CodeStates AI Bootcamp 15th Section4 Project

AI\_15\_정경재

Script

**Deep Learning based Alzheimer & MCI Diagnosis Model**

**딥러닝 기반 알츠하이머, 경도인지장애 진단 모델 개발**

**AI\_15\_정경재**

|  |  |
| --- | --- |
| Page | Script |
| 1 | 안녕하세요. AI 부트캠프 15기 정경재입니다. 발표를 시작하겠습니다. |
| 2 | 발표 목차입니다.  준비한 내용이 많아서 최대한 빠르게 진행해 보도록 하겠습니다. |
| 3 | 서론입니다. |
| 4 | 우선 데이터 직무를 데이터 사이언티스트와 데이터 분석가로 설정하고 진행하였습니다.  의료 및 헬스케어 분야에서는 분류문제에 가장 많은 관심이 집중 되어 있기 때문에  이러한 트렌드에 맞추어 주제를 분류모델로 정하게 되었습니다. |
| 5 | 그럼 치매에 대해 잠시 살펴보면, 치매는 건망증과는 다른 엄연한 질병이며, 가장 흔한 원인 질환은 알츠하이머입니다. |
| 6 | 알츠하이머 치매는 뇌세포의 퇴화로 인해 치매 증상이 유발되는 질병이고, MRI 사진처럼 뇌가 쪼그라들어 있는게 특징입니다.  여기서 경도인지장애라는 용어는 낯설은 분들도 계실텐데, 건망증과 알츠하이머의 중간 단계, 즉 치매 전단계라 보시면 됩니다. |
| 7 | 그렇다면 이러한 연구가 왜 필요하며, 왜 굳이 딥러닝을 적용해야하는지를 살펴보겠습니다.  경도인지장애를 예측하는 것은 예방 차원에서도 매우 중요합니다.  알츠하이머 진단에 있어서 가장 결정적인 것은 뇌 촬영 검사인데, 일반적인 통계나 머신러닝으로는 한계가 있습니다.  이러한 한계를 극복 할 수 있는 것이 바로 딥러닝입니다. |
| 8 | 따라서 이번 프로젝트에서는 의료영상이미지로 알츠하이머와 경도인지장애를 진단하는 모델 개발을 주제로 삼게 되었고,  이번 프로젝트의 목표 및 가설은  메타데이터 분석을 통한 트렌드를 파악  이미지 데이터를 통한 모델링  모델 개선,  그리고 검증과 예측을 진행하는 것입니다. |
| 9 | 이번 프로젝트의 딥러닝 파이프라인입니다.  다음과 같은 순서로 진행되었고, 이제 단계 별로 자세히 살펴보겠습니다. |
| 10 | 우선 글로벌 알츠하이머 종단 연구 사업인 ADNI의 데이터를 기반으로 진행하였고, |
| 11 | 추적관찰이 목적이 아니기 때문에 Baseline 데이터를 이용하였습니다.  이미지 데이터는 캐글에 이를 추려낸 데이터셋이 있어서 다운로드해서 진행했습니다. |
| 12 | 우선 메타 데이터는 특성 공학을 통해 가공하여 분석을 진행했습니다.  캐글에서 다운로드 한 MRI 이미지 데이터를 ADNI에서 직접 확인해본 결과, 간뇌 쪽의 데이터를 추려 냈음을 알 수 있었습니다. |
| 13 | 먼저 메타데이터로 트렌드를 파악해보았고, 분석 결과를 대시보드로도 만들어 보았습니다. |
| 14 | 우선 진단 그룹별로 살펴보면 경도인지장애의 비율이 가장 높은 비율을 차지합니다. |
| 15 | 성별분포는 거의 비슷했고 |
| 16 | 나이대 별로는 전반적으로 정규분포에 가까운 분포를 보입니다. |
| 17 | 이러한 결과들을 반영하여 대시보드로 구현하였습니다.  간단하게 살펴보면, (클릭)  스코어 카드들이 있고, 클릭해보시면 모든 차트들이 상호작용 하는 것을 확인 할 수 있지요.  자 그럼 다시 돌아와서… |
| 18 | 본격적으로 모델링을 진행해 보도록 하겠습니다. |
| 19 | 총 세가지 모델의 모델링은 코랩 환경에서 진행되었습니다. |
| 20 | 모델은 CNN으로 정하였는데, 그 이유는 CNN이 이미지 데이터 분석에는 가장 적합한 모델이라 판단했기 때문입니다. |
| 21 | 모델들은 전처리 부분부터 분류 부분까지 모델에 포함시키는 구조로 구성하였습니다. |
| 22 | 우선 구글 드라이브에 저장한 이미지 데이터들을 불러옵니다. |
| 23 | 불러온 데이터에 대한 비율을 확인하고, 이미지를 시각화 했습니다. |
| 24 | 그 다음으로 데이터 누수를 방지하기 위해 데이터를 분리합니다.  데이터수가 충분히 커서 교차검증이 아닌 3way 홀드아웃으로 8:1:1 비율로 분리해 검증 및 예측을 진행했습니다. |
| 25 | 모델의 앞부분은 전처리 레이어로 구성했습니다.  이미지 크기와 픽셀 범위를 정규화 했고  훈련 데이터셋에만 적용되는 데이터 증강기법을 적용하였습니다. |
| 26 | 그럼 첫번째로 가장 기본적인 바닐라 CNN 모델링을 진행하겠습니다. 구조 설명은 생략하겠습니다. |
| 27 | 평가지표는 어큐러씨를 기준으로 했습니다. |
