Review Article

J Korean Soc Radiol 2020;81(6):1305-1333
https://doi.org/10.3348/jksr.2020.0150
pISSN 1738-2637 / eISSN 2288-2928

대한영상의학회지

The Latest Trends in Attention Mechanisms and Their Application in Medical Imaging

어텐션 기법 및 의료 영상에의 적용에 관한 최신 동향

Hyungseob Shin, BS[†], Jeongryong Lee, BS[†], Taejoon Eo, BS, Yohan Jun, BS, Sewon Kim, BS, Dosik Hwang, PhD*

Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

Deep learning has recently achieved remarkable results in the field of medical imaging. However, as a deep learning network becomes deeper to improve its performance, it becomes more difficult to interpret the processes within. This can especially be a critical problem in medical fields where diagnostic decisions are directly related to a patient's survival. In order to solve this, explainable artificial intelligence techniques are being widely studied, and an attention mechanism was developed as part of this approach. In this paper, attention techniques are divided into two types: post hoc attention, which aims to analyze a network that has already been trained, and trainable attention, which further improves network performance. Detailed comparisons of each method, examples of applications in medical imaging, and future perspectives will be covered.

Index terms Deep Learning; Artificial Intelligence; Medical Imaging; Attention

Received August 13, 2020
Revised November 2, 2020
Accepted November 7, 2020

***Corresponding author**
Dosik Hwang, PhD
Department of Electrical and Electronic Engineering,
Yonsei University,
50 Yonsei-ro, Seodaemun-gu,
Seoul 03722, Korea.

Tel 82-2-2123-5771
E-mail dosik.hwang@yonsei.ac.kr

ORCID iDs
Hyungseob Shin
https://orcid.org/0000-0001-7936-5165
Jeongryong Lee
https://orcid.org/0000-0001-7251-2126
Taejoon Eo
https://orcid.org/0000-0002-3546-0184
Yohan Jun
https://orcid.org/0000-0003-4787-4760
Sewon Kim
https://orcid.org/0000-0002-3893-252X
Dosik Hwang
https://orcid.org/0000-0002-2217-2837

[†]These authors contributed equally to this work.

서론

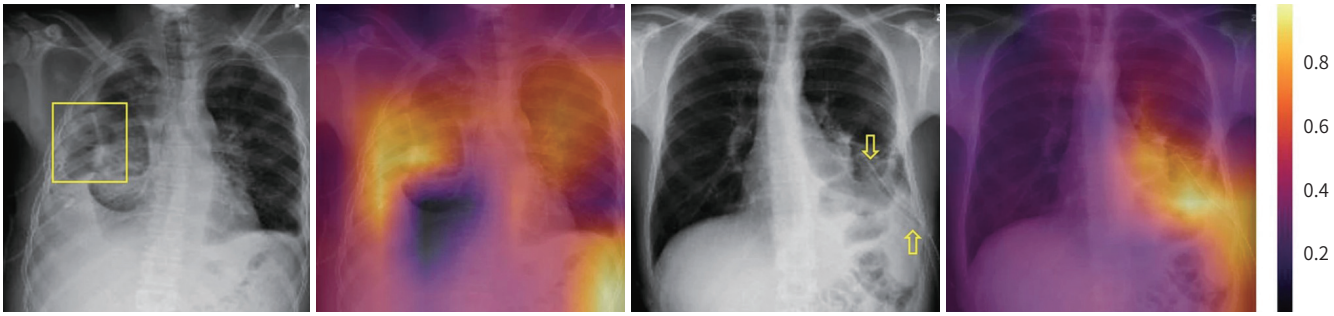
4차 산업 혁명의 핵심 기술 중 하나인 딥러닝은 기존의 규칙 기반(rule-based) 알고리즘 또는 딥러닝을 제외한 머신러닝 알고리즘 대비 뛰어난 특징 추출 능력(1)과 성능, 그리고 재현성(reproducibility)을 보인다. 이로 인해 자동 진단, 치료반응 평가(response assessment), 생존 예측(survival prediction) 등을 망라하는 의료 분야에 널리 적용되고 있으며(2-4) 최근에는 의료기기에 탑재되어 국내외적으로 인허가를 통과하는 단계까지 와 있다. 특히 의료 영상 재구성(reconstruction), 합성(synthesis), 자동 분석(analysis) 및 판독(diagnosis) 등의 영상의학(medical imaging) 분야에서 괄목할 만한 연구들이 나오고 있다(5-12). 의료

영상 재구성이란 의료 영상 촬영 기기에서 얻어진 획득 데이터(acquisition data 또는 raw data)를 영상으로 복원하는 기술인데, 이와 관련하여 하드웨어에서 얻어진 획득 데이터를 원래 얻어야 하는 양보다 적은 수로 희소 샘플링(undersampling 또는 sparse acquisition) 한 뒤, 이를 고해상 영상 복원(super-resolution)과 잡음(noise)/인공물(artifact) 제거 기술 등의 소프트웨어(software)적인 기술로 보완하는 딥러닝 기술의 개발이 근래 매우 활발히 이루어지고 있다. 촬영 시간이 길어 환자의 불편함을 초래하는 자기공명영상(이하 MRI)의 촬영 속도 가속화(5-7) 혹은 저선량 컴퓨터 단층촬영(이하 CT) 영상의 고품질 복원(8) 등이 이에 속하며, 많은 글로벌 의료기기 기업들이 앞다퉈 기술 개발에 투자하고 있을 정도로 매우 유망한 분야이기도 하다.

영상의학 관련 연구에서 딥러닝이 활발하게 이용되고 있는 또 다른 분야는 바로 의료 영상 자동 분석 및 판독 분야로서, 이는 영상의학과 전문의(radiologist)가 주로 담당하고 있는 업무[예컨대, 영상 판독, 환자 분류(triaging), 응급 질환 분류(referral suggestion) 등]의 일부(11, 12)를 수행하거나 그것을 보조하는 것을 목적으로 한다. 이것의 기반이 되는 기술은 기관(organ)이나 종양(tumor)과 같은 특이 부위 등의 자동 검출(detection) 및 분할(segmentation) 기술인데, 해당 기술은 패턴 추출(pattern recognition)에 특화된 데이터 중심(data-driven)의 딥러닝 알고리즘이 가장 강점을 드러내는 분야이기도 하다. 환자 맞춤형 정밀 의학(precision medicine)으로의 전환 추세에 따라 의료 영상 분석의 패러다임(paradigm) 역시 기존의 정성적 분석에서 정량적 분석으로 바뀌고 있으며(13), 의료 영상에서 견고한 특징(robust features)들을 추출하여 더 정교한 예측 및 진단을 추구하는 것이 해당 분야 내에서 대세로 자리 잡고 있다. 이 과정에서 머신 러닝 기반의 라디오믹스(radiomics) 기술(14)을 넘어 최근에는 딥러닝을 활용한 특징(deep feature) 추출, 병변 검출 및 분할 기술(3, 15-17) 등이 활발히 연구되고 있는데, 전문의를 보조하여 전문의의 진단 정확도를 향상시킬 수 있을 뿐 아니라(18) 이미 몇 가지 분과에 대해서는 이러한 인공지능 알고리즘이 전문의 수준의 분석 및 진단을 수행할 수 있는 것으로 보고되고 있다(19, 20). 또한, 딥러닝과 같은 인공지능 알고리즘은 인간에 비해 영상 분석 속도가 매우 빠르기 때문에 절대적인 판독량을 늘릴 수 있다는 점에서(예컨대, 환자 한 명을 판독하는 데 걸리는 시간은 고작 수 초 이내이며, 인간과 달리 24시간 내내 일할 수 있다) 영상의학과 전문의로의 과도한 업무 집중과 이에 기인한 의료 병목 현상 및 진단 부정확성을 낮출 수 있어 4차 산업 혁명 시대에 걸맞은 매우 유망한 기술이라 할 수 있다.

그러나 현재까지 개발된 대부분의 딥러닝 모델의 가장 큰 문제는 검출 및 분할 결과에서 그 판단의 근거를 명확히 확인하기 어렵다는 것이다. 즉, 딥러닝 모델이 영상의 어떤 부분을 보고 그러한 판단을 내렸는지 상세하게 알 수 없다는 것인데, 이로 인해 흔히 딥러닝 모델은 그 내부를 확인할 수 없다고 하여 '블랙박스'라고도 부른다. 딥러닝 모델의 판단 근거를 확인하는 것은 환자의 생명과 직결되는 의료분야에서 매우 중요한 문제이다. Fig. 1은 Baltruschat 등(21)의 연구에서 14가지의 병리학적 병변으로 구성된 공개된 흉부 엑스레이(X-ray) 영상 데이터셋(22)에 대하여 딥러닝으로 병변을 분류하도록 학습시킨 후, 테스트 영상(학습에 사용되지 않은 영상)에 대해 분류에 가장 핵심적이었던 영역을 히트맵으로 시각화한 영상이다. 해당 결과에 의하면, 딥러닝 모델이 환자의 엑스레이 영상을 기흉(pneumothorax)으로 분류함에 있어 실제 임상적으로 유의미한 곳(기흉

Fig. 1. Heatmap analysis for two example images.
The first image shows an X-ray image of a patient with pneumothorax, which is marked with a yellow box. The highest activation in the heatmap next to it is highlighting the correct area. On the other hand, the second example and its heatmap shows a negative example where the heatmap is highlighting the drain (marked with arrows) to be responsible for the final prediction of "pneumothorax." Adapted from Baltruschat et al. Sci Rep 2019;9:6381 (21) (https://doi.org/10.1038/s41598-019-42294-8), licensed under CC BY 4.0. Images have been rearranged.



