2024.04.15 (7주차) 미팅

2021105600 박지후
2018101819 김세한
2014110450 윤영근

1. 인간과의 일치성

Feature Map와 인간이 제공하는 Annotation(경계상자) 사이의 일치도를 평가
인간과의 일치성을 비교하기 위해, 논문에서는 두 가지 주요 지표를 사용

•EHR
임계값을 바탕으로 Feature map의 픽셀이 실제 객체 경계 내에 위치하는지 여부를 평가
다양한 임계값에서 모델의 일관성과 정확성을 검증

•EnergyPG
Feature Map 전체에서 객체 경계 상자 내외부에 할당된 중요도의 분포를 비교하여,
모델이 객체를 얼마나 중요하게 여기는지를 평가합니다.
(중요도 할당의 질적인 측면에 더 초점)

	GradCAM		IBA	
	EnergyPG	EHR	EnergyPG	EHR
Baseline	**0.566± 0.005**	**0.486 ± 0.005**	**0.645 ± 0.006**	**0.453 ± 0.004**
Cutout	0.544± 0.005	0.471 ± 0.005	0.575 ± 0.006	0.417 ± 0.004
Mixup	0.541± 0.005	0.479 ± 0.005	0.619 ± 0.006	0.412 ± 0.005
CutMix	0.528± 0.005	0.466 ± 0.005	0.570 ± 0.006	0.417 ± 0.004
SaliencyMix	0.524± 0.005	0.467 ± 0.005	0.561 ± 0.006	0.413 ± 0.004

기존의 베이스라인 모델이 인간이 그린 객체 경계 상자와 가장 일치

데이터의 중심 경향성을 나타내는 평균(또는 중앙값)을 계산하고, 데이터가 평균 주변에 얼마나 퍼져 있는지를 나타내는 표준편차를 계산

2. 모델의 충실도

충실도(Faithfulness)란 모델의 예측 결정에 대한 특징 할당(Feature Attribution) 방법의 정확성을
모델의 예측에 실제로 중요한 특징이 올바르게 강조되었는지를 평가하는 것

삽입(Insertion) 실험

삽입 실험에서는 초기에는 정보가 거의 없는 상태에서 시작하여
점차적으로 중요한 특징을 이미지에 추가하면서 모델의 예측 확률 변화를 관찰

중요한 특징을 추가할수록 모델의 예측 확률이 상승하는 것을 기대
MoRF와 LeRF의 두 가지 접근 방식을 사용하여
중요한 특징을 추가하는 순서에 따른 예측 확률의 변화를 평가

삭제(Deletion) 실험:

이미지의 일부를 순차적으로 삭제하면서 모델의 예측 확률이 어떻게 변화하는지를 관찰

이 과정에서 중요한 특징이 먼저 삭제될 때 예측 확률이 크게 떨어지는 것으로 보아,
해당 특징이 모델 예측에 중요하다고 판단

가장 중요한 특징부터 삭제하는 Most Relevant First (MoRF) 방식
가장 덜 중요한 특징부터 삭제하는 Least Relevant First (LeRF) 방식

특징의 중요도 순서에 따른 모델의 예측 성능 변화를 비교

	Inter-Model Insertion	
	GradCAM	IBA
Baseline	67.044 ± 0.0079	69.321 ± 0.0082
Cutout	**71.274** ± 0.0073	**73.595** ± 0.0073
Mixup	65.000 ± 0.0079	68.679 ± 0.0083
CutMix	63.625 ± 0.0073	67.276 ± 0.0075
SaliencyMix	64.443 ± 0.0070	68.079 ± 0.0071

	Inter-Model Deletion	
	GradCAM	IBA
Baseline	67.416 ± 0.0079	69.750 ± 0.0082
Cutout	**71.411** ± 0.0072	**73.679** ± 0.0072
Mixup	65.115 ± 0.0079	68.673 ± 0.0083
CutMix	63.908 ± 0.0072	67.890 ± 0.0074
SaliencyMix	64.711 ± 0.0070	68.348 ± 0.0071

