알츠하이머 중증도 진단을 위한 데이터 효율적인 CNN-ViT 하이브리드 모델

CUAI 7기 CV 1팀
성현우(전자전기공학부 19), 임현오(응용통계학과 20), 정현석(전자전기공학부 21)

[요약] 본 연구는 알츠하이머 병 중증도 진단을 위해 CNN과 Vision Transformer(ViT)를 결합한 하이브리드 이미지 분류 모델을 제안한다. ResNet-50을 기반으로 ViT를 결합한 하이브리드 아키텍처는 데이터 효율성을 극대화하기 위해 최적화되었다. MRI 이미지로 구성된 데이터셋을 사용한 실험 결과, 제안된 모델은 86.77%의 정확도와 클래스에 따른 가중화된 0.84의 F1 Score 평균치를 기록했으며, 일부 클래스에서는 1.00의 정밀도를 달성했다. 본 모델은 컴퓨팅 비용을 줄이면서도 적은 데이터로도 높은 성능을 유지하며, 의료 분야 응용의 실용성을 제시한다.

1. 서 론

최근 인공지능의 급격한 발전에 따라 다양한 분야에서 심층 신경망 이론 및 구조가 연구되고 있다. 특히 이미지 처리 분야에서는 AlexNet[1], EfficientNet[2], ResNet[3] 등 컨볼루션 뉴럴 네트워크(CNN)의 구조적인 부분에서 최적화가 되었으며, 대규모 이미지 학습을 통해 훈련된 모델이 배포되고 있다.

CNN은 주요 피처 맵(feature-map)을 추출하여 이미지 분류를 보다 직관적으로 수행할 수 있는 과정을 제공한다. 그러나 CNN은 지역적인 특성 추출에 뛰어나지만 이미지의 전역적인(long-range) 정보나 컨텍스트를 캡처하는 데 한계가 있고, 성능을 위해 네트워크 깊이를 증가시키거나 필터의 크기를 크게 해야 함으로써 계산 비용과 메모리 사용량이 급격하게 증가한다.

이에 따라 최근 등장한 self-attention 메커니즘 기반의 ViT(Vision Transformer)가 이미지 처리 분야의 새로운 아키텍처로써 제시되었다. 자연어 처리(NLP)에서 많이 사용되는 Transformer를 Vision Task에 적용하여 기존의 제한적인 Attetion 메커니즘에서 벗어나, CNN 구조 대부분을 Transformer로 대체하여 훨씬 적은 계산 리소스로, 우수한 결과를 얻었다는 점[4]에서 큰 의미가 있다. 다만 많은 데이터를 사전 학습해야 된다는 제한사항이 큰 도전과제이다.

데이터 효율적인, CNN-ViT 하이브리드 아키텍처가 DeiT 모델로써 주목을 받고 있다. 이로써 기존 사전 학습된 CNN 연산 기반의 입력 이후 self-attention을 적용한 student 모델에 연속적으로 처리됨으로써 연산 오버헤드를 줄이면서도 대용량 데이터가 필요하지 않다는 장점이 있다. 때문에 대용량 데이터 확보가 어려운 바이오, 메디컬 분야의 이미지 처리에서 하이브리드 아키텍처가 주목받는 이유이다.

본 논문에서는 이러한 바이오, 메디컬 분야의 어플리케이션을 위해 알츠하이머 중증도 이미지 분류 모델을 설계한다. 특히, 기존 CNN 중 ResNet 아키텍처와 ViT 모델을 구축하여 이들을 Base line으로써 이들간 융합 아키텍처인 ResNet Back-bone 기반의 ViT 하이브리드 아키텍처를 최적화하여 성능을 비교한다. 데이터 효율성이 중요한 알츠하이머 병 진단에 하이브리드 아키텍처를 제안함으로써 의료 모델로써의 성능과 컴퓨팅 리소스, 즉 연산 비용과 메모리 사용량을 최소화한 모델을 구축하여 의료 분야에서의 하이브리드 아키텍처의 가능성을 제시하고자 한다.