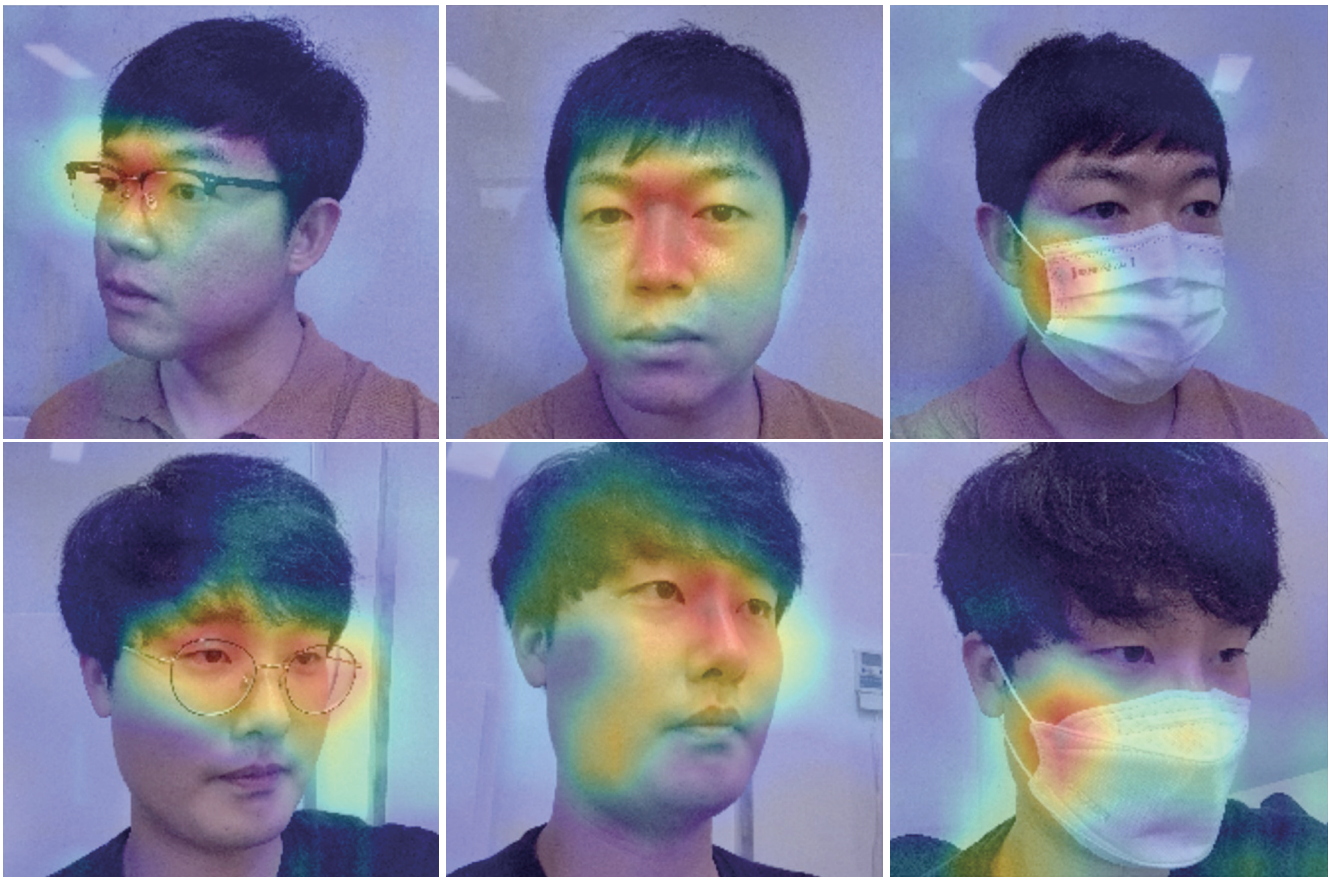
이 발생한 위치)이 아니라 기흉 치료를 위해 사용한 관(drain)을 보고 결정하기도 한다는 것을 알 수 있다. 이는 근본적으로는 학습 데이터 구성 시 이미 치료를 받은 환자의 영상까지 기흉으로 분류해놓은 것에 기인하지만, 높은 성능으로 작동하는 딥러닝 모델이 실상은 잘못된 판단을 하고 있을 수 있다는 사실을 암시한다. 또한, 의료 영상은 내재적인 불확실성, 표준화, 업데이트가 어려운 병원 인프라 등으로 인해 적대적 공격(adversarial attack)에 유난히 취약할 수 있는데(23), 이러한 문제들 때문에 딥러닝 모델이 의료 분야에서 안전하게 사용되기 위해서는 반드시 그 판단 근거를 확인하는 과정이 필요하다.

이를 위해 그 결정의 근거와 판단 과정을 일부나마 확인할 수 있도록 제시된 최근의 기법들을 설명 가능한 인공지능(explainable artificial intelligence, XAI) 기술이라 하며, 그중 하나가 바로 어텐션 기법(attention mechanism)이다. Fig. 2는 딥러닝 네트워크를 통해 사람의 안경 쓴 모습, 맨얼굴, 그리고 마스크 쓴 모습을 분류하도록 학습시킨 후 테스트 데이터에 대하여 그 히트맵을 시각화 한 것이다. 딥러닝 네트워크가 분류를 진행함에 있어 각 클래스와 관계되는 핵심 영역(안경, 미간 및 코와 같은 얼굴 전반, 그리고 볼과 마스크의 경계 부분)에 집중하고 있는 것을 확인할 수 있다. 어텐션 기법을 통해 이와 같이 딥러닝 모델의 판단 근거를 확인할 수 있을 뿐 아니라, 딥러닝 모델로 하여금 영상 전체가 아닌, 주어진 목적을 달성하는데 필요한(relevant) 정보가 포함되어 있는 특정 부분에 더 집중하도록 하여 성능 향상을 꾀할 수 있기 때문에 최근 어텐션 기법에 관한 연구진들의 관심이 매우 뜨겁다. 본 종설에서는 어텐션 기법의 기술적 측면을 먼저 살피고, 이를 의료 영상 분류 및 분할에 적용한 최신 동향을 토대로 앞으로의 전망과 발전 방향 등에 대해 다루고자 한다.

어텐션 기법 및 의료 영상에의 적용 가능성

어텐션 기법은 인공지능을 활용한 영상 분석의 성능을 한 차원 더 끌어올릴 수 있는 기법이다. 어텐션 기법은 '영상의 어느 영역에 집중하면 되는지 파악하는 능력'에 해당하는데, 구체적으로는 '달성해야 할 목적과의 연관성에 따라 추출한 특징에 가중치를 부여함으로써 네트워크 스스로 집중해야 할 영역을 선별하는 기술'이다(24).

Fig. 2. The resulting heatmaps after training a deep learning model to distinguish between a bare face, a masked face, and a face of a person wearing glasses. The regions that match the characteristics of each class are activated correctly. The photographs have been used after consent.



일반(디지털카메라) 영상과 비교하면, 의료 영상은 각 영상의 특성(대조도, 형태, 히스토그램 등)이 서로 상당히 유사하다. 즉, 의료 영상은 촬영 목적에 따라 MRI, CT, X-ray, 그리고 양성자방출단층촬영(PET)과 같은 촬영 장치에 유사한 촬영 매개변수를 사용하여 환자의 신체 부위 별로 표준화된 위치에서 획득하기 때문에, 촬영 부위의 방향, 범위 등에 있어 일반 영상에 비해 편차가 적다(25, 26). 이와 같이 영상의 특성이 유사하면서 탐색해야 할 영역의 범위가 상대적으로 제한적이고 이상 영역(병변)이 정상 영역에 비해 두드러지는 공통된 특성을 보이기 때문에, 딥러닝 모델로 하여금 집중해야 할 영역이 어디인지 학습시키는 것이 비교적 용이할 수 있으며, 이를 통해 정밀한 집중 영역 검출을 기대할 수 있다. 다시 말해, 딥러닝 모델로 하여금 어느 영역에 집중해야만 주어진 목적을 달성할 수 있는지를 학습시킴으로써 효과적인 성능 향상을 기대할 수 있는데(27), 이는 마치 전문의가 특정 병변을 찾기 위해 주로 어떤 특성의 부위를 주의 깊게 봐야 하는지에 관한 경험을 쌓아가면서 업무의 속도와 정확도를 점차 높일 수 있는 것과 비슷한 방법이다. 또한, 핵심 영역에 집중시킴으로써 의료 영상 데이터 부족 및 편중 현상으로 인한 성능 저하 문제를 이겨낼 수 있고, 학습 데이터와 분포가 다소 다른 데이터에 적용했을 때 발생하는 성능 저하 문제 또한 완화시킬 수 있다(24).

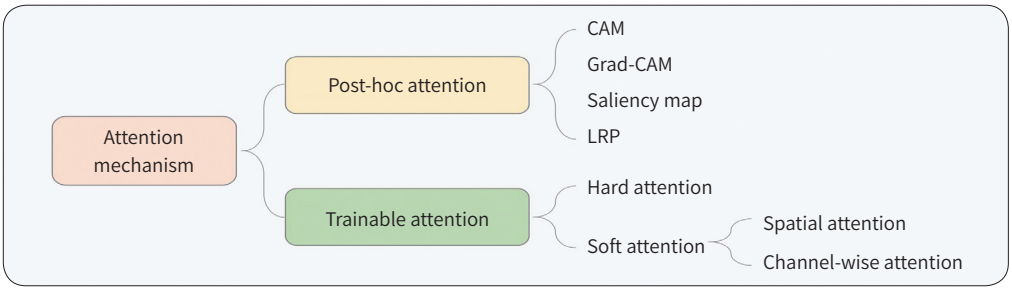
어텐션 기법의 구분

어텐션 기법은 크게 사후 네트워크 분석을 위한 어텐션(post-hoc attention for network analysis), 그리고 네트워크와 동시에 학습되는 학습 가능한 어텐션(trainable attention) 두 종류로 구분된다(Fig. 3). 먼저 "사후 네트워크 분석을 위한 어텐션"은 영상 판독이나 검출 등에서 활용되는 네트워크 추론(network reasoning) 기술로서, 학습이 모두 끝난 네트워크에 적용하여 분류나 검출의 근거로 삼은 영역이 어디인지를 파악하는 기술이다. 딥러닝 모델이 폐암 환자의 방사선 영상을 보고 폐암으로 분류했다면, 어느 부분을 보고 그런 결정을 내렸는지 표현하는 기술이다. 설명 가능한 인공지능 기술로서 자주 언급되는 클래스 활성화 지도(Class Activation Mapping; 이하 CAM), 경사 가중치 클래스 활성화 지도(Gradient-weighted Class Activation Mapping; 이하 Grad-CAM), 세일리언시 지도(Saliency Map), 계층별 관련도 전파법(Layer-wise Relevance Propagation; 이하 LRP) 등 다양한 기술이 여기에 포함된다(28-31). 두 번째로 "학습 가능한 어텐션"은 딥러닝 모델 안에 어텐션 기법을 삽입하여 내재적으로(intrinsically) 함께 학습되도록 함으로써 전체 영상이 아닌 특정 영역에 더 집중하도록 능동적으로 학습하는 기술이다. 학습을 다 마친 후에 딥러닝 모델이 어느 영역에 집중했는지 분석하는 것이 아니라, 어텐션 기법이 딥러닝 모델과 함께 학습되며 중요도가 높은 영역의 특징에 더 높은 가중치를 스스로 부여하도록 한다. 학습 가능한 어텐션 기술은 표현되는 어텐션 영역의 특성에 따라 크게 하드 어텐션(hard attention)과 소프트 어텐션(soft attention)으로 구분할 수 있으며, 최근에는 단순히 공간축(spatial dimension)을 따라 표현되는 어텐션과 다르게 딥러닝 모델 내 특징 지도(feature map)의 채널 축을 따라 표현되는 채널 방향의 어텐션(channel-wise attention) 기술도 많은 각광을 받고 있다.

사후 네트워크 분석용 어텐션 기법(Post-hoc Attention Mechanism)

딥러닝은 성능을 높이기 위해 모델을 깊게 설계할수록 그 복잡도가 기하급수적으로 증가하여 모델 내부에서 어떤 처리가 이루어지는지 직관적으로 해석하기 어렵다. 해석이 불가능하다는 점

Fig. 3. Overall tree of attention mechanisms, which fall into two main categories: post-hoc attention and trainable attention.
Post-hoc attention includes gradient-based CAM, Grad-CAM, saliency map, and propagation-based LRP, whereas trainable attention includes hard attention and soft attention.
CAM = Class Activation Mapping, Grad-CAM = Gradient-weighted CAM, LRP = Layer-wise Relevance Propagation



은 의료 분야, 특히 환자의 진단과 직결되는 영상의학 분야에서는 매우 치명적인데, 이를 해결하기 위해서 합성곱 신경망 기반 모델의 해석을 위한 다양한 알고리즘들이 개발되었으며, 그중 최근 연구에 많이 사용되는 대표적인 4가지 방법을 소개한다.

