2024 연세 디지털헬스케어 사이버 보안 경진대회

의료 영상자료 오염 탐지 및 모델 고도화

연세대학교 디지털 애널리틱스 융합협동과정
유민균, 서지혜, 강호승, 김시원, 서동혁

Contents

01

서론

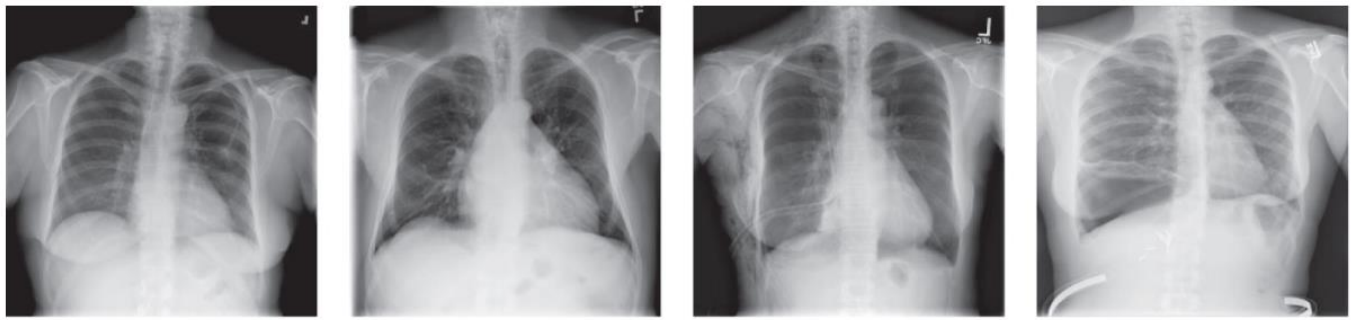
- 주제 개요
- 데이터 설명
- 접근 방법

주제 개요

목표

- 의료영상자료와 같은 의료 데이터의 오염 문제는 모델의 성능과 신뢰성을 낮추는 대표적인 문제. 이는 환자의 진단과 치료의 발향을 결정하는데 심각한 방해 요소.

- 자료 오염된 상태 에서도 강건한 예측이 가능한 인공지능 개발.
  - 오염된 이미지 있는 상태에서도 정상 / 질환, 질환 세부 **분류 성능 고도화**
  - 오염 종류 탐지 및 오염 데이터 **탐지 성능 고도화**



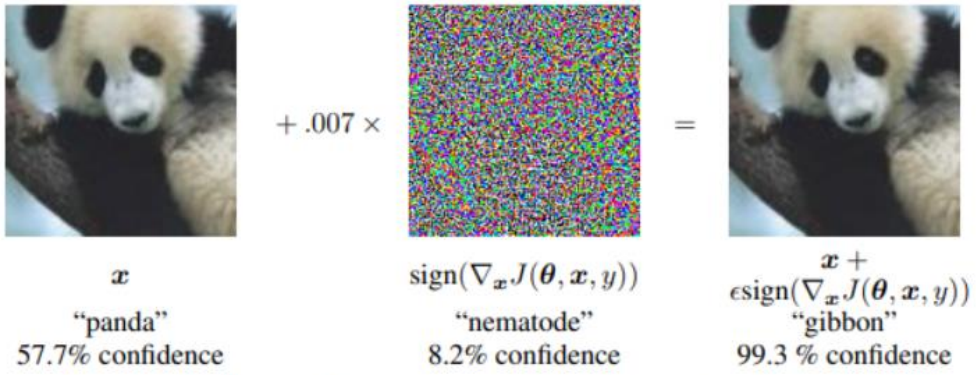
(a) "No Finding" **(b)** "Cardiomegaly" **(c)** "Pneumothorax" **(d)** "Pneumothorax"

주제 개요

데이터 오염의 유형

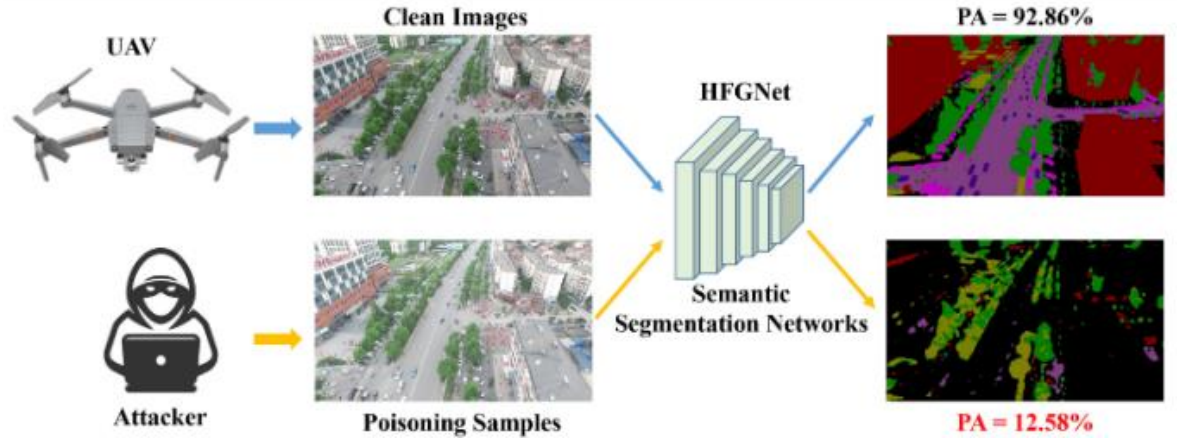
- 노이즈 추가 - 영상에 의도치 않은 노이즈가 추가되어 이미지의 품질 저하, 모델이 중요한 의료 정보 제대로 인식 x
- 라벨 오류 – 영상 자료가 잘못된 라벨과 연결되는 경우, 모델이 잘못된 예측을 할 가능성 커짐
- Poison Attack – 훈련 데이터에 고의적으로 악성 데이터를 삽입하여 모델의 성능 저하

