방사선 영상 이미지에서 Xception 모델을 이용한 퇴행성 관절염 등급 분류

Accuracy analysis of a convolutional neural network model based on radiographic imaging for diagnosing degenerative arthritis.

요약(Abstract)

본 논문은 퇴행성 관절염 등급 진단에 있어 기존의 사용되던 딥러닝 모델이 아닌 Xception 모델을 활용한 연구이며, 딥러닝 모델을 통한 진단 방법의 개선을 위한 연구이다.

1. Introdution(서론)

몸은 자유로운 움직임을 위해 수많은 관절들로 이루어져 있다. 이러한 관절에는 뼈와 뼈가 직접 부딫치는 것을 방지하기 위해, 사이에 연골(물렁뼈)이 있는데 이러한 연골은 정상적인 관절의 기능을 유지하는데 가장 중요한 조직으로, 매끄러우면서도 질기며 동시에 탄력성을 지니고 있어 뼈와 뼈 사이의 마찰을 줄여주고 충격을 흡수하여 관절이 움직일 때 잘 미끄러지도록 하는 작용을 하게 된다.

연골은 생활 속에서 계속 사용되어 나이가 들어갈수록 마모의 정도가 심해진다. 그 중에서도 가장 사용이 활발하여 마모에 대한 빈도가 높은 부분이 무릎의 연골이다. 나이가 들수록 마모된 연골에 의해 무릎의 통증이 증가하고, 아무리 새 차를 뽑아도 시간이 지나가면서 중고차가 되는 것처럼, 나이가 들면서 이런 중요한 연골도 점차 닳게 되어 정상적인 연골의 기능을 하지 못하게 되어 결국 관절염이 발생하게 된다. 물론, 젊은 사람의 경우에도 운동선수와 같이 무릎을 많이 사용하면 연골이 손상될 수 있다. 이렇게 발생하는 관절염을 퇴행성 관절염이라고 한다. [1] 퇴행성 관절염으로 진행이 심할 경우에는 인공 무릎 관절 수술을 진행해야하는 경우도 있다.

무릎의 연골 관련한 부분으로는 대퇴골(Femur), 경골(Tibia), 슬개골(Patella)로 구분되며, 대퇴골은 흔히 말하는 넓적다리뼈로, 골반에서 무릎까지의 뼈를 가리킨다. 인체에서 가장 긴 뼈라고 할 수 있다. 경골은 흔히 말하는 정강이 뼈로, 무릎에서 발목까지 이어지는 뼈이다. 대퇴골 다음으로 긴 뼈이다. 슬개골은 대퇴골과 경골 위를 덮고 있는 삼각형 모양의 뼈이다.

나이가 들면서 오래 사용된 무릎의 연골은 닳기 시작하며, 연골의 마모가 가장 많이 일어나는 곳은 움직임이 가장 활발한 대퇴골과 슬개골 사이 부분이다. 또한, 대퇴골과 슬개골 사이에는 활액이라는 윤활유 역할을 하는 액체가 있는데, 젊었을 때는 활액이 충분하지만, 나이가 들수록 활액도 점차 줄어들어 마모의 진행을 조금 더 가속화할 수도 있다.

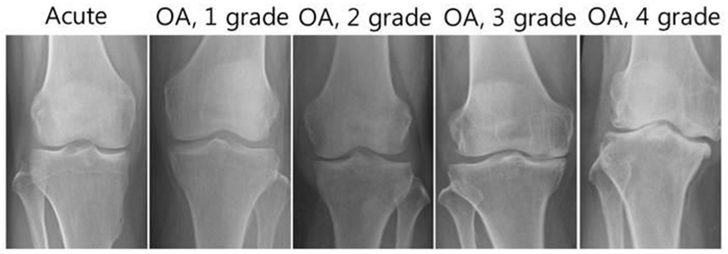
이러한 연골의 마모로 인한 연골의 손상은 Outerbridge, ICRS(International Cartilage Repair Society)와 Kellgren lawrence라는 세 가지 대표적인 지표(표1, 표2, 표3)를 활용하여 진단하며 각각 다른 지표지만 손상에 따른 진단 내용은 비슷하다고 볼 수 있다. [2]

본 연구에서는 세 가지 지표 중 K-L garde 지표를 기준으로 하였으며, K-L grade를 구분하기에는 방사선 이미지가 가장 효율적이다. 초기에는 정상 소견을 보일 수 있으나 점진적으로 관절 간격의 감소가 나타나며 연골 아래 뼈의 음영이 짙어지는 경화 소견을 볼 수 있고, 더욱 진행되면 그림 1과 같이 관절면의 가장자리에 뼈가 웃자란 듯한 골극이 형성되고 관절면이 불규칙해진다. [2]

이러한 시각적으로 나타나는 무릎 관절의 관절면 변화를 이용하여서 퇴행성 관절염의 진단이 가능하기 때문에 기존의 의사의 소견이 더욱 중요했던 관절염 진단 분야에 있어서 딥러닝 모델을 통해 기존의 방법보다 손쉽게 진단을 받을 수 있을 것이다.

따라서 본 연구에서는 무릎 관절 연골 마모에 따른 K-L grade를 구분하기 위해 무릎 방사선학 이미지를 사용하여 딥러닝 모델에 적용하여 그림 1과 같이 5개의 클래스로 퇴행성 관절염 진행 정도를 예측하였다.