클래스 활성화 지도(Class Activation Mapping)

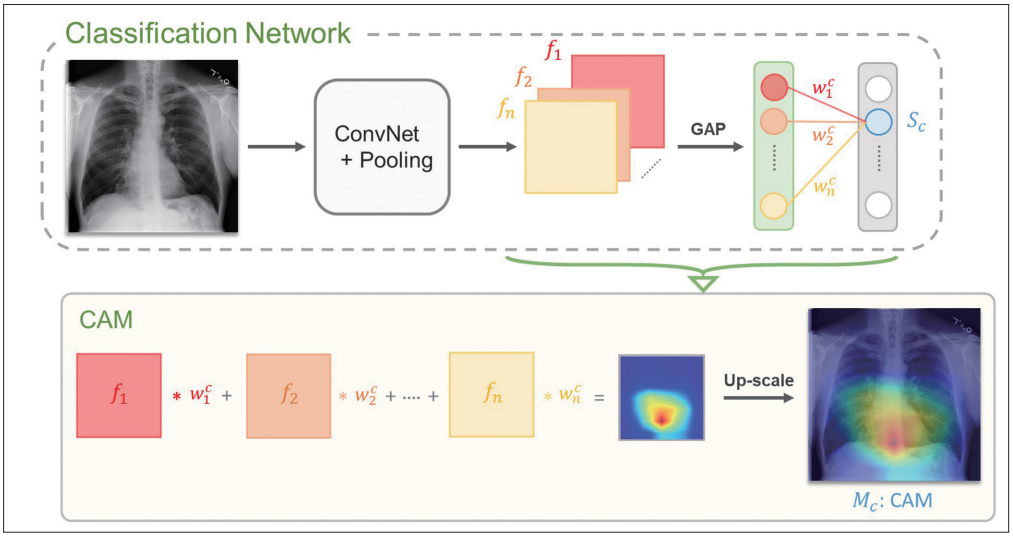
VGGNet, ResNet, GoogleNet 등 이름을 알린 거의 모든 합성곱 신경망 기반 딥러닝 모델은 그 층의 깊이가 매우 깊고 사용되는 파라미터(parameter)의 수가 많기 때문에 모델 내부의 계산 과정을 직관적으로 이해하기 힘들다. 클래스 활성화 지도(CAM)란 주로 이러한 분류 및 판별의 문제에서 각 클래스별 확률을 계산함에 있어 딥러닝 모델이 영상의 어느 부분에 집중했는지를 해석하기 위해서 고안된 방법이다.

우선 CAM은 Fig. 4와 같이 합성곱 신경망 층 이후 마지막 특징 지도(f_k)들을 각각의 평균값으로 압축하는 전역 평균 풀링(Global Average Pooling; 이하 GAP) 과정을 거치고 이에 한 개의 완전 연결 층(fully connected layer)이 따라오는 구조에 적용 가능하다. CAM의 기본 아이디어는, 특징 지도에 GAP을 적용하여 압축한 값(GAP를 통해 각 특징 지도를 대표적으로 표현하는 값)이 해당 클래스에 미치는 영향이 높을수록 높은 가중치가 형성되기 때문에 그 가중치를 특징 지도들에 반영하여 지도의 구역별로 특정 클래스를 활성화하는 정도를 시각화할 수 있다는 것이다. CAM을 추출하는 방법을 수식으로 보면 아래와 같다.

$$S_c = \sum_k W_k^c \left[\frac{1}{Z}\sum_{i,j} f_k(i,j)\right] \qquad 1)$$

$$M_c(i,j) = \sum_k W_k^c f_k(i,j) \qquad 2)$$

Fig. 4. Overall framework of CAM applied to an already trained classifier to highlight the class-related discriminative regions detected using the network.
CAM = Class Activation Mapping, GAP = Global Average Pooling



각 요소의 의미를 보면, S_c: 클래스 C에 대한 모델의 출력값, f_k: 모델 합성곱 층 최말단의 k번째 특징 지도, Z: 영상 안의 총 픽셀 개수, W^c_k: 클래스 C에 대한 K 번째 f_k에 적용되는 가중치, M_c: 클래스 C에 대한 CAM이다. 각 특징 지도(f_k)에 그에 상응하는 가중치(W^c_k)를 곱한 후 이를 합하여 CAM을 얻는다.

CAM은 고차원 특징이 반영된 최말단의 특징 지도(fk)를 사용하여 모델이 집중한 부분을 잘 지역화 한다는 장점이 있으나(28), 해상도가 낮다는 단점이 있다. 그 이유는 입력 영상이 네트워크를 통과하면서 수용 영역을 넓히고 더 많은 정보를 추출하기 위해 풀링(pooling)이 가해지기 때문에 최말단의 특징 지도(f_k)에 이르러서는 그 크기가 원본 영상보다 작아지는데, 이를 원본 영상에 대응시키기 위해서 CAM의 크기를 원본 영상만큼 업-샘플링(up-sampling) 하는 과정에서 해상도가 저하되기 때문이다.

의료 영상 분석에서의 CAM 활용

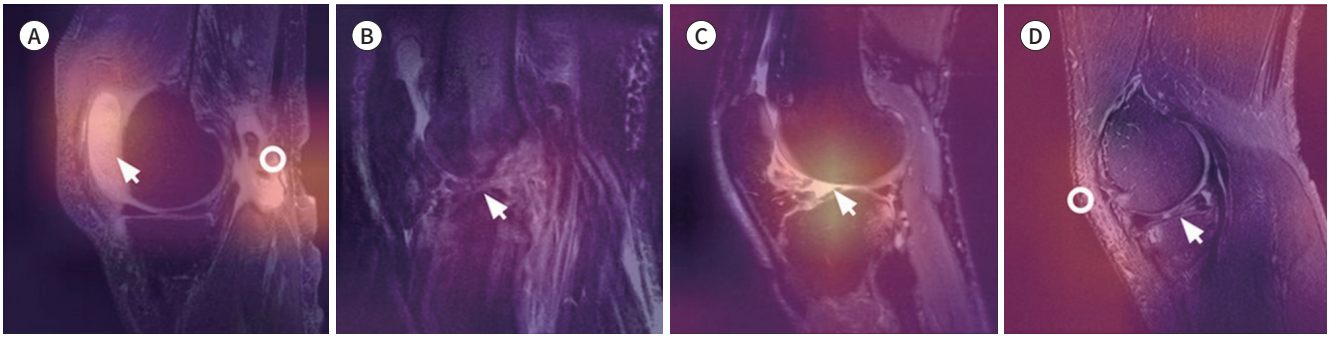
딥러닝 기반의 의료 영상 분석에 적용된 CAM의 예시로, PLOS Medicine 지에 게재된 Bien 등(32)의 "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet"이 있다. 이 논문의 연구진은 무릎 MRI에 자체 개발한 딥러닝 모델(MRNet)을 이용하여 3가지 (비특이적 이상, 전십자인대 파열, 반달연골 파열) 병명을 진단했고 CAM을 이용하여 그 진단의 근거를 확인하였다.

Fig. 5의 각 영상에 대한 설명은 아래와 같다.

Fig. 5A는 시상면(sagittal) T2 강조 무릎 영상(T2-weighted knee MRI)이며, 다량의 삼출(effusion)과 비복근 힘줄 파열(rupture of the gastrocnemius tendon)이 관찰되고 MRNet은 이를 이상이 있는 것으로 분류하였다. 모델의 CAM이 삼출 영역(화살표)과 비복근 힘줄 파열 부위(흰 고리)를 비교적 잘 나타내고 있다. 모델이 질병의 분류만을 학습했는데도 이상이 발생한 부위를 검출할 수 있다는 것을 보여준다.

Fig. 5B는 시상면 T2 강조 무릎 영상이며 환자의 움직임으로 인해 인공물이 심한 가운데 전십자인대 완전파열(complete anterior cruciate ligament tear)이 관찰된다. 모델은 이 영상을 전십자

Fig. 5. Examples of CAM-assisted medical image interpretation.
A-D. (A-C) shows representative images in which CAM correctly localizes the pathological regions (arrows and ring), whereas in (D) CAM is highlighting the wrong region (ring) instead of the pathologically correct region (arrow) for this decision. Adapted from Bien et al. PLoS Med 2018;15:e1002699 (32) (https://doi.org/10.1371/journal.pmed.1002699), licensed under CC BY 4.0. Images have been rearranged.
CAM = Class Activation Mapping



인대파열(화살표)로 진단했고 CAM 또한 알맞게 파열 부위를 활성화하고 있다.

Fig. 5C는 시상면 T2 강조 무릎 영상이며 b)와 마찬가지로 전십자인대 완전파열(화살표)이 관찰된다. CAM에서도 전십자인대 파열을 잘 찾아내고 있음을 확인할 수 있다.

Fig. 5D는 시상면 T2 강조 무릎 영상이며 Fig. 5A-C와 달리 CAM이 잘못된 부분을 활성화하고 있는 예시이다. 영상을 보면 외측 반달연골 후부 복합 파열(complex tear involving the posterior horn of the lateral meniscus)이 보이고 모델 역시 무릎에 이상이 있는 것으로 판단하였다. 하지만 CAM은 외측 반달연골이 아닌 전면 연조직을 활성화하고 있다. 이를 통해 모델이 환자를 이상이 있는 것으로 잘 분류했지만, 판단 기준은 잘못되었음을 확인할 수 있다.

위 결과를 보면 대부분은 모델의 판단 기준과 실제 진단의 이유가 일치하지만, 그렇지 않은 경우 또한 존재함을 알 수 있다. 이는 모델의 결과를 완전히 맹신할 수는 없다는 것을 나타낸다. 하지만 단순히 모델의 판단 결과만 보는 것을 넘어, 그러한 결과에 이르게 한 근거를 시각적으로 확인할 수 있기 때문에 실제 임상 진단에 효과적으로 사용될 수 있으며 때로는 영상의학 전문의가 놓치는 부분을 보조하는 역할도 기대할 수 있다.

경사 가중치 클래스 활성화 지도(Gradient-Weighted Class Activation Mapping)

앞서 본 CAM 기법의 경우, 모델이 반드시 GAP를 포함하여야 적용이 가능하다. 그러나 대부분의 딥러닝 모델들이 출력단에 다양한 구조를 사용하고 있다는 점에서 이러한 제한 사항은 큰 단점으로 작용한다. 따라서 모델의 판단 근거를 설명함에 있어서 모델의 구조에 구속되지 않는 유연한 방법이 요구되었으며, 이 필요성에 의해 고안된 것이 경사 가중치 클래스 활성화 지도(Grad-CAM)이다.

하지만 Grad-CAM 역시 모델의 구조에 약간의 제한이 있는데, 그것은 적어도 한 개 이상의 합성곱 신경망 층이 모델 내부에 포함되어 있어야 한다는 것이다. 하지만 영상처리에서 사용되는 딥러닝 모델의 경우 대부분의 경우에서 합성곱 신경망 층을 사용하기 때문에 이 제한은 실제로는 모델의 유연성에 거의 영향을 주지 않는다. Grad-CAM의 작동 구조는 Fig. 6에 나타나 있다.

Fig. 6을 수식을 통해 보면 다음과 같다.