<노이즈 추가에 대한 예시>



출처: Ian J. Goodfellow et al. "Explaining and Harnessing Adversarial Examples"

<Poison Attack에 대한 예시>



주제 개요

의료 분야에서 Poison Attack이 야기한 문제들?

- Ma et al. (2019)[1]의 연구에서는 공격자가 의료 영상 데이터의 라벨을 조작하여, 암 진단 모델이 정상 조직을 암으로 잘못 분류하거나, 반대로 암 조직을 정상으로 분류하도록 만들 수 있음을 실험적으로 증명
- 의료 시스템은 점점 더 많은 부분이 디지털화되고 있으며, 원격 진료와 같이 데이터의 전송 및 처리 과정이 복잡해지는 환경에서는 공격이 더욱 쉽게 이루어질 수 있음

의료 분야에서 Poison Attack 방어와 관련된 연구

- Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. "Certified defenses for data poisoning attacks." *Advances in neural information processing systems* 30 (2017).

→ 데이터의 일부가 공격에 의해 변조된 경우에도 안정적인 성능을 유지할 수 있는 알고리즘을 제안

- Alzubaidi, Laith, et al. "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging." Intelligent Systems With Applications 22 (2024)

-> 다양한 poison attack 을 적용한 이미지들을 만들고 그에 따른 여러 모델을 만들어 adversarial traing 을 진행하여 간건한 모델을 만듬

[1] Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." arXiv preprint arXiv:1804.05296 (2018).

Related work

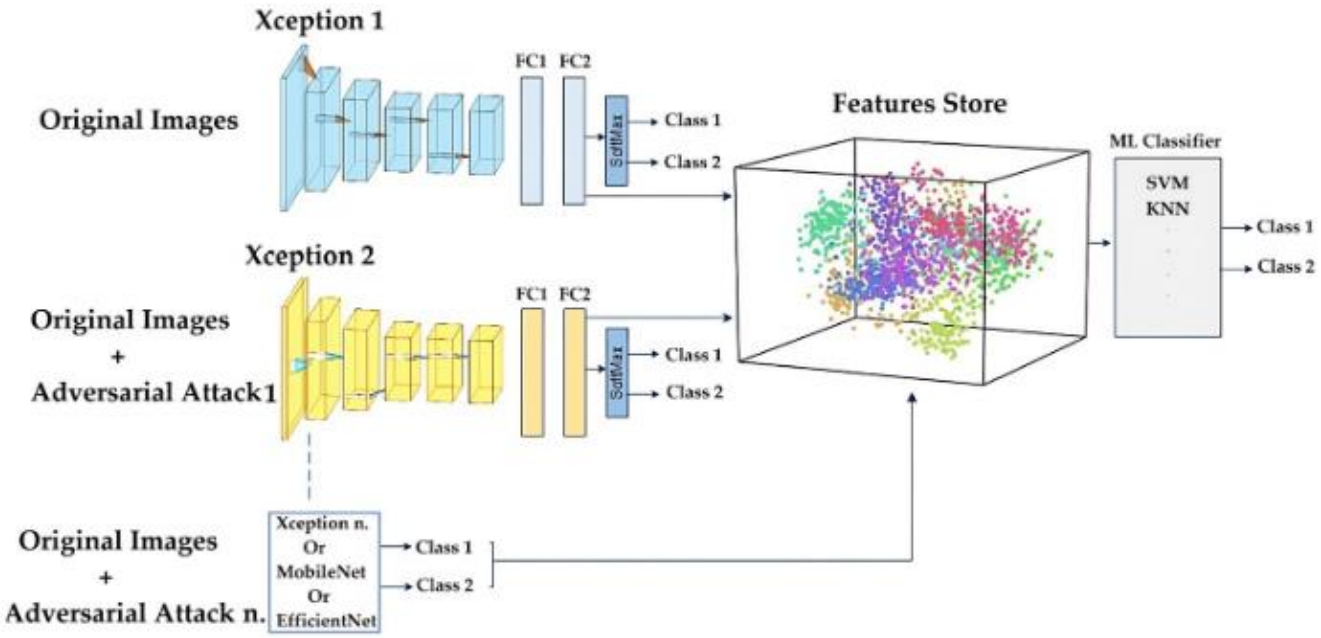


Fig. 2. Workflow of the proposed solution consisting of MEFF framework and inputs of original images and with various adversarial attacks.

- 직접 adversarial attack 을 만들어 내고 강건한 모델이 되도록 학습 (대부분의 선행연구 이런 식의 접근방법)
- 이런 접근 방식을 Adversarial Training 이라고 하고 poison data는 test셋에 있어서 성능 강건성을 평가함
- 우리 공모전 주제2는 처음 부터 poison 데이터가 훈련 데이터셋으로 존재.

