심층신경망의 분류 이유에 대한 시각적 설명 기술

고려대학교 패턴인식 및 머신러닝 연구실

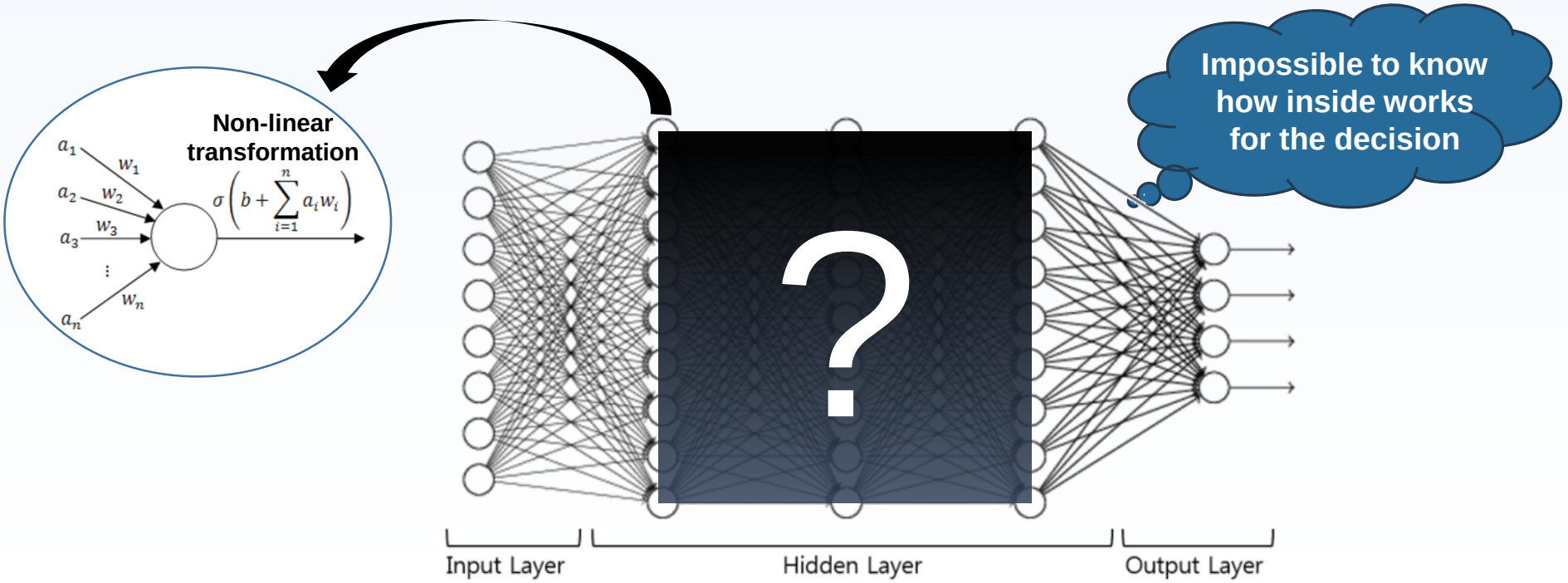
발표자: 박사과정 남우정, 정홍규

지도교수: 이성환

2020년 11월 20일

심층신경망 기반 인공지능의 문제점

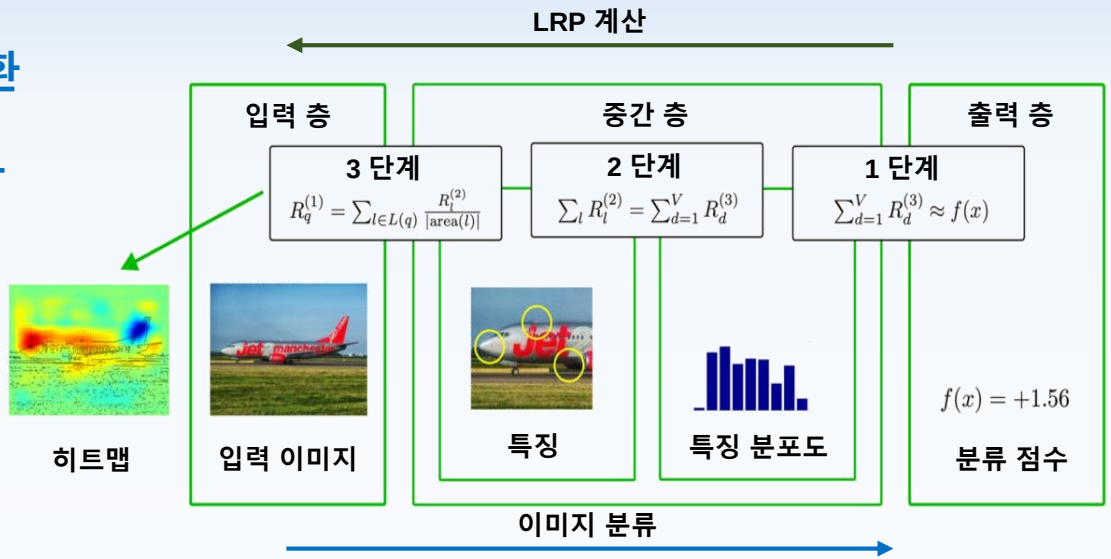
- 현재의 심층신경망은 네트워크 내부에 **무수히 많은 비선형적 연산**의 연결로 인해 의사결정에 대한 원인을 해석하기 어려움
- 이유설명이 불가능한 심층신경망 기술은 의료, 자율주행, 금융 등 높은 **신뢰성을 필요로 하는 분야에서 사용에 제약**이 존재



높은 복잡도의 비선형적 연산을 하는 심층신경망 예시

심층신경망의 시각적 이유설명 과정

- **이미지 분류 과정**
  - 입력 이미지는 특징 벡터로 변환
  - 분류기를 통해 분류 점수가 산출되고 특정 카테고리에 할당
- **픽셀 단위 분해 과정**
  - 분류 점수를 통해 각 층의 맵이 관련성 점수로 표현됨
  - 최종적으로 입력 층의 관련성 점수가 히트맵으로 시각화됨

