	8
# of experts	5270	5352	4886	4786
# of parameters	26.40B	26.79B	24.59B	24.12B
MMLU (acc)	0.7582	0.7608	0.7329	0.7316
Hellaswag (acc)	0.5883	0.5944	0.5710	0.5928
Winogrande (acc)	0.7048	0.7040	0.7127	0.7088
CLIcK (acc)	0.6130	0.6271	0.5924	0.5639
KMMLU (acc)	0.5492	0.5624	0.5214	0.5285

- Pruning과 Pruning+Merging을 각각 적용한 모델 중에서 비슷한 파라미터 수를 가진 모델들의 성능을 비교

Merging의 효과가 크지 않고, 다른 데이터셋에서의 성능 하락이 우려되어 Pruning만 적용

*Merging Coefficient를 학습시에는 불안정한 모습을 보였음

Mitigation of Reduced Performance

- Activated Expert 수에 따른 모델 성능 저하를 완화하기 위한 Training-Free 알고리즘 실험

두 방식 모두 성능 완화에 도움을 주지만, 특히 Method 2가 더 좋은 결과를 보여주어 경량화된 모델에 방식 적용

top6 pruining	Base	Our Method 1	Our Method 2
Activated experts	4	4	4
MMLU (acc)	0.6935	0.7281	0.7375
Hellaswag (acc)	0.5591	0.5451	0.5749
Winogrande (acc)	0.5880	0.6851	0.6803
CLIcK (acc)	0.5704	0.5890	0.5955
KMMLU (acc)	0.4866	0.4966	0.5162

Final Decision

- Threshold of Expert Pruning
- Activated Experts per
- Expert Pruning & SVD Merging
- Mitigation of Reduced Performance

→

- Top 6 Pruning with our method
- Activated Expert = 4

Future work

- Expert Merging 하이퍼파라미터 서치 및 Coefficient 학습 안정화
- 각 레이어별 활성화되는 전문가 수 조정

Thank you