2. 본 론

(1) 알츠하이머 중증도 분류 데이터

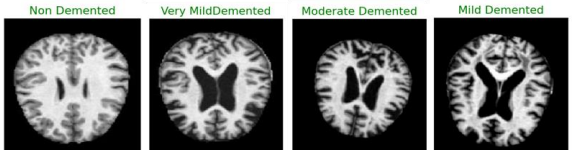

그림 1. 알츠하이머 중증도에 따른 각 클래스별 이미지

알츠하이머 병은 진행성 신경 질환으로, 기억력 상실, 인지 능력 저하, 행동 변화 등을 초래한다. 이 병은 치매의 가장 흔한 원인으로, 사람의 생각, 기억, 일상 활동 수행 능력에 영향을 미친다. 초기에는 경미한 기억 문제로 시작되지만, 시간이 지나면서 언어, 추론, 작업 수행 능력에까지 영향을 줄 수 있다. 정확한 원인은 아직 완전히 밝혀지지 않았지만, 유전적, 환경적, 생활 방식 요인이 복합적으로 작용한다고 알려져 있다. 현재로선 완치 방법이 없지만, 증상 관리를 위한 치료법들이 존재한다.

본 연구에서 사용하는 데이터는 알츠하이머 중증도를 분류하기 위한 4가지 클래스(Mild Demented, Moderate Demented, Non Demented, Very Mild Demented)로 나뉜 MRI 이미지들로 구성되어 있다. 각 클래스의 이미지 분포는 170개, 13개, 654개, 443개로, 총 1,280개의 이미지로 이루어져 있다. 각 클래스 데이터가 불균형하게 분포되어 있는 만큼, 데이터 증강을 통해서 이를 10배가량 보간하여 적절한 모델 학습이 가능하도록 하였다.

(2) 배경 지식 (CNN, ViT, Hybrid DeiT)

해당 논문에서 다루고자 하는 CNN, ViT 기반의 베이스 라인 대표 모델에 대해서 간단히 소개한 후에 제안하는 아키텍처에 대해서 자세히 후술한다.

① *ResNet*

CNN layer를 깊게 쌓을수록 성능이 더 좋아지지만 layer가 너무 깊어지면 Vanishing gradient 현상이 발생한다. 또한 20개 이상의 layer부터는 성능이 낮아지는 degradation problem도 발생한다. 이 두가지 문제를 잔차(Residual)을 통해 해결하고자 하는, 주요 모델이다.

ResNet은 기본적으로 VGG-19의 구조를 뼈대로 하며, 컨볼루션 연산 간 층들을 추가해서 깊게 만든 후에, shortcut들을 추가한 것이다. 본 논문에서는 CNN 기반 모델의 베이스 라인으로 사용한다.

② *ViT(Vision Transformer)*

Self-attention 기반 Transformer 모델이 자연어 처리(NLP)의 주요 모델로 자리 잡고 있는 상황에서 Transformer를 전체 아키텍쳐를 크게 변경하지 않은 상태에서 Vision Task에 적용하려는 모델이다. 기존의 제한적인 Attetion 메커니즘에서 벗어나, CNN구조 대부분을 Transformer로 대체한다.

ImageNet와 같은 Mid-sized 데이터셋으로 학습 시, ResNet보다 낮은 성능을 보이지만 JFT-300M으로 사전 학습 후, 전이 학습 시에 CNN구조 보다 매우 좋은 성능 달성하였다는 점에서 큰 주목을 받고 있다. 다만 CNN보다 더욱 큰 규모의 데이터가 요구된다는 문제점이 남아 있다.

③ *Hybrid DeiT(Date-Efficient Transformer)*[5]

앞서 설명한 것과 같이 ViT는 Data Efficiency가 낮은 모델이다. 비전 트랜스포머가 기존 inductive bias의 이점을 뛰어넘기 위해선 대규모 데이터셋으로 학습해야 하기 때문이다. DeiT는 ViT의 inductive bias가 얕다는 취약점을 Teacher 모델의 inductive bias를 전달받음(Distillation)으로써 보완한다.