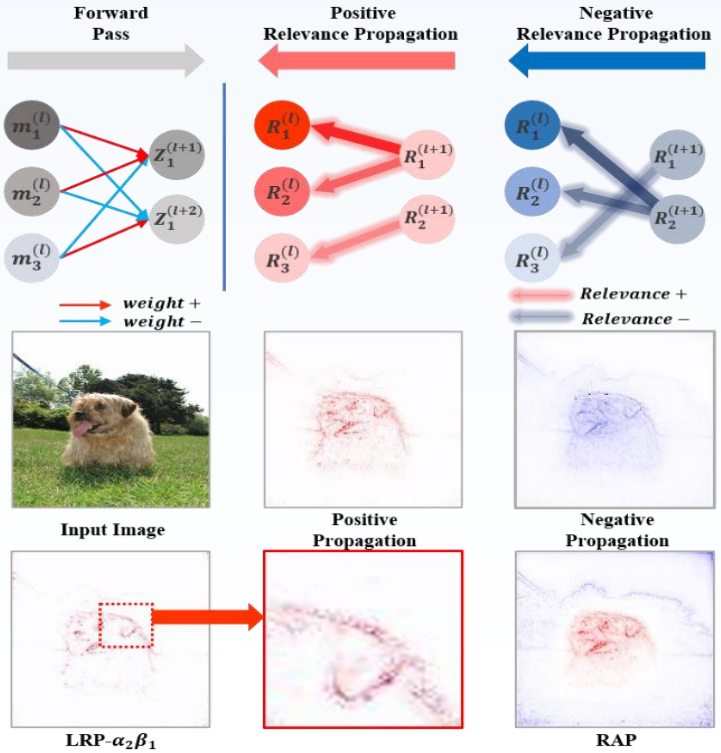


픽셀 단위 분해 과정 도식화

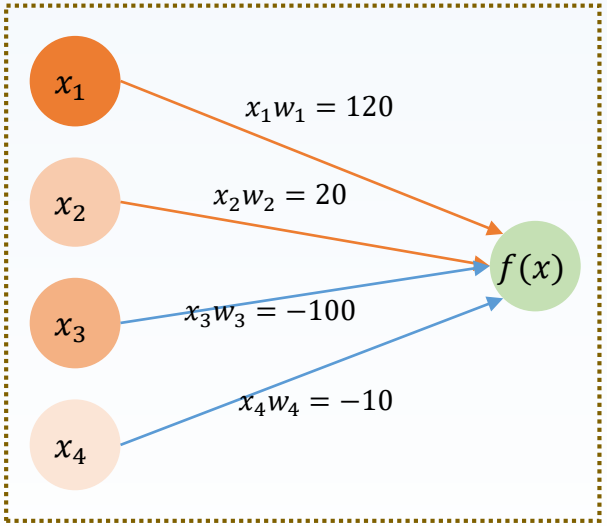
픽셀 단위 분해를 통한 시각화 과정

기존 심층신경망 시각적 이유설명 기술의 문제점

- 기존 시각적 이유설명 방식은 분류결정에 대한 양과 음의 관련성을 적절히 구분하지 못함
  - ➢ 기존 방법은 관련성의 값이 크기에 따라 전파되기 때문에 유사한 지역에 양과 음의 관련성이 겹쳐서 시각화됨
  - ➢ 이를 해결하기 위해 분류결정을 해석하는 관련성에 대한 관점을 실제 값의 영향력 정도로 전환하여 분석하는 방식이 필요



심층신경망의 의사결정과 뉴런들의 관련성을 분석하는 예시



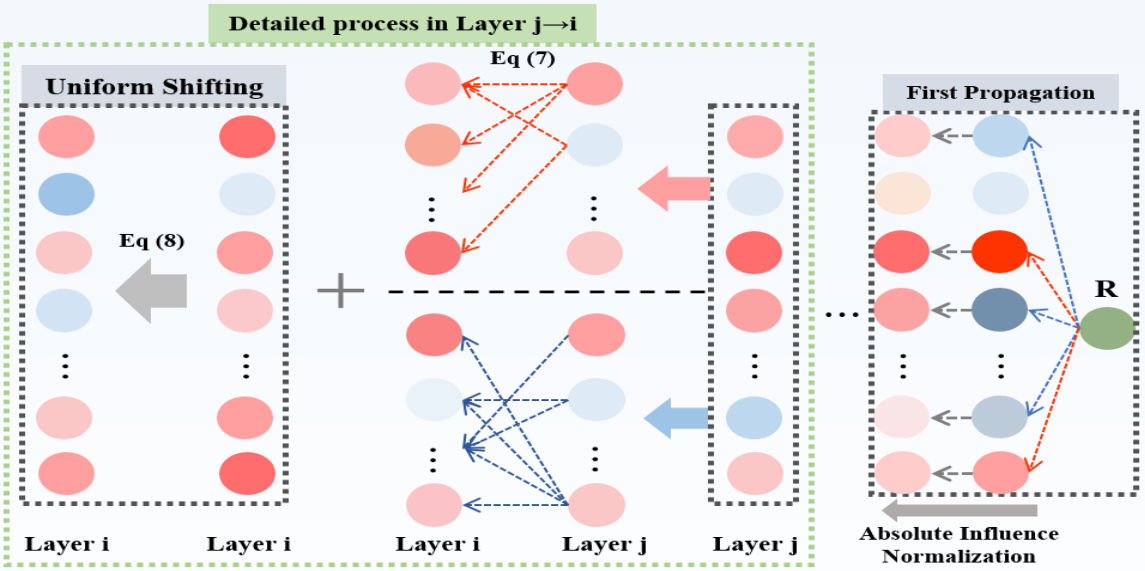
The relevance of neurons in different perspective

In terms of actual value? : $x_1w_1 > x_2w_2 > x_4w_4 > x_3w_3$
In terms of influence? : $x_1w_1 > x_3w_3 > x_2w_2 > x_4w_4$

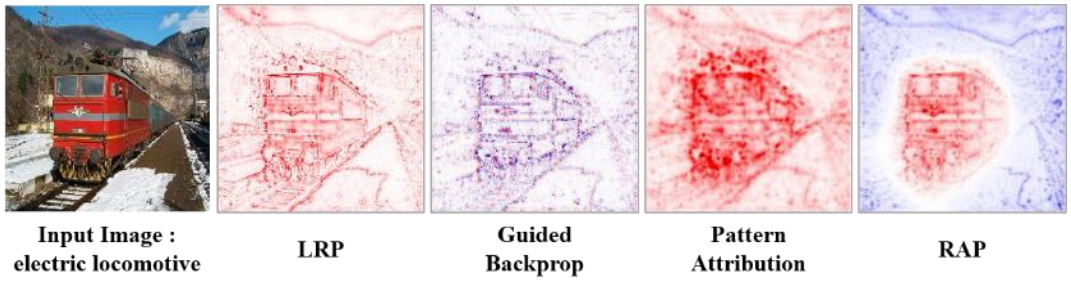
심층신경망의 의사결정과 뉴런의 관련성을 분석하는 관점

RAP 시각적 이유설명 기술 (1/4)

- 심층신경망의 결과를 뉴런의 중요도에 따라 관련성을 전파하는 Relative Attributing Propagation (RAP) 제안
  - 관련성은 가장 중요한 부분이 높은 양의 값(빨간색), 관련이 없는 부분이 가장 낮은 음수의 값(파란색)을 가짐



제안하는 RAP 기술의 관련성 전파 과정



기존 시각적 이유설명 기술과의 시각화 결과 비교

i) Absolute Influence Normalization
ii) Decide Criterion of Relevant Neuron
iii) Uniform Shifting