| 28 | 검증 스코어는 0.88 정도로 나왔고 모델을 파일로 저장했습니다. |
| 29 | 이젠 성능을 개선해 보도록 하겠습니다.  전이학습은 가장 많이 쓰이는 ResNet 사전학습모델로 진행했습니다. |
| 30 | 가중치가 학습되지 않게 설정했고 분류기 레이어를 설계하여 진행했습니다. |
| 31 | 학습을 진행해본 결과, 오히려 성능이 하락하는 네거티브 전이가 일어났기 때문에 모델 저장을 진행하지 않았습니다. |
| 32 | 네거티브 전이란, 관련성이 낮은 태스크로 전이학습을 시도하면 성능이 하락하는 현상을 네거티브 전이라고 합니다.  예를 들어 설명해 보겠습니다.  만약 라면을 요리하고 이게 라면이 맞는지 확인하는 태스크가 있다고 해봅시다.  요리 경험이 적은 초보자는 기본 조리법대로 끓일 것이고, 이는 라면이라 판단 하기 쉬울 겁니다.  그러나 다양한 요리에 대한 경험이 많은 요리사는 더 복잡한 레시피로 끓일 것이고, 그림과 같은 요리가 만들어 졌을 때는 이걸 라면이라 해야할지 판단하기 어려울 것입니다.  이처럼 때로는 전이학습을 진행하지 않은 경우가 오히려 더 좋은 결과로 이어 질 수 있다는 것이죠. |
| 33 | 튜닝 과정입니다.  튜닝할 파라미터는 컨볼류젼 층의 필터 수와 분류기 은닉층의 노드 수로 지정했습니다. |
| 34 | 튜닝은 케라스 튜너로 진행하였습니다. |
| 35 | 튜닝 후에는 더 이상 코랩 GPU를 이용할 수 없었기 때문에, 최적화 된 모델을 저장하였습니다. |
| 36 | 이제 최종 모델링을 마무리하겠습니다. |
| 37 | 로컬에서 진행이 되었고 |
| 38 | 저장했던 모델을 불러오고 컴파일은 동일하게 설정합니다.  이번에는 가장 성능 좋은 모델을 저장하고 성능이 5번이상 개선이 안되면 학습을 중단하도록 콜백 함수를 지정하였습니다. |
| 39 | 검증 스코어가 무려 0.996이 나왔습니다. 만족스럽네요. |
| 40 | 최종 모델에 대한 일반화 가능성을 살펴보니, 차이가 크지 않아 일반화도 잘 된 것 같습니다. |
| 41 | 이젠 예측을 통해 튜닝 전후 모델들을 비교 분석해보도록 하겠습니다. 분석을 마친 뒤엔 가상환경을 다른 환경에서도 사용할 수 있도록 텍스트 문서를 만들고, 테스트도 완료했습니다. |
| 42 | 학습에는 이용되지 않은 테스트 데이터셋을 예측에 이용했고, |
| 43 | 일단 12개의 이미지를 예측해보았습니다.  튜닝을 거치지 않은 모델에서는 틀리게 예측한 경우도 보이고, 맞게 예측을 해도 얼마나 확신을 가지고 예측 값을 판단했는지를 나타내는 수치인 Confidence값이 낮게 나오는 데이터들이 보입니다.  튜닝을 거친 최종 모델에서는 예측 정확도와 confidence 모두 개선 되었습니다. |
| 44 | 전체 테스트 데이터셋으로 예측을 진행해도  최종 모델이 모든 수치에서 높은 성능을 나타내고 있습니다. |
| 45 | 결론입니다. |
| 46 | 중요한 부분만을 요약해보면,  서론에서는 프로젝트의 필요성을 살펴봤고, 데이터셋을 선정하고 메타데이터를 분석하였습니다.  모델링은 CNN 모델로 진행했고, 전이학습 모델에서는 네거티브 전이가 일어났었습니다.  테스트 데이터로 예측한 결과 만족스러운 성능과 일반화 가능성이 나타났습니다. |
| 47 | 이번 프로젝트의 한계점과 추후 발전 방향을 살펴보면,  알츠하이머에 국한된 모델이기 때문에 다른 치매 유발 요인에 대한 추가분석이 이뤄질 수 있을 것 같고,  한정된 MRI 이미지 데이터를 사용했기 때문에 분석 범위를 넓혀서 분석해볼 수 있을 것이라 봅니다.  그리고 데이터 증강기법을 두가지만 썼는데, 좀더 다양한 증강 기법으로 모델을 더 굳건히 만들 수도 있을 것 같고,  학습에 이용하지 않은 테스트 데이터셋으로 예측을 진행했지만, 다른 기관에서 측정한 MRI 데이터로도 일반화가 되는지도 추가 분석 해볼수 있을 것같습니다. |
| 48 | 마지막으로,  본 프로젝트는 의료영상이미지로 알츠하이머, 경도인지장애, 정상으로 분류하는 딥러닝 모델을 개발하는 과정을 다루고 있으며,  이러한 모델을 알츠하이머 위험군을 선별하는데 활용하여 적극적인 예방을 실시 할 수 있을 것이라 생각하고  또한 이를 진단에 활용하여 좀 더 정확한 진단을 내리는 데에도 활용 할 수 있을 것이라 봅니다.  본 프로젝트 및 섹션에 대한 느낀점을 간략하게 말씀 드리자면  우선 많은 것들을 배울 수있어서 뿌듯했고, 성능도 높게 나타나서 만족스러웠습니다.  아쉬운점은 컴퓨터 사양의 한계로 코랩을 주로 이용했는데, 코랩이 생각보다 불편했어서  컴퓨터 사양에 대한 아쉬움이 남아 있는것을 꼽을 수 있을 것 같습니다. |
| 49 | 이상으로 발표를 마치도록 하겠습니다. 감사합니다. |