적은 수의 파라미터와 간단한 컴퓨팅 자원 그리고 높은 Data Efficiency을 목표로 하는 모델이다. 따라서 별도의 Distillation 토큰이 요구되며, Teacher 모델의 CLS 토큰에서 나온 hard label을 Student 모델에서 이 토큰을 기반으로 예측을 진행한다. 때문에 기존 Teacher 모델의 완성도가 있다면, 별도의 학습 과정에서 가중치 업데이트가 요구되지 않는다.[6]

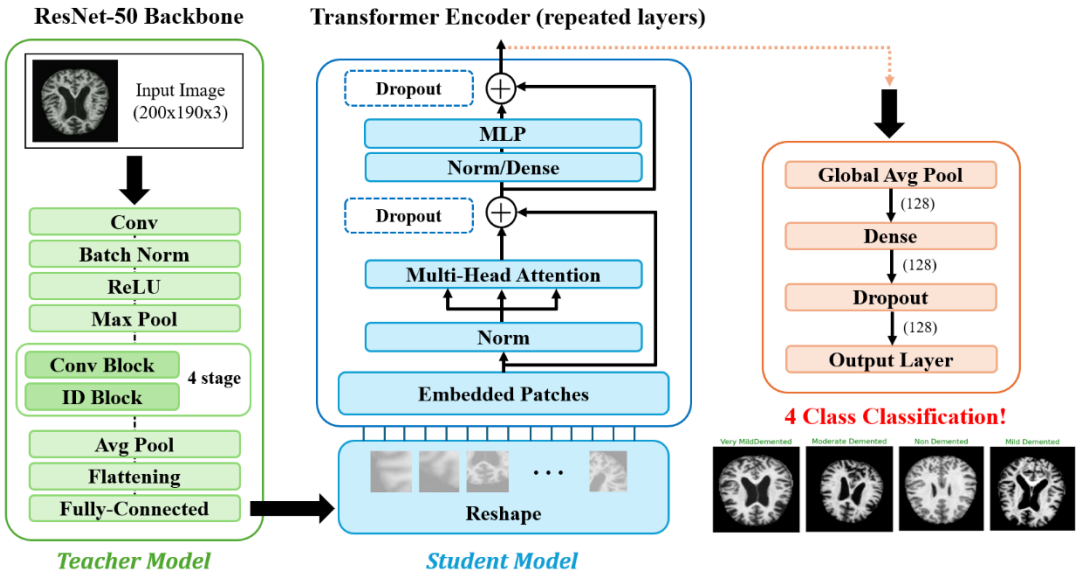

그림 2. 제안하는 CNN-ViT 하이브리드 아키텍처 기반 알츠하이머 중증도 진단 모델, ResNet-50 Backbone 기반으로 기존 ViT의 비선형성 추가, Dropout 레이어로 과적합 방지

(3) 제안하는 CNN-ViT 하이브리드 아키텍처

본 연구에서 제안하는 모델은 그림 2와 같다. 기본 DeiT 구조를 유지하며 Teacher 모델은 ResNet-50 Backbone에 기인하며 최종 컨볼루션 레이어의 출력은 Student 모델인 ViT의 입력으로 인가되어 패치 임배딩을 거치게 된다. 연산량 및 과적합을 줄이기 위한 Dropout 레이어 추가와 더불어 비선형성 강조를 위한 Dense 레이어를 통해 구현했다. 이를 통해 인코딩된 텐서는 최종 Global Average Pooling 이후 알츠하이머 중증도 4가지 클래스에 분류된다. 이로써 기존 CNN 모델의 추가적인 학습 및 가중치 조절이 생략 가능하므로 컴퓨팅 자원을 최소화하여 연산이 가능하며 메모리 사용량 또한 획기적으로 감축할 수 있을 것이 기대된다.

모델 학습 시, 동적 가중치 업데이트 과정에서 Adam Optimizer와 Learning-rate scheduler를 통해 동적으로 학습률을 제어하여 최적화에 힘쓰고자 했다.

다음 목차에서 제안된 CNN-ViT 하이브리드 아키텍처와 더불어 동일 데이터 셋으로 학습된 ResNet, ViT 모델의 Validation 데이터에 대한 에폭 당 Accuracy 및 다양한 성능 지표를 분석한다.