Absolute Influence Normalization

$$R_i^{(p)} = \left(\sum_i z_{ij}^{+} + \sum_i z_{ij}^{-}\right) * R_j^{(q)}$$

$$R_i'^{(p)} = |R_i^{(p)}| * \frac{\sum_i R_i^p}{\sum_i |R_i^p|}$$

Criterion of Relevant Neuron

$$v_j^{(l+1)} = R_j^{(l+1)} * \frac{\sum_i |z_{ij}^{-}|}{\sum_i (|z_{ij}^{+}| + |z_{ij}^{-}|)}$$

$$\bar{R}_{i\epsilon P,N}'^{(l)} = \sum_j \left(\frac{z_{ij}^{+}}{\sum_i (z_{ij}^{+})} R_j^{(l+1)} + \frac{z_{ij}^{-}}{\sum_i (z_{ij}^{-})} v_j^{(l+1)}\right)$$

Uniform Shifting

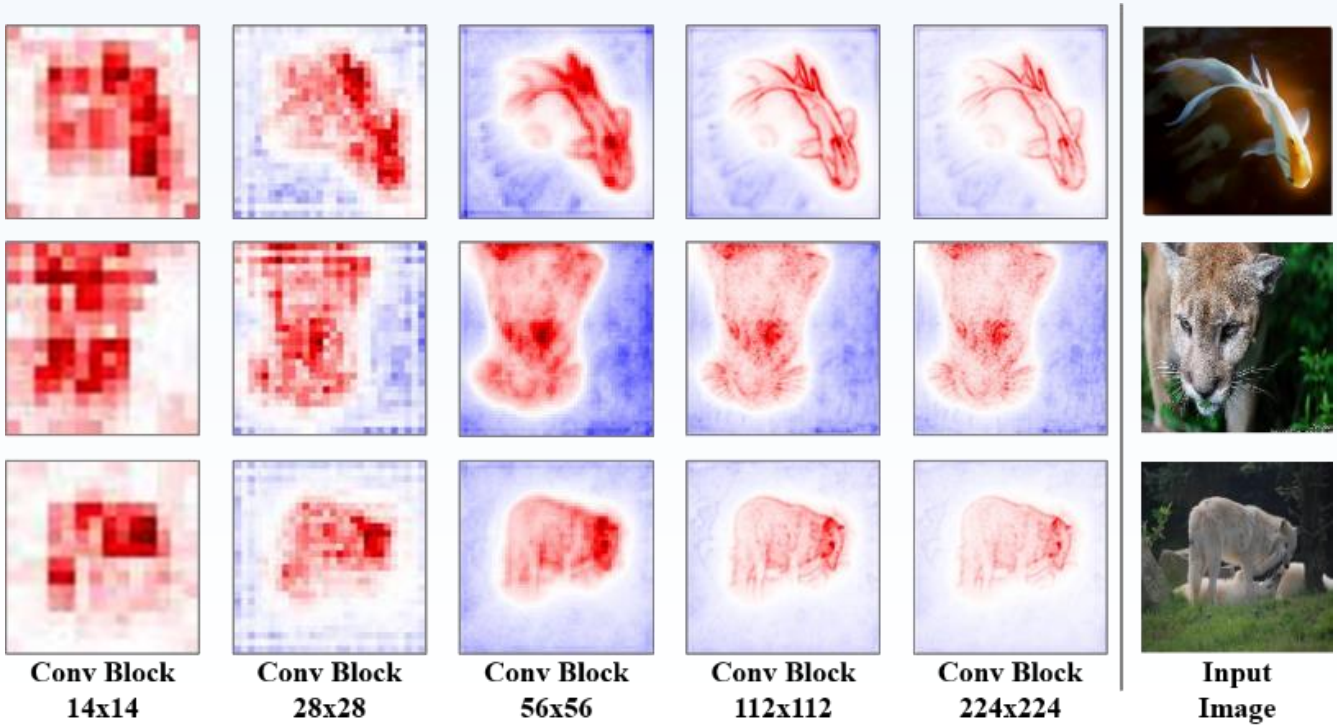
$$\Psi_i^l = \begin{cases} \sum_i \left(\bar{R}_{i\epsilon N}^{(l)}\right) * \frac{1}{\Gamma}, & m_i^l \text{ is activated} \\ 0, & \text{otherwise} \end{cases}$$

$$R_j^{(l)} = \bar{R}_{i\epsilon P \cup N}^{(l)} - \Psi_i^l$$

RAP 기술의 수식화

RAP 시각적 이유설명 기술 (2/4)

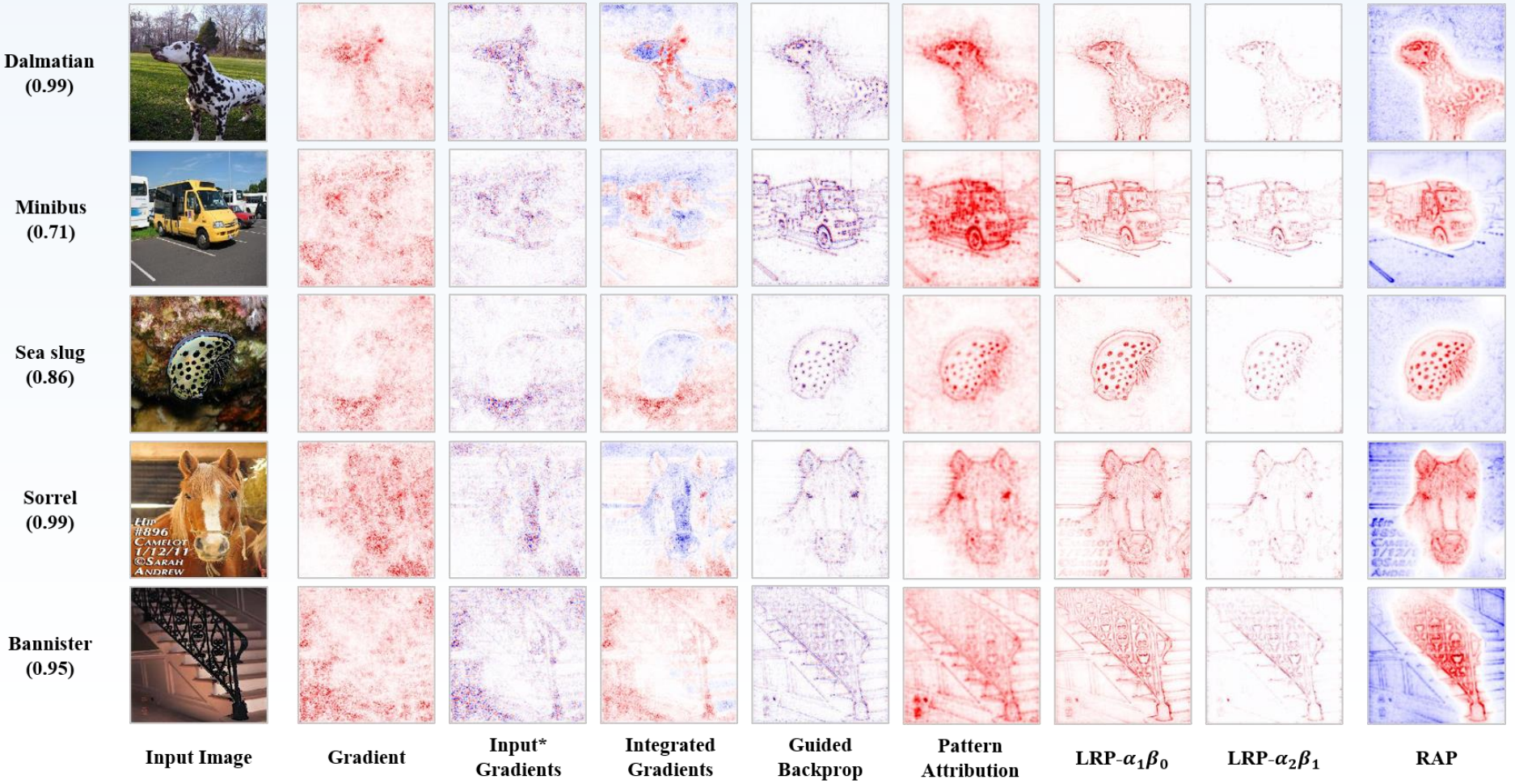
- 관련성의 전파과정 중 네트워크 중간계층에서의 분석
  - 네트워크의 의사결정 (왼쪽)으로 부터 입력 이미지 (오른쪽)까지 관련성이 전파되면서 중간계층의 변화를 관찰
  - 방법론의 설계대로 뉴런의 관련성 값은 서서히 중요한 부분이 높은 값(빨간색), 관련성이 낮은 부분은 음수의 값(파란색)을 갖음