Alzubaidi, Laith, et al. "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging." Intelligent Systems With Applications 22 (2024)

접근 방법

Detection

데이터 오염을 정확히 식별하여 모델의 신뢰성을 확보

Detection은 의료영상자료 내에 존재하는 오염된 데이터를 식별하고 제거하는 것을 목표로 한다. 이는 데이터의 신뢰성을 유지하기 위해 필수적인 단계이며, 오염된 데이터를 포함한 학습은 모델의 성능을 왜곡시키고 실제 적용 시 오류를 유발할 수 있다. 탐지 방법으로는 통계적 모델링을 하여 이상치를 탐지한다.

Robustness

오염된 데이터에도 모델의 일관된 성능을 유지하는 능력을 강화

Robustness는 인공지능 모델이 오염된 데이터나 예기치 않은 입력에 직면했을 때도 안정적이고 일관된 성능을 유지할 수 있도록 하는 과정이다. 우리는 ETF 학습 기법, Collaborate learning, TTE 등을 통해 달성한다. 이러한 기법들은 모델이 다양한 환경에서 발생할 수 있는 노이즈와 변동성에 대한 저항성을 키워주며, 결과적으로 모델의 실용성과 신뢰성을 크게 향상시킨다.

Contribution

- **Medical image에 ETF Classifier와 RECT의 최초 적용**: 본 연구는 메디컬 이미지 분류에 ETF Classifier와 RECT를 처음으로 적용한 연구이다. 이 두 가지 방법론을 통합하여 모델의 강건성을 향상시키고, 보다 안정적인 의료 의미지 분석을 가능하게 했다.
- **Collaborative Learning의 Poisoning Attack Defense 최초 적용**: 본 연구는 Poisoning Attack에 대한 방어 수단으로서 협업 학습(Collaborative Learning)을 최초로 적용한 연구이다. 이러한 접근은 Poisoning Attack 상황에서도 모델의 성능을 유지하는 데 효과적임을 보여준다.
- **새로운 Robustness 평가 방법 제안**: 우리는 기존의 평가 방법과는 다른, 새로운 성능 감소 평가 방법을 제안하였다. 기존에는 같은 poison attack 된 상황에서 모델 성능을 비교하였지만, 우리는 모델의 정확도, F1 스코어, 정밀도의 변화를 종합적으로 평가하여 모델의 강건성을 보다 직관적으로 이해할 수 있게 하였다.