그림 1. stage of knee osteoarthritis. [3]



(표 1) Outerbridge 평가 지표

|  |  |
| --- | --- |
| **등급** | **증상** |
| Grade 0 | 정상 연골 |
| Grade 1 | 연골의 연화(softening)와 부풀어짐(Swelling) |
| Grade 2 | 연골 표면에서 연골하골(Subchondral bone)까지 균열이 있는 상태 |
| Grade 3 | 직경 1.5cm 이상을 초과하는 결손과 연골하골(subchondral bone)까지 균열이 있는 상태 |
| Grade 4 | 연골하골(subchondral bone)이 노출된 상태 |

(표 2) ICRS(International Cartilage Repair Society) 평가 지표

|  |  |
| --- | --- |
| 등급 | 증상 |
| Normal | Grade 0 |
| Almost Normal | Grade 1a – superficial lesions/softening Grade 1b – As in 1a and/or superficial fissures or cracks |
| Abnormal | Grade 2 – extent< 50% of thickness |
| Severe lesion | Grade 3a – extent > 50% Grade 3b – down to the calcified layer Grade 3c – down to the surface of the subchondral bone(without penetration) Grade 3d – includes bulging of the cartilage around the lesion |
| Very severe lesion | Grade 4a – penetration of the subchondral bone but not across the entire diameter of the defect Grade 4b – penetraion across the full diameter of the defect |

(표 3) Kellgren Lawrence grade 평가 지표

|  |  |
| --- | --- |
| **등급** | **증상** |
| K-L grade 0 | 정상 |
| K-L grade 1 | 정상에 비해 관절간격이 좁아진 것으로 의심될 경우, 골극이 있을 수 있음 |
| K-L grade 2 | 관절 공간이 명확하게 좁아짐. 골극이 확실함 |
| K-L grade 3 | 관절 공간이 좁고 골경화와 골형태에 심한 변형이 관찰. 중증도의 다발성 골극 관찰 |
| K-L grade 4 | 관절 간격이 현저하게 좁아져 골형태에 심한 변형이 관찰 |

1. 관련 연구

그동안 퇴행성 관절염의 등급 진단을 위한 다양한 연구가 진행되었다. 이와 비슷한 연구에서는 먼저 학습에 사용할 데이터 생성을 위해 YoloV2 모델을 활용하여 무릎 연골의 탐지에 사용하였으며, 분류를 위한 학습에는 VGG19모델을 활용하여서 그림 2와 같은 전체 아키텍처를 구성하여 퇴행성 관절염 등급을 분류하였다. [5]

위 연구에서의 정확도는 그림 3에서 볼 수 있듯이, 올바르게 예측한 정확도는 전체 이미지의 69.58%에 해당하였고, 퇴행성 관절염 등급 분류 영역에 있어 SOTA 기록이 약 70% 인 것으로 보았을 때 최고 성능에 달하는 모델임을 알 수 있습니다. [5]

그림 2. VGG19 모델의 전체 아키텍처

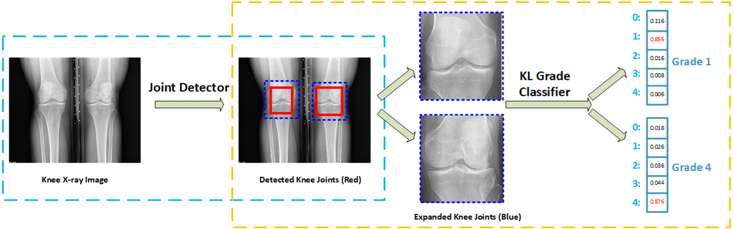
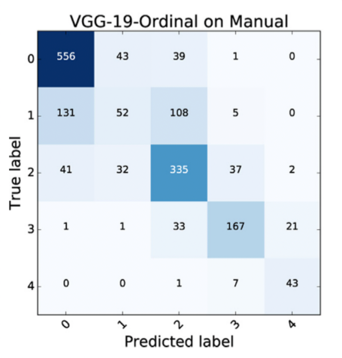


그림 3. VGG19 혼돈 매트릭스(Confusion Matrix)



또 다른 연구에서는 DenseNet-169, Inception V3 모델 두 가지를 사용하였으나, Inception V3 모델의 경우 낮은 성능을 보여 제외되었고, 최종적으로 DenseNet-169 모델을 그림 4와 같이 사용, 사전 학습으로는 ImageNet을 활용하였으며, 데이터 부족 문제를 해결하기 위해 약 80%의 학습 데이터에 그림 5와 같이 Crop, upscale, add noise, flip, randomize contrast를 단계적으로 적용하여 합성곱 신경망 모델 학습에 증강된 이미지를 사용하였습니다. [6]