관련성 전파과정에서 네트워크 중간 계층에서 변화를 분석

RAP 시각적 이유설명 기술 (3/4)

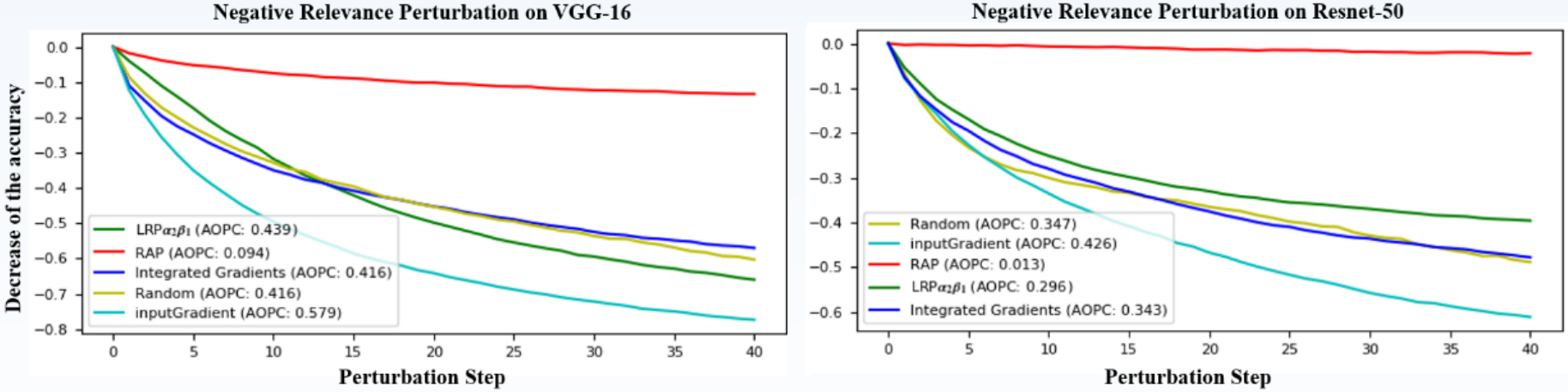
- 심층신경망의 시각적 설명의 정성적 평가
  - ➢ 기존의 설명기법들보다 직관적이고 분별력 있는 시각적 설명 제공



기존의 시각적 설명 기법들과 제안하는 RAP 방법론과의 정성적 비교

RAP 시각적 이유설명 기술 (4/4)

- 심층신경망의 시각적 설명의 정량적 평가
  - ➢ Negative Region Perturbation, Outside-Inside Ratio, mIOU 평가기법 활용
    - ✓ 입력 이미지에서 낮은 관련성 값에 해당하는 부분을 왜곡 시켰을 때 기존의 다른 방법론에 비해 RAP의 시각적 설명이 네트워크의 결과에 영향을 적게 미침
    - ✓ RAP의 시각적 설명이 네트워크의 결정에 해당하는 부분을 더 잘 찾음



Negative Relevance Perturbation 기법을 통한 평가

Outside-Inside Ratio		RAP	$LRP_{\alpha_1\beta_0}$	$LRP_{\alpha_2\beta_1}$	Gradient	Input* Gradient	Integrated Gradients	Pattern Attribution	Guided Backprop
VGG-16	ALL	**0.252**	-	0.616	-	0.989	1.230	-	1.069
	POS	**0.341**	0.474	0.524	0.619	0.691	0.827	0.415	0.427
Res-50	ALL	**0.164**	-	0.302	-	0.996	1.195	-	1.035
	POS	**0.166**	0.429	0.299	0.597	0.689	0.698	-	0.296
Segmentation Mask		RAP	$LRP_{\alpha_1\beta_0}$	$LRP_{\alpha_2\beta_1}$	Gradient	Input* Gradient	Integrated Gradients	Pattern Attribution	Guided Backprop
Imagenet	PIX ACC	**79.23**	75.40	72.95	70.01	66.38	66,52	76.84	71.98
	mIOU	**62.23**	55.78	50.86	49.30	44.01	45.90	58.05	49.87
Pascal VOC	PIX ACC	**73.91**	70.86	69.43	68.14	50.01	52.38	-	66.92
	mIOU	**55.60**	49.82	46.85	46.07	31.69	34.39	-	43.63