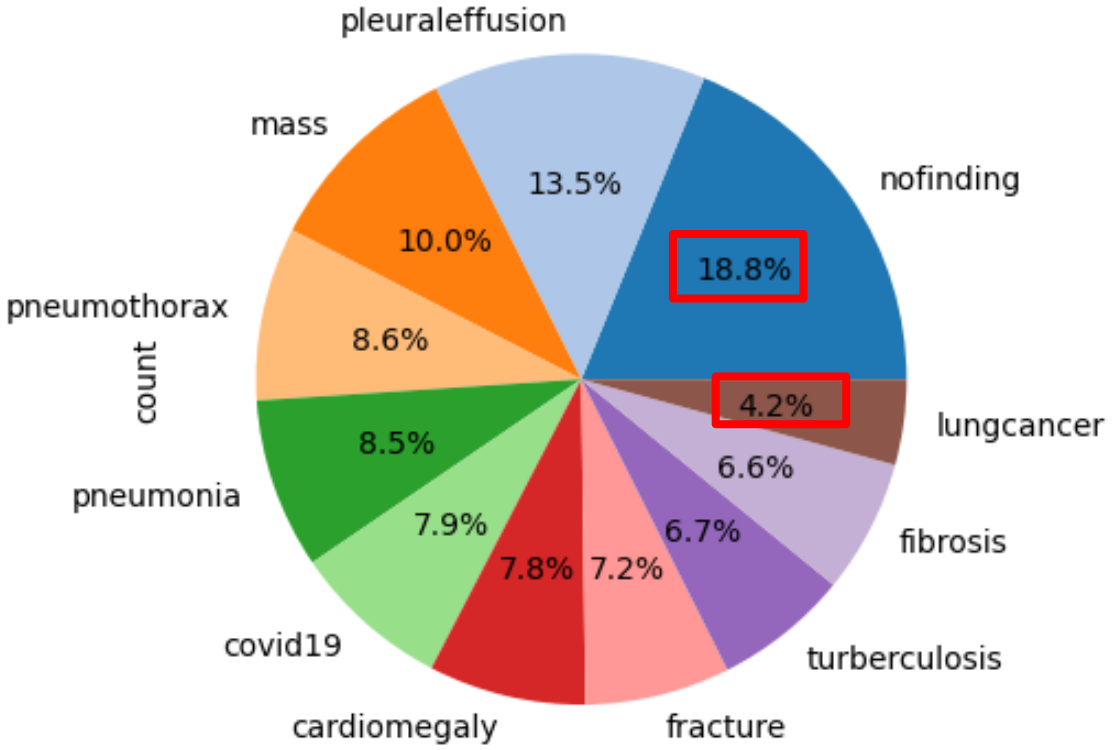
02

분석 프레임워크

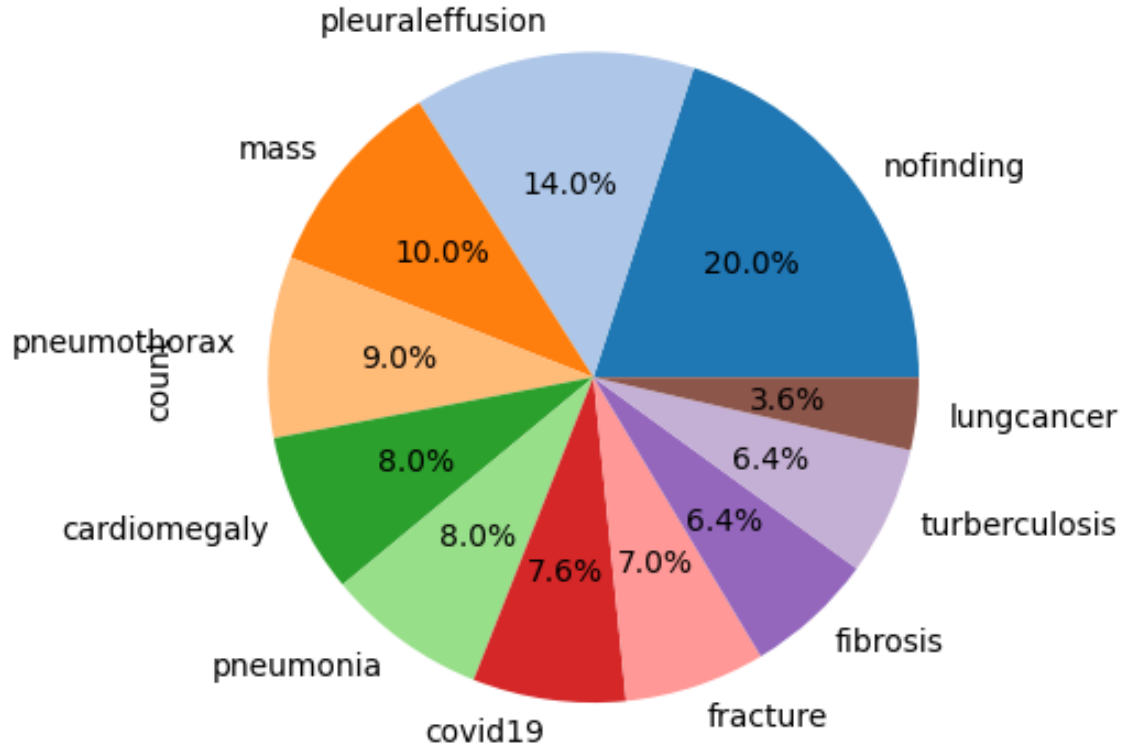
- EDA
- Modeling

EDA Label별 분포

- Data 1(Train)



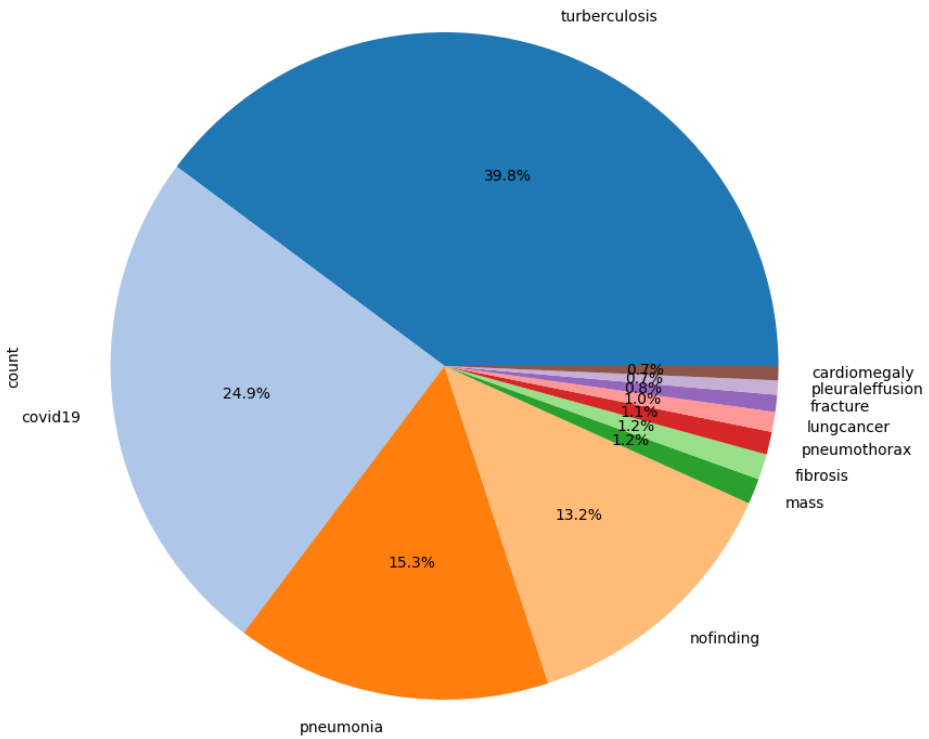
- Data 2(Test)