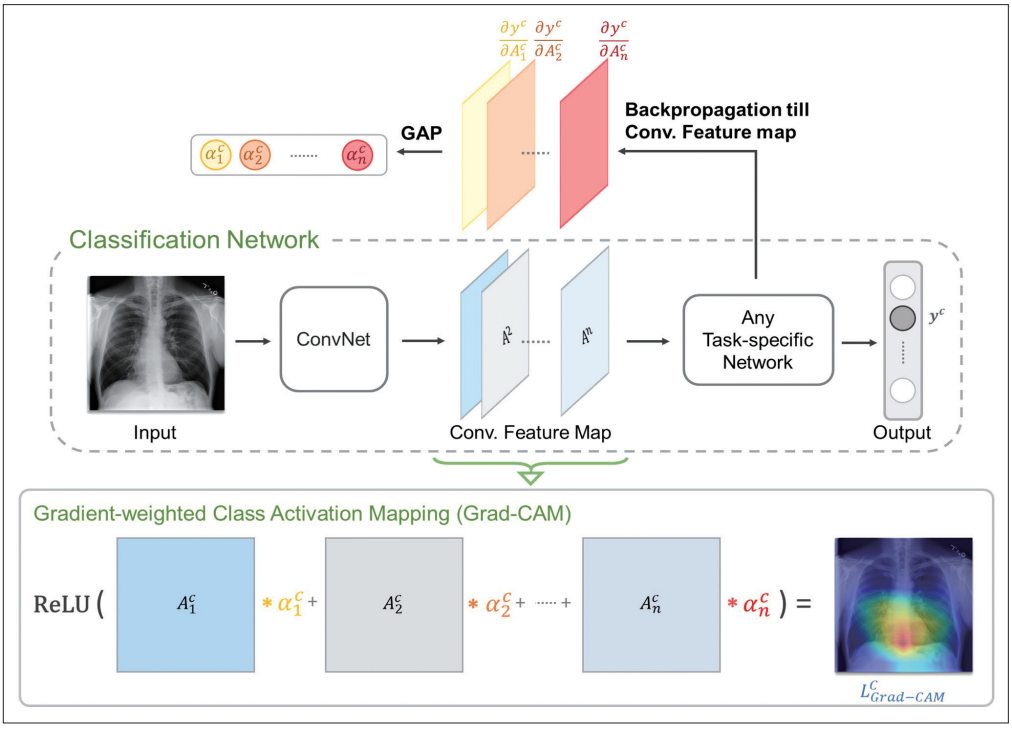
$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k} \qquad 3)$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^K\right) \qquad 4)$$

각 요소들의 의미를 보면, A^k: 합성곱 신경망의 k 번째 특징 지도, Z: 영상 내 총 픽셀 개수, $\frac{\partial y^c}{\partial A^k}$: 클래스 C에 대응되는 출력값($y^c$)에 대한 k 번째 특징 지도의 미분과 같다.

Grad-CAM 또한 CAM과 마찬가지로 합성곱 신경망 층의 특징 지도가 특정 클래스에 미치는 영향에 관련된 가중치들을 이용한다. 그러나 CAM에서는 완전 연결층의 가중치(Fig. 4의 W_k^c)를 그대로 사용한 것과 달리, Grad-CAM에서는 역전파(backpropagation)를 통해 얻을 수 있는 미분 지도들의 픽셀 별 그래디언트(gradient) 값($\frac{\partial y^c}{\partial A^k}$)의 전역 평균($\frac{1}{Z}\sum_i\sum_j$)을 구하여 가중치로 사용한다. 그 기반 아이디어는 특정 클래스와 밀접하게 연관된 특징 지도일수록 해당 특징 지도 내 픽

Fig. 6. Overall framework of Grad-CAM, which is more flexible than CAM in its application in neural networks.
CAM = Class Activation Mapping, Conv = convolutional layer, GAP = Global Average Pooling, Grad-CAM = Gradient-weighted CAM, ReLU = rectified linear unit



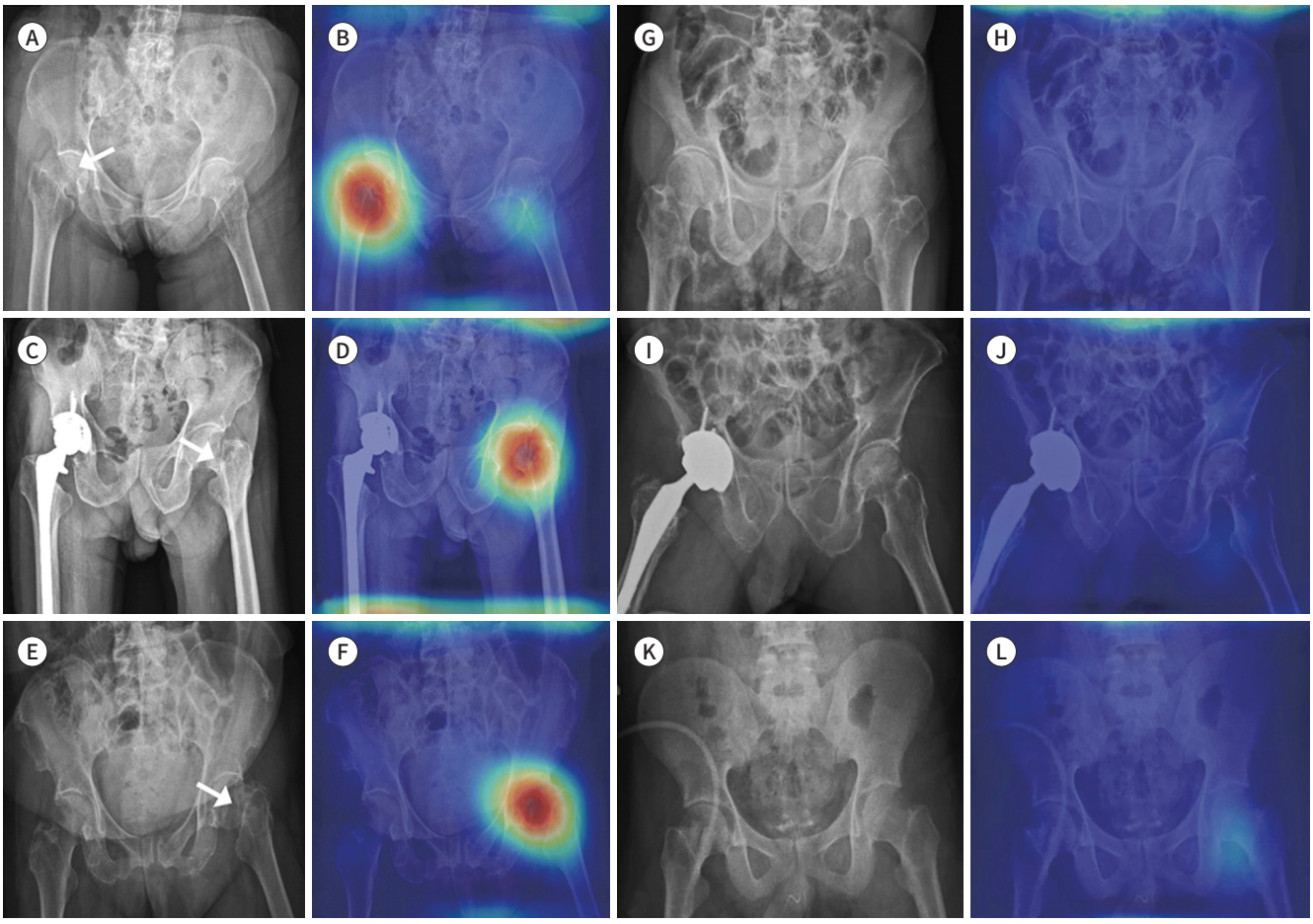
셀들에 대응하는 그래디언트 값이 클 것이며, 이를 평균하여 해당 특징 지도의 특정 클래스에 대한 가중치를 수치화 할 수 있다는 것이다. 미분 지도는 역전파를 통해 클래스에 대한 특징 지도(A^k)의 미분을 구하여 얻을 수 있는데, 대부분의 딥러닝 모델에서 1차 미분값은 쉽게 계산할 수 있으므로 모델 구조에 제한받지 않고 더 유연하게 적용이 가능하다. CAM과 마찬가지로 각 특징 지도에 그에 상응하는 가중치를 곱한 후 이를 합하여 Grad-CAM을 얻게 된다. 하지만 Grad-CAM 역시 합성곱 신경망 층의 특징 지도를 기반으로 만들어지는데, 효과적으로 수용 영역을 넓히기 위해 모델 구조에 풀링이 포함된 경우가 많아서 그 크기가 원본 영상보다 작아진다. 따라서 원본 영상과 대응시키기 위해서는 그 크기를 키워야 하고 그로 인해 CAM과 마찬가지로 해상도가 저하된다는 단점이 있다.

의료 영상 분석에서의 Grad-CAM 활용

딥러닝 기반 의료 영상 분석에 적용된 경사 가중치 활성화 지도의 예시로는, European Radiology에 게재된 Cheng 등(33)의 "Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs"이 있다. 이 모델이 만들어낸 Grad-CAM을 보면 Fig. 7과 같다.

Fig. 7의 좌측 두 열의 경우 골반 골절을 확인할 수 있으며(흰색 화살표, Fig 7A, C, E) 모델의 Grad-CAM 역시 골절 부분을 활성화하고 있는 것을 볼 수 있다. 우측 두 열의 경우 골반 골절이 없

Fig. 7. Examples of Grad-CAM-assisted image interpretation of hip fractures (A, C, E). Pelvic radiograph images with fractures indicated by white arrows (B, D, F). Respective Grad-CAM images of (A-C), which are correctly localizing the fracture sites (G, I, K). Normal pelvic radiograph images with no fractures (H, J, L). Respective Grad-CAM images of G, I, K, showing no activated regions for normal pelvic radiographs. Adapted from Cheng et al. Eur Radiol 2019; 29:5469-5477 (33) (https://doi.org/10.1007/s00330-019-06167-y), licensed under CC BY 4.0. Two figures have been concatenated.
Grad-CAM = Gradient-weighted Class Activation Mapping



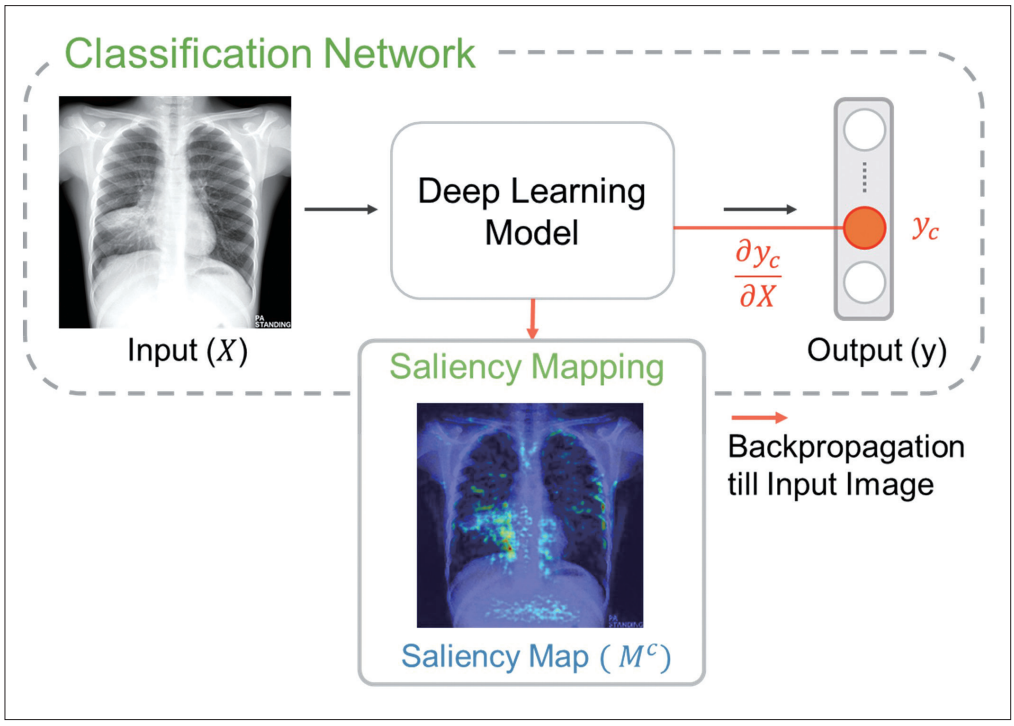
는 환자의 사진 및 Grad-CAM 영상인데, 이 경우에는 Grad-CAM이 특별히 활성화되지 않은 것을 확인할 수 있다. 이와 같이 Grad-CAM으로 모델의 판단 근거를 CAM에 비해 더 지역화된 형태로 정밀하게 시각화하여 효율적이고 안전하게 진료를 보조할 수 있다.

