MoE 를 활용한 Xray 객체분류방법론

이재건°,최장훈

경북대학교 데이터사이언스대학원

leejken530@naver.com, jhchoi09@knu.ac.kr

요 약

본 연구는 컴퓨터 비전 분야에서 늘어나는 모델 파라미터 수에 따른 연산 및 메모리 부담을 줄이면서도 높은 성능을 유지하기 위한 경량화 전략을 제안한다. 이를 위해 ConvNeXt 모델의 초기 구조 일부와 MoE(Mixture of Experts)를 결합하는 방식으로 모델을 재구성하였다. 제안한 모델은 원본 ConvNeXt-XXLarge 모델 대비 파라미터 수를 약 97% 이상 감소(약 844M→26.9M)시키면서도 유사한 분류 성능을 유지할 수 있음을 보였다. 이는 실시간 시스템 구현 및 자원 제한 환경에서의 컴퓨터 비전 응용 가능성을 높이는 데 기여할 것으로 기대된다.

Github: https://github.com/2JAE22/Xray_module_classification

1. 서론

컴퓨터 비전 분야는 이미지 분류, 객체 검출, 세그멘테이션 등 다양한 문제에서 뛰어난 성능을 보이는 대규모 모델의 발전과 함께 성장해왔다. 그러나 이러한 성능 향상 이면에는 막대한 파라미터 수 증가에 따른 학습 및 추론 비용 상승, 메모리 요구량 증가, 실시간 처리의 어려움 등의 문제가 존재한다

기존에는 성능 향상을 위해 파라미터를 단순히 늘리는 경향이 있었으나 실제 산업·실무 환경에서는 경량화된 모델이 중요하다. 본 연구는 이러한 필요성에 따라 ConvNeXt-XXLarge 모델에 MoE(Mixture of Experts)를 결합한 경량화 모델을 제안한다. 이 접근법은 일부 ConvNeXt 계층만 활용하고 MoE 를 통해 선택적으로 전문가 모듈만 활성화함으로써 파라미터 수를 크게 줄이면서도 기존 성능을 유지할 가능성을 모색한다.

2. 관련연구

2.1 MoE(Mixture of Expert)[1]

MoE 는 여러 전문가(Expert) 네트워크를 갖춘 구조로, 입력 특성에 따라 소수의 전문가모델만 활성화하여 처리하는 방식이다. 이는 파라미터를 효율적으로 분산시키고 특정 데이터 영역에 특화된 전문가를 통해 성능 및 효율성을 향상시킬 수 있다.

2.2 Convnext 모델[2]

ConvNeXt 는 현대적인 비전 트랜스포머의 디자인 요소들을 ConvNet 에 접목하여 개발된 모델로, 전통적인 ConvNet 의 효율성과 단순성을 유지하면서도 Transformer 수준의 성능을 목표로 한다. 다양한 컴퓨터 비전 작업에서 우수한 성능을 보이지만, 대규모 모델의 경우 파라미터 수가 매우 방대해지는 문제가 있다.

3. 연구방법론

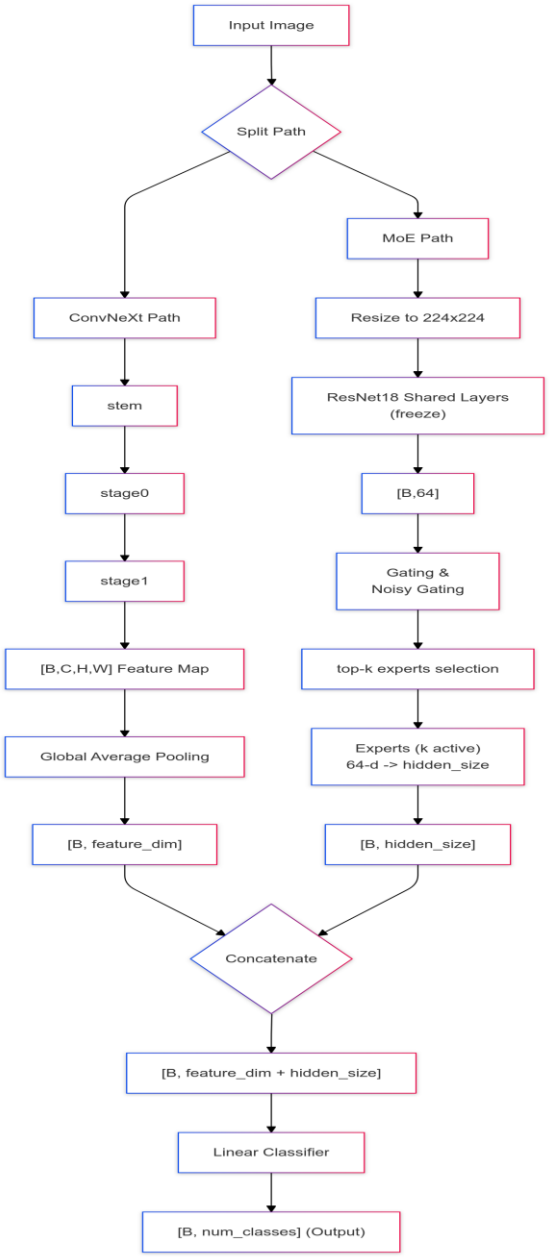
3.1 연구설계

본 연구는 AIhub 에서 제공하는 X-ray 다중 객체 인식 데이터셋[3]을 활용하였다. MoE 를 적용한 모델이 ConvNeXt 대형 모델 대비 파라미터 수를 크게 줄이면서도 유사한 성능(정확도, F1-score 등)을 유지할 수 있도록 다지안하였다. 이를 검증하기 위해 원본 ConvNeXt-XXLarge 모델과 제안한 MoE 기반 경량화 모델(ConvNeXt Custom)을 동일한 조건에서 학습 및 평가하였고 파라미터 수와 성능을 비교하였다.

