Enhancing Dermatological Diagnostics with Explainable AI: Integrating Vision Language Models for Improved Accuracy and Interpretability

A66041 박현호

서강대학교
SOGANG UNIVERSITY

Contents

Contents

1. 연구 배경 및 필요성

1. 의료 도메인의 딥러닝 활용 현황

- 의료 이미지 분석, 진단 지원 시스템 등 다양한 분야에서 딥러닝 기술이 활용되고 있음
- 의료 이미지 데이터 시장은 연간 5.4% (CAGR)의 지속적인 성장률을 보이며, 기술적 진보로 더 접근하기 쉽고 비용 효율적이 되고 있음
- 효과적인 활용을 위해 고정밀의 데이터가 필수적이며, 그 생성과 라벨링에 상당한 비용이 발생함

2. 의료 데이터의 특성과 문제점

- 의료 데이터는 전문적인 지식을 요하는 인력에 의해 라벨링되어야 하며, 이로 인한 **높은 인건비 발생**
- 데이터 수집의 어려움과 라벨링의 난이도로 인해 데이터 생성 비용이 증가
- 의료 진단의 정확성을 **넘어서 모델이 어떻게 결정을 내렸는지를 이해하는 해석력이 중요해짐**

3. 연구의 필요성 및 목적

- 현재 의료 데이터 활용의 어려움을 개선하고, **비용 효율적인 방법**을 모색하는 것이 시급함
- **모델의 해석력**을 높여, 의료 전문가가 AI 결정 과정을 이해하고 신뢰할 수 있도록 하는 것이 필수적임

Ref.
Your Medical Imaging Data Is Going To Waste—Here's How To Unlock Its Potential (Forbes, 2024.01)
https://www.globenewswire.com/en/news-release/2023/05/17/2670669/0/en/Medical-Imaging-Market-Generate-Revenue-Of-48-8-Billion-CAGR-Of-5-4-Worldwide-by-2032-Exclusive-Report-By-Market-us.html

1. 연구 배경 및 필요성

1. 의료 도메인의 딥러닝 활용 현황

- 의료 이미지 분석, 진단 지원 시스템 등 다양한 분야에서 딥러닝 기술이 활용되고 있음
- 의료 이미지 데이터 시장은 연간 약 5.4%의 지속적인 성장률을 보이며, 기술적 진보로 인해 점점 더 고 비용 효율적이 되고 있음
- 효과적인 활용을 위해 고정밀의 데이터가 필수적이며, 그 생성과 라벨링에 상당한 비용이 발생함

비용

해석력

2. 의료 데이터의 특성과 문제점

- 의료 데이터는 전문적인 지식을 가진 전문가에 의해 라벨링되어야 하며, 이로 인한 높은 비용이 발생함
- 데이터 수집의 어려움과 라벨링 작업의 복잡성으로 인해 생성 비용이 증가
- 의료 진단의 정확성을 넘어, 모델이 어떻게 결정을 내렸는지를 이해하는 해석력이 중요함

모델 선택 과정

Classification 모델 및 CAM 방법 선정

모델 평가 과정

선정된 모델을 VLM에 적용하여, 방법론 검증

3. 연구의 필요성 및 목적

- 현재 의료 데이터 활용의 어려움을 개선하고, **비용 효율적인 방법**을 모색하는 것이 시급함
- **모델의 해석력**을 높여, 의료 전문가가 AI 결정 과정을 이해하고 신뢰할 수 있도록 하는 것이 필수적임

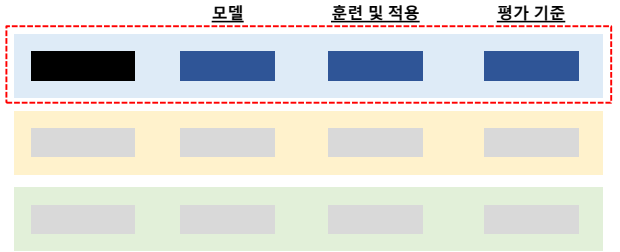
Ref.
Your Medical Imaging Data Is Going To Waste—Here's How To Unlock Its Potential (Forbes, 2024.01)
https://www.globenewswire.com/en/news-release/2023/05/17/2670669/0/en/Medical-Imaging-Market-Generate-Revenue-Of-48-8-Billion-CAGR-Of-5-4-Worldwide-by-2032-Exclusive-Report-By-Market-us.html

Contents

2. 이론적 배경 (Related works)

		모델	훈련 및 적용	평가 기준	
MODEL SEELCTION PROCESS	Classification Model	RESNET (18, 50, 101)	Transfer Learning	Accuracy	Class label 정보만을 가지고 모델을 훈련함
	Explainable AI	CAM (Grad-CAM, Grad-CAM++, Eigen-CAM, Layer-CAM)	Threshold & Coordinate	mIoU	CAM 결과를 바탕으로, 최적의 조건을 선정함
MODEL EVALUATION PROCESS	Vision Language Model	Honeybee (Kakao brain)	Prompt Engineering	G-Eval	선정된 모델과 조건의 추론 값을 VLM에 적용하여, VLM 성능 평가를 진행함

2. 이론적 배경 (Related works) – Classification Model

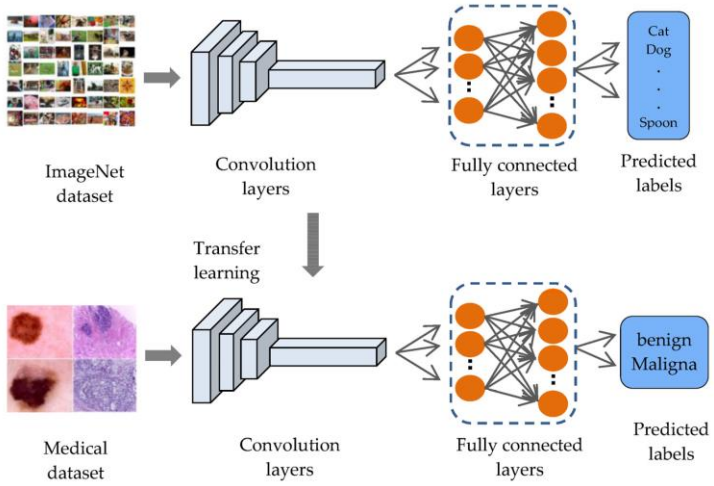


RESNET
(18, 50, 101)

- **Deep neural network 신경망 훈련 가능**: skip connection도입으로 네트워크를 통한 수백 층의 네트워크 훈련이 가능함
- **성능 개선:** 깊은 신경망을 훈련하여 복잡한 데이터셋 내 성능 향상함
- **그라디언트 소실 문제 극복:** 그라디언트 문제를 해결하여 안정적인 학습 가능

Transfer Learning

- **사전 훈련된 모델 활용:** 이미 훈련된 모델을 기반으로 특정 작업을 위해 추가적인 학습 가능함
- **학습 시간 단축:** 대규모 데이터셋에서 학습된 특징 추출 능력을 활용하여 새로운 데이터셋에 대한 학습 시간 단축 가능함
- **일반화 능력 향상:** 다양한 작업에 대한 모델의 일반화 능력 향상 가능함



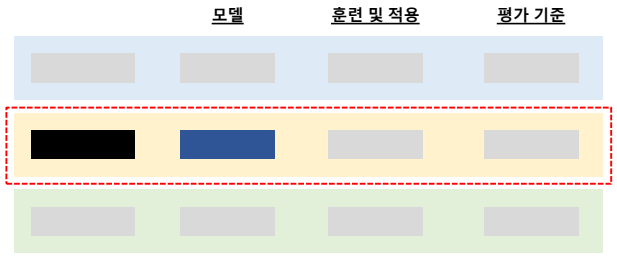
Accuracy

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(y_{pred,\,i} = y_{true,\,i})$$

- **n**: 샘플의 총 개수
- **i**: 샘플 번호
- $\mathbf{y_{pred}}$, **i**: i번째 샘플에 대한 예측 레이블
- $\mathbf{y_{true}}$, **i**: i번째 샘플에 대한 실제 레이블.
- **1()**: 지시 함수(indicator function)로, 예측 레이블이 실제 레이블과 일치할 때 1이 되고 그렇지 않을 때는 0으로 처리함

Ref. https://www.mdpi.com/1424-8220/23/2/570

2. 이론적 배경 (Related works)
– Explainable AI: CAM Methods



CAM

- Deep neural network의 해석 가능성을 높이기 위해 특정 분류 결정에 기여하는 이미지 영역을 강조함
- 마지막 컨볼루션 계층의 feature map과 global average pooling을 사용하여 의사 결정에 중요한 이미지 영역을 식별함

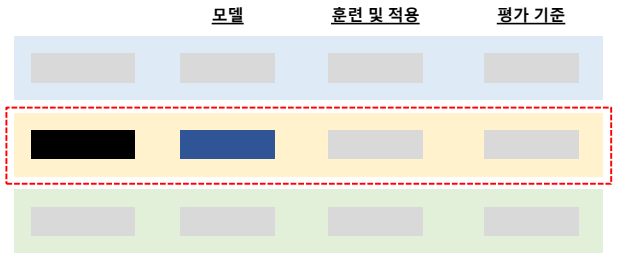
Grad-CAM

- CAM의 발전된 형태로, 마지막 컨볼루션 계층의 그라디언트 정보를 사용하여 상세한 클래스 활성화 맵을 생성함
- 더 세밀한 지역화 및 결정 경계를 제공하여, 다양한 질병 진단에 효과적임

Grad-CAM++

- Grad-CAM을 개선하여 클래스별 차별적 특징을 통해 공간적 정밀도 (spatial precision)를 높임
- Positive및 negative 그라디언트를 사용하여 feature map에 더 정확한 가중치를 부여함
- 상세한 feature map을 제공하지만, 노이즈에 민감하며 그라디언트에 크게 의존함

2. 이론적 배경 (Related works)
– Explainable AI: CAM Methods



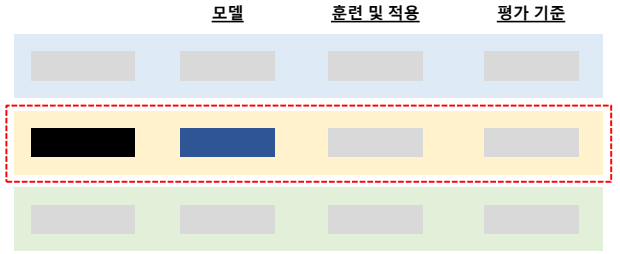
Layer-CAM

- 신경망의 다양한 계층에 걸쳐 특징을 분석하여 클래스 활성화 맵을 생성함
- 다중 계층 (multiple layer)에 걸쳐 중요한 영역을 강조하여 모델 예측에 대한 포괄적인 이해를 제공함

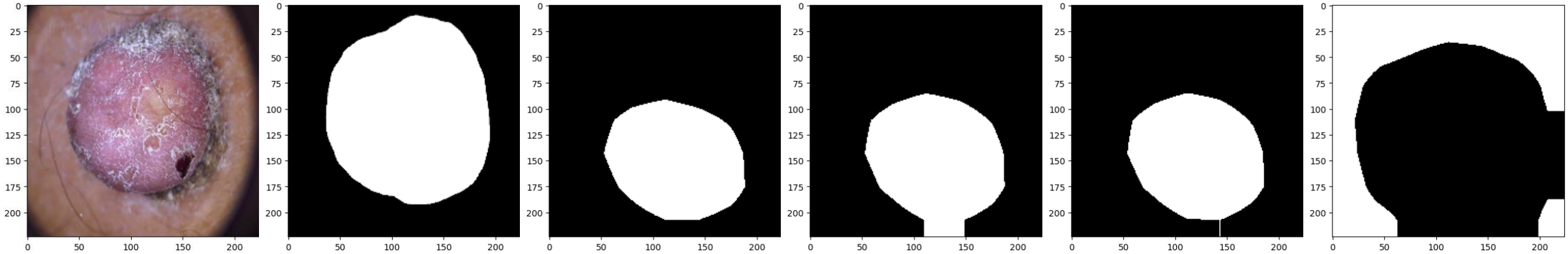
Eigen-CAM

- CNN의 feature map에서 주성분 분석(PCA)을 활용하여 모델 예측에 영향을 미치는 주요 구성 요소를 식별함
- 고차원 (high-dimension) 특징 공간을 명확하고 해석 가능한 시각화 가능함

2. 이론적 배경 (Related works) – Explainable AI: CAM Methods



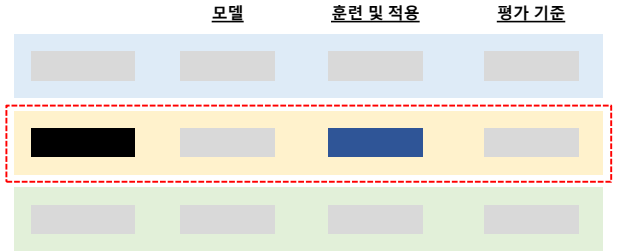
SAMPLE IMAGE (Dermatofibroma, 피부 섬유종)



ORIGINAL IMAGE | **ANSWER SEMANTIC AREA** | **Grad-CAM** | **Grad-CAM++** | **Layer-CAM** | **Eigen-CAM**

- **MODEL:** RESNET18 100 epochs Whole augmentation
- **Threshold :** 0.5

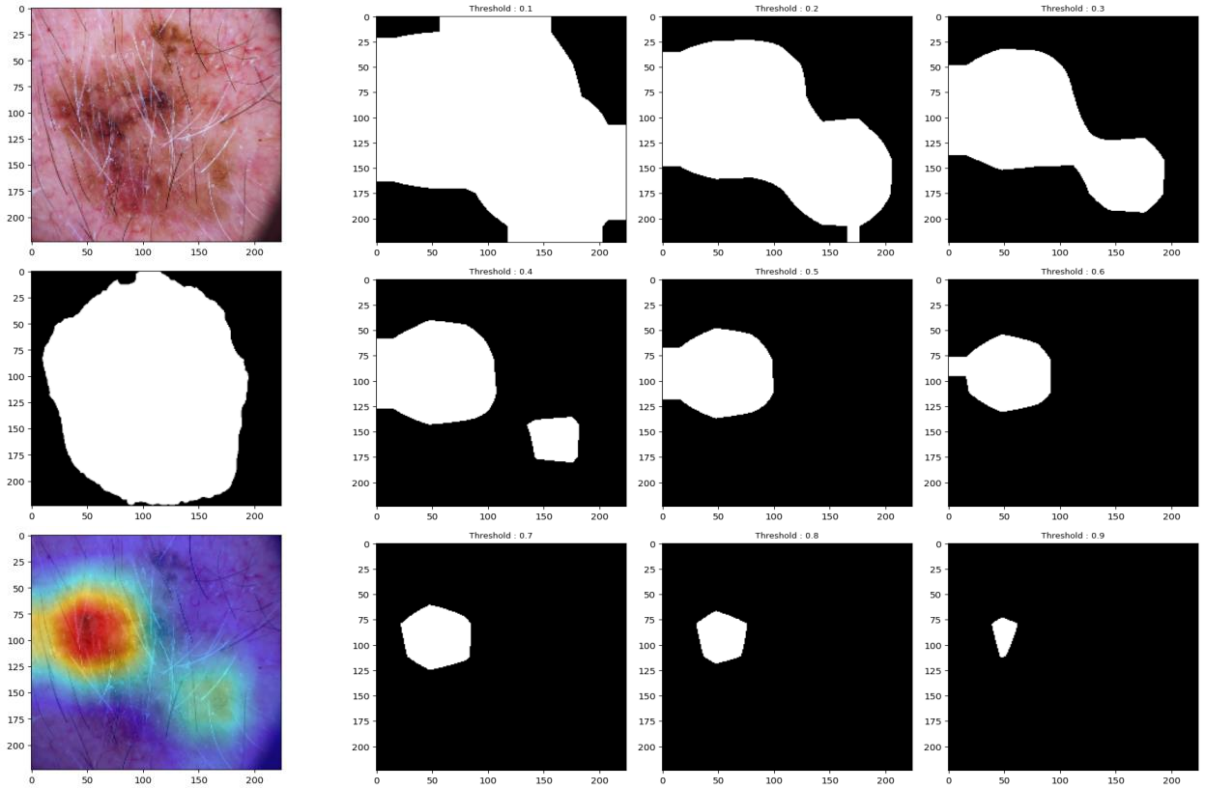
2. 이론적 배경 (Related works) – Explainable AI: Threshold



계산 방법

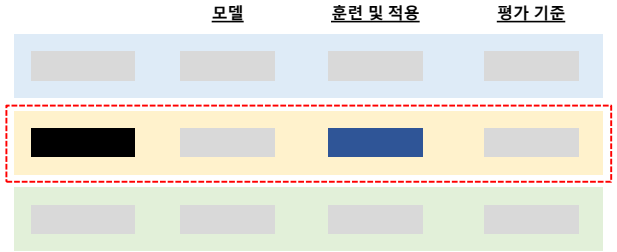
CAM으로 추출한 이미지 영역별로 0에서 1 사이의 값 [0, 1]을 가지며, threshold 이상에 해당하는 것을 semantic segmentation 적용함

$$CAM\ Segmentation\ =\ TRUE\ if\ \sum_{y}\sum_{x}(CAM\_output(y,x\ >\ threshold),\ otherwise\ FALSE$$



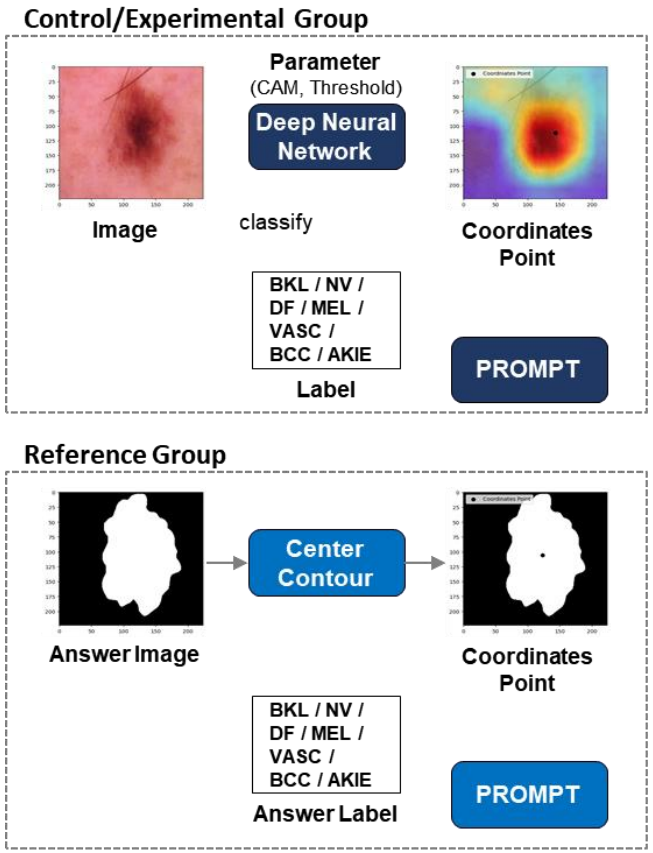
[Figure. 다양한 임계값에 따른 피부과 이미지에서의 Layer-CAM 비교 분석]
- **첫 번째 열 :** 위에서 아래로 원본 이미지 (MEL), 원본 정답 영역 그리고 최적 설정 하의 saliency map ; **3x3 matrix :** 0.1부터 0.9까지 변화하는 임계값 적용