그림 4. DenseNet-169 모델 학습의 전체 파이프라인

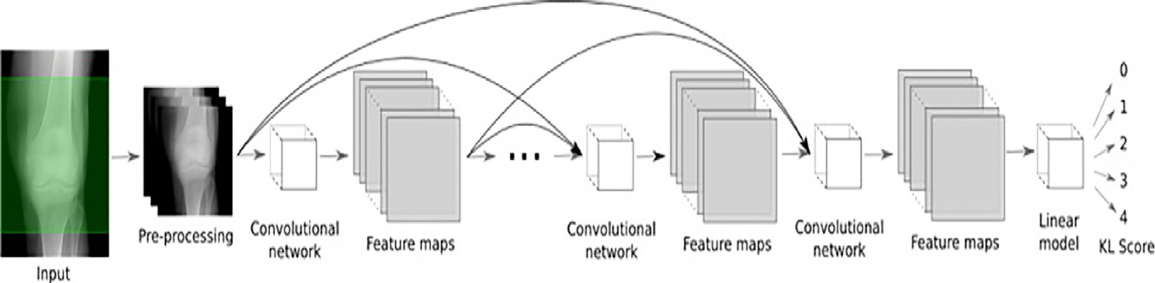
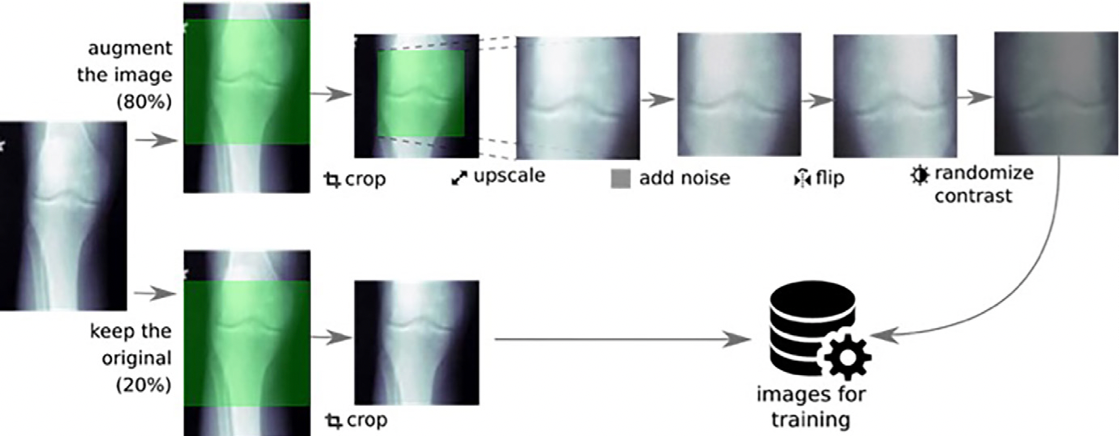


그림 5. 학습 데이터 증강 적용 단계



Yolo 모델을 Detection에 활용한 연구와 비슷하게 진행된 연구는 그림 7과 같은 FCN을 활용하여 방사선학 이미지에서 무릎 관절을 탐지한 후 그림 8, 9와 같이 결과 이미지를 딥러닝 모델의 학습을 통해 등급을 분류하는 연구가 진행되었으며, 그림 10처럼 예측하였습니다.

그림 6. 퇴행성 관절염 진단을 위한 전체 파이프라인

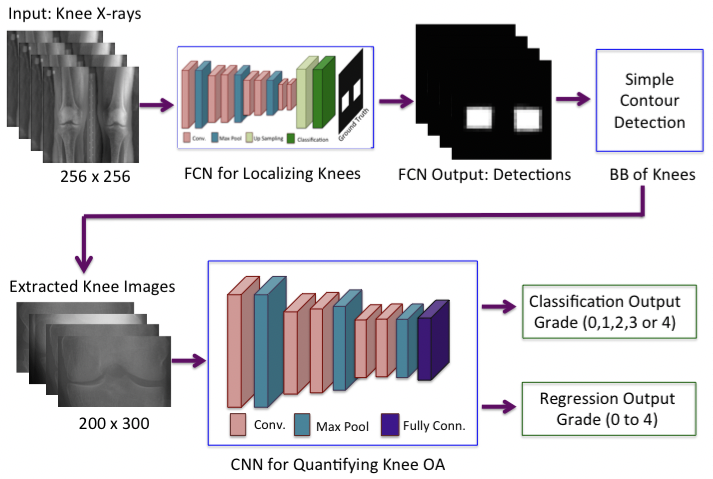


그림 7. 무릎 관절 탐지를 위한 Fully Convolution Network

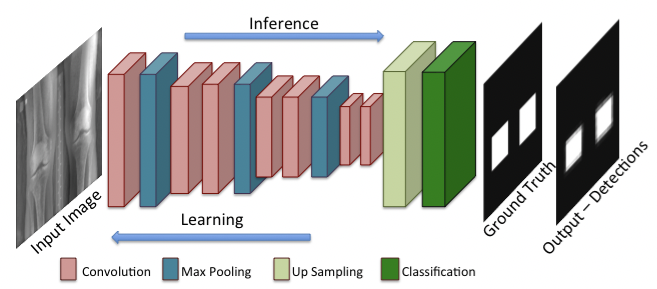


그림 8. 무릎 관절 이미지 분류를 위한 네트워크 아키텍처

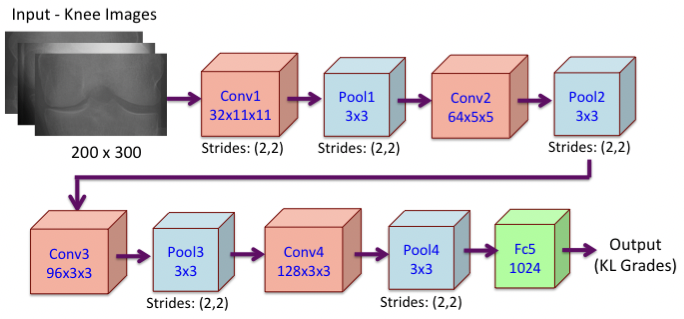


그림 9. 분류 및 회귀를 위한 네트워크 아키텍처

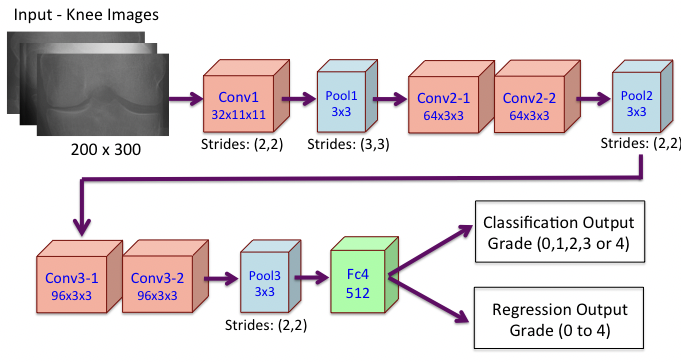
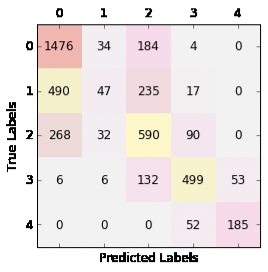