(4) 모델 성능 평가

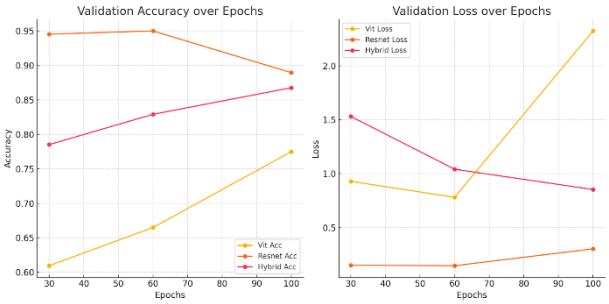

그림 3. 베이스 라인 (ResNet, BasicViT) 및 제안하는 모델(CNN-ViT Hybrid)의 에폭 당 성능 평가

알츠하이머 중증도 분류를 위한 증강된 데이터셋을 기준으로, 각 모델에 대한 학습을 진행했고, 검증용 데이터로 모델 성능을 평가한다. 그림 3에 따르면, 에폭에 따른 Top-1 Accuracy의 경우 ResNet이 88.95%로 가장 우월한 성능을 보이고 ViT가 가장 낮은 성능을 보인다. 이상적으로, 제안하는 CNN-ViT 하이브리드 아키텍처의 경우 에폭에 따라 성능 향상이 돋보였으며 최종적으로 100회 수행에 대해서 86.77%로 ResNet과 거의 동일한 성능을 유지한다. 이는 97.5%의 성능 유지를 의미한다. Loss 또한 ViT는 쉽게 과적합되어 특정 에폭 이후 급격하게 상승하는 경향을 보이나 제안하는 CNN-ViT 하이브리드 아키텍처의 경우 Dropout 레이어 및 Dense 레이어로써 과적합에 강인한 모델임을 보여준다.

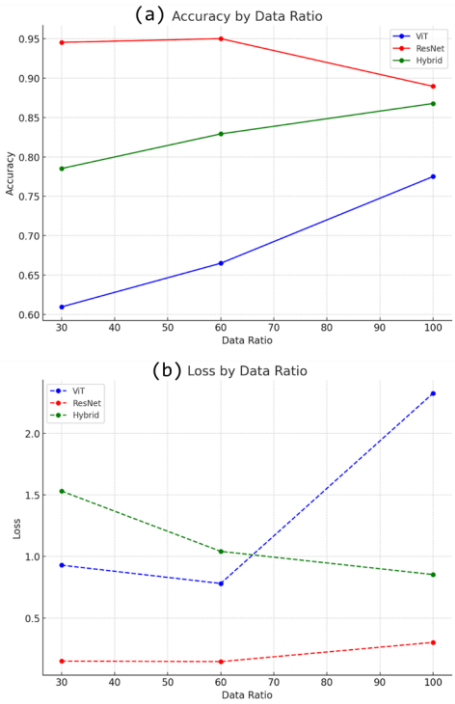

그림 4. 각 모델의 데이터 효율성 검증, (a) 학습 데이터 비율에 따른 Accuracy, (b) Loss

또한 제안하는 하이브리드 아키텍처는 기존 ViT 대비 Data-efficiency가 우수하다. 따라서 데이터 비율에 따른 Accuracy, Loss가 ViT 모델보다 훌륭해야 한다. 이를 검증하기 위해 동일 데이터 셋에 대해 각 클래스별로 30%, 60% 비율로 증강된 100% 비율의 데이터를 랜덤 샘플링하여 각 데이터에 대해 모델을 피팅 시켰을 때, 최종 Accuracy와 Loss를 평가했다(그림 4). 그 결과, 제안하는 CNN-ViT 하이브리드 아키텍처의 경우 ViT에 비교했을 때, 적은 데이터 비율로도 어느 정도 성능을 유지한다는 것을 알 수 있다. '데이터 효율적인' 학습이 가능하다는 것을 검증함으로써 이는 대규모 데이터 수집이 어려운 바이오, 메디컬 등 어플리케이션에 적합하다는 것을 증명한다.