세일리언시 지도(Saliency Map)

세일리언시 지도는 모델의 클래스별 출력값에 대한 입력 영상의 미분으로 계산이 되는데, 이는 입력 영상의 특정 픽셀값의 변화에 따른 출력의 변화가 상대적으로 크다면, 그 픽셀이 출력값에 기여하는 바가 크다는 것을 의미한다는 가정에서 고안되었다. 세일리언시 지도의 전체적인 설명도는 Fig. 8과 같다.

세일리언시 지도를 수식으로 보면 아래와 같다.

Fig. 8. Overall framework of saliency mapping.



$$M(i,j) = \frac{\partial y^c}{\partial X(i,j)} \qquad 5)$$

즉, 클래스 C에 상응하는 모델 출력에 대한 입력의 미분값이 세일리언시 지도가 된다.

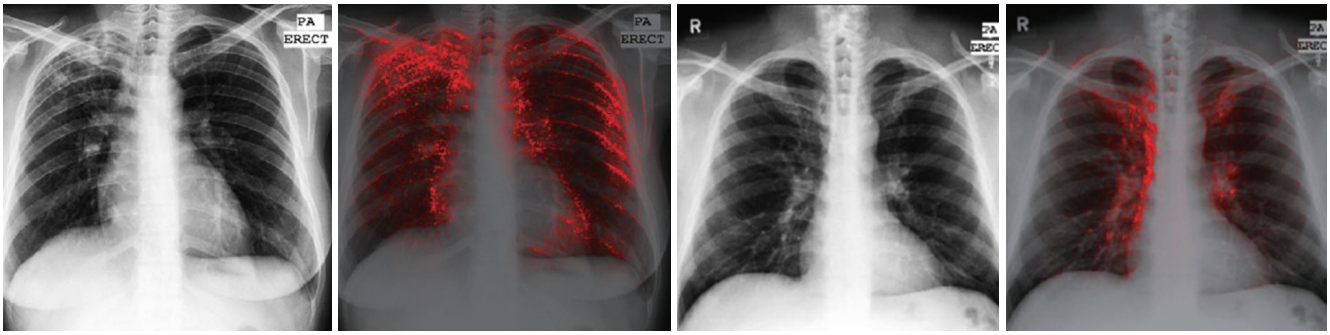
세일리언시 지도의 특징은 그 연산이 모델의 구조와는 완전히 무관함에 있다. 그렇기 때문에 다양한 모델에 유연하게 적용이 가능하다. 또한 출력에 대한 입력의 미분으로 계산되기 때문에 히트맵이 입력영상과 동일한 해상도를 유지할 수 있다. 따라서 다른 CAM 기반의 기법들에 비해 해상도의 저하 없이 판단의 근거가 되는 영역의 세밀한 부분을 시각화할 수 있다. 그러나, 그래디언트를 입력단으로 전파시키는 과정에서 비선형적인 활성화 함수 등으로 인한 그래디언트 정보 소실(shattered gradient problem)이 발생할 수 있어 표현되는 히트맵에 잡음이 나타나고(34), 고차원 정보가 압축된 특징 지도를 활용하지 않기 때문에 지역화 능력이 떨어진다는 단점이 있다.

의료 영상 분석에서의 Saliency Map 활용

2019년 Scientific Reports 지에 게재된 Pasa 등(35)의 연구에서는 흉부 엑스레이 영상에서의 결핵(tuberculosis) 검진을 위해 딥러닝이 사용되었고, 세일리언시 지도를 이용하여 모델을 이해하고 영상의학과 전문의의 육안 진단에 도움을 줄 수 있음을 확인했다. 실제로 세일리언시 지도가 적용된 영상을 보면 Fig. 9와 같다.

Fig. 9의 좌측의 두 영상은 결핵으로 진단된 환자의 흉부 엑스레이 영상 및 세일리언시 지도이다. 딥러닝 모델도 환자를 결핵으로 진단하였는데, 세일리언시 지도로 모델의 판단을 이해할 수 있다. 엑스레이 사진을 보면 우측 상엽에 흉막 정점의 비후(pleural apical thickening)로 인한 불

Fig. 9. Examples of saliency map-assisted X-ray interpretation.
The two images on the left show a positive example of saliency map-assisted interpretation where a chest X-ray image of a patient with tuberculosis is presented with the saliency map that is correctly highlighting the pathological regions. On the other hand, the two images on the right show a negative example of saliency map-assisted interpretation where a chest X-ray image of a person with no disease is presented with a saliency map that is incorrectly highlighting normal regions as abnormal. Adapted from Pasa et al. Sci Rep 2019;9:6268 (35) (https://doi.org/10.1038/s41598-019-42557-4), licensed under CC BY 4.0. Two figures have been slightly modified and concatenated.



투명함을 보이고 있으며 우측 폐문(right hilum)이 위쪽으로 편차를 보이고 있다. 세일리언시 지도에서 또한 우측 상엽이 강하게 활성화되어있는 것을 볼 수 있다. 이와 반대로, 우측의 두 영상은 질병이 없는 환자의 엑스레이 영상 및 세일리언시 지도이다. 딥러닝 모델은 이 환자를 결핵으로 잘못 진단하였는데, 세일리언시 지도를 보면 우측 상엽을 주목했음을 알 수 있다. 하지만 우측 상엽의 불투명함은 쇄골과 늑골의 중첩으로 인해 발생한 것이기 때문에, 결과적으로 모델의 판단이 틀렸음을 알 수 있다.

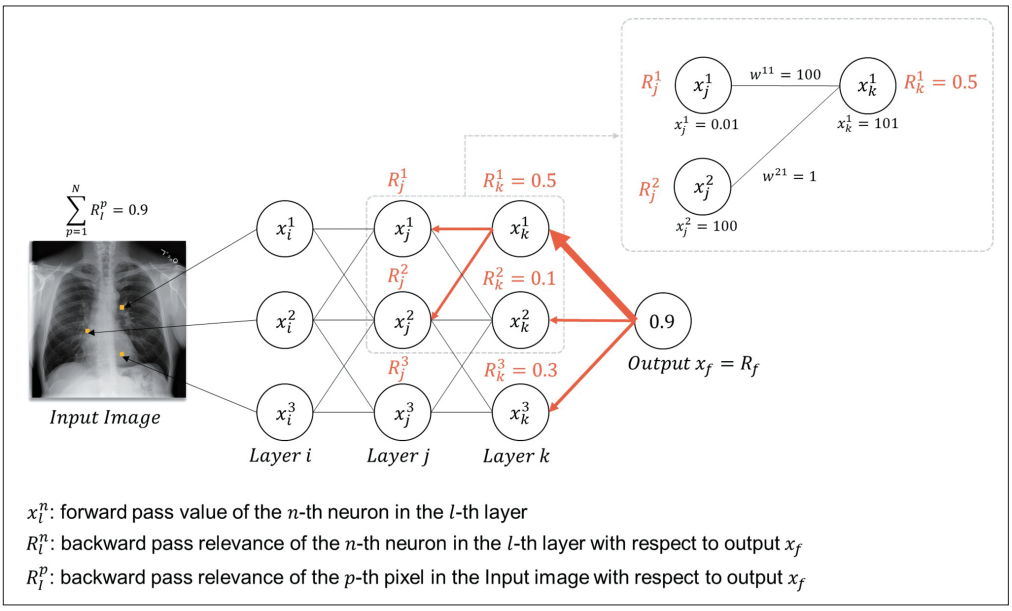
이 논문에서는 Grad-CAM과 세일리언시 지도를 비교하였는데, 세일리언시 지도의 경우 해상도에 변화가 없지만, Grad-CAM의 경우 영상 크기 확대로 인한 해상도 저하가 불가피하기 때문에 크기가 작은 결핵과 같은 질병의 경우 세일리언시 지도를 이용하는 것이 모델의 판단 근거를 정밀하게 시각화하는 데 있어 더 효과적이라고 서술하고 있다.

계층별 관련도 전파법(Layer-Wise Relevance Propagation)

앞서 소개한 CAM 기반의 방법 및 세일리언시 지도가 모두 그래디언트를 활용하여 각 특징 지도 및 입력 픽셀의 중요도를 결정한다면, LRP는 조금 다른 방식으로 모델을 해석한다. LRP에서는 추론 결과에 대한 '관련도(relevance)'를 입력 데이터까지 계층별(layer-wise)로 역산하여 전파(propagation) 시키는데, 이를 통해 특정 추론값을 결정하는데 있어 입력 픽셀 별 기여도(입력 영상에서 중요하게 작용한 영역이 어디였는지)를 직접적으로 표현한다. 세일리언시 지도와 비슷한 개념이지만, 그래디언트가 아니라 '관련도'를 입력 픽셀까지 전파시킨다는 점이 다르다. 때문에 그래디언트 기반의 방법에서 발생할 수 있는 그래디언트 소실 문제로부터 훨씬 자유롭다.

Fig. 10을 통해 LRP에 대한 쉬운 이해가 가능하다. 그림과 같이 입력 영상에 대하여 잘 훈련된 신경망이 특정 클래스로 분류를 했고 그 출력값으로 0.9를 얻었다면, 이를 출력단에서의 관련도로 설정한다. LRP에 내재된 아이디어는, 출력단 이전의 계층(layer k)에 존재하는 뉴런들에 출력값에 대한 기여도(관련도)를 분배할 수 있으며, 그 합은 0.9로 보존되어야 한다는 것이다. 같은 원리로 이를 입력단까지 반복하면, 각 입력 픽셀 별로 0.9라는 출력값에 대한 기여도를 정량화할 수 있으

Fig. 10. Overall framework of layer-wise relevance propagation. The relevance is distributed to each input pixel by reversely propagating the relevance layer-wise.



며, 특정 추론 결과를 창출함에 있어 모델이 집중한 영역을 표현할 수 있다.