그림 10. 모델의 성능 파악을 위한 혼돈 매트릭스 (Confusion matrix)



1. 문제 정의 및 실험 방법

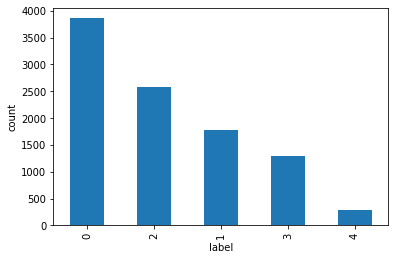
본 논문에서는 무릎 관절 방사선학 이미지들을 통해 관절의 손상 정도를 딥러닝 모델에 학습하여 그림 1과 같이 분류하고자 함에 있으며, DPhi의 Data Sprint #35: Osteoarthritis Knee X-ray [8]의 대회 진행 당시에 ‘akashkewar’ 유저가 Practice, Challenge score에서 Xception 모델을 활용한 이미지 분류 정확도가 96.8335%로 상당히 높은 점수를 기록했으나 DPhi의 대회 규칙이 외부 데이터를 사용하지 않아야 하기 때문에 대회에서 탈락하였다. 이것을 보고 기존의 알고 있던 이미지 분류 문제의 다른 딥러닝 모델들보다 Xception이 효과적인가에 대해서 검증하고자 하며, 추가 데이터 세트를 활용하고 모델의 세부 변수의 조정과 모델의 아키텍처 수정으로 더 높은 정확도를 갱신해보고자 실험을 진행하였으며, 정확도는 Confusion Matrix와 DPhi의 정확도를 기준으로 하였다.

1. 실험

실험에 사용한 주요한 데이터는 관련 연구에서 VGG19모델 학습에 사용되었던 ClsKLData 중 train, test, val, auto\_test 데이터[9] 모두를 학습 데이터로 활용하였으며, DPhi에서 제공하는 데이터[10] 중 학습 데이터를 검증 데이터로 사용하며, 테스트 데이터 셋으로는 DPhi에서 제공하는 테스트 데이터 셋과 추가 방사선학 이미지를 제작하여서 모델의 대한 성능 평가를 두 가지의 테스트 데이터 셋을 사용하였으며, 이미지 파일의 사이즈는 모두 224x224사이즈를 기준으로 하였다. 데이터 셋의 상세 구성은 그림 11, 그림 12, 그림 13과 같이 구성되어 있으며, 표 7에서 사용된 데이터 셋의 상세 구성 분포를 보면 ClsKLData와 KneeXray Train, KneeXray Test는 거의 동일한 구성을 가지고 있으나, 추가 Test 데이터 셋과는 다른 구성을 가지고 있는 것을 알 수 있다.

추가 방사선학 무릎 이미지는 DICOM파일로 구성되어 있는 것을 JPG로 변환 후 임의적으로 무릎의 연골에 해당하는 부분을 Crop하고 이미지의 사이즈를 맞추기 위하여 224x224 사이즈로 그림 14와 같이 Resize하여 구성하였다.

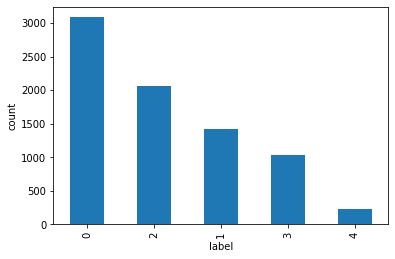
그림 11. ClsKLData 모든 데이터 셋 구성 분포도



(표 4) ClsKLData 상세 구성

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ClsKLData | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 |
| Auto\_test | 604 | 275 | 403 | 200 | 44 |
| Test | 639 | 296 | 447 | 223 | 51 |
| Train | 2286 | 1046 | 1516 | 757 | 173 |
| Val | 328 | 153 | 212 | 106 | 27 |
| Total | 3857 | 1770 | 2578 | 1286 | 295 |

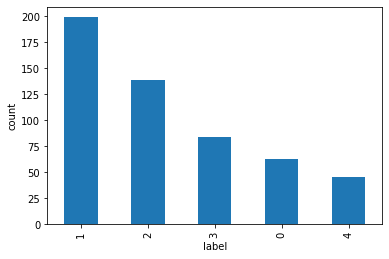
그림 12. KneeXray Train 데이터 셋 구성 분포도



(표 5) KneeXray 상세 구성

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| KneeXray | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Total |
| Train | 3085 | 1416 | 2062 | 1029 | 236 | 7828 |
| Test | 1958 | | | | | 1958 |