2. 이론적 배경 (Related works)
– Explainable AI: Coordinates Point



Coordinates Point

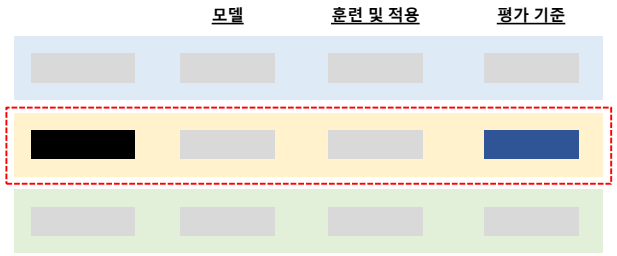
VLM 성능 향상 방안으로 CAM으로 추출한 이미지 영역 내에서 핵심 좌표 정보를 추출하여, 분류 결과와 좌표 정보로 활용함 (정답의 경우, OpenCV 내 contour를 활용함)




```python
# Construct the CAM object once, and then re-use it on many images:
with LayerCAM(model=pretrained_model, target_layers=target_layer) as cam:
    cam.batch_size = 1

    # You can also pass aug_smooth=True and eigen_smooth=True, to apply smoothing.
    grayscale_cam = cam(input_tensor=input_image, targets=targets)
    grayscale_cam = grayscale_cam[0, :]

    test_segmantation = np.where(grayscale_cam > iou_threshold, True, False)
    # Get the max gradcam pixel location
    max_activation_coord = np.unravel_index(grayscale_cam.argmax(), grayscale_cam.shape)

    visualization = show_cam_on_image(sample_image, grayscale_cam, use_rgb=True)
```

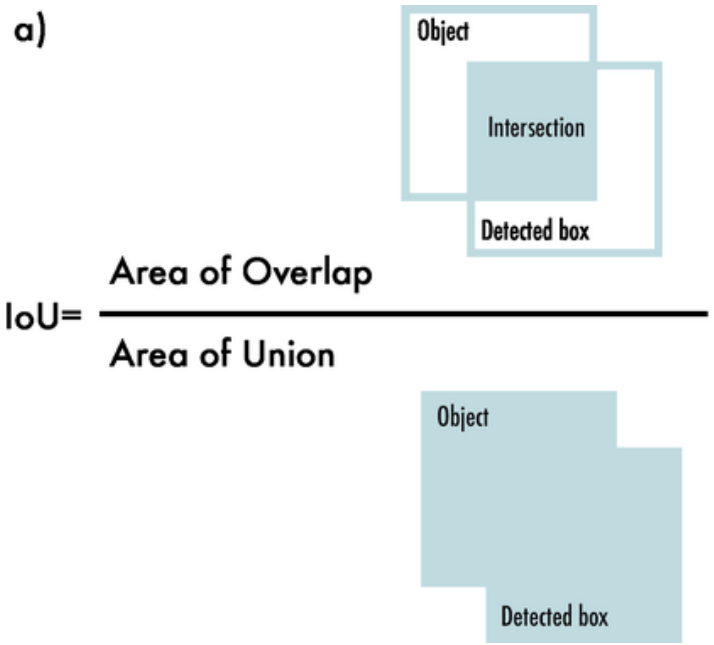
```python
def find_contour_centers(image):
    contours, _ = cv2.findContours(image, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)

    centers = []

    for contour in contours:
        M = cv2.moments(contour)
        if M['m00'] != 0:
            cx = int(M['m10'] / M['m00'])
            cy = int(M['m01'] / M['m00'])
            centers.append((cx, cy))
        else:
            centers.append((0, 0))  # 분모가 0인 경우를 처리

    return centers
```

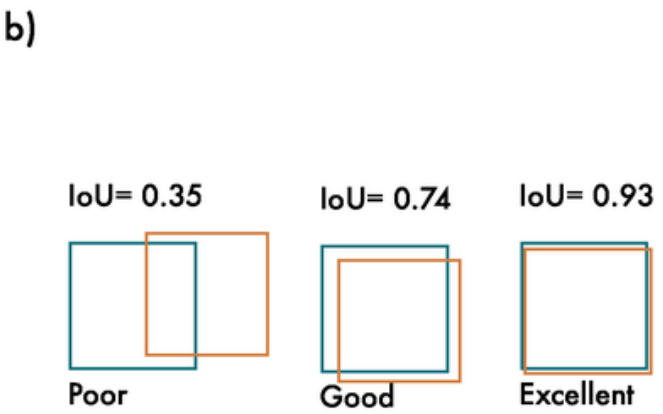
2. 이론적 배경 (Related works) – Explainable AI: IoU



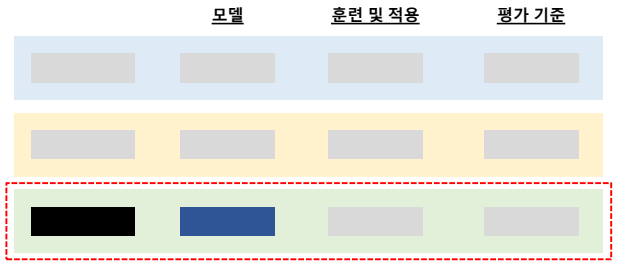
IoU (Intersection Over Union)

- **정의: 두 영역(실제 영역과 예측 영역)의 교집합과 합집합의 비율을 계산**하여, 모델이 얼마나 정확하게 객체를 추출했는지를 수치로 나타냄
- **계산 방법:** IoU는 교집합 영역을 합집합 영역으로 나눈 값으로 표현되며, **IoU = (예측 영역 ∩ 실제 영역) / (예측 영역 ∪ 실제 영역)** 공식을 사용



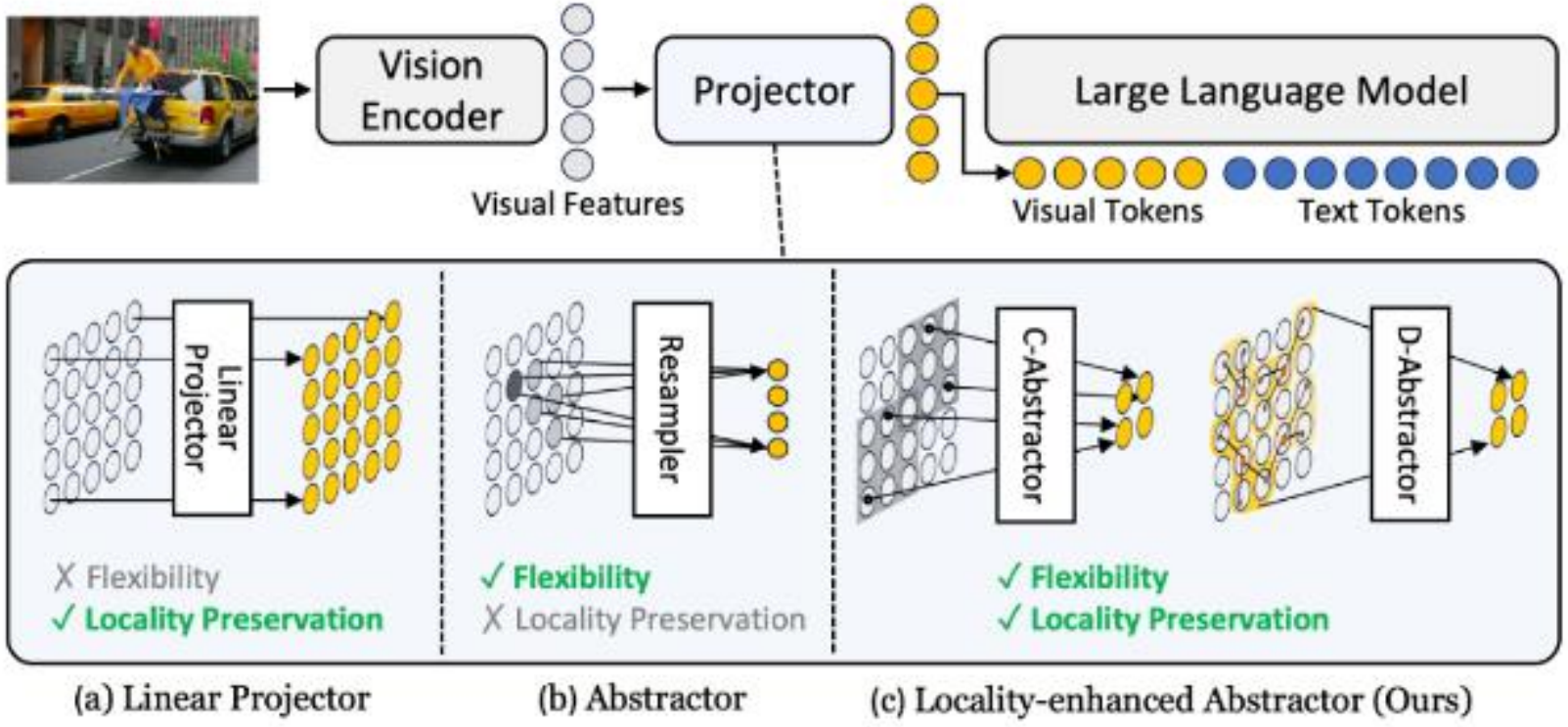


2. 이론적 배경 (Related works) – Vision Language Model: Honeybee



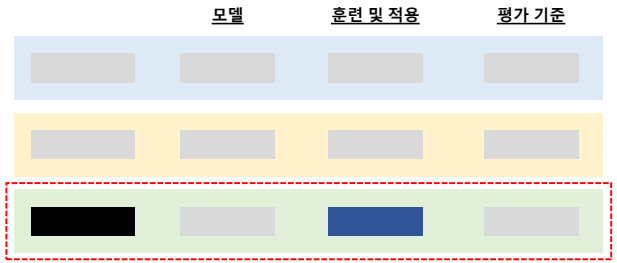
Honeybee (Kakaobrain, CVPR 2024)

- **모델 구조**: MLLM (Multi Large Language Model)로 시각적 특성을 '언어 모델이 이해할 수 있는 시각적 토큰으로 변환하는 프로젝터'를 활용함
- **지역 특징 활용 프로젝터:** 지역 정보를 강화한 프로젝터 (C-Abstractor/D-Abstractor)를 도입하여 성능과 효율성 사이의 균형을 개선함



Ref. Honeybee: Locality-enhanced Projector for Multimodal LLM (Junbum Cha, Wooyoung Kang, Jonghwan Mun, Byungseok Roh, 2023)

2. 이론적 배경 (Related works)
– Vision Language Model: Prompt Engineering



PROMPT

System prompt

- 자연어 처리 모델이 입력을 받아 작업을 수행하기 위해 사용하는 지시문으로, 이를 통해 모델은 주어진 작업에 대한 적절한 결과를 생성할 수 있음
- Cf., “이 이미지에 나타난 객체를 설명하시오" 또는 "이 이미지를 기반으로 스토리를 만드세요."

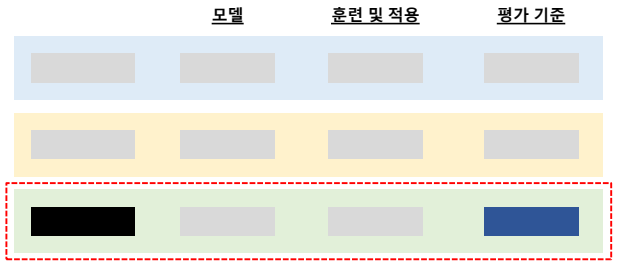
Image token

- 이미지를 분석하여 얻은 정보의 단위로, 이미지 내의 특정 객체나 특징을 기호화하여 모델이 이미지 내용을 이해하고 처리할 수 있도록 함
- Vision Language Models는 이미지를 여러 개의 이미지 토큰으로 분할하여 각 토큰이 이미지의 일부를 대표하도록 함

User prompt

- 최종 사용자가 모델에 입력하는 명령이나 요청으로, system prompt보다 더 구체적이거나 개인화된 정보를 요구할 수 있음
- Cf., 사용자가 사진을 업로드하고 입력하는 "이 사진 속 인물이 무슨 생각을 하고 있을까요?"와 같은 질문

2. 이론적 배경 (Related works)
– Vision Language Model: G-Eval



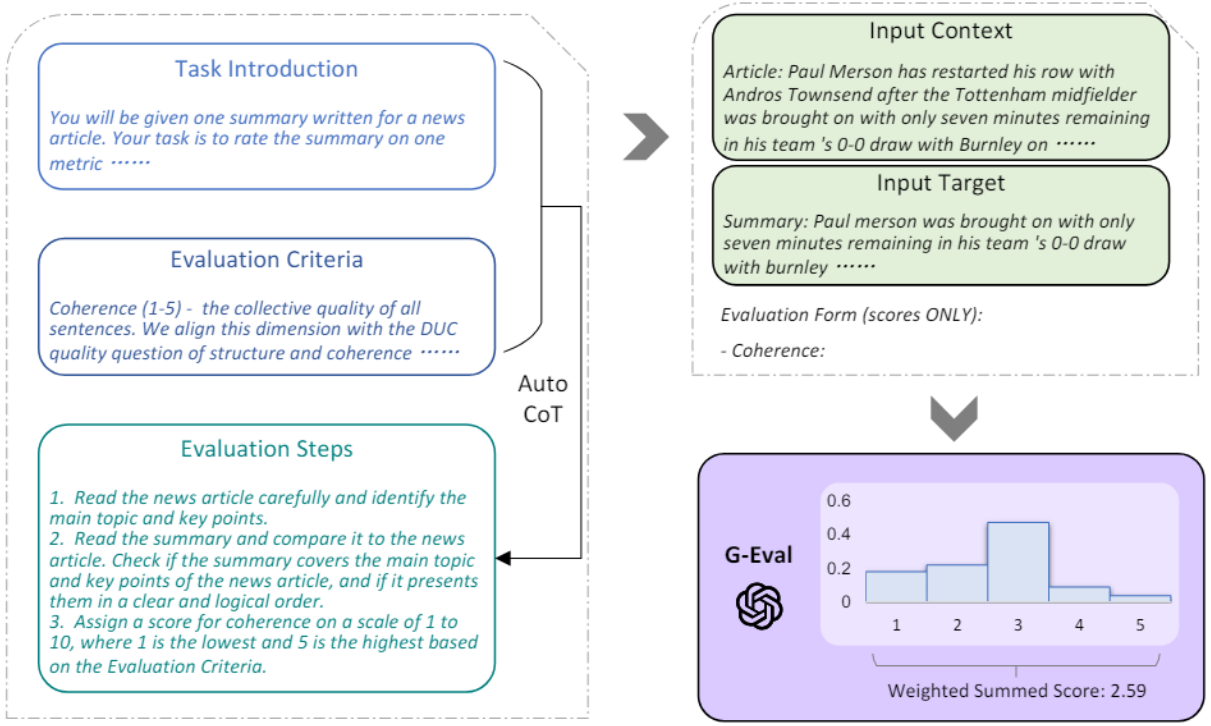
G-Eval

- **G-Eval 개념**
  - 자연어 생성(NLG) 시스템의 품질을 자동으로 평가하는 LLM 기반의 평가 시스템
  - 인간의 견해와 보다 일치하는 평가를 목표로 함

- **평가 메커니즘**
  - **Chain-of-Thought (CoT) 활용**: CoT 접근 방식을 사용하여 LLM이 생성한 중간 평가 단계를 자동으로 구성하며, 이를 통해 보다 상세한 컨텍스트와 지침을 제공함
  - **Form-Filling Paradigm**: 평가 과정에서 LLM은 평가하려는 태스크의 세부 사항을 기반으로 형식을 채워 점수를 산출함

- **평가의 유효성**: 다양한 NLG 작업에서 기존의 reference-control 및 reference-free 평가 방법을 능가하는 성능을 달성함

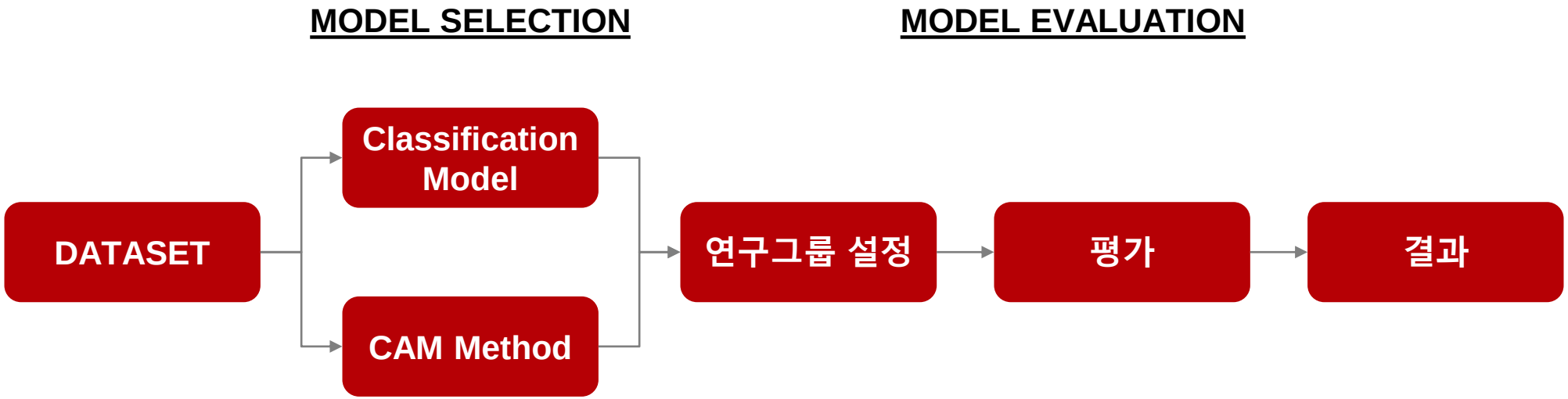
- **상관계수 및 성능**:
  - 사람의 평가와의 상관계수가 높으며, 평가 도구로 사용될 수 있음
  - GPT-4를 사용한 실험에서 특히 높은 상관성을 달성함



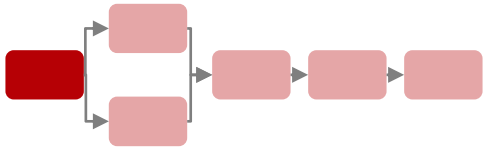
Ref. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment (Yang Liu. et. al. 2023)

Contents

3. 실험 구성


MODEL SELECTION
MODEL EVALUATION
DATASET
Classification Model
CAM Method
연구그룹 설정
평가
결과

3. 실험 구성 (Dataset)



1. **HAM10000 소개**
   - 10,015장의 피부 질환 이미지 (해상도는 450 x 600, 7가지 피부 질환 이미지)
   - **Train / Evaluation / Test : 0.8 (8,012) / 0.1 (1,001) / 0.1 (1,002)비율로 구분함**