기여도를 분배하는 원리는 Fig. 10의 박스 안과 같다. Layer k에서 $x_k^1 = x_j^1 \times w^{11} + x_j^2 \times w^{21}$이 성립하고, x_k^1를 형성함에 있어 $x_j^1 \times w^{11}$보다 $x_j^2 \times w^{21}$의 값이 훨씬 더 크게 기여한다. 이 비율에 따라 x_f에 대한 x_k^1의 기여도 R_k^1을 다음과 같이 R_j^1, R_j^2에 분배할 수 있다.

$$R_j^1 = \frac{x_j^1 \times w^{11}}{x_j^1 \times w^{11} + x_j^2 \times w^{21}} \times R_k^1 = \frac{0.01 \times 100}{0.01 \times 100 + 100 \times 1} \times 0.5 \approx 0.0049 \quad 6)$$

$$R_j^2 = \frac{x_j^2 \times w^{21}}{x_j^1 \times w^{11} + x_j^2 \times w^{21}} \times R_k^1 = \frac{100 \times 1}{0.01 \times 100 + 100 \times 1} \times 0.5 \approx 0.4950 \quad 7)$$

다만, R_j^1과 R_j^2에 대한 계산은 끝난 것이 아니며, R_j^1의 경우 x_k^2, 그리고 R_j^2의 경우 x_k^2 및 x_k^3에서 전파되어온 관련도까지 같은 원리에 따라 모두 합쳐야 x_f에 대한 뉴런 x_j^1, x_j^2의 관련도 계산이 끝나게 된다. 이를 입력 데이터의 모든 픽셀까지 반복적으로 시행했을 때, 출력값에 대한 입력 픽셀 별 기여도를 표현할 수 있게 된다.

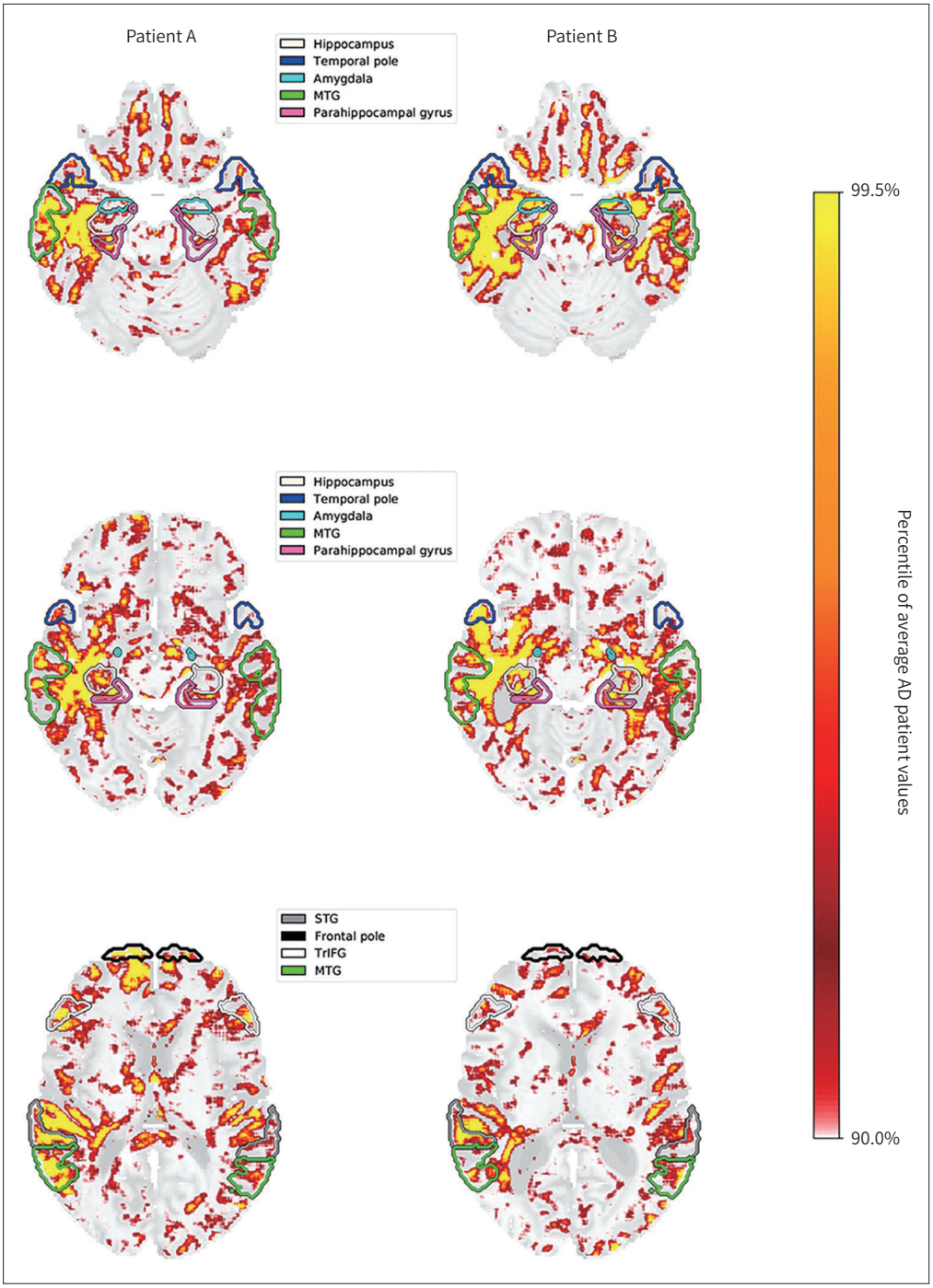
의료 영상 분석에서의 LRP 활용

2019년 Frontiers in Aging Neuroscience 지에 게재된 Böhle 등(36)의 연구에서는 환자의 뇌(brain) 자기공명영상을 Convolutional Neural Network (이하 CNN) 기반의 신경망에 통과시켜 알츠하이머병(Alzheimer's Disease) 여부를 판별하였고, LRP를 통해 추론값에 크게 기여한 입력 영역을 히트맵으로 시각화하였다.

Fig. 11은 알츠하이머로 분류된 두 환자 A, B의 뇌에서 네트워크의 결정에 기여한 영역이 다소

다른 것을 나타내고 있다. 환자 B의 경우 측두엽(temporal lobe)의 영역에 크게 영향을 받는 반면, 환자 A의 경우 전두부(frontal area) 및 상측두회(superior temporal gyrus) 영역이 가장 크게 기여하는 것으로 나타났다. 계산되는 관련성 지도가 준수한 지역화 능력과 함께 CAM 기반의 히

Fig. 11. Relevance heatmaps for two patients, A and B, who were classified as having AD. In the case of patient A, the frontal area and STG were informative, whereas the temporal lobe was important in the model decision for patient B. Adapted from Böhle et al. Front Aging Neurosci 2019;11:194 (36) (https://doi.org/10.3389/fnagi.2019.00194), licensed under CC BY 4.0.
AD = Alzheimer's disease, MTG = medial temporal gyrus, STG = superior temporal gyrus



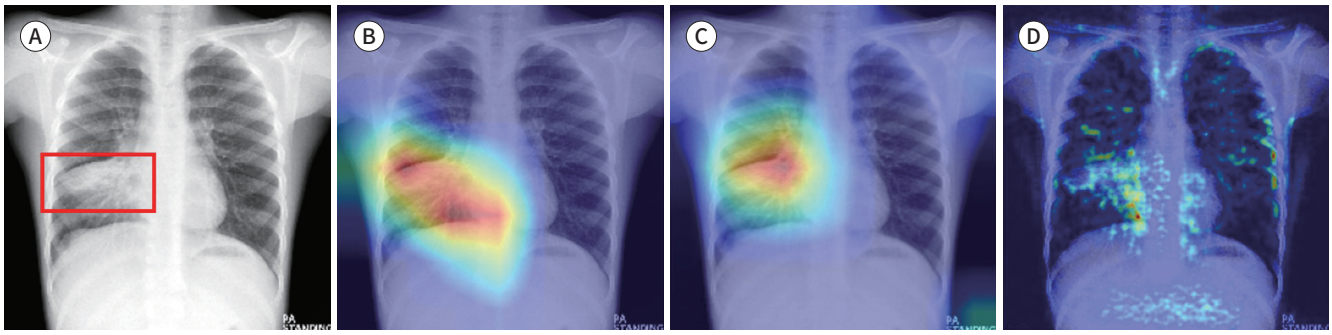
트맵에 비해서 높은 해상도로 각 픽셀 별 기여도를 표현하고 있는 것을 알 수 있다.

의료 영상 분석에 있어 사후 분석용 어텐션 기법 간의 비교

Fig. 12는 폐렴 환자 및 정상인의 X-ray 사진을 분류하는 인공지능 모델에 본 연구진이 직접 CAM, Grad-CAM, 세일리언시 지도 등의 사후 분석용 어텐션을 적용하여 각 기법을 비교한 결과이다. Fig. 12A는 원본 X-ray 영상으로서(37), 폐렴으로 인해 폐의 우측 중반에 불투명함을 관찰할 수 있다. Fig. 12B-D는 분류 모델의 학습이 끝난 후 각각 CAM, Grad-CAM, 세일리언시 지도를 적용한 영상이다. 3가지 어텐션 기법 모두 대체로 병변 부분을 잘 활성화하고 있지만, 자세히 관찰하면 그 특성이 서로 다르다. 세일리언시 지도의 경우 CAM이나 Grad-CAM보다 더 높은 해상도의 활성화 지도를 보여주고 있지만, 왼쪽 갈비뼈 및 하단 복부 부분도 활성화를 하고 있어 CAM이나 Grad-CAM에 비해 지역화 능력(localization)은 떨어지는 것을 확인할 수 있다. CAM과 Grad-CAM의 경우, 비슷한 활성화 지도를 보여주고 있지만 Grad-CAM은 모델에 대한 제한이 더 적다는 장점으로 인해 더 고성능의 모델을 사용하였기에 CAM에 비해 실제 병변 위치에 더 가까운 지역을 활성화하고 있는 것을 확인할 수 있다.

지금까지 블랙박스인 딥러닝 모델을 설명하기 위해 고안된 CAM 기반 활성화 기법 및 세일리언시 지도, 그리고 LRP의 이론 및 적용 사례를 알아보았다. CAM은 가장 직관적으로 이해가 가능하지만 모델의 구조에 제한이 있으며, 여러 층의 합성곱 층을 지난 후에 얻어지는 특징 지도를 통해서 계산되기 때문에 그 크기를 입력 영상과 대응시키기 위해 영상의 크기를 확대하는 과정에서 해상도가 낮아지는 단점이 있다. Grad-CAM은 단순 CAM 기법에 비해서는 모델의 형식에 제한이 적지만 여전히 해상도에 대한 문제를 극복하지 못했다. 그러나 이러한 CAM 기반의 기법들은 해상도와는 별개로 모델 내부의 특징 지도에서 얻어지는 고차원적인 정보를 이용하기 때문에 그만큼 모델의 판단 근거가 되는 영역의 검출에 뛰어나다는 강점이 있다. 이에 반해 세일리언시 지도의 경우 모델의 구조에 제한이 전혀 없고, 그 지도의 크기가 입력 영상과 동일하기 때문에, 해상도의 저하가 일어나지 않는다는 장점을 가지고 있다. 하지만 세일리언시 지도는 단순히 입력에 대한

Fig. 12. Comparison between CAM, Grad-CAM, and saliency map.
A-D. (A) chest X-ray image of a patient with pneumonia, (B) CAM, (C) Grad-CAM, (D) saliency map overlaid on the original image. Adapted from Rad_doc, rID 47997, "Childhood Pneumonia". Available at. https://radiopaedia.org/cases/childhood-pneumonia-1?lang=us, with permission of Radiopaedia (37).
CAM = Class Activation Mapping, Grad-CAM = Gradient-weighted CAM



출력의 미분이기 때문에 그만큼 고차원적인 정보가 이용되지 않으며, 그에 따라 CAM 기반의 기법들에 비해 영역 검출의 정확성(accuracy)이 떨어지고, 잡음 등의 외적 요인에 민감하다는 단점이 있다. LRP의 경우, CAM이나 세일리언시 지도 등의 그래디언트 기반과는 다른 접근의 모델 해석 방식으로서, 입력 픽셀 별 기여도를 높은 해상도로 관찰할 수 있으면서도 그래디언트 소실 문제로 인한 잡음에서 비교적 자유롭다. 그러나, 고차원 특징 지도를 활용하지 않기에 CAM 기반의 방식에 비해서는 역시 지역화 능력이 떨어진다. Table 1을 통해 각 방식의 특성을 핵심적으로 비교할 수 있으며, 본 저자들이 직접 구현한 다음 링크의 코드를 통해 폐렴 데이터를 분류(정상 혹은 폐렴) 하는 데 있어 딥러닝 모델이 집중한 영역을 시각화하고 CAM 기반 방법 및 세일리언시 지도의 특성을 쉽게 비교해 볼 수 있다(https://github.com/mongeoroo/git_pneumonia).