Outside-Inside metric 와 mIOU 를 사용하여 기존의 방법과 수치적 비교

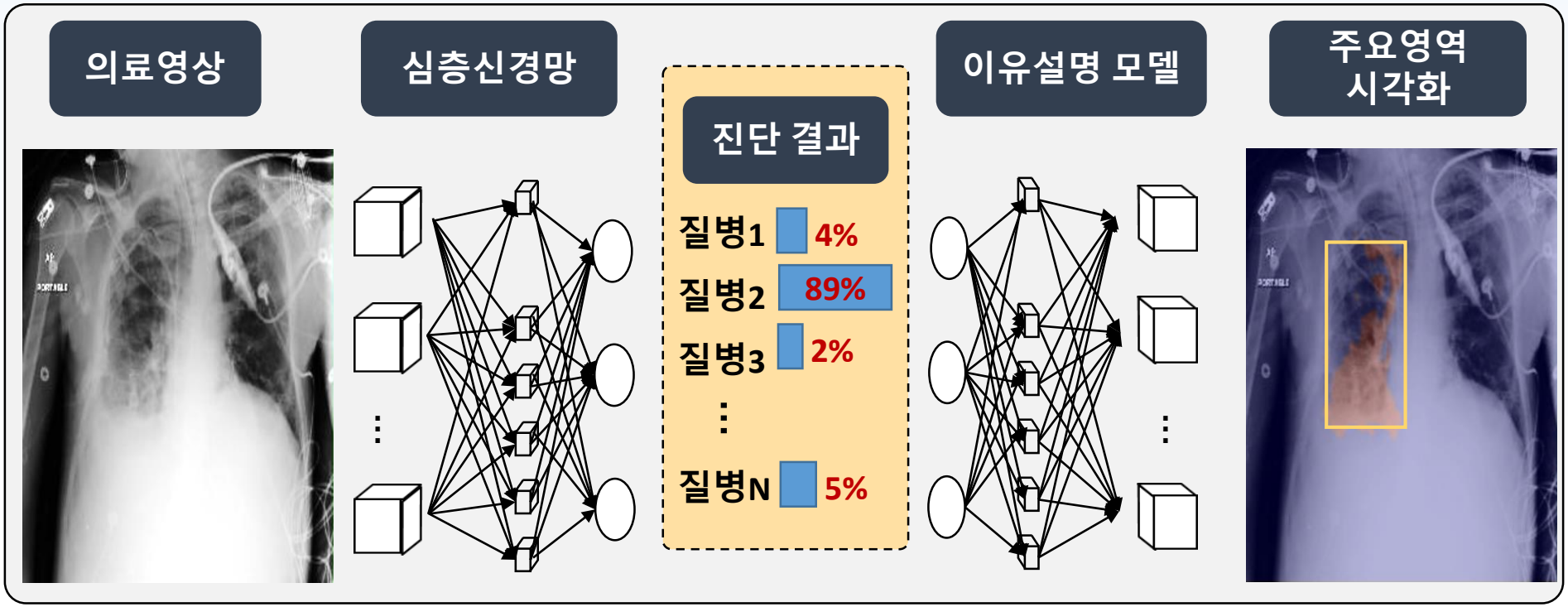
Method	mIOU
Guillaumin et al. (Guillaumin, Küttel, and Ferrari 2014b)	57.30
DeepMask (Pinheiro, Collobert, and Dollár 2015)	58.69
DeepSaliency (Li et al. 2016)	62.12
Xiong et al. (Xiong, Jain, and Grauman 2018)	64.22
Ours	62.23

mIOU를 통해 기존의 Weakly supervised segmentation 기법들과 수치적 비교

의료영상 기반 인공지능 의사결정에 대한 시각적 이유설명 기술

*CAD: Computer-Aided Diagnosis

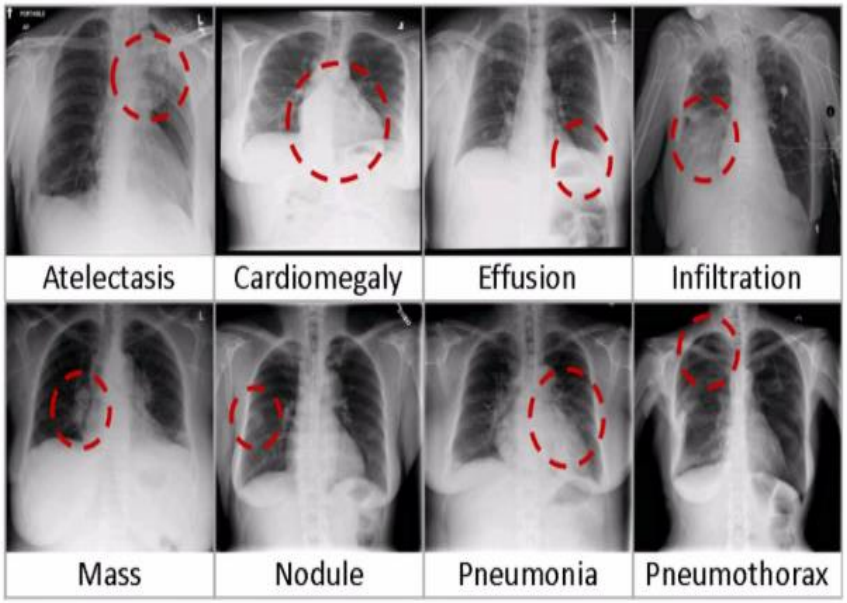
- 연구 필요성
  - 의료영상의 경우 인간의 생명과 직결되기 때문에 단순히 AI 기반 진단 결과를 보여주는 것은 전문가/일반인에게 큰 **신뢰도를 제공**하지 못함
  - 현 AI 기술 및 의료영상 데이터 구성의 한계를 보았을 때 완벽한 AI 기반 의료진단 기술을 개발하기 보다는 **CAD* 관점에서의 접근**이 필요



의료영상 기반 인공지능 의사결정에 대한 시각적 이유설명 과정

기계학습을 위한 의료영상 데이터 구성

- **NIH Chest X-ray Dataset [X. Wang, CVPR 2018] 예시**
  - ➢ **대상 질병 클래스**
    - ✓ **1 : Atelectasis, 2 : Cardiomegaly, 3 : Effusion,**
      4 : Infiltration, 5 : Mass, 6 : Nodule,
      7 : Pneumonia, 8 : Pneumothorax, 9 : No findings
  - ➢ **데이터 개수**
    - ✓ **30,805 환자를 대상으로 한 112,120 정면 X-ray 영상에 대해 질병 Label 존재**
    - ✓ **질병 위치에 대한 Bounding Box Label은 약 1000장의 영상에서 존재**