Model		Precision	Recall	F1-Score
Hybrid Architecture DeiT	Macro Average	0.86	0.73	**0.77**
	Weighted Average	0.84	0.84	**0.84**

Class	Class Count	Precision	Recall	F1-Score
Mild Demented	170	0.77	0.75	**0.76**
Moderate Demented	13	1.00	0.46	**0.63**
Non Demented	654	0.89	0.88	**0.88**
Very Mild Demented	443	0.80	0.83	**0.81**

표 1. 최종 제안하는 모델의 성능 지표 정리

마지막으로 해당 모델의 목적은 알츠하이머 병 중증도 진단에 초점을 맞춘다면, 결국 모델의 실용화를 위해서는 정확도(Accuracy) 외에도 정밀도(Precision)가 주요한 성능 지표가 될 것이다. 정밀도, 재현율 사이의 관계를 묘사하는 F1 Score를 분석하는 것 또한 주요 과제이다. 제안하는 하이브리드 아키텍처의 경우 Precision이 Macro Average에 대해 0.86, Weighted average에 대해 0.84로 평가되며 중증도에 따른 4가지 클래스의 증강되지 않은, 즉 원본 데이터 비율에 따른 가중치가 적용된 경우 F1 Score는 0.84이다. 특히 Moderate Demented 클래스에 대해서 1.00의 매우 우수한 정밀도를 구현하기에 알츠하이머 병 초기 진단에 대한 추후 연구의 적절한 최적화가 이루어지면 제안하는 아키텍처 기반의 모델로 초, 중기 병 진단에 유리한 위치를 가져갈 수 있을 것이다.

3. 결 론

본 논문에서는 알츠하이머 병 중증도 진단을 위한 이미지 분류 모델을 설계했다. 무엇보다 최신 연구에서 주목하는 컴퓨팅 비용 및 메모리 사용과 관련되어 어려움을 가지는 전통적인 합성곱 심층 신경망(CNN)과 이에 대한 해결책으로 제시되는 비전 트랜스포머(ViT)의 데이터 의존적인 한계를 동시에 극복하고자 하는 아키텍처를 제안한다. ResNet-Backbone 기반의 하이브리드 CNN-ViT 아키텍처를 구성하여, 기존 Teacher 모델인 ResNet의 추가적인 가중치 업데이트 없이, Student 모델에 최적화된 ViT가 이를 상속받아 Self-attention을 통해 보다 컴퓨팅 자원을 적게 소비하며, 데이터 의존도를 줄인 하이브리드 아키텍처를 최적화하는 연구를 진행했다. 특히, 트랜스포머의 비선형성을 위한 Dense 레이어와 더불어 과적합 방지를 위한 Dropout 레이어를 통해 이상적인 손실 함수 수치를 구상할 수 있었다.

알츠하이머 중증도 분류에서는 100 epoachs에 대해 86.77%의 Accuracy와 0.853의 Loss function 값을 얻었으며, Data-efficient 측면에서도 ViT에 비해 30% 비율의 데이터 셋으로 학습된 모델은 128% 개선된 정확도를 보장한다. 각 Class 분류에서 의료 진단의 주요 평가 지표인 정밀도 측면에서도 최소 0.77, 최대 1.00으로 높은 값을 기록하였다. 각 Class의 데이터 수에 따른 가중치가 적용된 F1 Score의 평균값은 0.84이다.

해당 연구를 통해 알츠하이머 중증도 진단에 최적화된 이미지 분류 모델을 설계하였으며, 하드웨어 리소스를 적게 소비한다는 점과 대규모 데이터 수집의 중요성이 덜 강조된다는 점은 주목할만하다. 이에 따라 알츠하이머 중증도 초기 진단에 해당 모델을 응용함으로 사회적 순기능이 기대되며, 제안하는 CNN-ViT 하이브리드 아키텍처를 후속 연구로써 의료 분야에 최적화된 모델로 발전시킬 수 있을 것이라 기대한다.

참고 문헌

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In Advances in Neural Information Processing Systems, vol. 25, 1097-1105. 2012.

[2] Tan, Mingxing, and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." In Proceedings of the 36th International Conference on Machine Learning (ICML), vol. 97, 6105-6114. 2019.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778

[4] Alexey, D. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929

[5] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (pp. 10347-10357).

[6] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.