2. **질병 라벨 및 분포**
   a. **양성 각화증류 병변 (Benign Keratosis-like Lesions, BKL)**: 노화나 햇빛 노출과 관련된, 양성의 피부 병변 (1,099장)

   b. **멜라닌성 모반 (Melanocytic Nevi, NV)**: 흔히 점으로 알려져 있으며, 대부분 양성이지만 일부는 흑색종으로 발전할 위험이 있어 정기적인 검진이 필요한 상태 (6,705장)

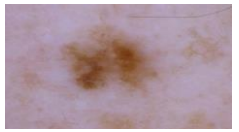
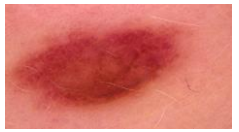
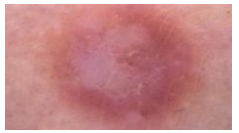
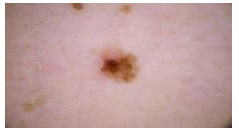
   c. **피부섬유종 (Dermatofibroma, DF):** 주로 팔이나 다리에 나타나는 작고 단단한 결절 (115장)

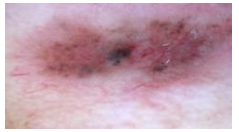
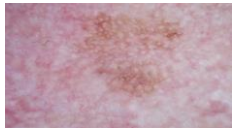
   d. **흑색종 (Melanoma, MEL):** 멜라닌을 생성하는 세포에서 발생하는 가장 치명적인 형태의 피부암으로, 빠르게 전이되어 조기 발견 및 치료가 중요함 (1,113장)

   e. **혈관성 병변 (Vascular Lesions, VASC):** 혈관의 비정상적 성장이나 반응으로 발생하는 다양한 유형의 병변으로, 피부에 붉은 반점이나 종양으로 나타날 수 있음 (142장)

   f. **기저세포암 (Basal-Cell Carcinoma, BCC):** 피부의 기저세포층에서 발생하는 가장 흔한 형태의 피부암 (514장)

   g. **햇빛 각화증 및 상피내암 / 보웬병 (Actinic Keratoses and Intraepithelial Carcinoma / Bowen's Disease, AKIEC):** 얼굴, 귀, 손에 나타나며 햇빛 노출로 인한 비정상적인 세포 성장과 관련이 있으며, **광선각화증**은 암으로 발전할 가능성이 있는 선암성 상태이며, **보웬병**은 피부암의 초기 단계인 상피암(327장)

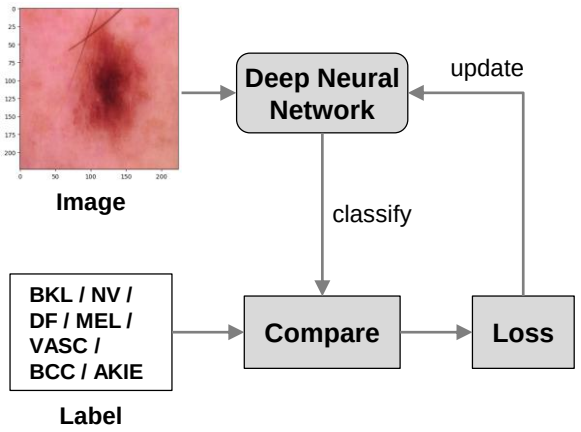



3. 실험 설계 (MODEL SEELCTION)

MODEL SELECTION PROCESS

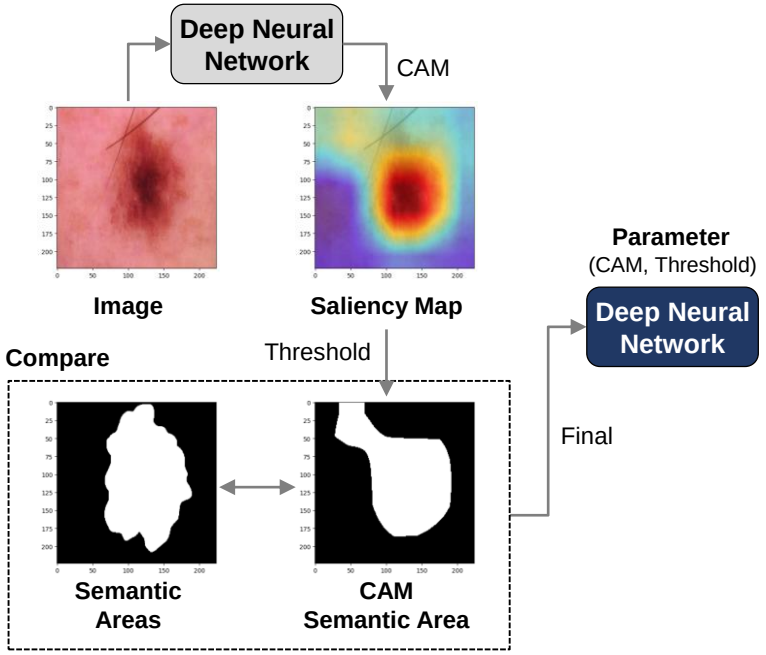
Train Classification Model

Model training using image-level labels

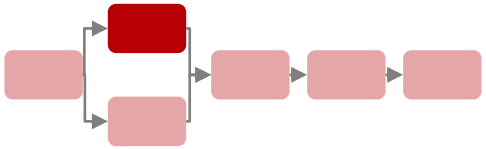


Select CAM Method

Model evaluating using semantic segmentation

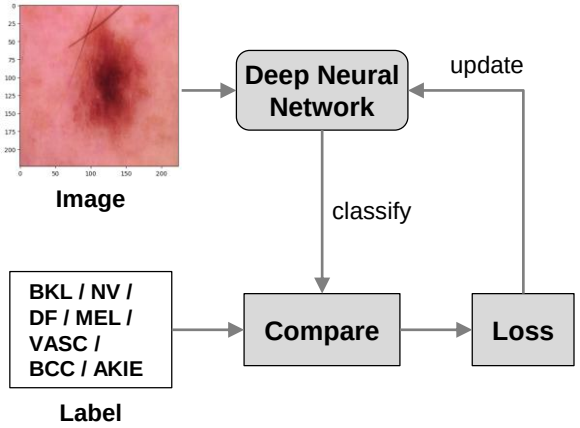


3. 실험 설계 (MODEL SEELCTION)
Classification Model 훈련



Train Classification Model

Model training using image-level labels



1. 이미지 크기 조정

- 기본 이미지 크기: 224 픽셀 ([Honeybee 모델 기본 설정]

2. 데이터 증강

- 데이터 증강 기법: 데이터셋의 다양성을 향상시키고 모델의 과적합에 대한 견고성을 강화하기 위해 다음과 같은 기법을 사용함
  - 30도 회전
  - 무작위 수평 뒤집기
  - 무작위 수직 뒤집기

- 데이터 증강 조건:
  - 증강 없음: 증강 없이 원본 데이터셋에서 모델을 학습하여 기준 성능을 설정
  (Initial images: 8,012; Post-augmentation: 8,012)
  - 무작위 증강: 데이터셋의 각 이미지에 무작위로 세 가지 증강 기법 중 하나를 적용하여 변이를 도입
  (Initial images: 8,012; Post-augmentation: 16,024)
  - 전체 증강: 데이터셋의 각 이미지에 모든 세 가지 증강 방법을 적용하여 다양성을 극대화함
  (Initial image: 8,012; Post-augmentation: 32,048)

3. 하이퍼파라미터

- 학습 epoch 범위: 10, 50, 100
- 학습률: 0.001
- 가중치 감쇠: 0.0001

4. 모델 아키텍처

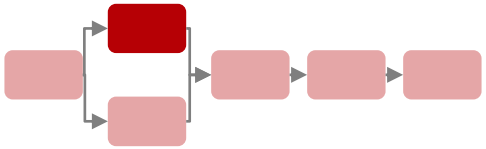
- ResNet18, ResNet50, ResNet101

5. 평가 방법

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(y_{pred,i} = y_{true,i})$$

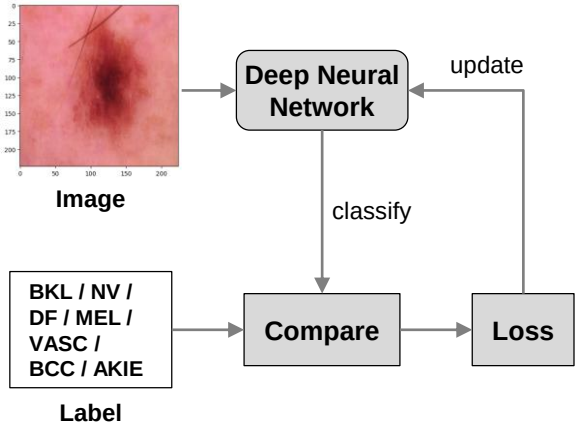
Ref. https://github.com/kakaobrain/honeybee/blob/main/configs/default.yaml

3. 실험 설계 (MODEL SEELCTION)
Classification Model 훈련



Train Classification Model

Model training using image-level labels



1. 이미지 크기 조정

- 기본 이미지 크기: 224 픽셀 ([Honeybee 모델 기본 설정]

2. 데이터 증강

- 데이터 증강 기법: 데이터셋의 다양성을 향상시키고 모델의 과적합에 대한 견고성을 강화하기 위해 다음과 같은 기법을 사용함
  - 30도 회전
  - 무작위 수평 뒤집기
  - 무작위 수직 뒤집기

- **데이터 증강 조건:**
  - 증강 없음: 증강 없이 원본 데이터셋에서 모델을 학습하여 기준 성능을 설정
  (Initial images: 8,012; Post-augmentation: 8,012)
  - 무작위 증강: 데이터셋의 각 이미지에 무작위로 세 가지 증강 기법 중 하나를 적용하여 변이를 도입
  (Initial images: 8,012; Post-augmentation: 16,024)
  - 전체 증강: 데이터셋의 각 이미지에 모든 세 가지 증강 방법을 적용하여 다양성을 극대화함
  (Initial image: 8,012; Post-augmentation: 32,048)

3

3. 하이퍼파라미터

3

- 학습 epoch 범위: 10, 50, 100
- 학습률: 0.001
- 가중치 감쇠: 0.0001

4. 모델 아키텍처

3

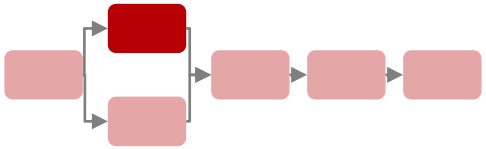
- ResNet18, ResNet50, ResNet101

5. 평가 방법

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(y_{pred,i} = y_{true,i})$$

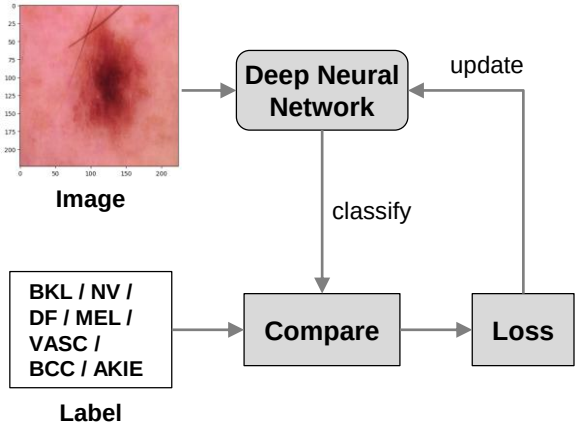
Ref. https://github.com/kakaobrain/honeybee/blob/main/configs/default.yaml

3. 실험 설계 (MODEL SEELCTION)
Classification Model 훈련



Train Classification Model

Model training using image-level labels

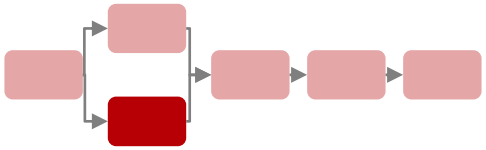


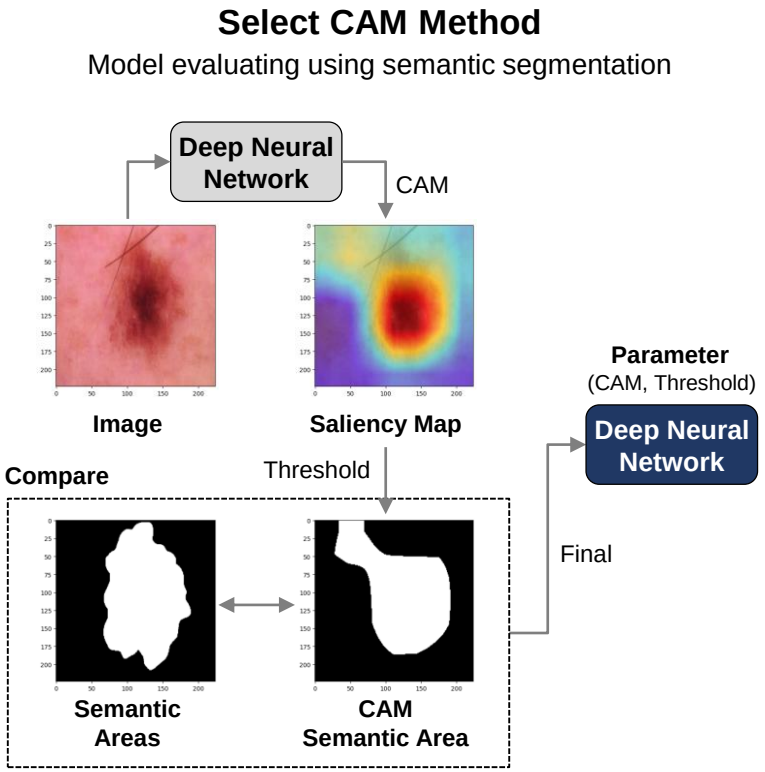
3 × 3 × 3 =

		Augmentation		
Model	Epoch	#None	#Random	#Whole
resnet18	10	74%	82%	84%
	50	77%	**85%**	83%
	100	83%	82%	**85%**
resnet50	10	78%	79%	77%
	50	81%	81%	79%
	100	80%	81%	80%
resnet101	10	77%	78%	72%
	50	77%	78%	73%
	100	77%	75%	79%

Ref. https://github.com/kakaobrain/honeybee/blob/main/configs/default.yaml

3. 실험 설계 (MODEL SELECTION)
CAM Method 선정



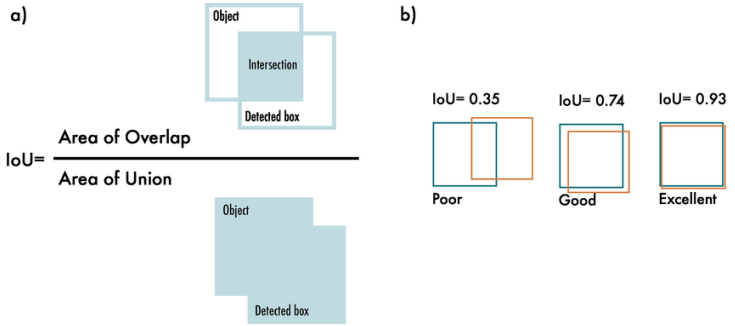


1. 적용 범위: 27개 분류 모델에 적용

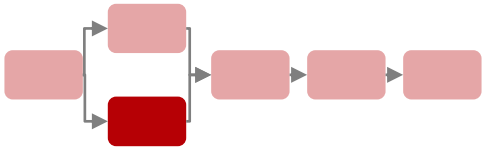
2. XAI 방법:
- Grad-CAM
- Grad-CAM++
- Layer-CAM
- Eigen-CAM

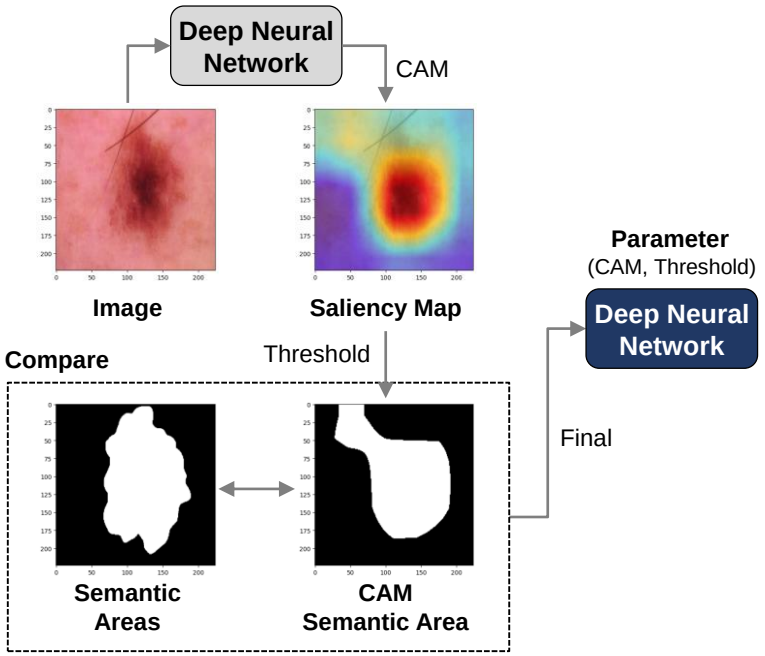
3. 임계값 설정:
- [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

4. 평가 기준:
- Intersection Over Union (IoU)



3. 실험 설계 (MODEL SELECTION)
CAM Method 선정





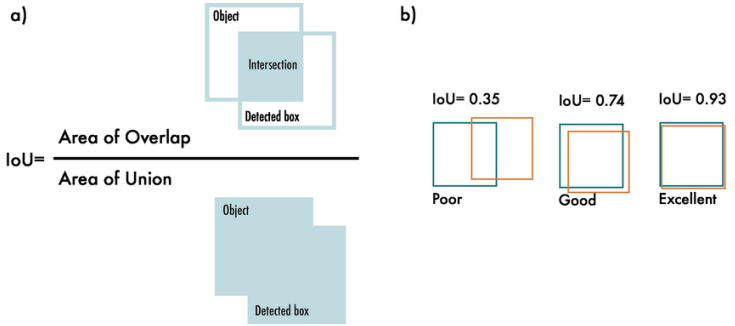
27 × 4 × 9

1. 적용 범위: 27개 분류 모델에 적용

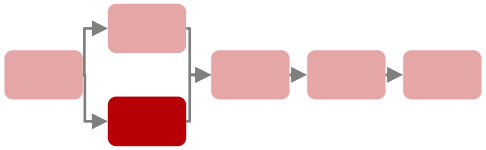
2. XAI 방법:
- Grad-CAM
- Grad-CAM++
- Layer-CAM
- Eigen-CAM