NIH Chest X-ray Dataset의 질병 별 영상 예시

- **Atelectasis**
- **Effusion**
- **Infiltration**

- **Cardiomegaly**
- **Effusion**
- **Infiltration**

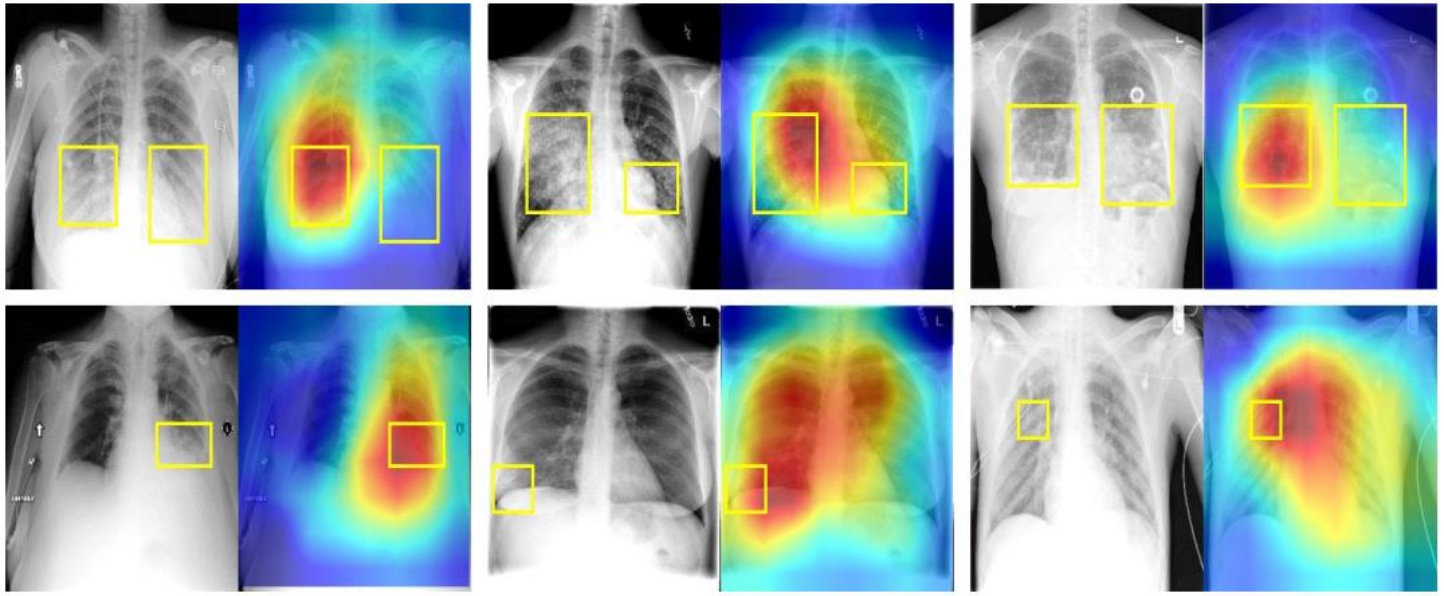
의료영상 학습 Label

의료영상 별 클래스 Label 예시

X. Wang et al., "TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays," in CVPR 2018

현 의료영상 데이터의 문제점 및 기존 시각화 기술 적용 결과

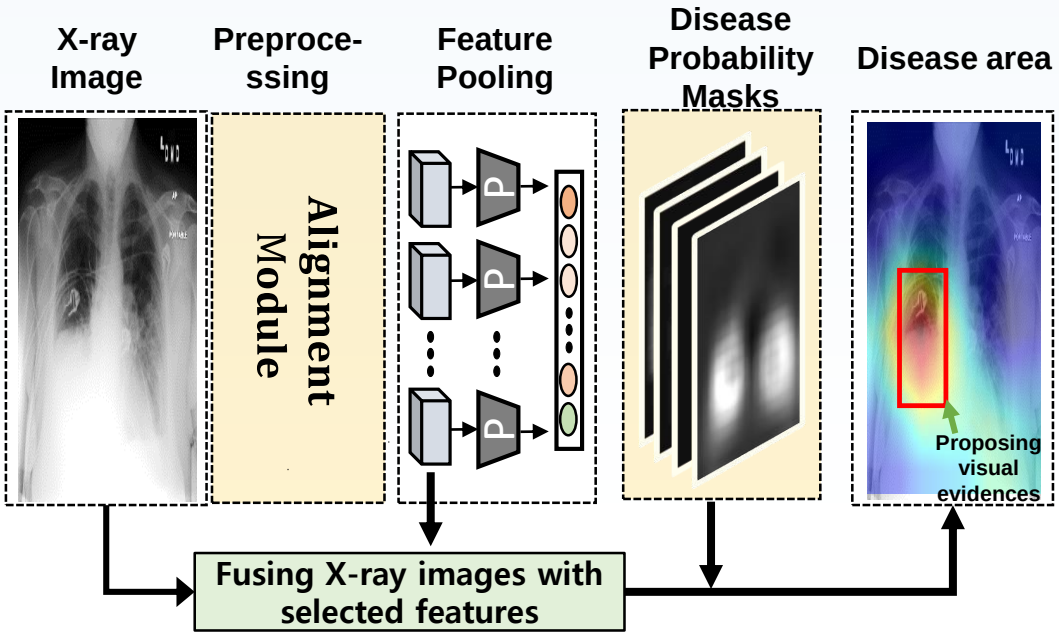
- 기계학습 적용에 있어 현재 의료영상 데이터 문제점
  - ➢ 하나의 의료영상에 대해 정확한 정답 Label이 존재하지 않음
    - ✓ 의사의 추측으로 가능성 있는 모든 질병에 대해 정답 Label을 지정
  - ➢ 다중 클래스 분류 문제로 기존 일반 영상(ImageNet) 분류 대비 어려움
    - ✓ 다중 클래스 분류 문제임에도 불구하고 매우 극소량의 영상에 대해서만 Bounding Box를 제공
- 기존 시각화 기술 적용 결과
  - ➢ 위와 같은 의료영상 데이터의 문제에 의해 기존 일반 영상(ImageNet)에 상대적으로 잘 적용되던 시각적 이유설명 기술의 성능이 급격히 저하



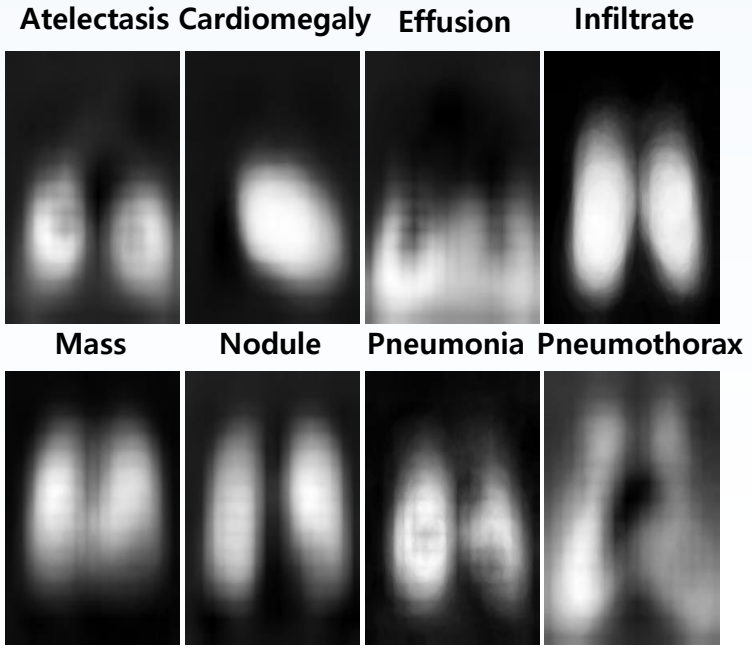
기존 시각화 기술(CAM)을 의료영상에 적용한 결과 예시

질병 확률 마스크 기반 의료영상 분류 이유 시각화 기술 (1/2)

- 질병 확률 마스크를 생성하고 심층신경망 특징맵에 Attention 적용
  - 다중 클래스 문제 및 질병이 다는데도 의료영상이 유사한 특성을 해결하기 위해 특정 질병이 나올 확률이 높은 마스크를 생성하여 심층신경망을 학습
- 영상 정렬 모듈을 활용한 의료영상 전처리
  - 촬영 환경에 따라 의료영상의 회전, 확대 정도가 다르며 이를 해결하기 위해 영상 정렬 모듈 개발 및 전처리 시 적용
  - 본 연구는 특정 질병이 나올 확률이 높은 마스크를 학습에 활용하므로 일정한 회전, 확대율을 갖는 영상이 학습에 사용되어야 함