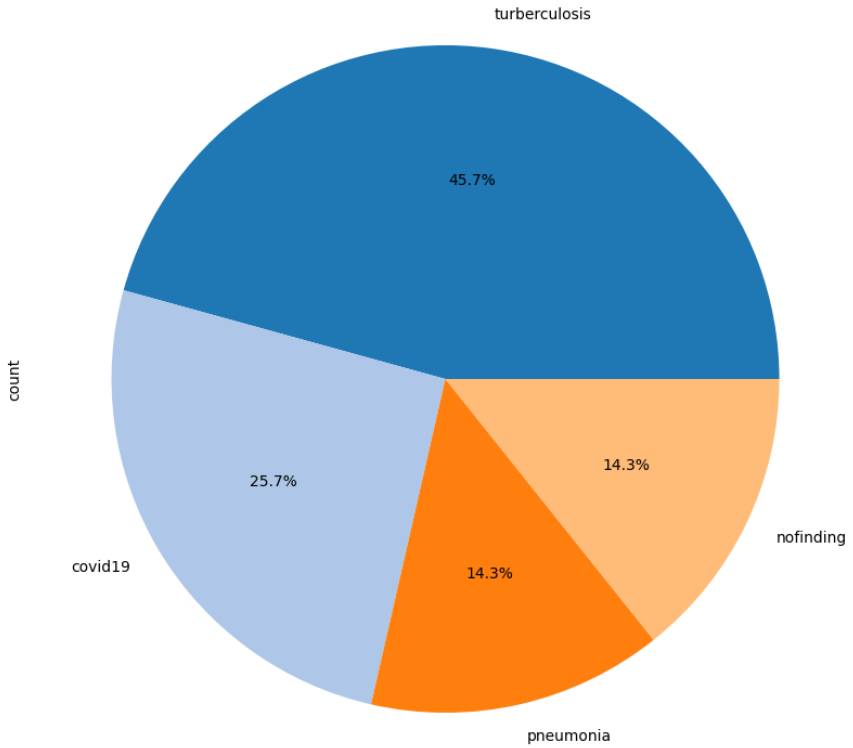
✓ 학습용 데이터와 평가용 데이터의 Label 분포의 차이는 거의 없음
✓ 하지만 각 데이터 내부에서 일부 Label에 대해 차이가 존재.

EDA Label별 분포 (성별, AP_PA가 Null인 값들)

- Data 1(Train)



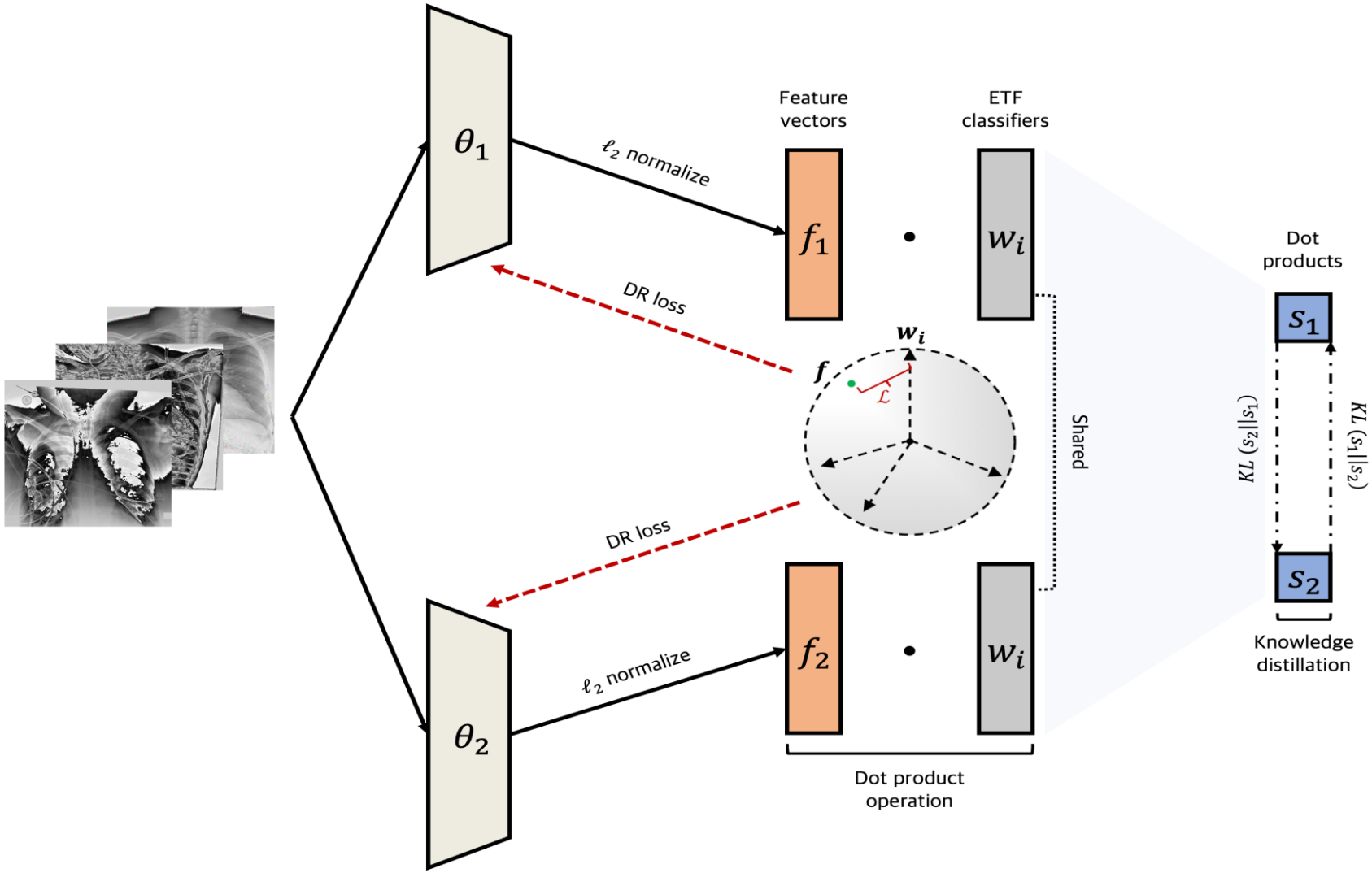
- Data 2(Test)



✓ 학습용과 평가용 데이터 중 성별, AP_PA가 Null인 데이터의 대부분은 turberculosis, covid19, pneumonia, nofinding이 차지.