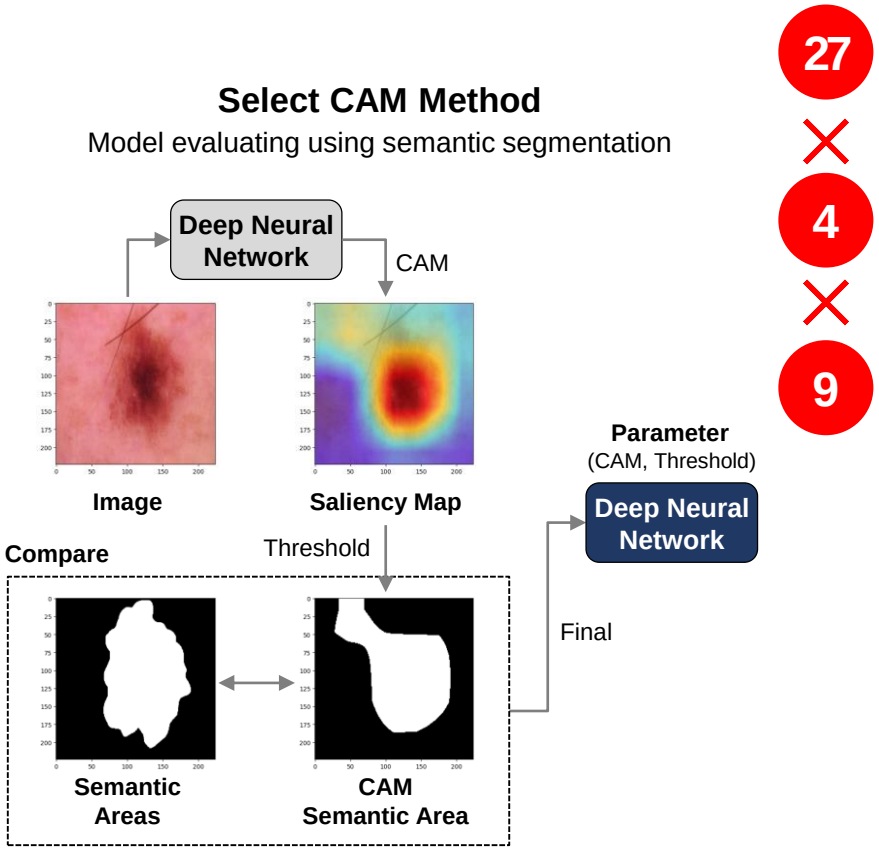
3. 임계값 설정:
- [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

4. 평가 기준:
- Intersection Over Union (IoU)



3. 실험 설계 (MODEL SELECTION)
CAM Method 선정 결과





27 × 4 × 9 = 729

Grad-CAM Semantic Segmentation Best Result (mIoU); **0.5107**

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_100_aug_Whole	0.3614	0.4275	0.4807	**0.5107**	0.5091	0.4699	0.3895	0.2703	0.1199

Grad-CAM++ Semantic Segmentation Best Result (mIoU); **0.5248**

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_100_aug_Whole	0.3684	0.4396	0.4952	**0.5248**	0.5207	0.4775	0.3993	0.2708	0.1179

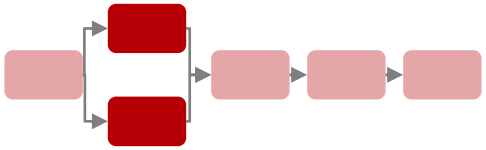
Layer-CAM Semantic Segmentation Best Result (mIoU); **0.5287**

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_100_aug_Whole	0.3707	0.4431	0.4994	**0.5287**	0.5231	0.4787	0.3931	0.2699	0.1165

Eigen-CAM Semantic Segmentation Best Result (mIoU); **0.4021**

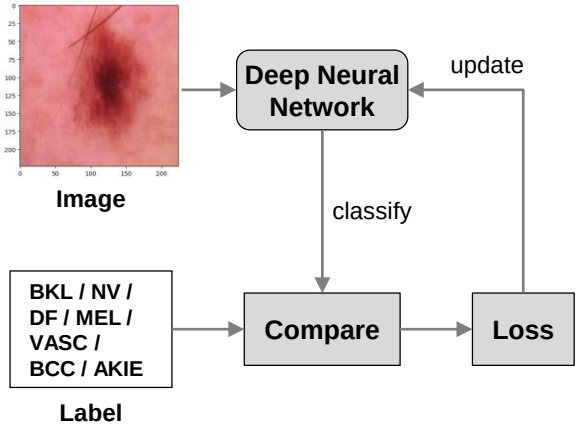
Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet50_iter_10_aug_Whole	**0.4021**	0.3578	0.3029	0.2444	0.1878	0.1352	0.0870	0.0464	0.0146

3. 실험 설계 (MODEL SELECTION)
최종 모델 선정

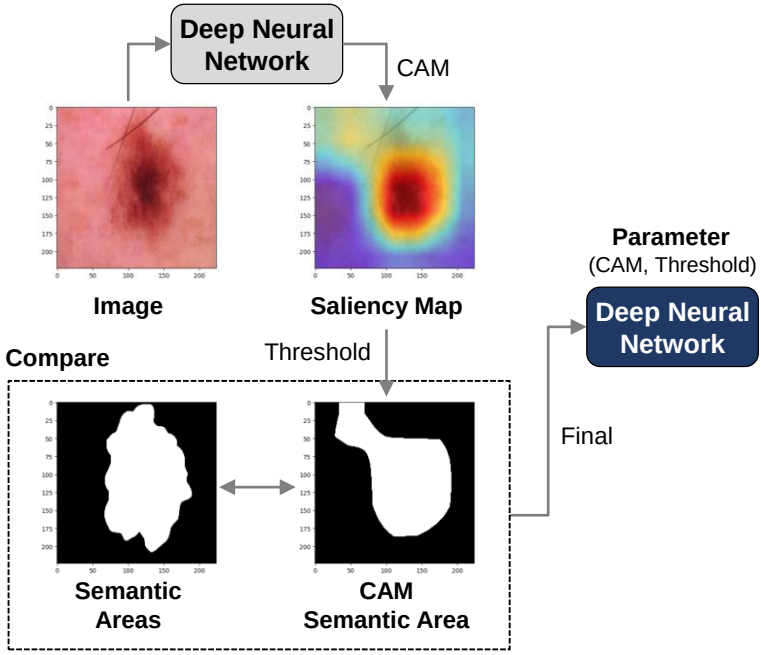


MODEL SEELCTION PROCESS

Train Classification Model
Model training using image-level labels



Select CAM Method
Model evaluating using semantic segmentation



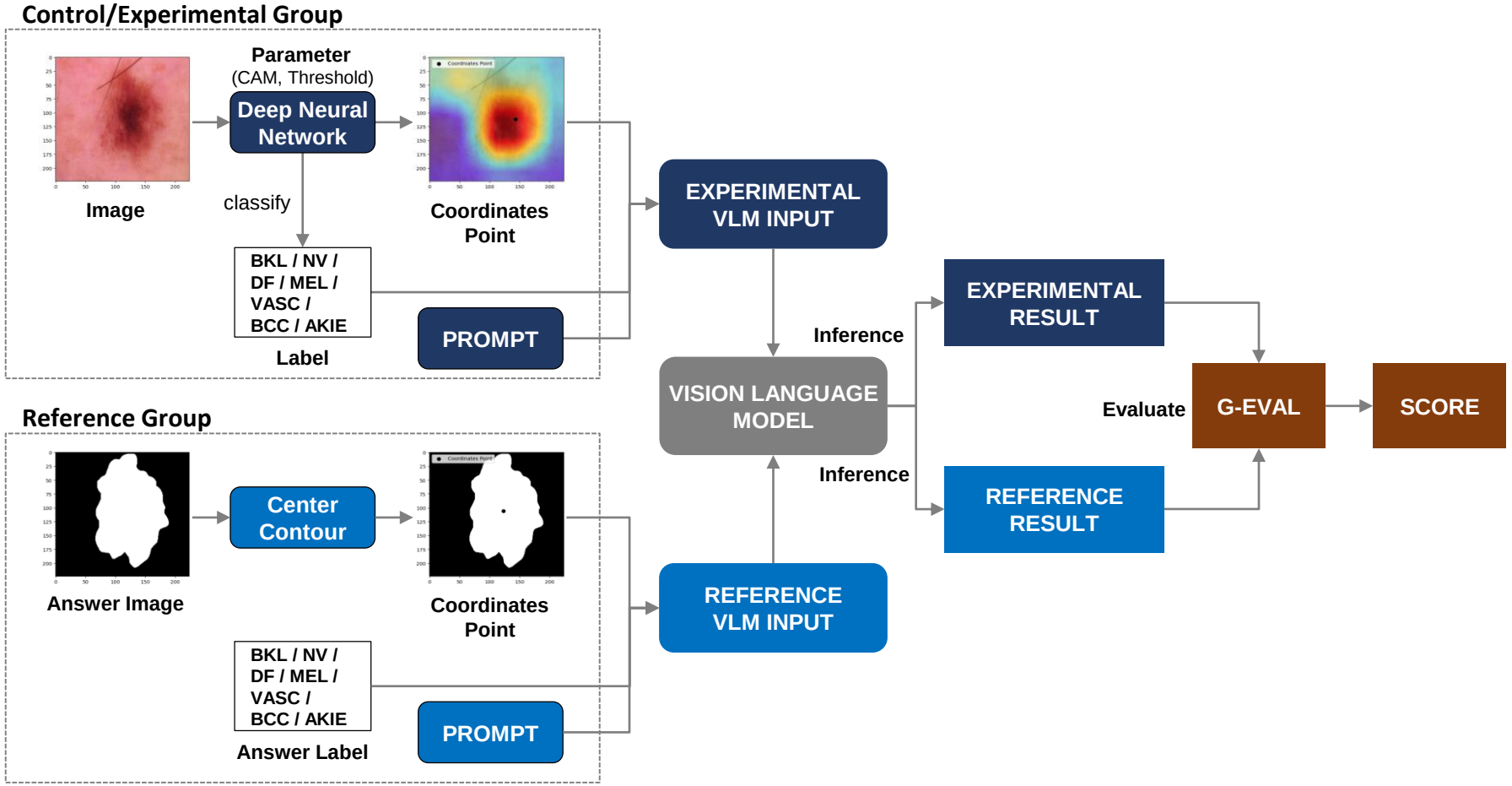
최종적으로 **resnet18_iter_100_aug_Whole** 모델과 **Layer-CAM, 0.4 임계값 (threshold)**를 선정함

3. 실험 설계 (MODEL EVALUATION)

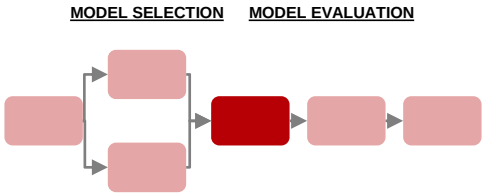
MODEL EVALUATION PROCESS

Evaluate Vision Language Model Performance

Utilizing the coordinates point with VLM



3. 실험 설계 (MODEL EVALUATION) 프롬프트 가이드



PROMPT

System prompt

- 자연어 처리 모델이 입력을 받아 작업을 수행하기 위해 사용하는 지시문으로, 이를 통해 모델은 주어진 작업에 대한 적절한 결과를 생성할 수 있음

Image token

- 이미지를 분석하여 얻은 정보의 단위로, 이미지 내의 특정 객체나 특징을 기호화하여 모델이 이미지 내용을 이해하고 처리할 수 있도록 함
- Vision Language Models는 이미지를 여러 개의 이미지 토큰으로 분할하여 각 토큰이 이미지의 일부를 대표하도록 함

User prompt

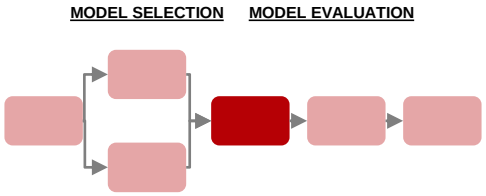
- 최종 사용자가 모델에 입력하는 명령이나 요청으로, system prompt보다 더 구체적이거나 개인화된 정보를 요구할 수 있음

SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot.
Your purpose is to diagnose skin diseases using images.
- BKL: Benign Keratosis-like Lesions
- NV: Melanocytic Nevi
- DF: Dermatofibroma
- MEL: Melanoma
- VASC: Vascular Lesions
- BCC: Basal Cell Carcinoma
- AKIEC: Actinic Keratosis and Intraepithelial Carcinoma
You will receive an image with dimensions 224 width and 224 height for diagnosis.
The classification model identified the image as {inference_label} disease with 85% accuracy.'‘’

IMAGE_TOKEN = f"\nHuman: <image>"

USER_PROMPT = f"\nHuman: Please diagnose the image"

3. 실험 설계 (MODEL EVALUATION)
CASE 구분



VLM 실험 설계

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

1. **Version 1** : System prompt 내 7가지 클래스의 정보를 사용
2. **Version 2 :** System prompt 내 단일 클래스 정보를 사용
3. **Version 3 :** Version 2의 system prompt와 유사하지만, user prompt에도 좌표 정보를 사용
4. **Version 4 :** Version 1의 system prompt와 유사하지만, user prompt에서 좌표 정보를 사용

'정답지'로 활용할 데이터셋

Structured Answer Prompts

각 version 1, 2, 3, 4 활용: 해당 케이스에서는 각 version 별로 생성된 데이터셋을 정답지로 활용함

Focused Answer Prompts

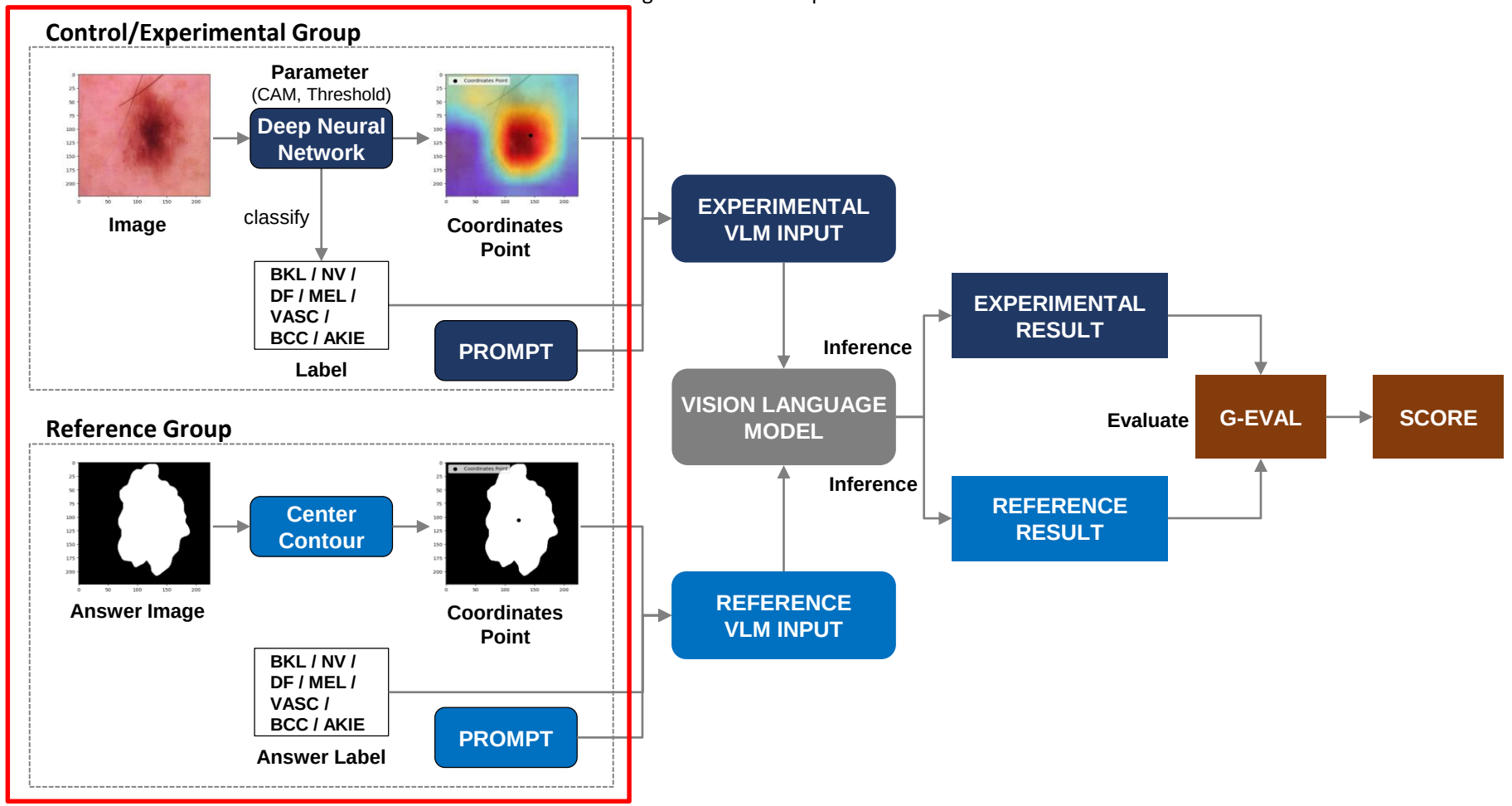
Version 3: system prompt에 해당 클래스 정보만 활용하며, LLM 답변 품질을 위해 좌표 정보를 user prompt에도 사용하여, 전체적인 답변 품질을 보장함. 또한 동일한 답변으로 비교하여, 각 version별 성능 비교도 가능함

3. 실험 설계 (MODEL EVALUATION)

MODEL EVALUATION PROCESS

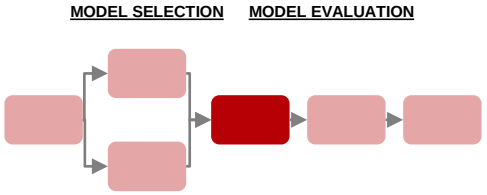
Evaluate Vision Language Model Performance

Utilizing the coordinates point with VLM



3. 실험 설계 (MODEL EVALUATION)
연구 그룹(통제, 실험, 정답) 설정

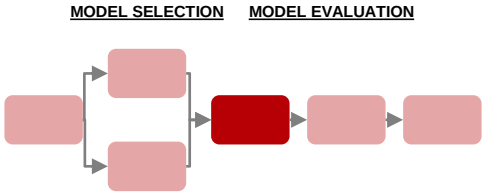
USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3



Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosing {y_label} skin disease. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions calculated by contours centroid.'"
IMAGE_TOKEN = f""\nHuman: <image>""	IMAGE_TOKEN = f""\nHuman: <image>""	IMAGE_TOKEN = f""\nHuman: <image>""
USER_PROMPT = f""\nHuman: Please diagnose the image""	**USER_PROMPT = f""\nHuman: Please diagnose the image""**	**USER_PROMPT = f""\nHuman: Please diagnose the image""**