그림 13. 추가 테스트 데이터 셋 구성 분포도



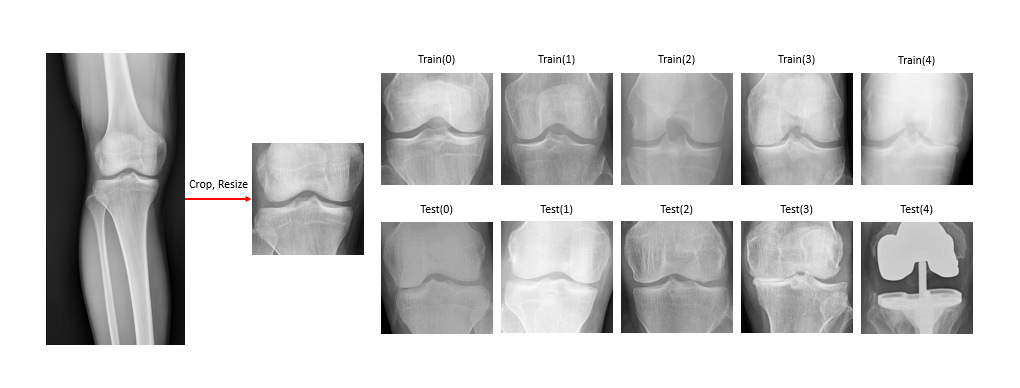
(표 6) 추가 테스트 데이터 셋 상세 구성

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Total |
| Plus Data | 63 | 199 | 139 | 84 | 46 | 531 |

(표 7) 사용된 전체 데이터 셋의 라벨 별 구성(%)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Total Data | Grade 0 | Grade 1 | Grade 2 | Grade 3 | Grade 4 | Total |
| ClsKLData | 39.41(%) | 18.09(%) | 26.34(%) | 13.14(%) | 3.01(%) | 100(%) |
| KneeXray train | 39.42(%) | 18.07(%) | 26.35(%) | 13.12(%) | 3.01(%) | 100(%) |
| KneeXray Test | 39.43(%) | 18.08(%) | 26.35(%) | 13.13(%) | 3.01(%) | 100(%) |
| Plus Data | 11.86(%) | 37.48(%) | 26.18(%) | 15.82(%) | 8.66(%) | 100(%) |

그림 14. 학습과 테스트에 사용된 데이터셋 예시



본 논문에서의 실험은 표 8, 9와 같은 환경에서 진행되었으며, 학습 진행에 참조한Xception을 활용한 학습 방법은 본래의 google의 구성되어 있는 Xception 모델 끝단의 3개의 Filter Layer를 추가하였으며, 그림 16와 같이 전체 아키텍처를 나타낼 수 있으며, 해당 실험에서의 Batch size는 32로 진행되었으며, 기존의 Xception 모델 끝단의 Filter Layer를 추가하였지만, 본 논문의 실험에 사용된 모델은 그림 17과 같이 나타낼 수 있으며, 기존 Xcepion 모델의 끝단 Layer를 수정하여 그림 17과 같이 구성하여 학습을 진행하였다.

학습 방법은 두 가지 방법 모두 데이터셋 구성의 편차가 심하기 때문에 가중치 부여를 통해서 학습이 진행되었다.

학습 과정에서는 우선 ClsKLData 데이터 셋을 학습 데이터로 사용하고 학습 과정에 있어서 검증을 위해 KneeXray Train을 검증 데이터로 사용하여 학습이 1차적으로 진행되며, 이후로 2차적 학습으로 1차 학습에서 검증에 사용되었던 KneeXray Train 데이터 셋을 추가 학습에 사용하여 모델 학습을 완료하였다.

(표 8) 실험 환경

|  |  |
| --- | --- |
| **Experiment** | |
| OS | Windows 10 |
| GPU | GeForce RTX 3090 |
| CUDA | 11.2 |
| cuDNN | 8.1 |
| Tensorflow | 2.6.0 |
| Tensorflow-gpu | 2.6.0 |
| Keras | 2.6.0 |

(표 9) 세부 매개 변수

|  |  |
| --- | --- |
| **Hyperparameter** | |
| Model | Xception |
| Batch size | 8 |
| Optimizer | Adam (learning rate = 0.00001, decay = 0.0001) |
| Train Epochs | 135 (Early stopping) |
| Plus Train Epochs | 50 |
| ImageDataGenerator | Rescale = 1./255 |
| Input image size | 224x224 |
| Pre tarin | Imagenet |