Cutout을 사용한 모델이 다른 데이터 증강 전략을 적용한 모델들에 비해 더 높은 충실도를 보임

Cutout 증강을 사용한 모델이 모델 출력에 대한 특징 할당 방법의 정확성이 더 높다는 것을 의미

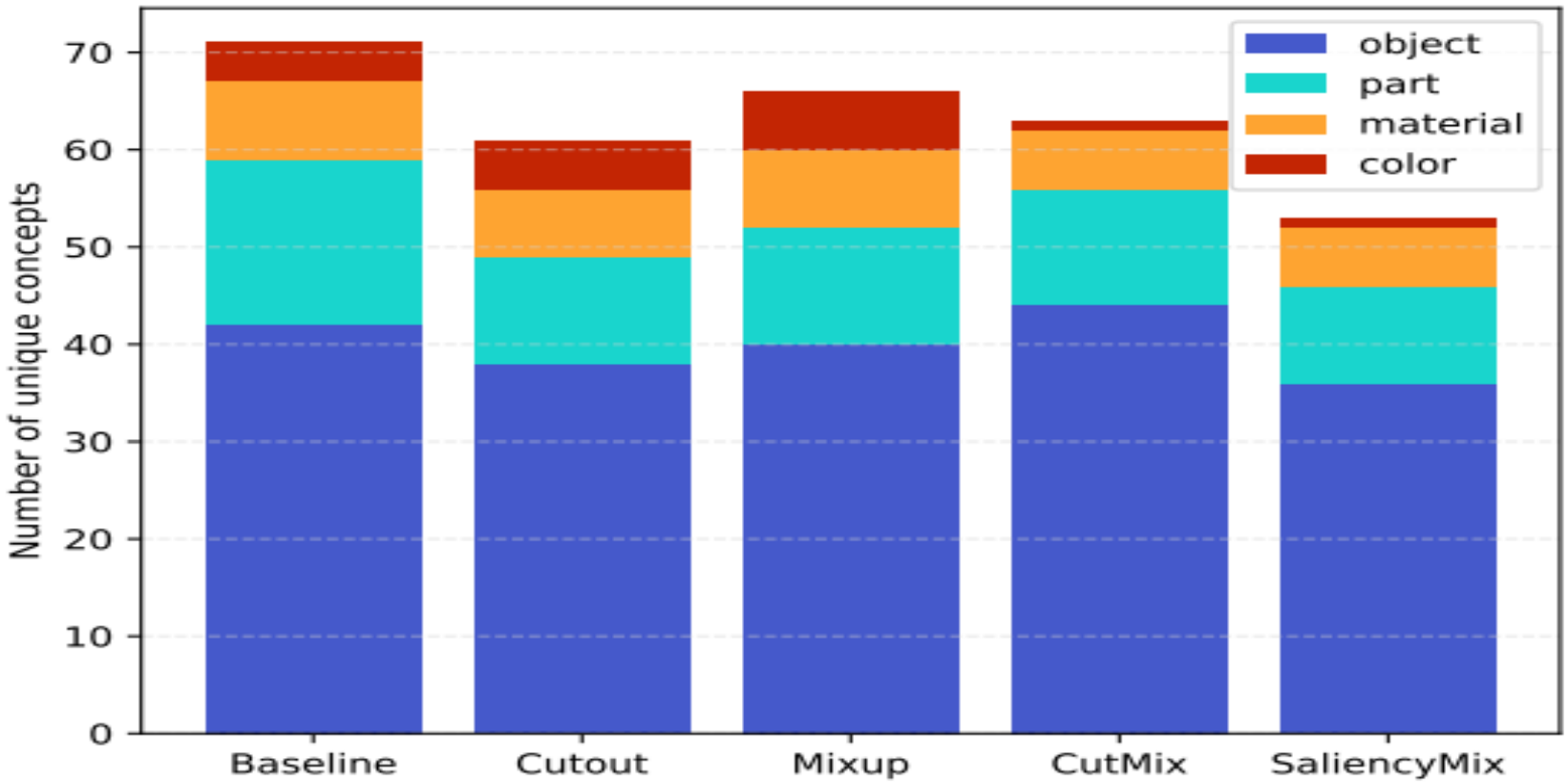
3. 인간이 인식 가능한 개념의 수

딥러닝 모델, 특히 컨볼루션 신경망(CNN)에서 개별 유닛(뉴런)이
어떤 시각적 개념을 인식하는지를 정량적으로 평가하는 데 사용됩니다.