3. 실험 설계 (MODEL EVALUATION)
연구 그룹(통제, 실험, 정답) 설정

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3



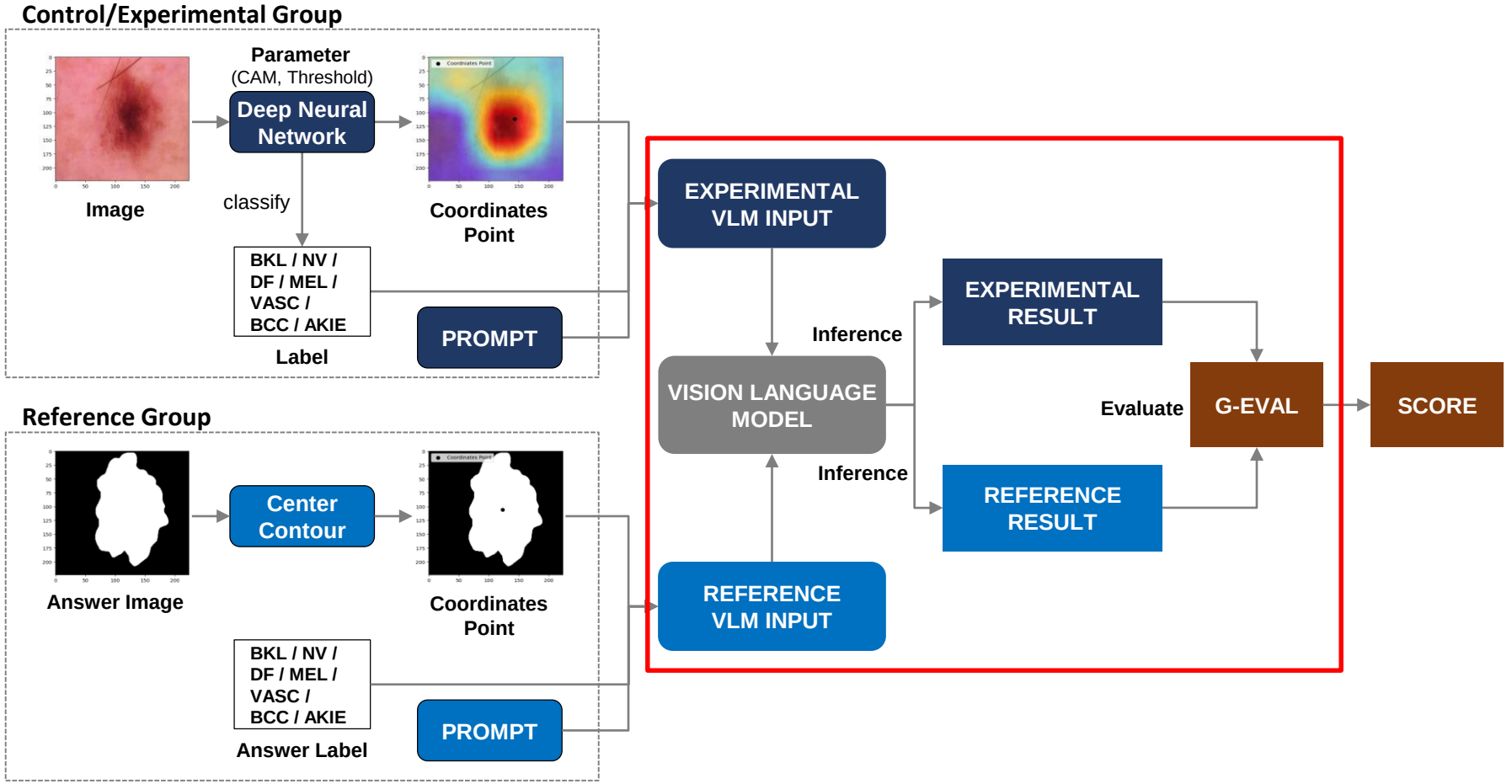
Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosing the {y_label} skin disease. Pay close attention to the pixel at coordinates {max_activation_coord}, as this point indicates a region calculated by the centroid of contours.'"
IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"
USER_PROMPT = f"\nHuman: Please diagnose the image"	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**

3. 실험 설계 (MODEL EVALUATION)

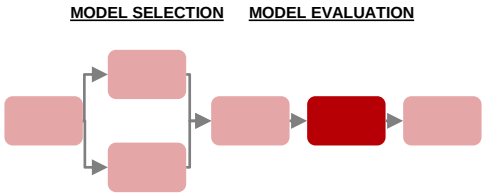
MODEL EVALUATION PROCESS

Evaluate Vision Language Model Performance

Utilizing the coordinates point with VLM



3. 실험 설계 (MODEL EVALUATION)
Vision Language Model: G-Eval



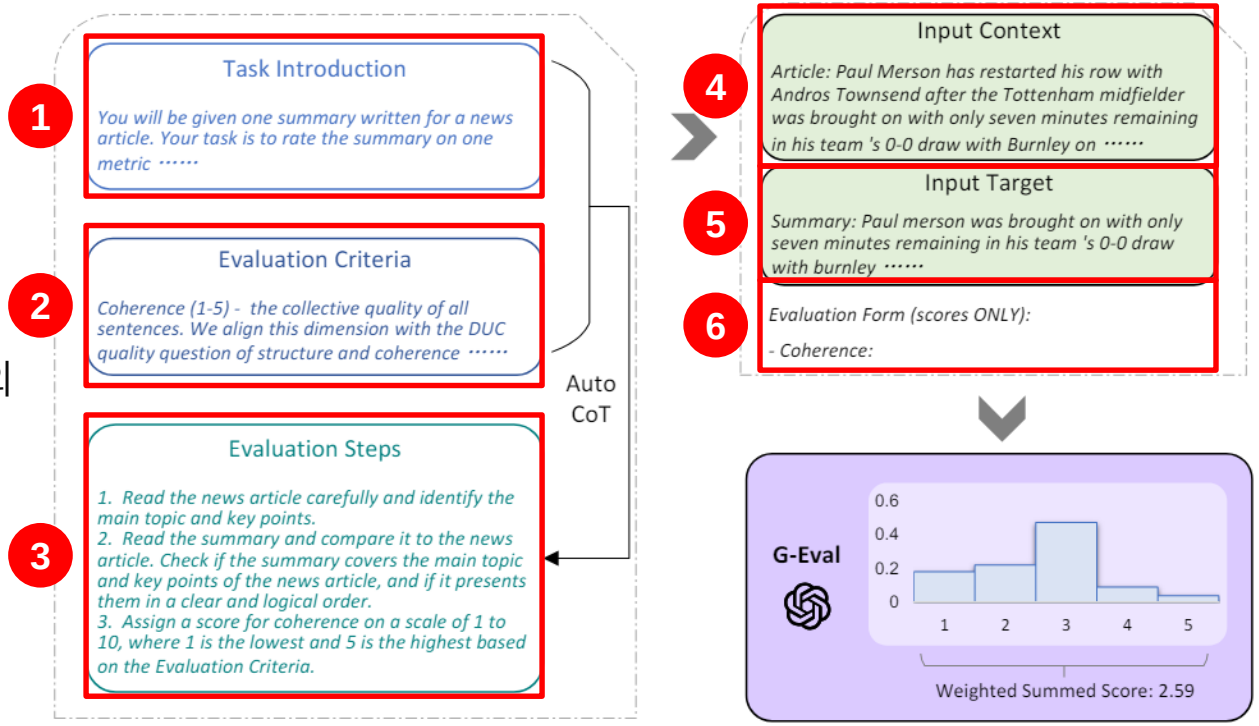
G-Eval

- **G-Eval 개념**
  - 자연어 생성(NLG) 시스템의 품질을 자동으로 평가하는 LLM 기반의 평가 시스템
  - 인간의 견해와 보다 일치하는 평가를 목표로 함

- **평가 메커니즘**
  - **Chain-of-Thought (CoT) 활용**: CoT 접근 방식을 사용하여 LLM이 생성한 중간 평가 단계를 자동으로 구성하며, 이를 통해 보다 상세한 컨텍스트와 지침을 제공함
  - **Form-Filling Paradigm**: 평가 과정에서 LLM은 평가하려는 태스크의 세부 사항을 기반으로 형식을 채워 점수를 산출함

- **평가의 유효성**: 다양한 NLG 작업에서 기존의 reference-control 및 reference-free 평가 방법을 능가하는 성능을 달성함

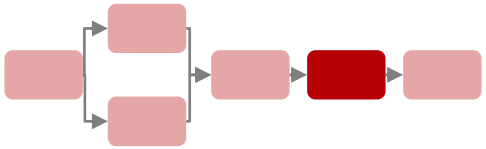
- **상관계수 및 성능**:
  - 사람의 평가와의 상관계수가 높으며, 평가 도구로 사용될 수 있음
  - GPT-4를 사용한 실험에서 특히 높은 상관성을 달성함



Ref. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment (Yang Liu. et. al. 2023)

3. 실험 설계 (MODEL EVALUATION)
Vision Language Model: G-Eval prompt

MODEL SELECTION MODEL EVALUATION



1 Task Introduction

You will be provided with one AI model-diagnosed disease report provided based on the answer document. Your task is to evaluate the AI Model-Diagnosed Disease Report based on specific criteria. Please carefully review and follow the instructions below. Keep the answer document open while reviewing, and refer to it as needed.

2 Evaluation Criteria

Evaluation Criteria:

1. Accuracy and Completeness (1-10):
- Correct Diagnosis: Assess whether the diagnosed disease accurately matches the symptoms presented in the answer document.
- Coverage of Symptoms: Determine if the diagnosed result covers all relevant symptoms mentioned in the answer document.
- Informativeness: Evaluate the level of detail and completeness of information provided in the diagnosed result.

2. Interpretability and Explanation (1-10):
- Clarity of Diagnosis: Evaluate how clearly the diagnosed disease and associated information are presented.
- Explanation of Reasoning: Assess whether the AI model provides clear explanations for its diagnosis, including the reasoning behind each prediction.
- Ease of Interpretation: Determine how easily healthcare professionals can interpret and understand the diagnosed result.

3. Safety and Reliability (1-10):
- Risk Assessment: Assess potential risks associated with incorrect diagnoses provided by the model and suggest safeguards to minimize adverse outcomes.
- Reliability of Predictions: Evaluate the reliability of the model's predictions in real-world clinical settings.

4. Adaptability and Generalization (1-10):
- Ability to Generalize: Assess how well the model performs across different skin disease datasets and variations in input data.
- Adaptability to New Cases: Determine the model's ability to adapt to new cases and provide accurate diagnoses without human supervision.

5. Sensitivity and Specificity (1-10):
- Sensitivity: Evaluate the true positive rate of the AI model in correctly identifying positive cases.
- Specificity: Assess the true negative rate of the AI model in correctly identifying negative cases.

3 Evaluation Step

Evaluation Steps:
1. Read the diagnosed disease result carefully and identify the possible disease.
2. Evaluate the diagnosed result based on each criterion provided above.
3. Assign a score for each criterion on a scale of 1 to 10, where 1 is the lowest and 10 is the highest.

4 Input Context

Provided Document:
<answer document>
{source}
</answer document>

5 Input Target

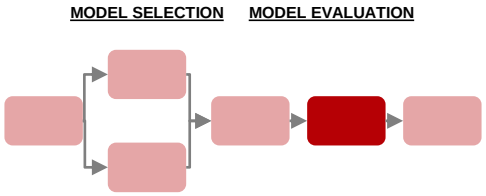
<AI Model-Diagnosed Disease Report>
{report}
</AI Model-Diagnosed Disease Report>

6 Output Format

Evaluation Form (No explanation, scores ONLY):
1. Accuracy and Completeness:
2. Interpretability and Explanation:
3. Safety and Reliability:
4. Adaptability and Generalization:
5. Sensitivity and Specificity:

서강대학교 SOGANG UNIVERSITY

3. 실험 설계 (MODEL EVALUATION)
Vision Language Model: G-Eval 평가 기준



1. Accuracy and Completeness 정확도와 완성도 (1-10):
- 올바른 진단: 진단 결과가 소스 문서에 기술된 증상과 정확히 일치하는지 확인합니다.
- 증상의 포괄성: 진단 결과가 소스 문서에 언급된 모든 관련 증상을 포함하는지 판단합니다.
- 정보 제공의 정확성: 진단 결과의 정보 제공 수준과 완전성을 평가합니다.

2. Interpretability and Explanation 해석 가능성과 설명 (1-10):
- 진단의 명확성: 진단된 질병과 관련 정보가 명확하게 제시되는지 평가합니다.
- 추론의 설명: 모델이 각 예측의 근거를 포함하여 진단에 대한 명확한 설명을 제공하는지 평가합니다.
- 해석 용이성: 의료 전문가들이 진단 결과를 해석하고 이해하기 쉬운지 판단합니다.

3. Safety and Reliability 안전성과 신뢰성 (1-10):
- 위험 평가: 모델이 제공한 부정확한 진단으로 인한 잠재적인 위험을 평가하고 부작용을 최소화하기 위한 보호장치를 제안합니다.
- 예측의 신뢰성: 모델의 예측이 현실 세계의 임상 환경에서의 신뢰성을 평가합니다.

4. Adaptability and Generalization 적응성과 일반화 (1-10):
- 일반화 능력: 모델이 다양한 피부 질환 데이터셋과 입력 데이터의 변화에 대해 얼마나 잘 수행하는지 평가합니다.
- 새로운 케이스에 대한 적응성: 모델이 새로운 케이스에 대해 적응하고 인간 감독 없이 정확한 진단을 제공하는 능력을 판단합니다.

5. Sensitivity and Specificity 민감도와 특이도 (1-10):
- 민감도: 모델이 양성 사례를 올바르게 식별하는 데에 대한 신뢰성을 평가합니다.
- 특이도: 모델이 음성 사례를 올바르게 식별하는 데에 대한 신뢰성을 평가합니다.

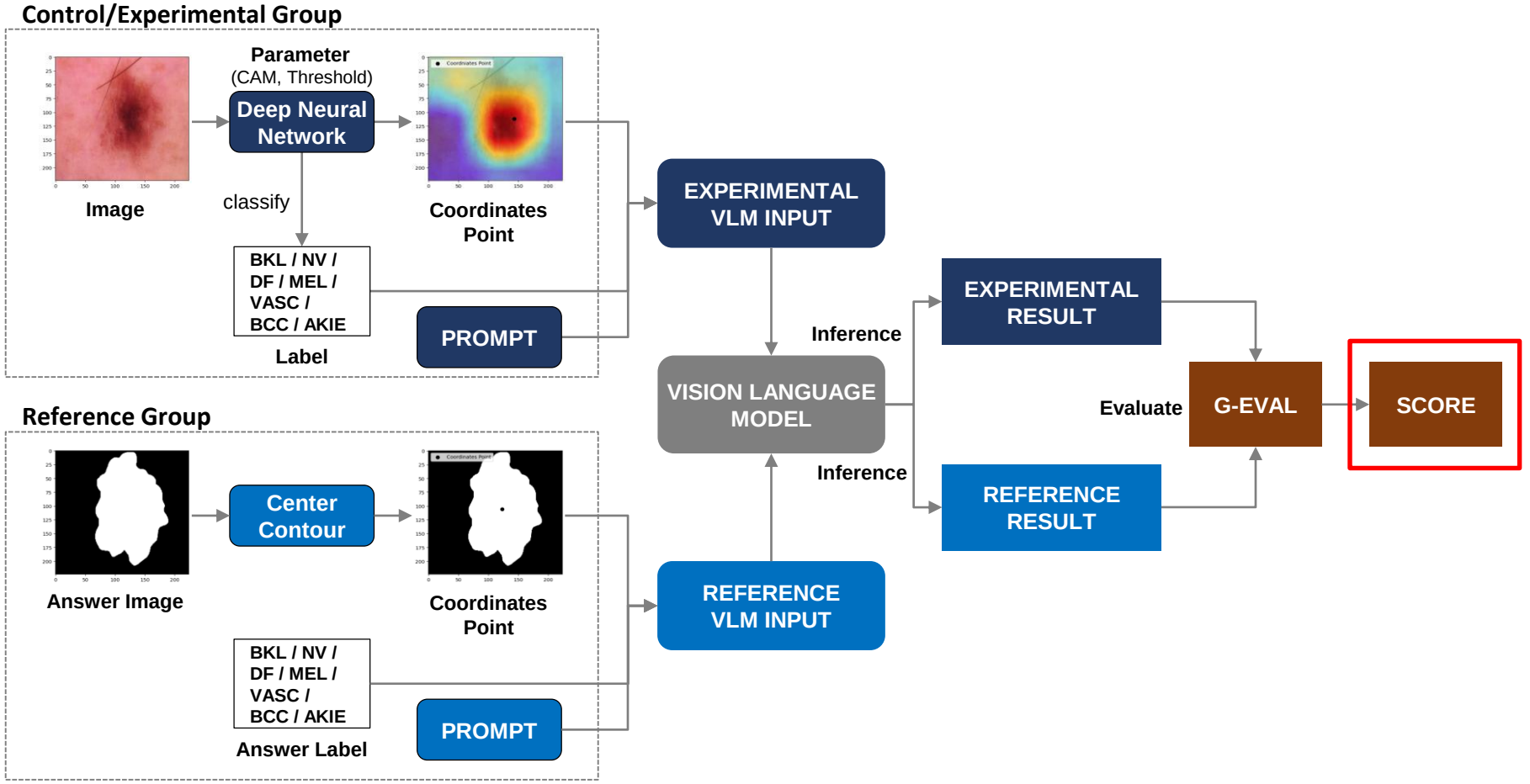
Ref. Evaluating large language models for use in healthcare: A framework for translational value assessment

3. 실험 설계 (MODEL EVALUATION)

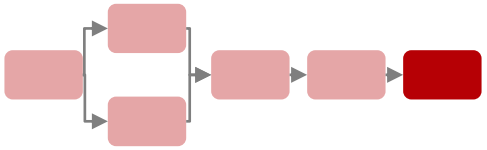
MODEL EVALUATION PROCESS

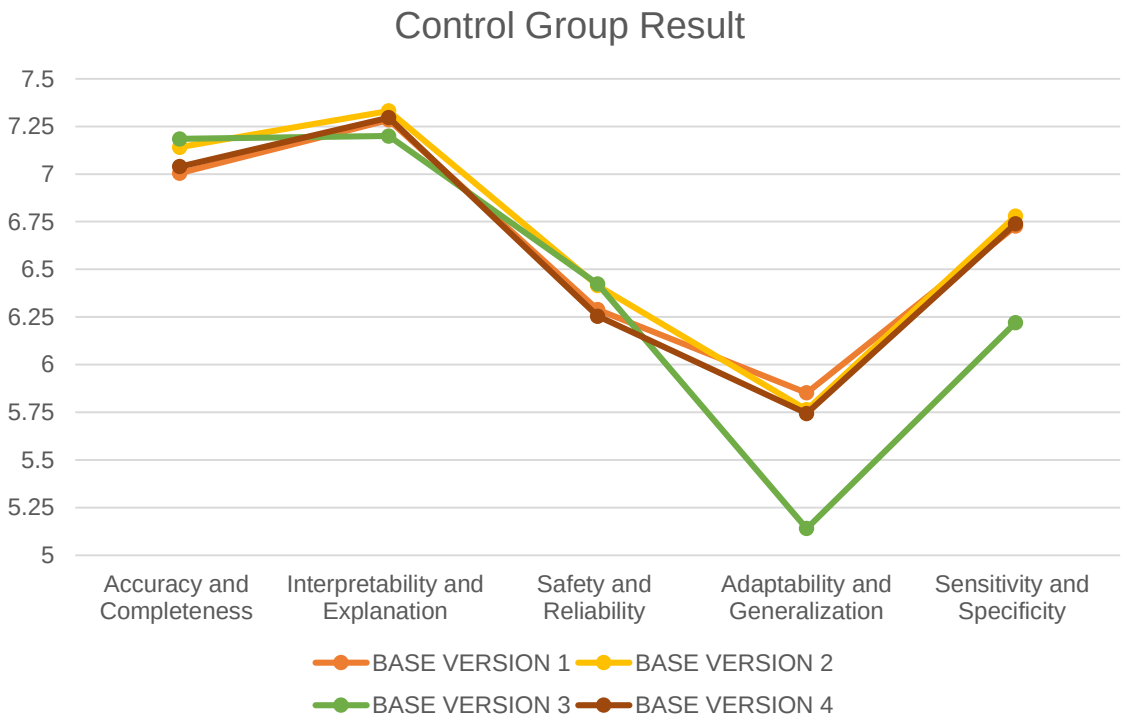
Evaluate Vision Language Model Performance