3.2 모델 Architecture

제안된 모델의 전체적인 아키텍처는 다음과 같

다.



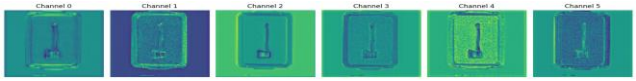
(Figure 1) Model Architecture flowchart

원본 ConvNeXt-XXLarge 는 stem, stage0, stage1, stage2, stage3 및 classification head 로 구성된다. 그러나 본 연구에서는 ConvNeXt 의 **stem, stage0, stage1 까지만 사용**하고, 이후 단계(stage2, stage3, head)는 제거하거나 사용하지 않는다. 그 대신 stage1 까지의 출력을 전역 평균 풀링(Global Average Pooling)을 통해 [B, C] 형태의 벡터로 축소한다. 이 특징 벡터는 주요 저수준 및 중간 수준 특징을 반영한다(Figure 2).

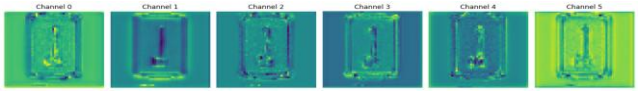
한편, 원본 이미지 입력은 MoE 모듈에도 전달된다. MoE 내부에서는 ResNet18 의 초기 일부 계층(conv1, bn1, act1, maxpool, layer1)만 freeze 상태로 활용하여 [B, 64] 크기의 단순한 특징 벡터를 추출한다. 이후 게이트(gate)와 노이즈(noisy gating) 메커니즘을 통해 상위 k 개의 전문가만 활성화하고, 각 전문가(expert)는 이 64 차원 벡터를 hidden_size(예: 128 차원)로 변환한다. 모든 활성화 전문가의 출력을 Combine 하여 최종 [B, hidden_size] 벡터를 생성한다.

최종적으로 ConvNeXt 를 통한 [B, feature_dim] 벡터와 MoE 를 통한 [B, hidden_size] 벡터를 합쳐 [B, feature_dim+hidden_size] 차원의 벡터를 형성한 뒤, 이를 최종 Linear Classifier 에 통과시켜 분류 결과를 얻는다. 이로써 ConvNeXt 의 거대한 뒷단(head) 없이도 효과적인 분류 성능을 유지하며, 파라미터 수를 비약적으로 줄일 수 있다.

Stage[0] : 이미지의 전체적인 형태와 텍스처가 보존됨



Stage[1] : 이미지의 윤곽선과 구조를 포괄적으로 표현하는 중간 계층의 출력.



(Figure 2) Stage 별 특징예시

4. 실험 결과 및 분석

4.1 파라미터 수 비교

원본 ConvNeXt-XXLarge 모델은 총 844M 파라미터(이 중 학습 가능한 파라미터 약 115M)를 가지며, 약 3,378MB 에 달하는 대규모 모델이다. 반면 제안한 MoE 기반 ConvNeXt Custom 모델은 파라미터를 약 26.9M 까지 줄여 전체 용량이 약 109.6MB 수준으로 감소하였다. 이는 파라미터 수 기준으로 약 97% 이상의 감축을 달성한 것이다.

Convnext

115M	**Trainable params**
729M	Non-Trainable params
844M	Total Params
3,378.417	Total estimated model param size(MB)
517	Modules in train mode
0	Modules in eval mode

Convnext_with_MoE

24.4M	**Trainable params**
157K	Non-Trainable params
24.5M	Total Params
98.105	Total estimated model param size(MB)
143	Modules in train mode
0	Modules in eval mode

(Figure 3) 파라미터수 비교

4.2 성능 비교

모델 파라미터가 크게 줄었음에도 불구하고, 특정 분류 작업에서 유사한 정확도(accuracy)와 F1-score 를 유지하는 결과를 확인하였다.향후 파라미터 튜닝, 데이터 증강, 전문가 수(k 값) 조정 등을 통해 추가적인 성능 개선 여지가 있음을 확인하였다.

Model	Base/convnext_xlarge	Moe_epoch100
val_acc	0.76214	0.89853
val_f1_score	0.68424	0.68746

5. 결론

본 연구에서는 ConvNeXt-XXLarge 모델 구조 중 일부만 활용하고 MoE 를 결합하는 방법을 통해, 파라미터 수를 대폭 줄이면서도 성능을 유지할 수 있는 경량화 전략을 제시하였다. 특히 844M 파라미터 규모의 원본 모델을 약 26.9M 파라미터로 축소함으로써, 실시간 시스템 구현과 자원 제한 환경에서의 활용 가능성을 제시하였다. 향후 연구에서는 다양한 데이터셋과 작업에 대한 일반화, 전문가 선택 전략 최적화, 파라미터 튜닝을 통해 성능 및 효율성을 더욱 개선할 수 있을 것이다.

감사의 글

감사합니다.

참고문헌

[1] Noam Shazeer et al., “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer,” arXiv:1701.06538, 2017.

[2] Zhuang Liu et al., “A ConvNet for the 2020s,” arXiv:2201.03545, 2022.

[3] AIhub, “X-ray 다중 객체 인식 데이터,” [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?dataSetSn=71442