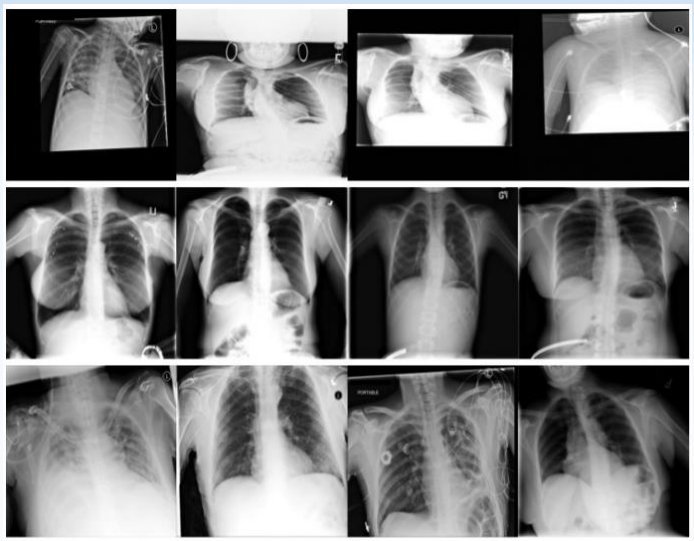


질병 확률 마스크 기반 의료영상 분류 이유 시각화 과정



생성된 질병 확률 마스크 예시

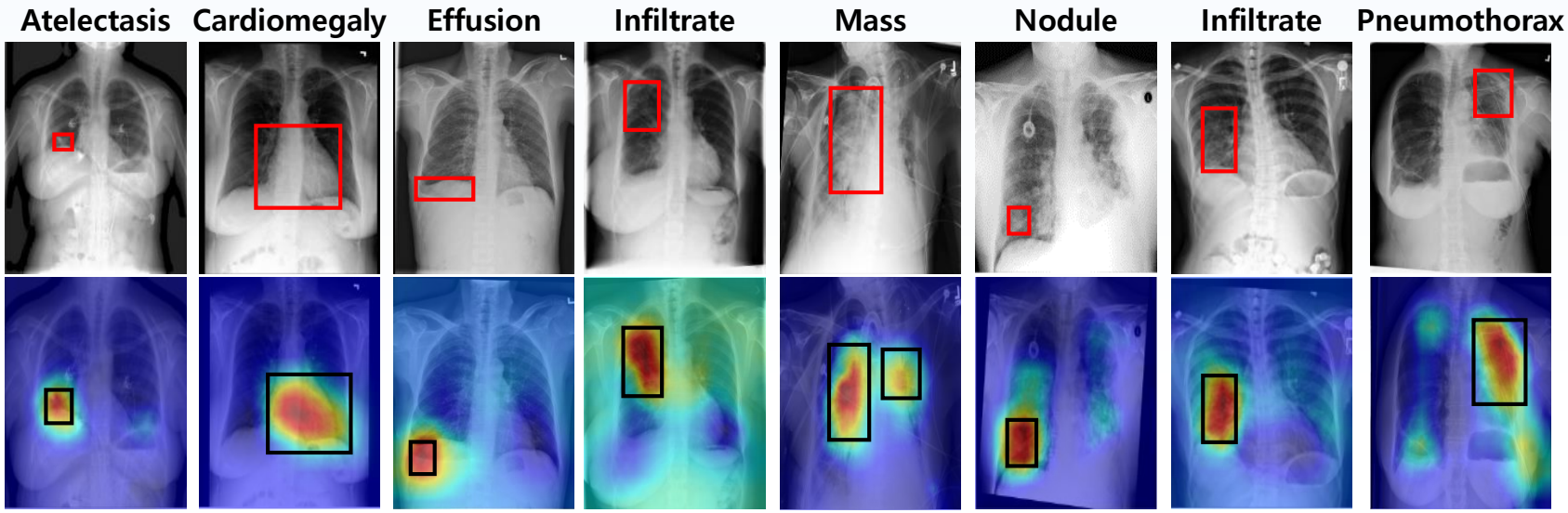
질병 확률 마스크 기반 의료영상 분류 이유 시각화 기술 (2/2)



데이터셋에서 제공하는 X-ray 영상

딥러닝 학습을 위한 정렬된 X-ray 영상

의료 영상 Raw 데이터 및 전처리 결과 예시

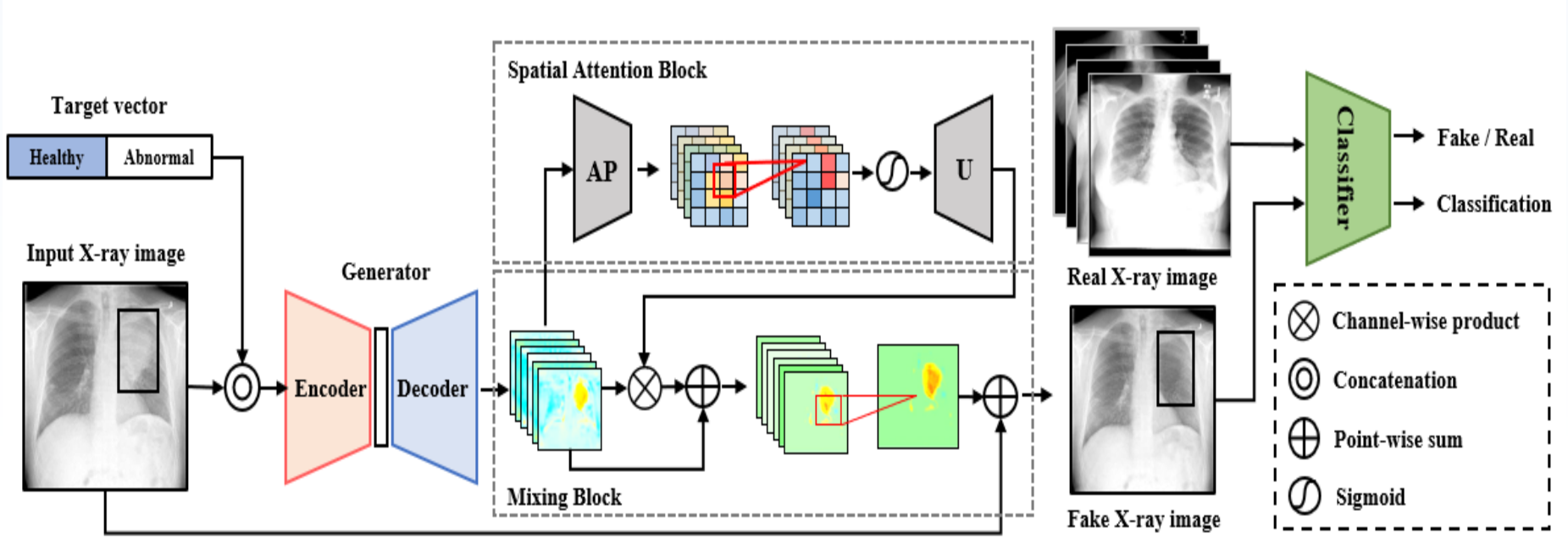


제안 기술을 통한 분류 이유 시각화 및 질병 부위 Localization 수행 결과

대조군 의료영상을 생성하는 Counterfactual 설명 기술 (1/2)