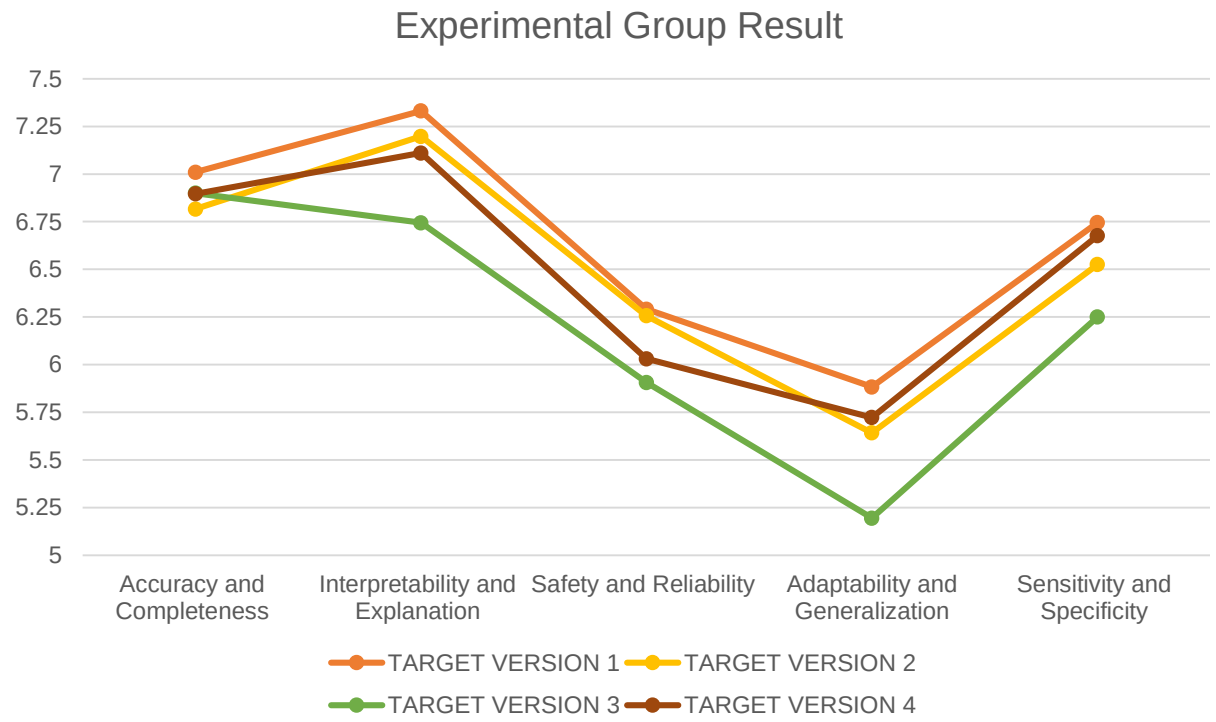
Utilizing the coordinates point with VLM



3. 실험 설계 (MODEL EVALUATION)
Structured answer prompts G-Eval 결과

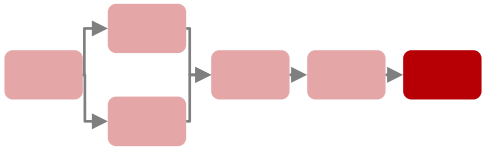






Version	Accuracy and Completeness			Interpretability and Explanation			Safety and Reliability			Adaptability and Generalization			Sensitivity and Specificity		
	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%
Ver 1	7.00	7.01	0.07%	7.28	7.33	0.64%	6.29	6.29	0.05%	5.85	5.88	0.53%	6.73	6.75	0.28%
Ver 2	7.14	6.82	-4.54%	7.33	7.20	-1.82%	6.41	6.26	-2.46%	5.76	5.64	-2.10%	6.78	6.52	-3.75%
Ver 3	7.18	6.90	-3.96%	7.20	6.74	-6.32%	6.42	5.91	-8.05%	5.14	5.19	1.07%	6.22	6.25	0.47%
Ver 4	7.04	6.90	-2.03%	7.30	7.11	-2.54%	6.25	6.03	-3.57%	5.74	5.72	-0.36%	6.74	6.68	-0.93%

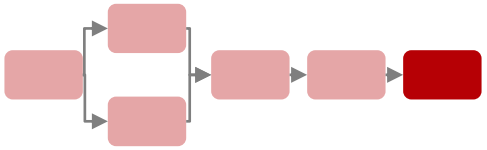
3. 실험 설계 (MODEL EVALUATION)
Structured answer prompts G-Eval 결과



Version	Accuracy and Completeness			Interpretability and Explanation			Safety and Reliability			Adaptability and Generalization			Sensitivity and Specificity		
	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%
Ver 1	7.00	7.01	0.07%	7.28	7.33	0.64%	6.29	6.29	0.05%	5.85	5.88	0.53%	6.73	6.75	0.28%
Ver 2	7.14	6.82	-4.54%	7.33	7.20	-1.82%	6.41	6.26	-2.46%	5.76	5.64	-2.10%	6.78	6.52	-3.75%
Ver 3	7.18	6.90	-3.96%	7.20	6.74	-6.32%	6.42	5.91	-8.05%	5.14	5.19	1.07%	6.22	6.25	0.47%
Ver 4	7.04	6.90	-2.03%	7.30	7.11	-2.54%	6.25	6.03	-3.57%	5.74	5.72	-0.36%	6.74	6.68	-0.93%

- 대부분의 version에서 성능 향상이 크게 미미하며, 오직 **version 1에서만 경미한 개선이 관찰됨**
- 전체 클래스 정보를 통합하는 것 (version 1)은 모델 성능을 약간 개선하는데 도움이 되지만, 전반적인 영향은 미미함
- 오히려 전체적인 성능 하락이 나타나고 있음
- 다른 버전 간에 직접 점수를 비교할 수 없는 한계점 나타남

3. 실험 설계 (MODEL EVALUATION)
Structured answer prompts: Version 1 결과



VERSION 1 STRUCTURED

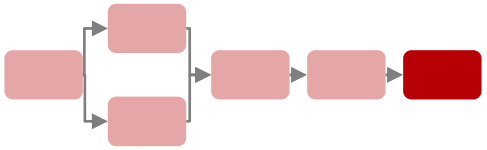
BKL

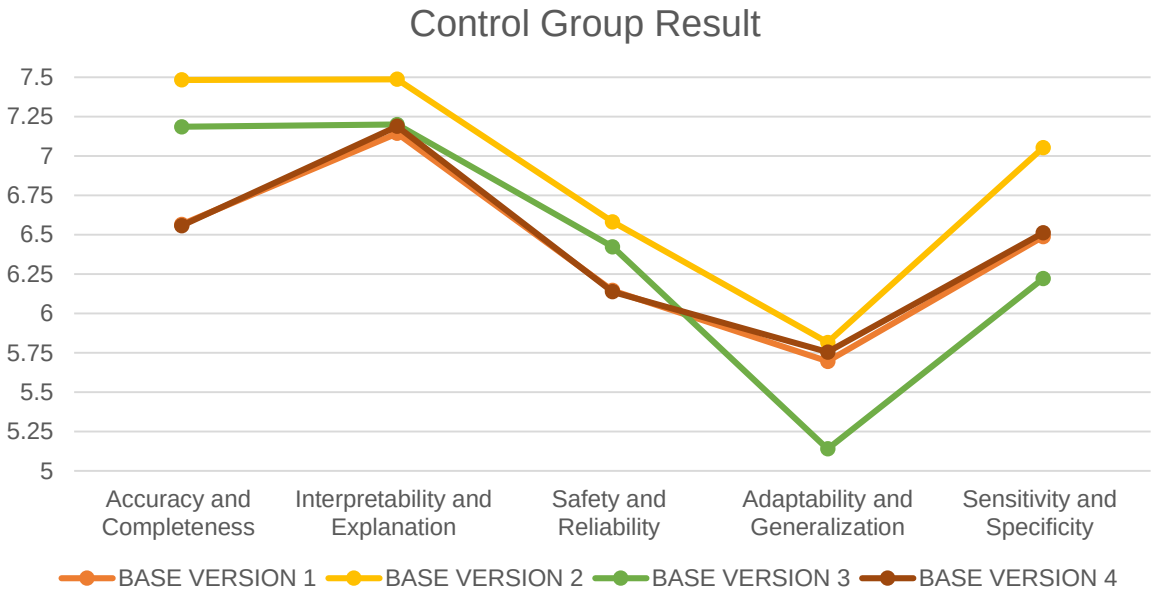
Control	The image shows a close-up view of a skin lesion, which is a benign keratosis-like lesion (BKL). The lesion appears to be pink and has a slightly rough texture. It is located on the upper part of the skin, covering a significant portion of the area. The diagnosis is based on the visual appearance of the lesion and the fact that it is a benign growth.
Experimental	The image shows a close-up view of a lesion on a person's skin, which is indicative of a BKL (Benign Keratosis-like Lesions) disease. The lesion appears to be a crusty, pink, and potentially inflamed area, with a size of 224x224 pixels. The diagnosis is based on the Class Activation Map (CAM) generated by the classification model, which highlighted the region of interest at the pixel coordinates (111, 111) as being indicative of BKL disease.
Reference	Based on the image, the lesion appears to be a BKL (Benign Keratosis-like Lesion). The image shows a close-up view of a pink, raised, and possibly crusty lesion on the skin. The lesion's size is approximately 1.5 cm in diameter, and it is located on the upper part of the skin. The presence of a hair follicle nearby suggests that the lesion might be related to the hair follicle. However, without further information or a larger sample, it is difficult to determine the exact diagnosis with certainty.

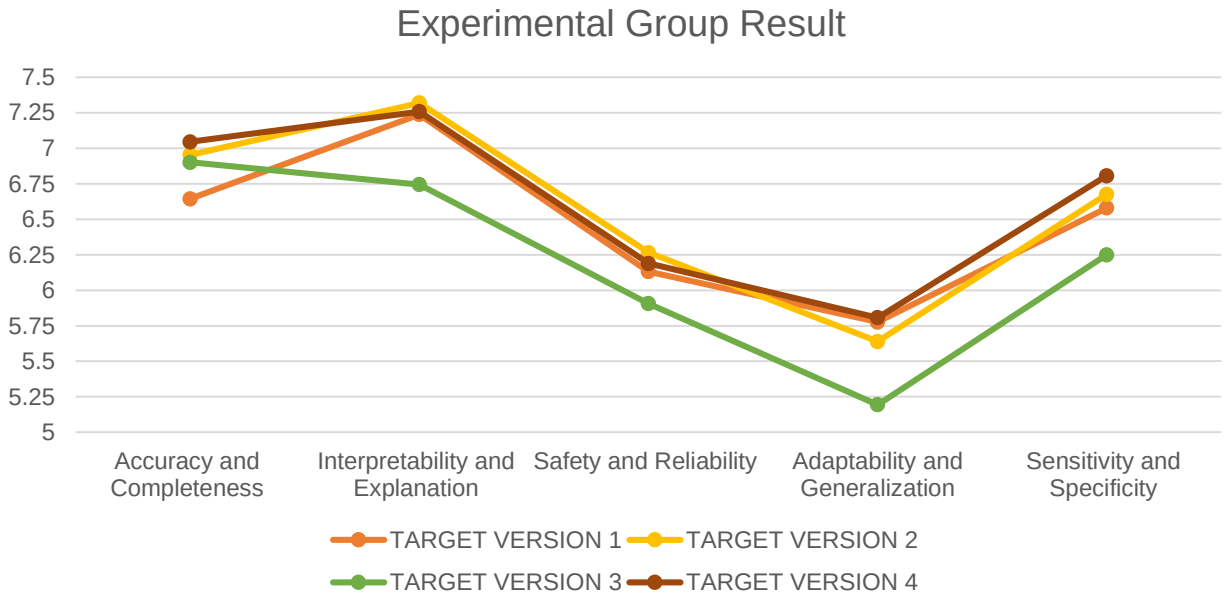
Ref. ISIC_0033488

3. 실험 설계 (MODEL EVALUATION)
Focused answer prompts G-Eval 결과

MODEL SELECTION MODEL EVALUATION



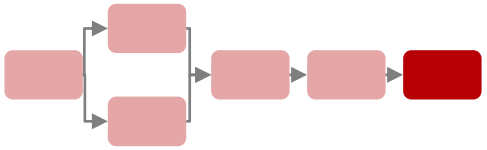




Version	Accuracy and Completeness			Interpretability and Explanation			Safety and Reliability			Adaptability and Generalization			Sensitivity and Specificity		
	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%
Ver 1	6.56	6.64	1.20%	7.14	7.24	**1.31%**	6.15	6.13	-0.23%	5.70	5.78	**1.42%**	6.49	6.58	1.43%
Ver 2	7.48	6.95	-7.07%	7.49	7.32	-2.25%	6.58	6.27	-4.81%	5.82	5.64	-3.04%	7.05	6.68	-5.34%
Ver 3	7.18	6.90	-3.96%	7.20	6.74	-6.32%	6.42	5.91	-8.05%	5.14	5.19	1.07%	6.22	6.25	0.47%
Ver 4	6.56	7.04	**7.44%**	7.19	7.26	0.97%	6.14	6.19	**0.83%**	5.75	5.81	0.92%	6.51	6.81	**4.54%**

서강대학교 SOGANG UNIVERSITY

3. 실험 설계 (MODEL EVALUATION)
Focused answer prompts G-Eval 결과

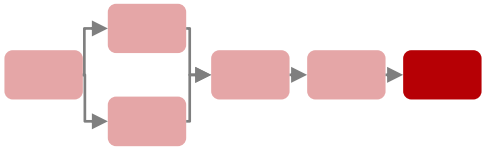


Version	Accuracy and Completeness			Interpretability and Explanation			Safety and Reliability			Adaptability and Generalization			Sensitivity and Specificity		
	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%	Con.	Exp.	%
Ver 1	6.56	6.64	1.20%	7.14	7.24	1.31%	6.15	6.13	-0.23%	5.70	5.78	1.42%	6.49	6.58	1.43%
Ver 2	7.48	6.95	-7.07%	7.49	7.32	-2.25%	6.58	6.27	-4.81%	5.82	5.64	-3.04%	7.05	6.68	-5.34%
Ver 3	7.18	6.90	-3.96%	7.20	6.74	-6.32%	6.42	5.91	-8.05%	5.14	5.19	1.07%	6.22	6.25	0.47%
Ver 4	6.56	7.04	7.44%	7.19	7.26	0.97%	6.14	6.19	0.83%	5.75	5.81	0.92%	6.51	6.81	4.54%

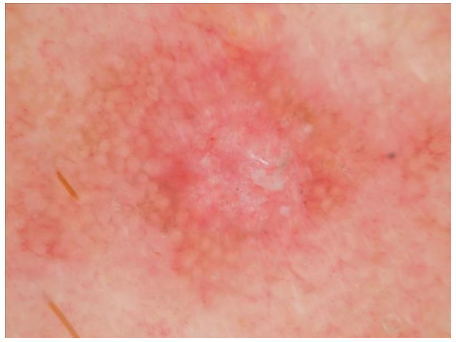
- Structed answer prompts 경우와는 다르게 **version 1과 4에서 성능 향상을 보여줌**

- **Version 4는 모든 평가 지표에서 상당한 개선**을 보였으며, 통제 그룹 대비 다음과 같은 성능 향상됨
  - Accuracy and Completeness 정확도와 완성도: +7.44%
  - Interpretability and Explanation 해석 가능성과 설명: +0.97%
  - Safety and Reliability 안전성과 신뢰성: +0.83%
  - Adaptability and Generalization 적응성과 일반화: +0.92%
  - Sensitivity and Specificity 민감도와 특이도: +4.54%

3. 실험 설계 (MODEL EVALUATION)
Focused answer prompts: Version 4 결과



VERSION 4 FOCUSED



AKIEC

Control	The image shows a close-up view of a skin lesion, which is a red, raised, and slightly crusty patch on the skin. The lesion appears to be a combination of a benign keratosis-like lesion and an intraepithelial carcinoma. The diagnosis is consistent with Actinic Keratosis and Intraepithelial Carcinoma (AKIEC). This type of lesion is commonly found on sun-exposed areas of the skin and can be removed through surgical excision or other treatments.
Experimental	The image shows a close-up view of a skin lesion, possibly an Actinic Keratosis and Intraepithelial Carcinoma (AKIEC) on the skin of a person. The lesion appears to be red and has a rough texture. The region of interest, around coordinates (112, 111), shows signs of inflammation and abnormal cell growth, which are characteristic of AKIEC. The diagnosis suggests that the lesion is likely benign, but it is essential to monitor its progression and consult a dermatologist for further evaluation and treatment.
Reference	The image shows a close-up view of a skin lesion, possibly an actinic keratosis or intraepithelial carcinoma. The lesion is located on the upper right side of the image, with a size of 224x224 pixels. The region of interest around coordinates (109, 97) displays a pink, raised, and slightly crusty patch of skin, which is indicative of a skin disease. The surrounding area appears to be slightly inflamed, and the overall appearance suggests a potential skin cancer or precancerous condition. It is important to consult a dermatologist for further evaluation and appropriate treatment.