학습 가능한 어텐션 기법(Trainable Attention Mechanism)

앞서 소개한 사후 분석용 어텐션 기법은 이미 학습이 완료된 모델에 적용하여 그 판단의 근거를 확인하기 위해 사용되는 기법이다. 위 기법들은 학습된 모델이 영상의 어떤 특징들에 집중하고 있는지 분석하기 위한 용도로 사용될 순 있어도, 성능 향상에는 직접적으로 기여하지 못한다. 이와 달리, 학습 과정에서 네트워크로 하여금 중요한 특징들에 더 집중하고 그렇지 않은 특징에는 덜 집중하도록 능동적으로 학습하게 하는 기법이 바로 학습 가능한 어텐션 기법이다.

학습 가능한 어텐션 기법은 크게 하드 어텐션(hard attention)과 소프트 어텐션(soft attention)으로 나뉜다. 하드 어텐션과 소프트 어텐션의 가장 큰 차이점은 바로 생성되는 어텐션 지도(attention map)의 형태에 있다. 하드 어텐션의 경우 생성되는 어텐션 지도가 중요 특징 영역은 1, 나머지 부분은 0으로 구성된 이진 마스크 형태이며, 소프트 어텐션의 경우 어텐션 지도 전반에 걸쳐 값이 존재하되 중요 영역의 값이 나머지 영역에 비해 훨씬 큰 값을 가지는 형태이다(Fig. 13). 어떤 종류의 어텐션을 선택하느냐에 따라 상충 관계(trade-off)가 존재하는데, 하드 어텐션의 경우 처리 과정에서 전체 영상이 아닌 특정 영역만(예컨대, 값이 1인 부분만) 저장되기에 계산과 메모리 사용량을 효과적으로 줄일 수 있는 반면, 이진 마스크를 잘라내는 (cropping)과정이 미분 가능하지 않아(non-differentiable) 일반적인 딥러닝의 역전파 알고리즘으로는 학습할 수 없고 강화 학습(reinforcement learning)과 같은 좀 더 까다로운 방법으로 학습해야 한다는 단점이 있다(27, 38). 그에 반해 소프트 어텐션은 종종 하드 어텐션보다 더 많은 메모리와 계산을 요구하지만 그 생성 과

Table 1. Comparison between the Characteristics of CAM, Grad-CAM, Saliency Map, and LRP. CAM-Based Methods Have Strength in Localization, Whereas Saliency Map and LRP Produce High Resolution Heatmaps with Flexible Application

	CAM	Grad-CAM	Saliency Map	LRP
Localization	++	+++	-	+
Flexibility	+	++	+++	+++
Resolution	-	-	+	+

CAM = Class Activation Mapping, Grad-CAM = Gradient-weighted CAM, LRP = Layer-wise Relevance Propagation

Fig. 13. Comparison between hard attention map (left) and soft attention map (right).

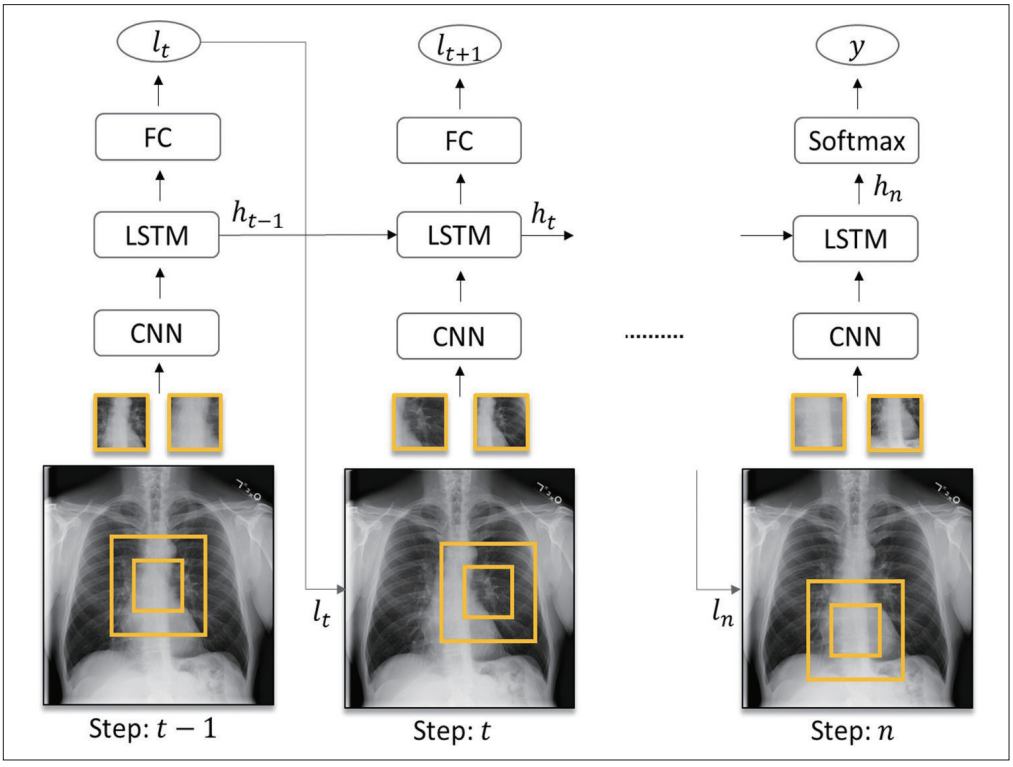
정이 미분 가능하여 일반적인 딥러닝 모델과 함께 역전파 알고리즘으로 쉽게 종단 간(end-to-end) 학습이 가능하다는 장점이 있어, 최근에는 하드 어텐션보다 더 활발히 쓰이는 추세이다.

하드 어텐션(Hard Attention)

하드 어텐션의 경우, 역전파 알고리즘을 활용한 종단 간 학습이 어렵고, 강화 학습 등을 사용해야 한다는 점에서 소프트 어텐션 대비 기술 자체의 복잡도가 높다. 본 종설에서는, 강화 학습 기반의 기본적인 하드 어텐션 및 강화 학습 없이 비교적 쉽게 구현한 하드 어텐션 연구를 소개하여 독자들의 기초적인 이해를 돕고자 한다.

강화 학습 기반의 하드 어텐션은 순환 어텐션 모델(recurrent attention model; 이하 RAM)을 이용하여 학습된다. RAM은 Fig. 14와 같이 핵심적인 정보(informative area)를 포함하고 있는 영역(Fig. 14의 노란 박스)을 순환적으로 탐색해가는 모델로서, 순환 신경망(recurrent neural network; 이하 RNN) 및 강화 학습을 활용한다. 즉, Fig. 14의 t 단계를 기준으로 설명하면, 현재 위치 l_t를 기준으로 서로 다른 크기 및 상황 정보(context)를 가지는 두 개의 패치(patch)를 추출한 후, 이를 신경망에 투입하여 정보를 압축한 뒤 이전 단계(step t-1)에서 추출된 메모리 정보(h_{t-1})와 함께 RNN 기반의 Long Short Term Memory (이하 LSTM)에 투입하여 현재까지 탐색한 정보를 종합한다. 이후, LSTM의 출력을 기반으로 완전 연결층을 통해 다음 탐색 위치를 선정한다. 이 과정을 강화 학습으로 학습하여 미리 정해진 n 단계 동안 반복적으로 시행한 후 최종적으로는 마지막 단계에서 추출된 지역의 영상 정보만을 토대로 분류를 진행한다. 이와 같은 접근을 의료 영상 분석에 활용한 첫 연구가 Ypsilantis와 Montana (38)의 “Learning what to look in chest X-rays with a recurrent visual attention model”이다. 해당 방법론을 통해 훨씬 적은 파라미터를 사용하고도 기존의 합성곱 신경망 기반의 방법 대비 필적할 만한 분류 성능을 보였으며, 분류에 핵심적인 영역을 추출하여 그 부분에 집중할 수 있도록 하였다. 그러나, 해당 방법은 패치 기반의 방법이므로 추출되는 정보가 제한적이며, 이 때문에 핵심 영역으로 접근하기 위해 굉장히 많은 학습 반복(training iteration)이 필요할 수 있고, 이러한 단점은 분류하고자 하는 클래스의 수가 많고 다양할수록 더 심화될 것으로 예상된다.

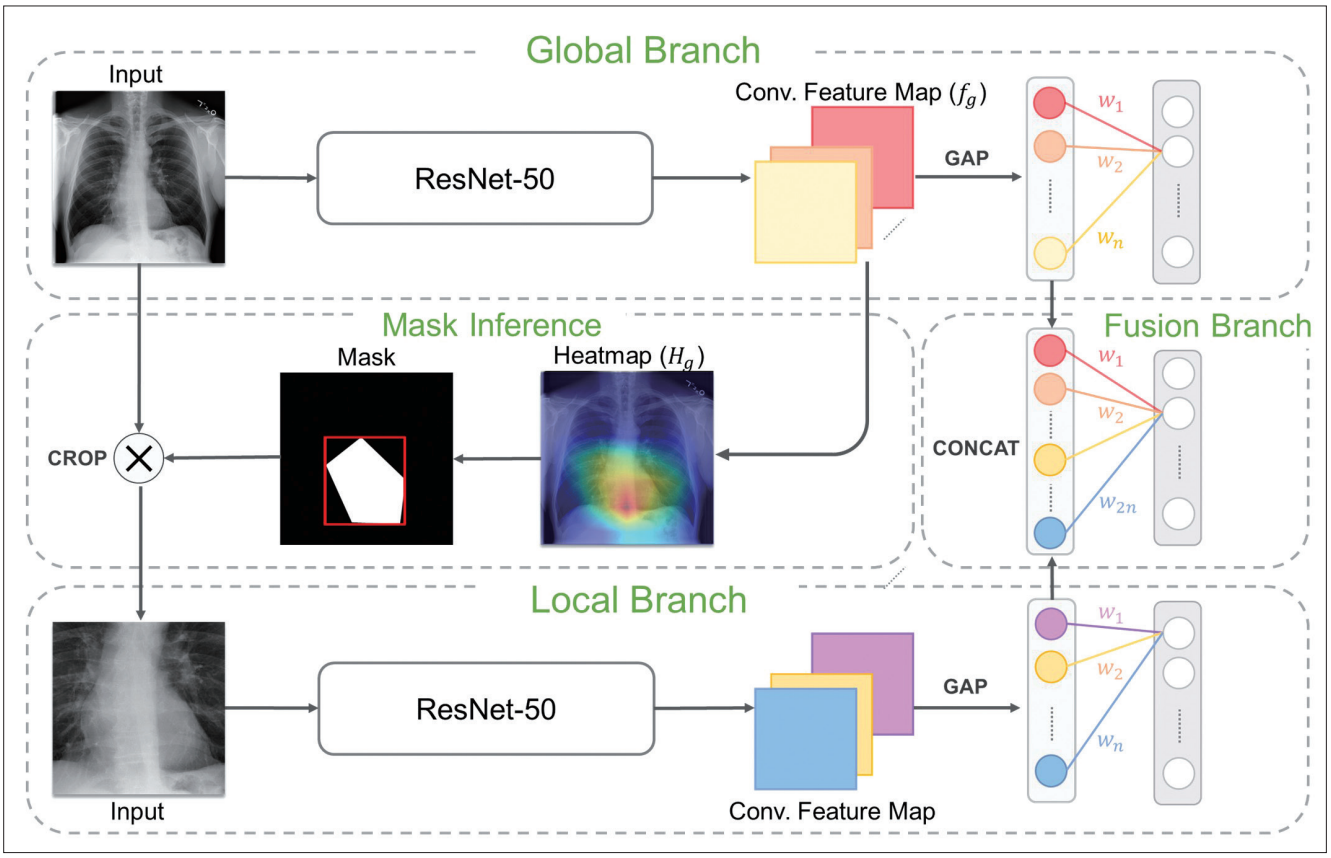
Fig. 14. Recurrent attention model. At time step, t, the next position (l_{t+1}) to be attended is sampled and this process is recursively repeated until step n, where the output of LSTM is used for classification. Figure inspired by (38).
CNN = Convolutional Neural Network, FC = fully connected network, LSTM = Long Short Term Memory