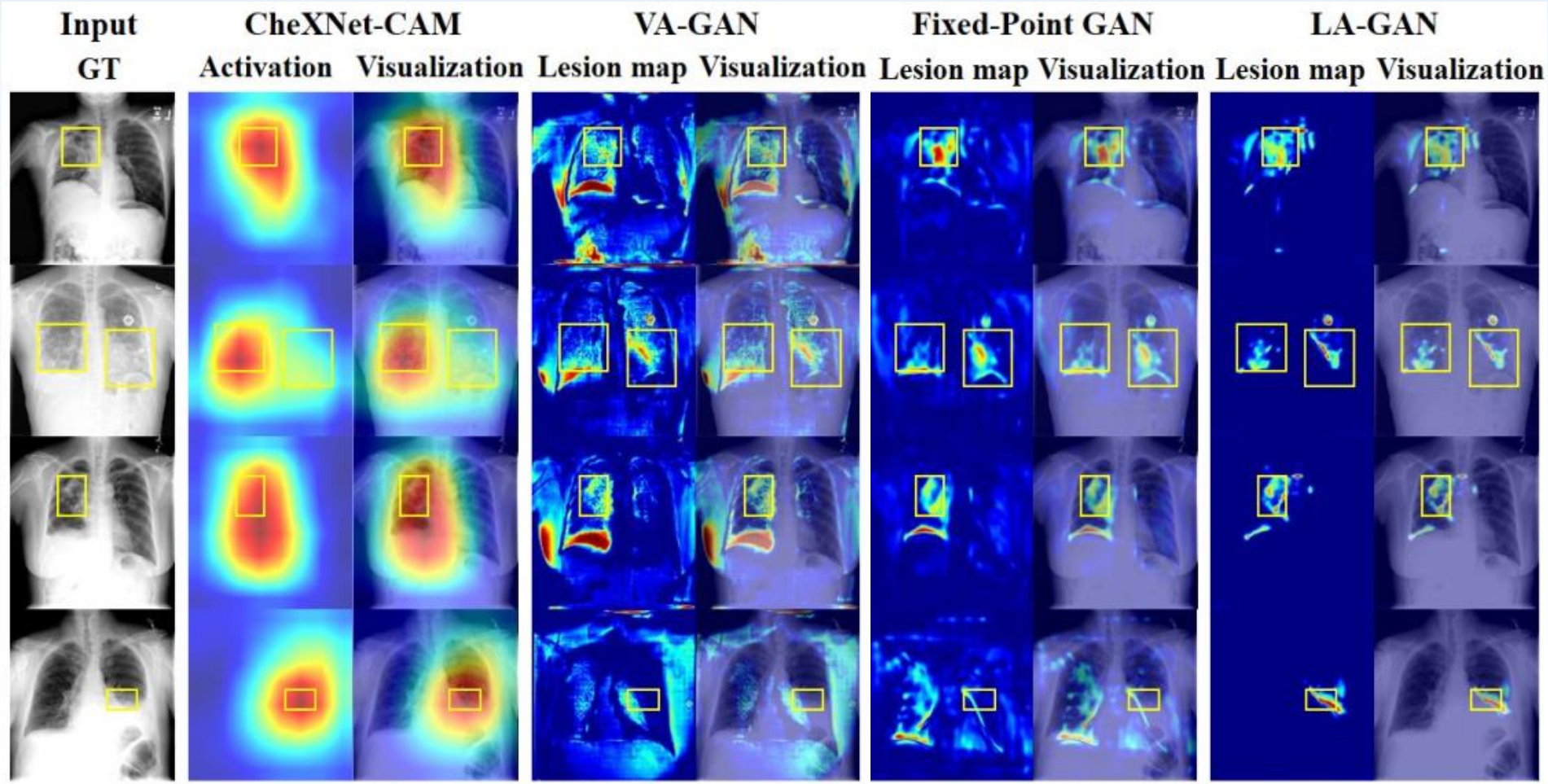
*GAN: Generative Adversarial Network

- Counterfactual 설명 기술 의미
  - 심층신경망의 분류 이유를 설명하는 기존 기술과 달리 다른 클래스로 분류되기 위해 변경되어야 하는 특징을 설명하는 기술
- GAN* 기반 대조군 의료영상 생성 기술 제안
  - 현재 진단된 결과 (예: Healthy or Abnormal)에 대비한 대조군 의료영상을 생성하기 위해 Attention 기반 GAN 구조를 적용
  - 입력영상 대비 생성된 영상의 구조가 크게 변하는 것을 억제하기 위해 Residual Connection 기법 적용



질병 확률 마스크 기반 의료영상 분류 이유 시각화 과정

대조군 의료영상을 생성하는 Counterfactual 설명 기술 (2/2)



대조군 의료영상 생성을 통한 질병 영역 시각화 기술 예시 (LA-GAN: 제안기술)

고려대학교 PRML 연구실 2020년 주요 연구실적 (XAI 사사포함 기준)

- W.-J. Nam, J. Choi, S.-W. Lee, "Interpreting Deep Neural Networks with Relative Sectional Propagation by Analyzing Comparative Gradients and Hostile Activations," Submitted to AAAI 2021
- H.-G. Jung and S.-W. Lee, "Few-Shot Object Detection using Bayesian Feature Hallucination and Class-Adaptive Regression," Submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence
- S.-H. Kang, H.-G. Jung, D.-O. Won, S.-W. Lee, "Counterfactual Explanation Based on Gradual Construction for Deep Networks," Under Revision at Pattern Recognition
- J.-W. Seo, H.-G. Jung, S.-W. Lee, "Self-Augmentation: Generalizing Deep Networks to Unseen Classes for Few-Shot Learning," Under Revision at Neural Networks
- P. Bertens and S.-W. Lee, "Network of Evolvable Neural Units can Learn Synaptic Learning Rules and Spiking Dynamics," Nature Machine Intelligence, 2020 (Accepted)
- D.-O. Won, K.-R. Müller, and S.-W. Lee, "An Adaptive Deep Reinforcement Learning Framework Enables Curling Robots with Human-Like Performance in Real World Conditions," Science Robotics, 2020
- H.-G. Jung and S.-W. Lee, "Few-Shot Learning with Geometric Constraints," IEEE Trans. on Neural Networks and Learning Systems, 2020
- W.-J. Nam, S. Gur, J. Choi, L. Wolf, S.-W. Lee, "Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks," in AAAI 2020