Ref. ISIC_0031108

Contents

4. 성능 고도화 방안

- CAM의 semantic segmentation 결과에서 50%의 정확도에 불과했던 점을 고려하여, 성능 향상 시 VLM 품질 변화 검토
- 성능이 향상된 **“Focused Answer Prompts”**를 바탕으로, 성능 향상을 이룬 **version 1과 4 사이에서 성능 변화를 검증함**

Coordinates point group	L2 Distance Score
Initial L2 Distance for Experimental Group	28.97395612
Average L2 Distance to Reference	14.54104287
Adjusted L2 Distance with Reference Integration	0

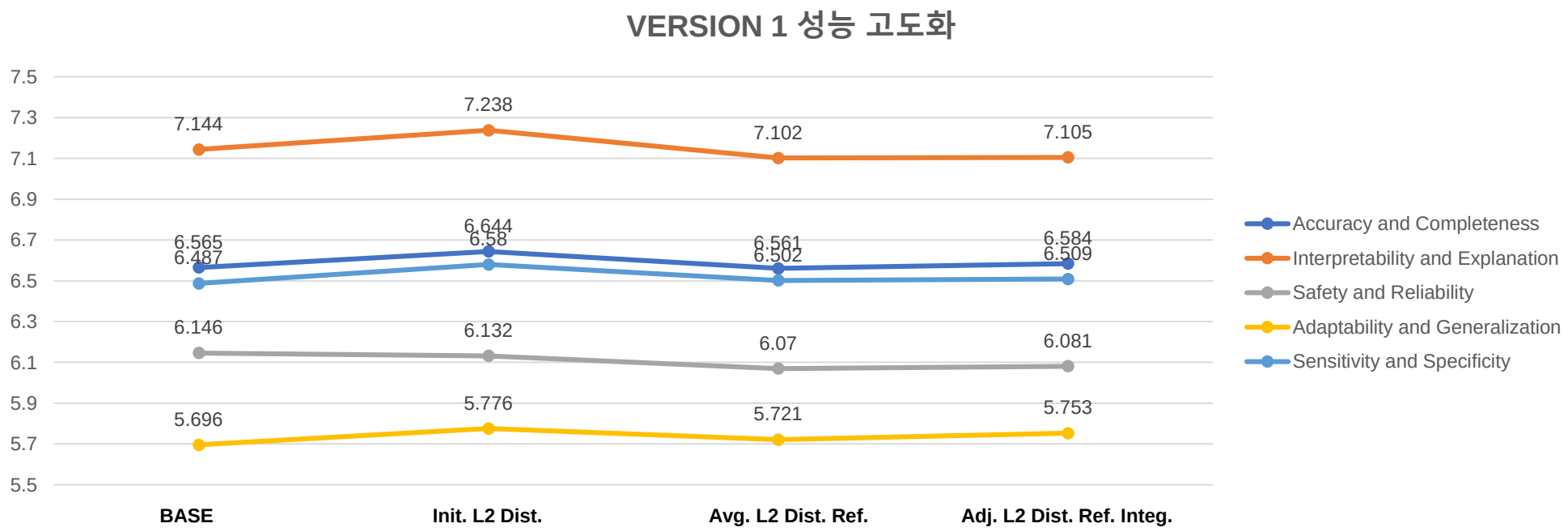
- **L2 distance 지표 사용:** L2 distance 지표를 사용하여 모델의 정확성을 평가하며, L2 거리 값이 작을수록 모델 예측이 정답과 밀접하게 일치함을 나타냄

- **Initial L2 Distance for Experimental Group**: 실험 그룹의 L2 거리 초기 측정값으로, 참조 데이터 없이 모델 예측의 정확성을 나타냄
- **Average L2 Distance to Reference**: 실험 그룹 결과와 참조 데이터 사이에 계산된 평균 거리로, 실제 데이터에 대비한 개선된 측정을 제공함
- **Adjusted L2 Distance with Reference Integration**: 참조 그룹 데이터를 사용하여, 이상적인 시나리오에서 L2 거리가 일치하는 경우

4. 성능 고도화 방안 (Version 1)

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Metric	Base	Init. L2 Dist.	% Imp. from Base	Avg. L2 Dist. Ref.	% Imp. to Ref.	Adj. L2 Dist. Ref. Integ.	% Imp. with Integ.
Acc. & Comp.	6.565	6.644	1.20%	6.561	-0.06%	6.584	0.29%
Interp. & Expl.	7.144	7.238	1.31%	7.102	-0.59%	7.105	-0.54%
Safety & Rel.	6.146	6.132	-0.23%	6.07	-1.23%	6.081	-1.06%
Adapt. & Gen.	5.696	5.776	1.42%	5.721	0.44%	5.753	1.02%
Sens. & Spec.	6.487	6.58	1.43%	6.502	0.23%	6.509	0.34%



4. 성능 고도화 방안 (Version 1)

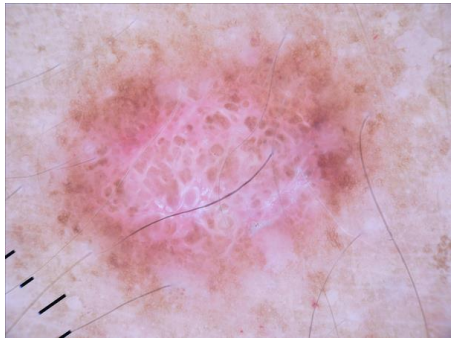
USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Metric	Base	Init. L2 Dist.	% Imp. from Base	Avg. L2 Dist. Ref.	% Imp. to Ref.	Adj. L2 Dist. Ref. Integ.	% Imp. with Integ.
Acc. & Comp.	6.565	6.644	1.20%	6.561	-0.06%	6.584	0.29%
Interp. & Expl.	7.144	7.238	1.31%	7.102	-0.59%	7.105	-0.54%
Safety & Rel.	6.146	6.132	-0.23%	6.07	-1.23%	6.081	-1.06%
Adapt. & Gen.	5.696	5.776	1.42%	5.721	**0.44%**	5.753	**1.02%**
Sens. & Spec.	6.487	6.58	**1.43%**	6.502	0.23%	6.509	0.34%

- **성능 변화**: 버전 1에서는 성능에 유의미한 변화가 관찰되지 않았습니다. 오히려 일부 영역에서는 성능 저하가 나타남
- **안전성 및 신뢰성 지표 하락**: 특히 안전성과 신뢰성 지표에서 성능 저하가 감지되었으며, 이는 의료 진단에서 신뢰성과 설명 품질에 중대한 영향을 미치는 것으로 개선이 필요함

4. 성능 고도화 방안 (Version 1 답변)

VERSION 1 | FOCUSED



DF
(ISIC_0030427)

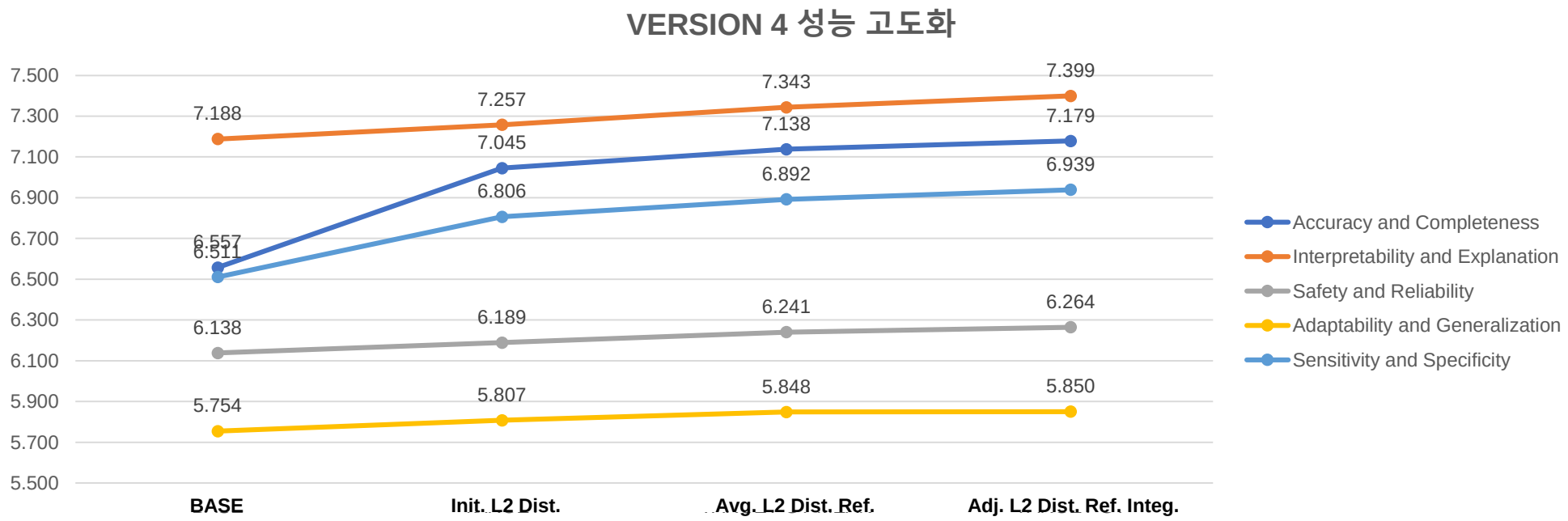
Control	The image shows a close-up view of a skin lesion, which is a Dermatofibroma. Dermatofibromas are benign growths that are usually brown or black in color and can be found on the face, neck, or other parts of the body. They are usually slow-growing and may cause symptoms such as itching or discomfort. In some cases, they may be removed if they become problematic or if there is a concern for malignancy.
Experimental	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, with a pink and brown color. The lesion is surrounded by a thin, white border. The diagnosis is based on the presence of a single, well-defined, and relatively flat lesion, which is indicative of a Dermatofibroma. The image does not show any other skin lesions or abnormalities.
Average L2 Distance to Reference	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, on the upper chest area of a person. The lesion appears to be brown and has a slightly rough texture. The size of the lesion is approximately 1.5 centimeters in diameter. The surrounding skin appears normal and there are no other visible abnormalities in the image.
Adjusted L2 Distance with Reference Integration	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, on the upper chest area of a person. The lesion appears to be brown and has a slightly rough texture. The size of the lesion is approximately 1.5 centimeters in diameter. The surrounding skin appears normal, without any visible abnormalities or signs of infection. The diagnosis is based on the visual examination of the image and the presence of typical features of a Dermatofibroma.
Reference	The image shows a close-up view of a skin lesion, possibly a dermatofibroma, with a pink and brown color. The lesion is surrounded by a thin, white border. The region of interest, around coordinates (108, 102), shows a clear, well-defined, raised, and somewhat rough border. This indicates that the lesion is likely a dermatofibroma, a common benign skin growth. The diagnosis is based on the visual examination of the image, and a formal diagnosis by a medical professional is recommended for a definitive conclusion.

좌표 정보를 제공하였지만, 실제 LLM이 답변을 추론하는 과정에서는 좌표 정보를 활용하지 못 하는 경우

4. 성능 고도화 방안 (Version 4)

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	**Version 4**	Version 3

Metric	Base	Init. L2 Dist.	% Imp. from Base	Avg. L2 Dist. Ref.	% Imp. to Ref.	Adj. L2 Dist. Ref. Integ.	% Imp. with Integ.
Acc. & Comp.	6.557	7.045	**7.40%**	7.138	**8.90%**	7.179	**9.50%**
Interp. & Expl.	7.188	7.257	1.00%	7.343	2.20%	7.399	2.90%
Safety & Rel.	6.138	6.189	0.80%	6.241	1.70%	6.264	2.10%
Adapt. & Gen.	5.754	5.807	0.90%	5.848	1.60%	5.85	1.70%
Sens. & Spec.	6.511	6.806	4.50%	6.892	5.90%	6.939	6.60%



서강대학교 SOGANG UNIVERSITY

4. 성능 고도화 방안 (Version 4)

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	**Version 4**	Version 3

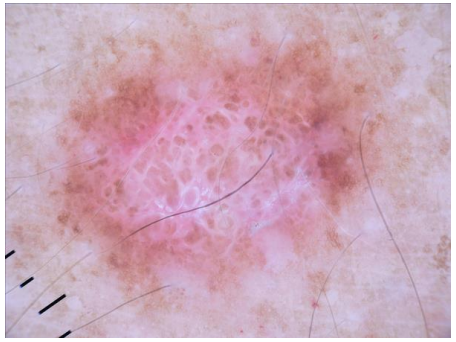
Metric	Base	Init. L2 Dist.	% Imp. from Base	Avg. L2 Dist. Ref.	% Imp. to Ref.	Adj. L2 Dist. Ref. Integ.	% Imp. with Integ.
Acc. & Comp.	6.557	7.045	**7.40%**	7.138	**8.90%**	7.179	**9.50%**
Interp. & Expl.	7.188	7.257	1.00%	7.343	2.20%	7.399	2.90%
Safety & Rel.	6.138	6.189	0.80%	6.241	1.70%	6.264	2.10%
Adapt. & Gen.	5.754	5.807	0.90%	5.848	1.60%	5.85	1.70%
Sens. & Spec.	6.511	6.806	4.50%	6.892	5.90%	6.939	6.60%

- Version 4는 version 1에 비해 L2 distance가 개선됨에 따라 모든 주요 성능 지표에서 상당히 개선함

- **정확성 및 완전성**
  - 통제 그룹에서 실험 그룹으로 이동하면서 7.4% 향상
  - 실험 그룹에서 참조 평균 그룹으로 8.9% 향상
  - 참조 그룹에서 9.5% 향상
- **안전성 및 신뢰성**
  - 각 단계별로 0.8%, 1.7%, 2.1%의 점진적인 개선을 보임
- **민감성 및 특이성**
  - 각 단계별로 각각 4.5%, 5.9%, 6.6% 향상됨
- **해석 가능성 및 설명, 적응성 및 일반화**
  - 좌표가 답변으로 수렴함에 따라 향상되었으며, user prompt에서 좌표 정보를 명시적으로 포함하는 것이 LLM이 답변에 가까운 정보를 효과적으로 활용하는 데 크게 도움이 된다는 것을 확인함

4. 성능 고도화 방안 (Version 4 답변)

VERSION 4 | FOCUSED



DF
(ISIC_0030427)

Control	The image shows a close-up view of a skin lesion, which is a Dermatofibroma. Dermatofibromas are benign growths that are usually brown or black in color and can be found on the face, neck, or other parts of the body. They are usually slow-growing and may cause symptoms such as itching or discomfort. In some cases, they may be removed if they become problematic or if there is a concern for malignancy.
Experimental	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, with a pink and brown color. The lesion is surrounded by a mix of pink and brown tissue, which might be indicative of a benign growth. The region of interest around coordinates (143, 112) shows a more prominent pink and brown color, which could be a sign of inflammation or abnormal cell growth. However, without further information or a larger image, it is difficult to provide a definitive diagnosis. It is important to consult a medical professional for a proper diagnosis and treatment plan.
Average L2 Distance to Reference	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, with a pink and brown color. The lesion is surrounded by a mix of pink and brown tissue, which is indicative of a benign growth. The region of interest, around coordinates (126, 107) shows a more prominent pink color, which is consistent with the overall appearance of the lesion. The diagnosis suggests that the lesion is likely a benign growth, such as a Dermatofibroma, which is a common skin condition characterized by the growth of fibrous tissue under the skin.
Adjusted L2 Distance with Reference Integration	The image shows a close-up view of a skin lesion, possibly a Dermatofibroma, with a pink and brown color. The lesion is surrounded by a mix of pink and brown tissue, which might be indicative of a benign growth. The region of interest around coordinates (108, 102) shows a more prominent pink and brown color, which could be a sign of inflammation or abnormal cell growth. However, without further information or a larger image, it is difficult to provide a definitive diagnosis. It is important to consult a medical professional for a proper diagnosis and treatment plan.
Reference	The image shows a close-up view of a skin lesion, possibly a dermatofibroma, with a pink and brown color. The lesion is surrounded by a thin, white border. The region of interest, around coordinates (108, 102) shows a clear, well-defined, raised, and somewhat rough border. This indicates that the lesion is likely a dermatofibroma, a common benign skin growth. The diagnosis is based on the visual examination of the image, and a formal diagnosis by a medical professional is recommended for a definitive conclusion.

실제 LLM이 답변을 추론하는 과정에서 좌표 정보를 활용하는 경우

Contents

5. 해석 및 결론

1. VLM 잠재력
- VLM 활용하여, 의료 이미지 분석의 정확성과 해석 가능성을 크게 향상시킬 수 있음
- 이미지 라벨 데이터셋에서 훈련된 모델을 활용하여, CAM 방법을 활용하여 모델의 설명력을 개선할 수 있음

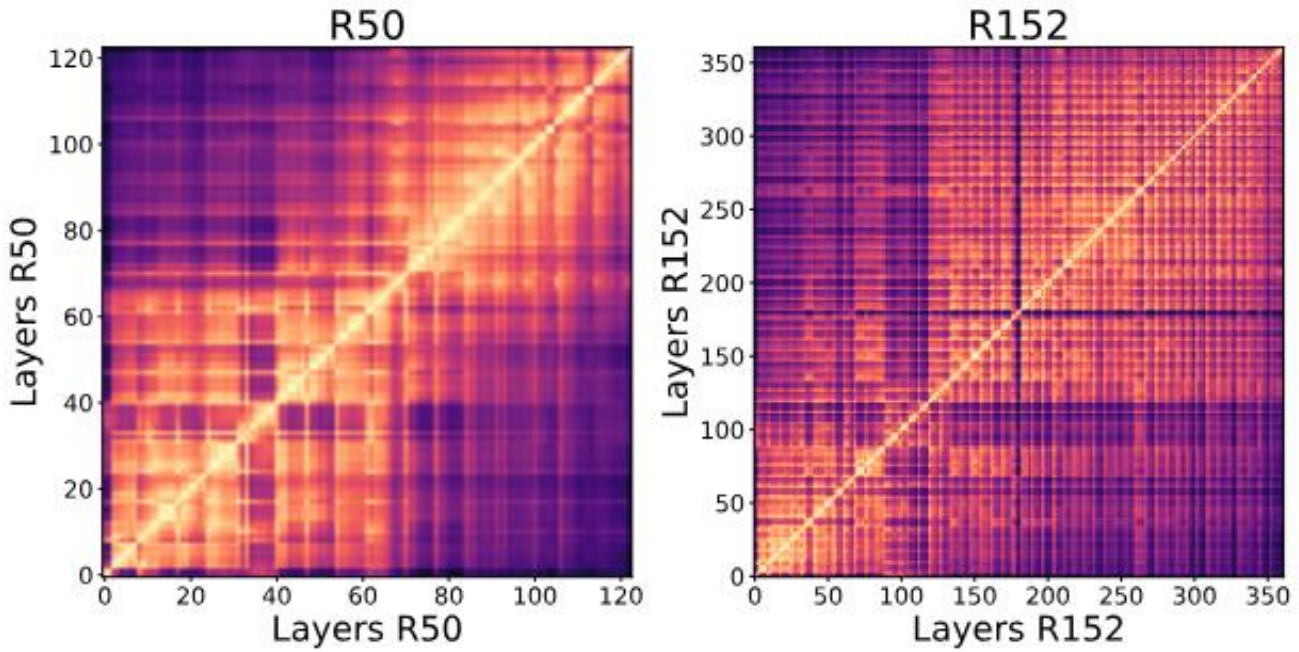
2. Focused answer prompt 방법론 효과
- 특정 의료 훈련 없이도 의료 환경에서 모델을 효과적으로 활용할 수 있음
- 모델은 불완전하거나 부정확한 의미론적 정보를 사용하여도 적응성을 보여줌
- 더 정확한 좌표 정보가 제공될 경우, 성능이 크게 향상됨

3. 성능 향상 전략
- System prompt, user prompt tuning을 통해 모델 성능을 크게 향상시킴

4. 주요 한계 및 개선 사항
- HAM10000 데이터셋은 주로 밝은 피부 톤의 이미지로 구성되어 일반화 가능성이 제한됨
- 글로벌 인구 통계를 더 잘 대표하는 다양한 데이터셋 필요함
- Semantic segmentation 성능 한계로, 최신 Capsule Networks (CapsNet)와 같은 기술로 해결하려고 함

Contents

Appendix 1. Resnet18 outperforms resnet50, 101



The layer similarity heatmap of resnet models showing reduced similarity between the early and late layers.
(Adapted from Fig. 1, Do Vision Transformers See Like Convolutional Neural Networks?")]