Modeling

Figure 1 _ Overview of the Proposed Method

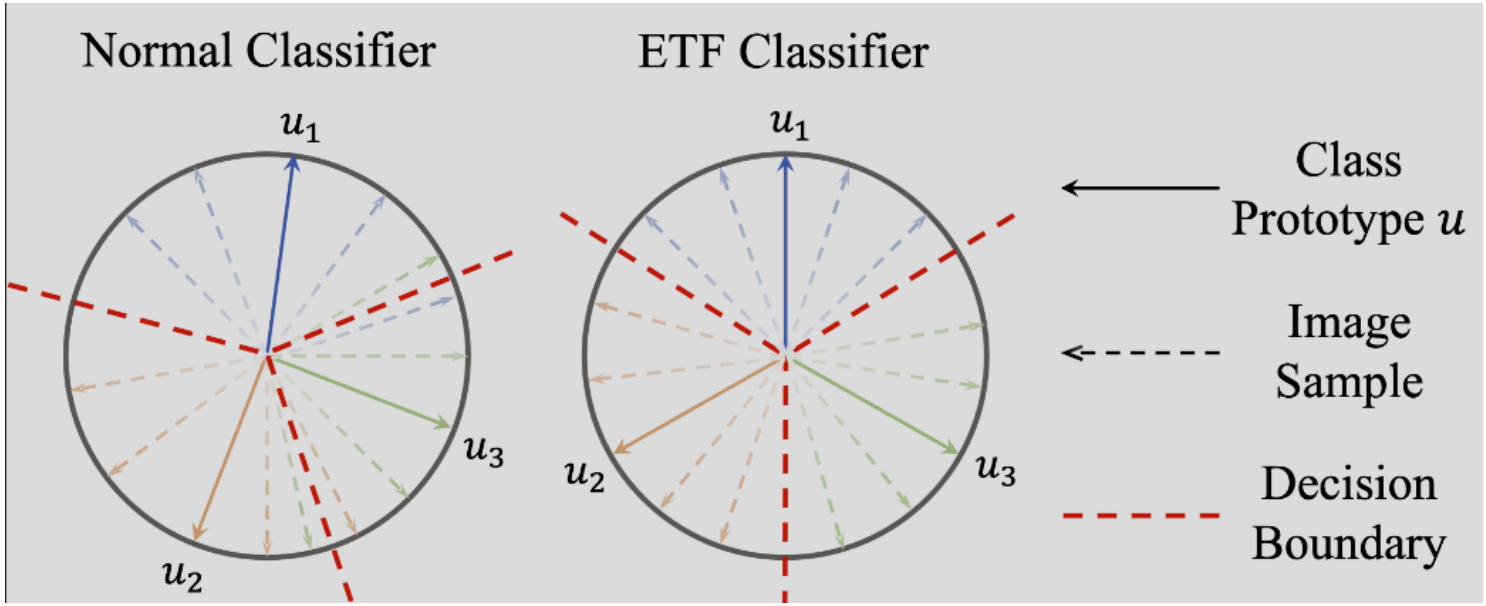


- 이 Figure는 두 개의 ETF Classifier를 사용한 DML 학습 과정을 보여준다.
- 두 모델은 각각의 특징 벡터를 ETF 분류기에 전달하며, DR 손실을 통해 더 나은 학습을 유도한다.
- 지식 증류를 통해 두 모델 간의 학습을 공유한다.

Modeling

ETF Classifier

- ETF (Equiangular Tight Frame) 분류기는 서로 다른 클래스의 특징 표현이 잘 분리되고, 특징 공간에서 균일하게 분포되도록 하는 기술이다. 이는 클래스 간의 혼동을 최소화하는 데 유용하며, 특히 서로 가까운 클래스들 사이의 오류를 줄이는 것이 중요한 환경에서 사용된다.



Yang, Yibo, et al. "Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?." *Advances in neural information processing systems* 35 (2022)

Modeling

Rectification (Rect)

- Rect는 학습 중 클래스의 특징 벡터를 조정하는 방법이다. 이 방법은 ETF 벡터에 연산을 적용하여 클래스 간의 분리를 더욱 세밀하게 조정한다. Class 수가 적은 거에 더 긴 벡터 길이가 되도록 조정하였다. 이를 통해 Class imbalance 문제를 해결한다.