그림 15. Xception 구조

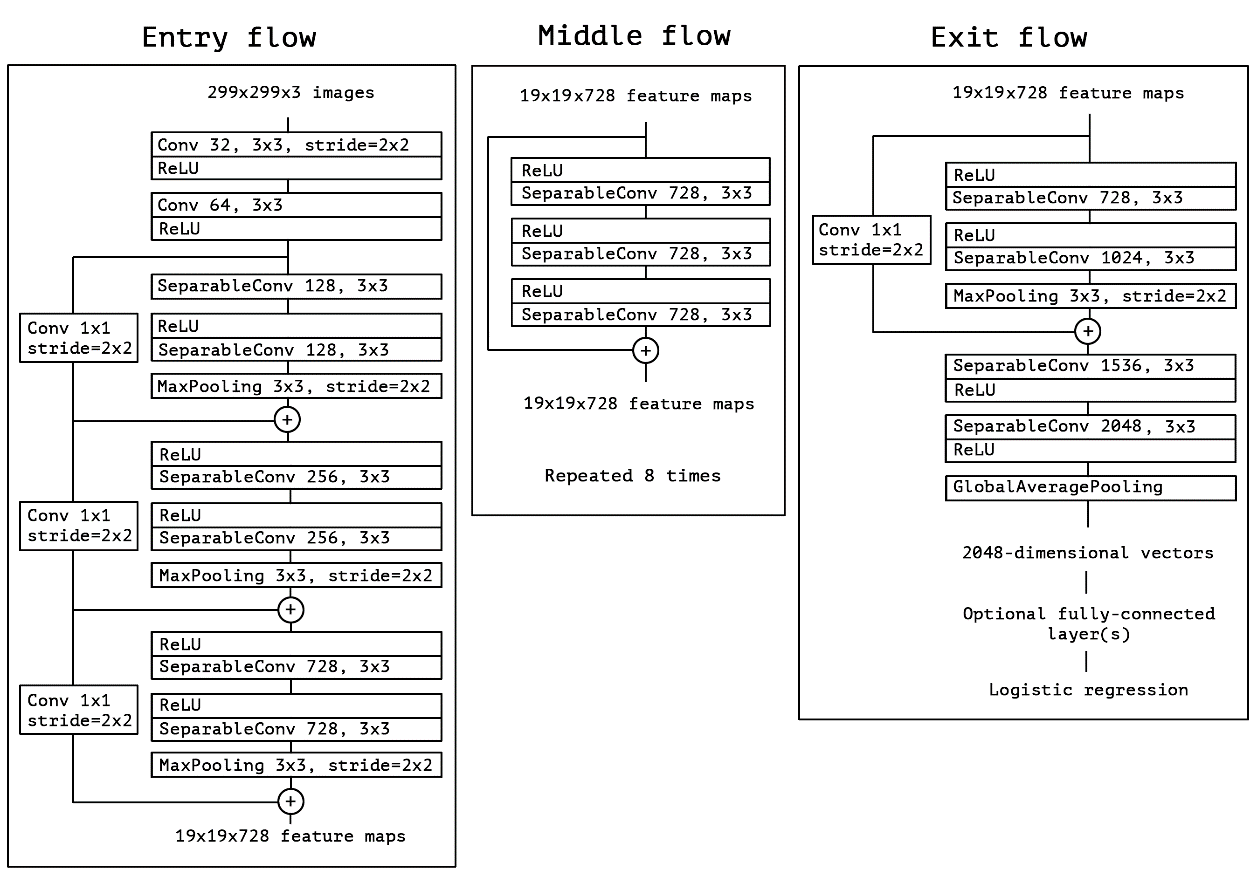


그림 16. 참조한 Xception 모델 구조

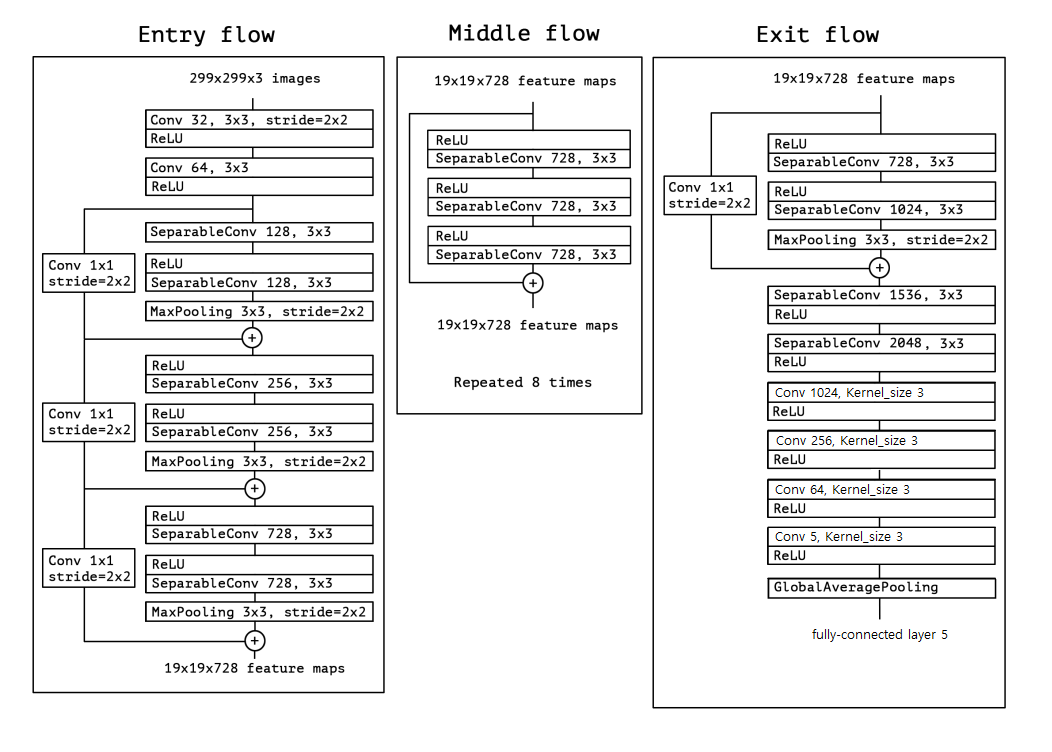
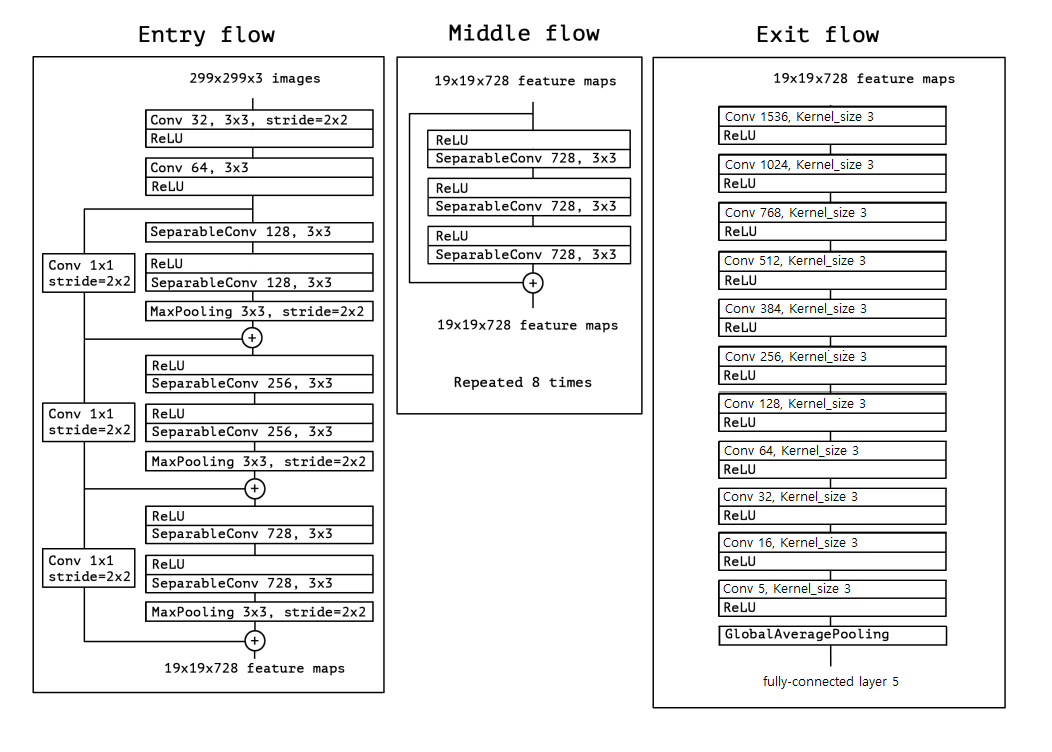