Appendix 2. Version 1 prompt

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosing {y_label} skin disease. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions calculated by contours centroid.'"
IMAGE_TOKEN = f""\nHuman: <image>""	IMAGE_TOKEN = f""\nHuman: <image>""	IMAGE_TOKEN = f""\nHuman: <image>""
USER_PROMPT = f""\nHuman: Please diagnose the image""	**USER_PROMPT = f""\nHuman: Please diagnose the image""**	**USER_PROMPT = f""\nHuman: Please diagnose the image""**

Appendix 3. Version 2 prompt

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosing the {y_label} skin disease. Pay close attention to the pixel at coordinates {max_activation_coord}, as this point indicates a region calculated by the centroid of contours.'"
IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"
USER_PROMPT = f"\nHuman: Please diagnose the image"	**USER_PROMPT = f"\nHuman: Please diagnose the image"**	**USER_PROMPT = f"\nHuman: Please diagnose the image"**

Appendix 4. Version 3 prompt

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **{short name label} : {full name label}** You will receive an image with dimensions 224 width and 224 height for diagnosing the {y_label} skin disease. Pay close attention to the pixel at coordinates {max_activation_coord}, as this point indicates a region calculated by the centroid of contours.'"
IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"
USER_PROMPT = f"\nHuman: Please diagnose the image"	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**

Appendix 5. Version 4 prompt

USER PROMPT	SYSTEM PROMPT	
	7 CLASS	1 CLASS
No Coordinates Point	Version 1	Version 2
With Coordinates Point	Version 4	Version 3

Control Group	Experimental Group	Reference Group
SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy.'"	SYSTEM_PROMPT = f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosis. The classification model identified the image as {inference_label} disease with 85% accuracy. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions of interest identified by the model's Class Activation Map (CAM) for better visual performance.'"	SYSTEM_PROMPT= f"'System: You are a skin disease diagnosis explainable AI chatbot. Your purpose is to diagnose skin diseases using images. - **BKL: Benign Keratosis-like Lesions** - **NV: Melanocytic Nevi** - **DF: Dermatofibroma** - **MEL: Melanoma** - **VASC: Vascular Lesions** - **BCC: Basal Cell Carcinoma** - **AKIEC: Actinic Keratosis and Intraepithelial Carcinoma** You will receive an image with dimensions 224 width and 224 height for diagnosing {y_label} skin disease. Please focus on the pixel at coordinates {max_activation_coord}, as this point is indicative of regions calculated by contours centroid.'"
IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"	IMAGE_TOKEN = f"\nHuman: <image>"
USER_PROMPT = f"\nHuman: Please diagnose the image"	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**	**USER_PROMPT = f"\nHuman: Diagnose the image, paying attention to the region around coordinates {max_activation_coord}."**

Appendix 6. Grad-CAM RESULT

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_10_aug_None	0.3056	0.3225	0.3380	0.3454	0.3388	0.3136	0.2654	0.1887	0.0824
resnet18_iter_10_aug_Random	0.3305	0.3772	0.4151	0.4373	0.4384	0.4128	0.3557	0.2589	0.1186
resnet18_iter_10_aug_Whole	0.3311	0.3796	0.4281	0.4663	0.4792	0.4564	0.3905	0.2785	0.1293
resnet18_iter_50_aug_None	0.3289	0.3569	0.3786	0.3876	0.3745	0.3355	0.2729	0.1882	0.0835
resnet18_iter_50_aug_Random	0.3235	0.3697	0.4176	0.4503	0.4601	0.4410	0.3881	0.2939	0.1438
resnet18_iter_50_aug_Whole	0.3358	0.3915	0.4448	0.4842	0.4932	0.4646	0.3967	0.2831	0.1242
resnet18_iter_100_aug_None	0.3230	0.3636	0.4083	0.4484	0.4673	0.4559	0.4010	0.2923	0.1335
resnet18_iter_100_aug_Random	0.3322	0.3811	0.4216	0.4472	0.4517	0.4283	0.3700	0.2693	0.1234
resnet18_iter_100_aug_Whole	0.3614	0.4275	0.4807	**0.5107**	0.5091	0.4699	0.3895	0.2703	0.1199
resnet50_iter_10_aug_None	0.3132	0.3323	0.3161	0.2823	0.2390	0.1904	0.1362	0.0797	0.0278
resnet50_iter_10_aug_Random	0.3130	0.3223	0.3084	0.2692	0.2114	0.1483	0.0905	0.0431	0.0123
resnet50_iter_10_aug_Whole	0.3149	0.3240	0.3198	0.2949	0.2468	0.1834	0.1204	0.0641	0.0203
resnet50_iter_50_aug_None	0.3279	0.3467	0.3625	0.3731	0.3628	0.3172	0.2419	0.1462	0.0509
resnet50_iter_50_aug_Random	0.3080	0.3126	0.2942	0.2563	0.2085	0.1557	0.1042	0.0563	0.0179
resnet50_iter_50_aug_Whole	0.3002	0.2951	0.2815	0.2576	0.2223	0.1749	0.1197	0.0648	0.0209
resnet50_iter_100_aug_None	0.3246	0.3450	0.3626	0.3751	0.3663	0.3217	0.2462	0.1497	0.0527
resnet50_iter_100_aug_Random	0.3084	0.3138	0.2959	0.2594	0.2123	0.1592	0.1062	0.0574	0.0185
resnet50_iter_100_aug_Whole	0.3011	0.3009	0.2822	0.2492	0.2063	0.1565	0.1041	0.0560	0.0189
resnet101_iter_10_aug_None	0.3148	0.3254	0.3274	0.3148	0.2854	0.2322	0.1582	0.0811	0.0249
resnet101_iter_10_aug_Random	0.3092	0.3077	0.2973	0.2721	0.2306	0.1772	0.1167	0.0604	0.0182
resnet101_iter_10_aug_Whole	0.2970	0.2969	0.2909	0.2745	0.2421	0.1913	0.1299	0.0662	0.0188
resnet101_iter_50_aug_None	0.3110	0.3199	0.3220	0.3140	0.2912	0.2486	0.1856	0.1071	0.0350
resnet101_iter_50_aug_Random	0.3164	0.3205	0.3072	0.2804	0.2370	0.1820	0.1216	0.0628	0.0185
resnet101_iter_50_aug_Whole	0.3005	0.3017	0.2974	0.2814	0.2486	0.1985	0.1354	0.0698	0.0206
resnet101_iter_100_aug_None	0.3119	0.3224	0.3292	0.3294	0.3172	0.2882	0.2335	0.1483	0.0515
resnet101_iter_100_aug_Random	0.3006	0.2947	0.2762	0.2439	0.1966	0.1388	0.0846	0.0409	0.0114
resnet101_iter_100_aug_Whole	0.3101	0.3138	0.3078	0.2870	0.2505	0.1972	0.1365	0.0737	0.0235

Appendix 7. Grad-CAM++ RESULT

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_10_aug_None	0.3123	0.3366	0.3556	0.3618	0.3527	0.3243	0.2714	0.1892	0.0795
resnet18_iter_10_aug_Random	0.3350	0.3874	0.4247	0.4441	0.4423	0.4138	0.3539	0.2531	0.1133
resnet18_iter_10_aug_Whole	0.3376	0.3935	0.4466	0.4856	0.4944	0.4667	0.3932	0.2748	0.1236
resnet18_iter_50_aug_None	0.3340	0.3675	0.3912	0.3996	0.3839	0.3418	0.2769	0.1910	0.0848
resnet18_iter_50_aug_Random	0.3257	0.3794	0.4291	0.4601	0.4664	0.4438	0.3887	0.2921	0.1428
resnet18_iter_50_aug_Whole	0.3394	0.4027	0.4576	0.4958	0.5030	0.4712	0.3989	0.2805	0.1225
resnet18_iter_100_aug_None	0.3300	0.3779	0.4276	0.4680	0.4821	0.4607	0.3982	0.2843	0.1274
resnet18_iter_100_aug_Random	0.3364	0.3898	0.4297	0.4527	0.4547	0.4271	0.3655	0.2630	0.1192
resnet18_iter_100_aug_Whole	0.3684	0.4396	0.4952	**0.5248**	0.5207	0.4775	0.3933	0.2708	0.1179
resnet50_iter_10_aug_None	0.3209	0.3483	0.3397	0.3058	0.2596	0.2065	0.1470	0.0857	0.0297
resnet50_iter_10_aug_Random	0.3283	0.3219	0.2896	0.2423	0.1872	0.1323	0.0814	0.0392	0.0112
resnet50_iter_10_aug_Whole	0.3276	0.3546	0.3537	0.3141	0.2535	0.1855	0.1205	0.0635	0.0204
resnet50_iter_50_aug_None	0.3310	0.3553	0.3707	0.3794	0.3714	0.3281	0.2511	0.1512	0.0534
resnet50_iter_50_aug_Random	0.3213	0.3310	0.3082	0.2682	0.2185	0.1653	0.1117	0.0607	0.0203
resnet50_iter_50_aug_Whole	0.3086	0.3062	0.2921	0.2666	0.2277	0.1769	0.1180	0.0635	0.0203
resnet50_iter_100_aug_None	0.3273	0.3522	0.3700	0.3812	0.3741	0.3319	0.2539	0.1530	0.0542
resnet50_iter_100_aug_Random	0.3222	0.3328	0.3110	0.2712	0.2218	0.1682	0.1132	0.0618	0.0205
resnet50_iter_100_aug_Whole	0.3106	0.3140	0.2948	0.2586	0.2120	0.1591	0.1052	0.0558	0.0187
resnet101_iter_10_aug_None	0.3153	0.3134	0.3035	0.2865	0.2607	0.2149	0.1472	0.0749	0.0227
resnet101_iter_10_aug_Random	0.3145	0.3200	0.3126	0.2868	0.2426	0.1863	0.1222	0.0627	0.0184
resnet101_iter_10_aug_Whole	0.3058	0.3082	0.2999	0.2779	0.2373	0.1822	0.1203	0.0597	0.0165
resnet101_iter_50_aug_None	0.3114	0.3256	0.3278	0.3162	0.2890	0.2448	0.1803	0.1033	0.0334
resnet101_iter_50_aug_Random	0.3219	0.3270	0.3155	0.2894	0.2467	0.1897	0.1261	0.0643	0.0186
resnet101_iter_50_aug_Whole	0.3088	0.3142	0.3095	0.2892	0.2505	0.1952	0.1304	0.0664	0.0197
resnet101_iter_100_aug_None	0.3141	0.3306	0.3419	0.3410	0.3222	0.2854	0.2225	0.1351	0.0452
resnet101_iter_100_aug_Random	0.3093	0.3082	0.2890	0.2485	0.1941	0.1342	0.0811	0.0396	0.0112
resnet101_iter_100_aug_Whole	0.3204	0.3291	0.3207	0.2953	0.2524	0.1941	0.1317	0.0696	0.0211

Appendix 8. Layer-CAM RESULT

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_10_aug_None	0.3152	0.3425	0.3628	0.3680	0.3575	0.3263	0.2715	0.1877	0.0784
resnet18_iter_10_aug_Random	0.3421	0.3977	0.4354	0.4545	0.4513	0.4204	0.3556	0.2507	0.1110
resnet18_iter_10_aug_Whole	0.3402	0.3987	0.4542	0.4936	0.5020	0.4721	0.3966	0.2752	0.1224
resnet18_iter_50_aug_None	0.3350	0.3699	0.3940	0.4027	0.3871	0.3453	0.2803	0.1939	0.0868
resnet18_iter_50_aug_Random	0.3278	0.3845	0.4356	0.4670	0.4732	0.4495	0.3932	0.2948	0.1425
resnet18_iter_50_aug_Whole	0.3428	0.4086	0.4652	0.5023	0.5066	0.4733	0.4000	0.2808	0.1225
resnet18_iter_100_aug_None	0.3326	0.3824	0.4336	0.4728	0.4834	0.4589	0.3941	0.2809	0.1257
resnet18_iter_100_aug_Random	0.3432	0.4000	0.4405	0.4638	0.4636	0.4326	0.3657	0.2589	0.1164
resnet18_iter_100_aug_Whole	0.3707	0.4431	0.4994	**0.5287**	0.5231	0.4787	0.3931	0.2699	0.1165
resnet50_iter_10_aug_None	0.3273	0.3572	0.3466	0.3131	0.2654	0.2111	0.1505	0.0869	0.0304
resnet50_iter_10_aug_Random	0.3305	0.3164	0.2804	0.2380	0.1862	0.1320	0.0816	0.0392	0.0116
resnet50_iter_10_aug_Whole	0.3349	0.3732	0.3677	0.3197	0.2554	0.1875	0.1216	0.0643	0.0207
resnet50_iter_50_aug_None	0.3300	0.3582	0.3794	0.3900	0.3774	0.3278	0.2471	0.1489	0.0525
resnet50_iter_50_aug_Random	0.3248	0.3370	0.3158	0.2761	0.2254	0.1716	0.1165	0.0635	0.0208
resnet50_iter_50_aug_Whole	0.3110	0.3103	0.2979	0.2745	0.2384	0.1907	0.1312	0.0730	0.0241
resnet50_iter_100_aug_None	0.3264	0.3549	0.3786	0.3921	0.3801	0.3321	0.2502	0.1508	0.0539
resnet50_iter_100_aug_Random	0.3255	0.3393	0.3192	0.2798	0.2293	0.1747	0.1185	0.0646	0.0208
resnet50_iter_100_aug_Whole	0.3121	0.3150	0.2953	0.2580	0.2111	0.1595	0.1070	0.0573	0.0189
resnet101_iter_10_aug_None	0.3137	0.3093	0.3018	0.2896	0.2644	0.2215	0.1538	0.0780	0.0239
resnet101_iter_10_aug_Random	0.3144	0.3215	0.3139	0.2887	0.2435	0.1873	0.1228	0.0637	0.0192
resnet101_iter_10_aug_Whole	0.3096	0.3147	0.3112	0.2944	0.2590	0.2034	0.1372	0.0690	0.0197
resnet101_iter_50_aug_None	0.3122	0.3268	0.3270	0.3152	0.2897	0.2482	0.1873	0.1098	0.0360
resnet101_iter_50_aug_Random	0.3216	0.3226	0.3099	0.2812	0.2367	0.1809	0.1205	0.0621	0.0189
resnet101_iter_50_aug_Whole	0.3117	0.3199	0.3196	0.3064	0.2710	0.2165	0.1467	0.0760	0.0222
resnet101_iter_100_aug_None	0.3145	0.3325	0.3446	0.3450	0.3278	0.2909	0.2308	0.1416	0.0495
resnet101_iter_100_aug_Random	0.3103	0.3099	0.2874	0.2440	0.1905	0.1329	0.0808	0.0394	0.0111
resnet101_iter_100_aug_Whole	0.3209	0.3320	0.3272	0.3064	0.2672	0.2101	0.1439	0.0770	0.0243

Appendix 9. Eigen-CAM RESULT

Model	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
resnet18_iter_10_aug_None	0.2383	0.2275	0.2151	0.2006	0.1829	0.1596	0.1293	0.0903	0.0446
resnet18_iter_10_aug_Random	0.1517	0.1334	0.1181	0.1041	0.0901	0.0750	0.0578	0.0389	0.0184
resnet18_iter_10_aug_Whole	0.1483	0.1295	0.1134	0.0982	0.0830	0.0673	0.0505	0.0330	0.0148
resnet18_iter_50_aug_None	0.1806	0.1603	0.1429	0.1265	0.1099	0.0919	0.0706	0.0457	0.0200
resnet18_iter_50_aug_Random	0.1764	0.1649	0.1545	0.1429	0.1288	0.1107	0.0874	0.0569	0.0229
resnet18_iter_50_aug_Whole	0.1566	0.1421	0.1301	0.1171	0.1024	0.0852	0.0651	0.0429	0.0186
resnet18_iter_100_aug_None	0.1447	0.1258	0.1109	0.0972	0.0837	0.0694	0.0535	0.0361	0.0171
resnet18_iter_100_aug_Random	0.1503	0.1319	0.1169	0.1028	0.0890	0.0743	0.0576	0.0389	0.0186
resnet18_iter_100_aug_Whole	0.1285	0.1095	0.0946	0.0812	0.0685	0.0554	0.0421	0.0278	0.0137
resnet50_iter_10_aug_None	0.3733	0.3518	0.3211	0.2840	0.2388	0.1887	0.1341	0.0769	0.0255
resnet50_iter_10_aug_Random	0.3014	0.2670	0.2299	0.1917	0.1522	0.1112	0.0714	0.0366	0.0108
resnet50_iter_10_aug_Whole	**0.4021**	0.3578	0.3029	0.2444	0.1878	0.1352	0.0870	0.0464	0.0146
resnet50_iter_50_aug_None	0.3894	0.3732	0.3426	0.3025	0.2557	0.2024	0.1430	0.0820	0.0282
resnet50_iter_50_aug_Random	0.3411	0.3122	0.2771	0.2375	0.1942	0.1490	0.1024	0.0575	0.0196
resnet50_iter_50_aug_Whole	0.3113	0.2815	0.2481	0.2123	0.1748	0.1352	0.0927	0.0516	0.0181
resnet50_iter_100_aug_None	0.3945	0.3790	0.3496	0.3097	0.2623	0.2076	0.1476	0.0860	0.0302
resnet50_iter_100_aug_Random	0.3446	0.3159	0.2807	0.2402	0.1962	0.1507	0.1044	0.0589	0.0204
resnet50_iter_100_aug_Whole	0.3334	0.3116	0.2852	0.2547	0.2206	0.1828	0.1397	0.0907	0.0382
resnet101_iter_10_aug_None	0.3036	0.2912	0.2691	0.2382	0.1986	0.1525	0.1045	0.0574	0.0196
resnet101_iter_10_aug_Random	0.3264	0.2971	0.2594	0.2170	0.1723	0.1257	0.0800	0.0405	0.0121
resnet101_iter_10_aug_Whole	0.3345	0.3117	0.2824	0.2467	0.2038	0.1533	0.1014	0.0526	0.0162
resnet101_iter_50_aug_None	0.3028	0.2860	0.2657	0.2394	0.2064	0.1655	0.1189	0.0695	0.0236
resnet101_iter_50_aug_Random	0.2851	0.2601	0.2309	0.1964	0.1582	0.1180	0.0782	0.0421	0.0140
resnet101_iter_50_aug_Whole	0.3444	0.3249	0.2976	0.2616	0.2169	0.1641	0.1074	0.0558	0.0168
resnet101_iter_100_aug_None	0.3539	0.3440	0.3259	0.2992	0.2644	0.2180	0.1589	0.0925	0.0328
resnet101_iter_100_aug_Random	0.3092	0.2853	0.2579	0.2280	0.1964	0.1636	0.1262	0.0877	0.0442
resnet101_iter_100_aug_Whole	0.3204	0.2895	0.2529	0.2136	0.1727	0.1307	0.0903	0.0513	0.0181