$$\gamma = \frac{B}{U} \qquad \mathrm{rect}(i) = \sqrt{\frac{\gamma}{c_i}}$$

B: Batch size, U: Batch에 있는 클라스 수

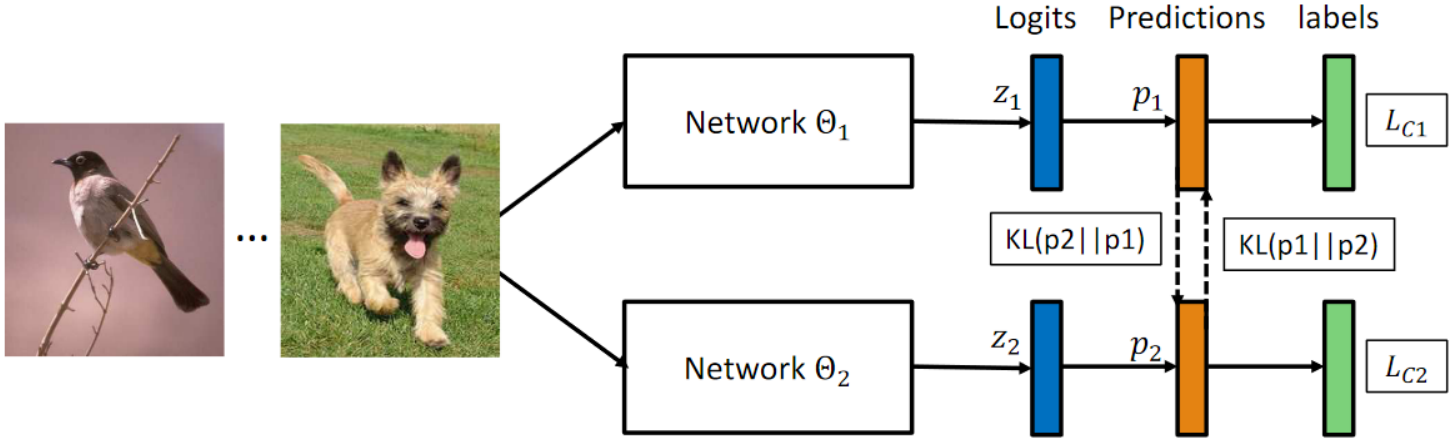
C_i: i클라스의 수이다.

- 이 수식은 클래스 클라스 의 빈도 C_i 에 반비례하는 가중치를 부여함으로써, 클래스가 적게 나타날수록 더 큰 가중치를 주어 모델 학습 시 모든 클래스가 균형 있게 학습되도록 유도하는 역할을 한다.

Yang, Yibo, et al. "Neural collapse inspired feature-classifier alignment for few-shot class incremental learning." *arXiv preprint arXiv:2302.03004* (2023).

Deep Mutual Learning

- 딥 뮤추얼 러닝(DML)은 여러 모델이 동시에 학습하며 서로의 지식을 공유하는 기법이다. 모델들이 독립적으로 학습하는 대신, DML은 서로의 예측을 학습 과정에 포함시켜, 모든 모델의 일반화 능력을 향상시킨다. DML은 앙상블 효과 극대화를 위해 훈련데이터에 과적합 되도록 학습
- ETF Classifier를 DML과 함께 사용하기 위해 우리는 모델에서 출력되는 벡터와 정답 벡터간의 내적 값 간에 Knowledge distillation 을 진행한다



Zhang, Ying, et al. "Deep mutual learning." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

평가방법

Combined Performance Drop Metric

- 기존 연구에서의 단순히 Poison 상황에서 모델 간의 성능 비교는 Robustness 측정하는데 적합하지 않음. [3][4]
- 성능이 높은 건 그냥 제안한 모델의 이미지 분류 성능이 높은 거 일 수 있음
- 우리는 poison attack 된 상황과 poison attack 안된 상태에서 모두 실험하고 모델의 성능이 얼마나 떨어지는지 비율을 계산
- Accuracy 뿐만 아니라 Precision, f1 score 등의 drop rate을 가중 평균 계산

[3]Kim, W. J., Cho, Y., Jung, J., & Yoon, S.-E. (2023). *Feature Separation and Recalibration for Adversarial Robustness*. 8183–8192
[4] Perez, J. C., Alfarra, M., Jeanneret, G., Rueda, L., Thabet, A., Ghanem, B., & Arbelaez, P. (2021). Enhancing Adversarial Robustness via Test-time Transformation Ensembling. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 81–91

평가방법

Combined Performance Drop Metric

- 각 성능 지표 M에 대해, 모델이 정상 데이터에서의 성능 M_{normal}과 공격 또는 노이즈가 적용된 데이터에서의 성능 M_{attack} 사이의 백분율 감소를 다음과 같이 계산.

$$\text{Drop}_M = \frac{M_{\text{normal}} - M_{\text{attack}}}{M_{\text{normal}}} \times 100$$

- 모든 성능 지표 $M_1, M_2, \ldots, M_n$ 에 대해 각각의 백분율 감소를 계산한 후, 이들의 평균을 구해 복합 성능 감소 지표를 도출.

$$\text{Combined Drop} = \frac{1}{n} \sum_{i=1}^{n} \text{Drop}_{M_i}$$

03

실험 결과

- 실험 셋팅
- 실험 결과

실험 설정

설정 요약

- 데이터 분할: 7:2:1 (Training:Validation:Test)
- Optimizer: Adam (LR: 0.0001, Weight Decay: 0.001)
- Early Stopping: Patience = 20
- Scheduler: ReduceLROnPlateau
- Experiment 1: 원래 데이터로 학습 및 테스트
- Experiment 2: 원래 데이터로 학습 후 테스트 데이터에 가우시안 노이즈 추가 (Mean: 0.0, Std: 0.01)

실험 결과 – Experiment 1

Model	Accuracy	F1 Score	Precision
ResNet-50 (Baseline)	0.384	0.314	0.288
ResNet-50 + ETF	0.443	0.365	0.357
ResNet-50 + ETF +Rect	0.463	0.399	0.420
ResNet-50 + ETF +Rect +DML (Ours)	**0.467**	**0.449**	**0.469**

Table1. Performance comparison of different models on medical image classification task

우리가 제안한 방법이 기존 Baseline 대비 모든 측면에서 성능이 높은 것을 확인 하였다. Accuracy 기준으로는 9% point 증가하였고 F1, Precision등은 15% point 정도 증가하였다.