그림 17. 변경된 Xception model 구조



모델의 성능 평가를 위해서 검증 데이터와 추가 테스트 데이터에 대한 혼돈 매트릭스와 DPhi에서 제공한 KneeXray test 데이터 셋에 대한 정확도 검증을 하였다.

그림 18, 19, 20, 21, 22를 통해 전체적인 성능을 시각화로 나타냈으며, 참조한 모델과 개선한 모델들의 성능이 기존의 테스트 데이터에서는 엄청나게 높은 성능을 보였지만,

추가 테스트 데이터 세트에 대해서는 기대치에 미치지 못한 성능을 보이는데 표 7을 참조에서 전체적인 분포를 볼 수 있듯 학습 데이터의 분포는 0 > 2 > 1 > 3 > 4 로 구성 되어있으며, DPhi의 데이터 셋 또한 분포는 동일하게 0 > 2 > 1 > 3 > 4 로 구성 하였다.

학습, 검증 테스트 데이터 세트가 동일한 분포를 가지고 있었고, 분포가 거의 같기 때문에 Data leakage 현상으로 인해 정확도가 높게 나왔으며, 추가 테스트 데이터 셋에 대한 검증에 있어 추가 테스트 데이터 셋의 분포가 1 > 2 > 3 > 0 > 4 순으로 학습, 검증에 사용된 데이터 셋과 다른 분포를 가지고 있으나, 표 11를 보면 0 > 2 > 3 > 1 > 4로 학습 데이터의 분포와 유사하게 예측을 진행하는 것을 보아 Data leakage 현상이 발생하였다는 것을 발견하였습니다.