이 방법은 모델의 각 레이어에서 활성화되는 특징 맵(feature map)을 분석하여,
특정 유닛이 특정 개념(예: 색상, 질감, 형태, 물체 등)에 반응하는지를 평가합니다.
모델을 대규모 이미지 데이터셋에 대해 실행하여 각 유닛의 활성화 패턴을 수집합니다.

이러한 활성화 패턴과 사전에 정의된 시각적 개념(예를 들어, ADE20K 데이터셋에서 제공하는 물체 분류, 색상, 재질 등) 사이의 일치성을 비교 분석

특정 유닛이 일정 임계값 이상으로 특정 개념과 일치할 때, 그 유닛을 해당 개념을 "인식"하는 것으로 간주


Number of unique concepts
0
10
20
30
40
50
60
70
Baseline
Cutout
Mixup
CutMix
SaliencyMix
object
part
material
color

4. ViT – GRADCAM 적용

```
# 모델 평가 및 Grad-CAM 설정 ###
model.eval()
grad_cam = GradCAM(model=model, target_layers=[model.blocks[-1].norm1])

# 테스트 데이터 로드 및
test_images, test_labels = next(iter(testloader)) #테스트셋에서 배치 로드
test_images = test_images.to(device)
```

4. ViT – GRADCAM 적용

```
    # 모델 예측 및 가장 확률이 높은 클래스를 대상으로 Grad-CAM 적용
    with torch.no_grad():
        outputs = model(test_images)
        predicted_classes = outputs.argmax(dim=1)  # 가장 높은 확률을 가진 클래스 인덱스
        targets = [ClassifierOutputTarget(predicted_classes[idx]) for idx in
range(test_images.size(0))]

    cams = grad_cam(test_images, targets=targets)  # Grad-CAM 적용
```

axis 3 is out of bounds for array of dimension 3

해당 오류는 Grad-CAM 라이브러리가 모델에서 반환된 그래디언트 차원을 제대로 처리하지 못하고 있음
Norm외에 attn 같은 다른 layer을 적용 해봐도 해결 안됨.

개발 현황

1. 흉부 X-ray MAE -> 어깨 X-ray(CNN, ViT)

-

2. 흉부 X-ray MAE -> 어깨 X-ray proxy -> 어깨 X-ray(CNN, ViT)

3. 흉부 X-ray MAE -> 어깨 X-ray MAE(center/random) -> 어깨 X-ray(CNN, ViT)

질문 1

1. 어깨 x-ray 이미지가 700~1000정도 사이즈. Densenet은 224 사이즈...

2. 이미지 정규화 어떻게?

```
mean = (0.5056, 0.5056, 0.5056)
std = (0.252, 0.252, 0.252)
```

3. 흉부 x-ray로 pretraining한 densenet121의 input channel=3인데 어떻게 해결? 어깨 x-ray image가 3channel이길래 일단 그대로 둠

질문 2

질문1. 저희 조는 이미 MAE와 PROXY방식을 통하여 데이터를 분석하기로 방법을 정하였는데
위의 방법들이 이와는 또 다른관점인지 궁금합니다.
(판단건데. 모델 충실도 방법에서 Deletion방법은 MAE와 방식이 상당히 비슷하여 보이고
Human-recognizable Concept의 경우 X-ray사진상 다른 사진에 비하여 인간이 인식할 수있는 개념(색상, 질감 등)이 적어보이는데
적용이 적정한지 의견이 듣고 싶습니다.)

질문2. 인간이 제공하는 주석(Annotation) 사이의 일치도를 평가하는 방식의 경우
모델이 집중해야할 특정부분을 사전학습시켜 놓은 데이터셋이 필요한걸로 보입니다.
골다공증의 진단에 필요한 정보가 사전데이터셋이 학습이 되어있는건지 궁금합니다.

질문3. 무엇보다 최종 결론으로 유의미한 절대적 수치는 없이 대조군간의 수치로만 비교를 진행하고 있습니다.
저희의 프로젝트에서도 기본 모델과의 수치비교를 통하여 통계를 내면 될지 궁금합니다.

향후 계획

4/23(화) : X
4/29?
4/30(화) : 중간 보고서 제출 마감
- Proxy / 어깨 MAE를 시도해보겠지만, 여력이 안 될 경우
  MAE없이 어깨 task(CNN, ViT)
  흉부x-ray MAE -> 어깨 task(CNN, ViT)
  비교 결과를 내서 MAE의 효과를 먼저 보여주는 게 어떨지