실험 결과 – Experiment 2 (Poison attack)

Model	Accuracy	F1 Score	Precision
ResNet-50 (Baseline)	0.240	0.171	0.260
ResNet-50 + ETF	0.314	0.280	0.390
ResNet-50 + ETF +React	**0.332**	0.293	**0.283**
ResNet-50 + ETF +React +DML	0.315	**0.376**	0.287

Table1. Performance comparison of different models on medical image classification task

모든 Test data에 Poison attack을 하였을 때 Baseline 보다 우리가 제안한 알고리즘이 강건성이 보였다. 그러나 DML 적용한 것과 안 한것의 성능 차이가 지표별로 결과가 달라서 어떤 것이 강건한 모델인지 명확하지 않다. 그렇기에 우리게 제안하는 Combination drop rate 성능평가방법이 필요했다.

실험 결과 – Combined Performance Drop Metric 사용

Model	Combination Drop	Accuracy Drop	F1 Score Drop	Precision drop
ResNet-50 (Baseline)	12.557%	37.5%	14.3%	2.8%
ResNet-50 + ETF	9.72 %	29.12%	8.5%	**- 3.3 %**
ResNet-50 + ETF +Rect	**9.51%**	**28.29%**	10.6%	13.7%
ResNet-50 + ETF +Rect +DML	10.93%	32.55%	**7.3%**	18.2%

Table1. Performance comparison of different models on medical image classification task

DML을 사용하지 않는것이 combination drop이 가장 적은 drop rate을 가졌다.
우리가 제안한 방법으로 평가하면 명확히 Robustness는 DML을 사용 안 하는게 가장 좋다는 것을 보여준다.

04

결론 및 한계점

- 결론
- 한계점

결론 및 한계점

결론

- 이번 연구에서는 메디컬 이미지 분석에서의 효율적인 분류 성능과 강건한 모델을 개발하기 위해 새로운 방법론을 제안하고 평가했다.
- ETF Classifier와 RECT 접근법을 결합한 모델은 기존의 분류 알고리즘에 비해 더욱 강건한 성능을 보여주었다.
- ETF에 DML을 추가함으로써 모델의 전반적인 성능과 내성을 개선할 수 있었다.
- 특별히, 우리는 Poisoning Attack과 같은 악의적인 데이터 변형에 대해 모델이 얼마나 강건한지를 평가하기 위한 새로운 지표를 제안했다. 이 지표는 단순한 성능 저하 측정 뿐만 아니라, 다양한 측면에서의 모델의 내성을 평가할 수 있도록 하였다.

결론 및 한계점

한계점

- 본 연구는 메디컬 이미지 데이터셋에서의 모델 성능과 강건함을 중점적으로 다루었으나, 몇 가지 한계점이 존재한다.
- 첫째, 제안된 접근법은 차이가 미세한 Medical image 에 특화되어 있으며, 다른 종류의 이미지 데이터에 대해서는 추가 실험이 필요하다.
- 둘째, Poisoning Attack에 대한 방어 측면에서 제안된 DML과 같은 방법론이 효과적이었으나, 실제 임상 환경에서의 다양한 공격 시나리오에 대해 추가적인 평가가 필요히다.
- 마지막으로, 제안된 평가 지표가 모델의 강건함을 잘 측정할 수는 있으나, 다른 연구에서 널리 사용될 수 있도록 표준화 과정이 필요하다.

- End -
의료 영상자료 오염 탐지 및 모델 고도화

연세대학교 디지털 애널리틱스 융합협동과정
유민균, 서지혜, 강호승, 김시원, 서동혁

참조논문

- Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." arXiv preprint arXiv:1804.05296 (2018).
- Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang. "Certified defenses for data poisoning attacks." *Advances in neural information processing systems* 30 (2017).
- Alzubaidi, Laith, et al. "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging." Intelligent Systems With Applications 22 (2024)
- Kim, W. J., Cho, Y., Jung, J., & Yoon, S.-E.. Feature Separation and Recalibration for Adversarial Robustness. 8183–8192 (2023)
- Perez, J. C., Alfarra, M., Jeanneret, G., Rueda, L., Thabet, A., Ghanem, B., & Arbelaez, P. (2021).
- Zhang, Ying, et al. "Deep mutual learning." Proceedings of the IEEE conference on computer vision and pattern recognition. (2018).
- Yang, Yibo, et al. "Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network?." Advances in neural information processing systems 35 (2022)
- Yang, Yibo, et al. "Neural collapse inspired feature-classifier alignment for few-shot class incremental learning." arXiv preprint arXiv:2302.03004 (2023).