그림 18. 참조한 모델의 혼돈 매트릭스

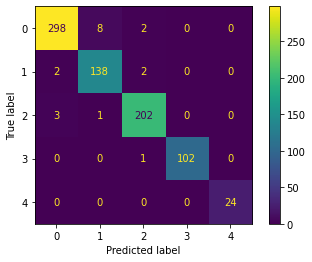


그림 19. 참조한 모델에 추가 테스트 데이터에 대한 혼돈 매트릭스

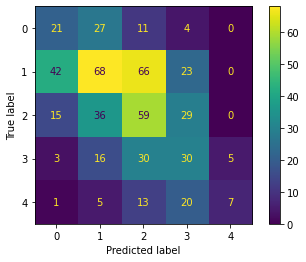


그림 20. 실험이 진행된 모델의 혼돈 매트릭스

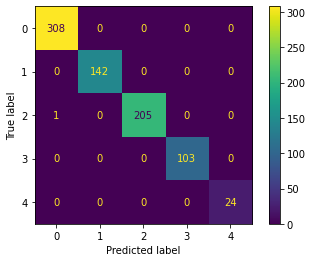


그림 21. 추가 테스트 데이터에 대한 혼돈 매트릭스

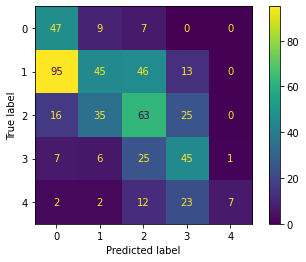
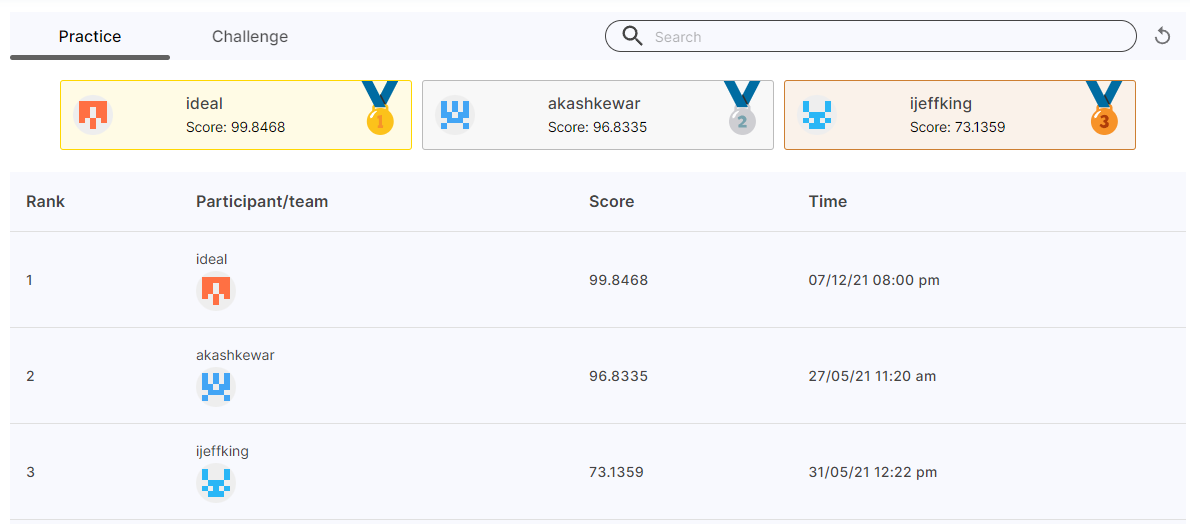


그림 22. DPhi Pratice score Recode



(표 10) 전체적인 성능

|  |  |  |
| --- | --- | --- |
|  | 정답(%) | 오답(%) |
| 그림 18 | 96.8335(%) | 3.1665(%) |
| 그림 19 | 34.8(%) | 65.2(%) |
| 그림 20 | 99.8468(%) | 0.1532(%) |
| 그림 21 | 38.98(%) | 61.02(%) |

(표 11) 학습된 모델이 예측한 결과 값의 분포

|  |  |  |
| --- | --- | --- |
| Label | 사용된 데이터의 개수 | 모델이 예측한 개수 |
| 0 | 63 | 167 |
| 1 | 199 | 97 |
| 2 | 139 | 153 |
| 3 | 84 | 106 |
| 4 | 46 | 8 |

결과적으로, 99.80%로 엄청나게 높은 정확도를 보였지만, 높은 정확도에 대한 원인은 Data Leakage 현상과 학습 데이터와 테스트 데이터의 분포가 동일하기 때문이다.

이러한 현상에 대한 검증을 진행하기 위해 기존의 학습 데이터는 유지하고 학습에 사용하고 추가 테스트 데이터에 대해 데이터 증강을 진행 후 구성 분포를 기존과 다르게 구성 한 후 검증 데이터로 사용하였으나, 학습 과정에서 validation loss가 계속해서 40% 내외를 유지하며 학습이 이루어지지 않았습니다.

그림 19와 같이 추가 테스트 데이터에 대해 검증을 위해서 참고하였던 모델에도 적용해 보았으나, 오히려 정확도가 더 떨어지는 것으로 보았을 때 추가 테스트 데이터 셋의 제작이 잘 못 되었고 보기는 어렵기에, Data Leakeage 현상으로 인해 야기된 정확도라고 추론하였다.

1. 결론

기존의 다른 연구에서 사용되었던 VGG19 모델과 같은 이미지 분류에서 저명한 모델들에 비해 Xception 모델의 성능이 앞서는가에 대한 모델 성능 비교를 진행하였다. 결과 적으로 추가 데이터 셋에 있어서 VGG19 모델의 성능을 앞서지 못하였다고 할 수 있으며, 이는 다른 무릎 관절 방사선학 이미지 데이터 셋을 활용하여서 더욱 테스트를 해보아야 명확하게 검증이 가능하기에 추후에 진행할 예정이며, 같은 데이터 셋의 구성으로 VGG19 모델에 적용하여 추가 테스트 데이터 셋에 대한 검증도 진행할 예정이다.

1. Acknowledgement

이 논문은 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2020R1A6A1A03040583)

`

참고 문헌

[1] https://www.snubh.org/dh/main/index.do?DP\_CD=JRC&MENU\_ID=002008024

[2]https://globaljnt.tistory.com/entry/%EB%AC%B4%EB%A6%8E-%EC%97%B0%EA%B3%A8%EC%9D%98-%EC%86%90%EC%83%81-%EB%8B%A8%EA%B3%84

[3] https://m.blog.naver.com/anzyme/221149697273

[4] http://www.snuh.org/health/nMedInfo/nView.do?category=DIS&medid=AA000196

[5] Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss,

https://github.com/PingjunChen/GradingKneeOA

[6] Automated Classification of Radiographic Knee Osteoarthritis Severity Using Deep Neural Networks

[7] Automatic Detection of Knee Joints and Quantiﬁcation of Knee Osteoarthritis Severity

[8] <https://dphi.tech/challenges/data-sprint-35-osteoarthritis-knee-x-ray/81/overview/about>

[9] <https://data.mendeley.com/datasets/56rmx5bjcr/1>

[10] <https://drive.google.com/file/d/1NdDqPK4NLn2aV8ZdF5ilux1sfG6IyebC/view>

[11] https://dphi.tech/notebooks/1680/akashkewar/xception-architecture-with-imagenet-finetuned-on-kaggle-dataset

[github] https://github.com/d9249/MDL/tree/main/Term%20project/%EC%B5%9C%EC%A2%85%20%EB%B3%B4%EA%B3%A0

[학습 Log] <https://github.com/d9249/MDL/blob/main/Term%20project/%EC%B5%9C%EC%A2%85%20%EB%B3%B4%EA%B3%A0/train.ipynb>

[data leakage] https://m.blog.naver.com/hongjg3229/221811766581