이러한 이유 때문인지 최근에 발표되는 하드 어텐션 기반의 방법들은 강화 학습 없이도 하드 어텐션을 적용할 수 있는 방법을 택하는데, 대표적으로 소개하고자 하는 연구가 Pattern Recognition Letters에 출판된 Guan 등(39)의 “Thorax disease classification with attention guided convolutional neural network”이다. 해당 연구는 전체 영상으로 이미 학습된 모델의 특징 지도로부터 모델이 집중한 핵심 영역을 뽑아낸 뒤, 전체 영상에서 해당 영역만 추출하여 이어지는 네트워크로 하여금 핵심 영역만 관찰하도록 강제하는 방식으로 하드 어텐션 기반 학습을 구현하였다. 즉, 마치 영상의학 전문의가 진단하는 방식대로 1) 전체 영상을 우선 관찰하고, 2) 하드 어텐션을 통해 추출된 국소적인 집중 영역을 관찰한 후, 3) 최종적으로는 전체 영상 및 집중 영역을 종합적으로 활용하여 최종 판단(분류) 하는 과정을 통해 기존 기술 대비 더 향상된 성능을 도출하였다고 주장한다.

모델은 Fig. 15와 같으며, 전역 모듈(global branch), 지역 모듈(local branch), 그리고 융합 모듈(fusion branch)로 구성된다. 먼저 전역 모듈에서는 전체 의료 영상을 입력으로 받아 어떤 질환에 속하는지 분류를 학습한다. 학습이 완료된 이후, 마지막 합성곱 신경망 층에서 획득된 특징 지도에서 히트맵을 산출하고, 이를 문턱값 처리(thresholding) 하여 분류 과정에서 모델이 집중한 핵심 영역만 남기는 이진 마스크를 생성한다. 지역 모듈에서는 전역 모듈에서 생성된 이진 마스크를 이용하여 전체 영상에서 그에 상응하는 부분을 잘라낸 후 해당 국소 영역만을 입력으로 받아

Fig. 15. Overall architecture of hard attention-guided chest X-ray classification network. Figure inspired by (39).
Conv = convolutional layer, GAP = Global Average Pooling



또 한 번의 분류를 학습하는데, 이렇게 핵심적인 국소 영역을 선정함에 있어 하드 크라핑(hard-cropping)을 활용하기 때문에 이 연구를 하드 어텐션 기반의 연구로 분류할 수 있다. 마지막의 융합 모듈에서는 전역 모듈과 지역 모듈에서 압축한 정보가 내장되어 있는 노드(nodes)를 결합(concatenation) 한 뒤, 이를 기반으로 최종 분류를 학습한다. 즉, 융합 모듈에서는 전체 영상과 국소 영상 모두를 반영한 질환 분류를 수행한다.

$$H_g(x, y) = \max_k (|f_g^k(x, y)|), k \in \{1, \ldots, K\} \quad 8)$$

해당 논문에서 하드 어텐션 기반의 집중 영역 추출 과정(mask inference)은 위 식처럼 전역 모듈의 최종 합성곱 신경망 층의 특징 지도에 절대값을 취한 뒤 채널 축을 따라 각 픽셀 별 최대값을 추출하는 비교적 단순한 방식이다. 이러한 접근은 CAM 기반의 히트맵 추출과는 다소 상이한 접근으로서, CAM 기반의 방법이 특정 클래스의 가중치를 사용해 특정 클래스에 대해 모델이 집중한 영역을 시각화하는 방법인데 반해, 위 방식은 특정 클래스에 대한 가중치를 사용하지 않고 모델이 계산 과정에서 전반적으로 주목한 부분을 표현한다는 점이 차이점이다. 실험 결과, 전체 영상과 국소 영상 모두를 활용하는 융합 모듈의 분류 정확도가 전역 모듈, 지역 모듈 각각의 정확도

보다 높게 측정됐다. 즉, 추출된 집중 영역이 네트워크로 하여금 전체 영상에 존재하는 불필요한 노이즈보다는 분류에 핵심적인 정보가 내포되어 있는 병변 영역에 집중하게 함으로써 성능을 향상시켰음을 알 수 있다. 해당 논문에서는 위와 같은 방법을 통해, 기존의 최고 성능 방법 대비 학습 데이터를 10% 덜 사용하고도 평균 ROC 곡선 하위 영역 수치(area under the ROC; AUC)를 2.9% 향상시켰다. 폐 결절(nodule)과 같이 매우 국소적으로 분포하는 병변의 경우, 전체 영상에 비해 굉장히 작은 영역을 차지하는 핵심 영역을 추출하여 집중적인 분석을 진행하는 것이 커다란 성능 향상으로 이어졌다. 반면, 무기폐(atelectasis)나 심장 비대(cardiomegaly)와 같이 병리학적 영역이 다소 넓은 질환의 경우 지역 모듈을 결합하는 것이 간혹 성능 저하로 이어지기도 했는데, 이는 강제적인 마스킹(masking)으로 집중 영역을 추출하는 과정에서 넓은 영역에 걸쳐 존재하는 질병에 관한 정보가 일부 손실되어 정확한 분류에 오히려 악영향을 미쳤기 때문으로 추측된다.

이와 같이 하드 어텐션 기법을 통해 모델이 어느 영역에 집중했는지 분석 가능할 뿐 아니라, 모델로 하여금 핵심 영역에 더 집중하도록 함으로써 추가적인 성능 향상을 도출할 수 있다. 집중해야 할 영역을 지도(guide) 하기 위한 경계 박스(bounding box)를 굳이 만들지 않아도 네트워크로 하여금 집중 영역을 스스로 생성하게 함으로써 해당 영역에 대한 더 정밀한 분석을 할 수 있다는 점이 하드 어텐션과 같은 학습 가능한 어텐션 기법의 장점으로 꼽을 수 있다. 그러나, 강화 학습을 이용하거나 딥러닝 기반으로 구현하기 위해서는 위 논문처럼 네트워크의 각 모듈별로 끊어서 학습을 진행해야 하기 때문에 구현이 불편하다는 점이 하드 어텐션 기법의 단점이다.

소프트 어텐션(Soft Attention)

소프트 어텐션은 하드 어텐션과 달리 학습 과정이 미분 가능하기 때문에 딥러닝 네트워크에 쉽게 결합할 수 있다. 즉, 유넷과 같은 기존의 딥러닝 모델에 소프트 어텐션을 위한 모듈(이 역시 CNN과 같이 학습 가능한 파라미터로 구성된다)을 결합한 후, 역전파 알고리즘을 이용하여 어텐션 모듈 및 그 외 부분이 함께 종단 간 학습되는 형태로 구현이 가능하다. 하드 어텐션처럼 강화 학습이나 모듈 별 독립된 학습이 필요 없는 이러한 장점으로 인해 현재 하드 어텐션보다 더 활발하게 연구되고 있다.

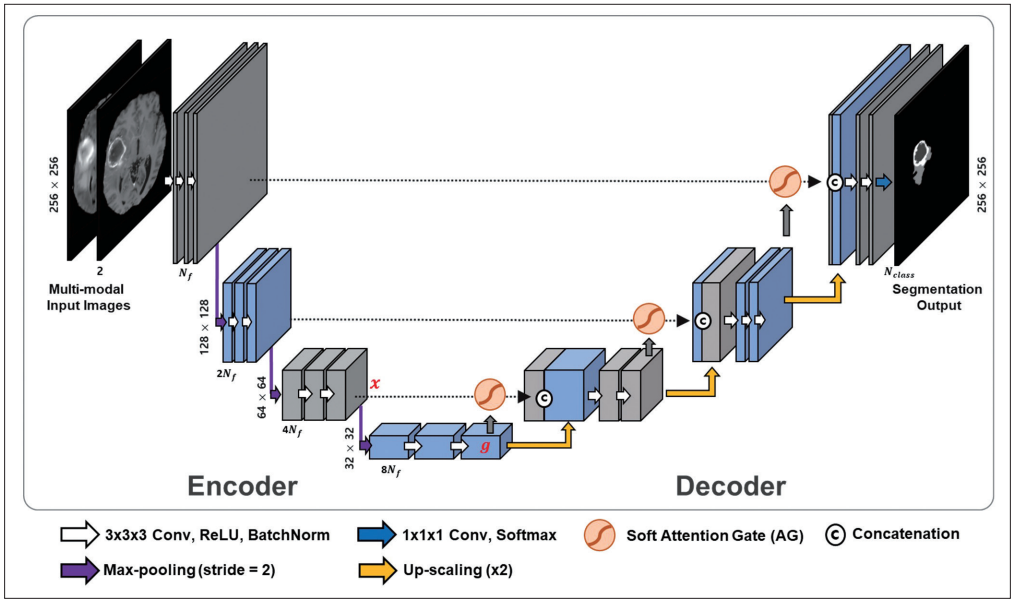
소프트 어텐션을 의료 영상 분석에 적용한 연구는 Medical Image Analysis지에 게재된 Schlemper 등(40)의 “Attention gated networks: learning to leverage salient regions in medical images”가 대표적이다. 해당 논문은 소프트 어텐션을 의료 영상 분석에 활용한 최초의 연구 사례 중 하나로서, 소프트 어텐션 기법을 유넷에 결합하였을 때(이를 어텐션 유넷이라 한다) 3차원 복부 CT 영상 분할에 있어 일반 유넷에 비해 굉장히 적은 파라미터만을 추가로 활용하고도 훨씬 향상된 분할 결과를 도출할 수 있다고 보고하였다.

어텐션 유넷의 구조는 Fig. 16과 같은데, 추가된 어텐션 게이트(attention gate)만 제외하면 나머지 부분의 구조는 일반 3D 유넷과 거의 동일하다. 즉, 일반 유넷을 이해하고 있다면, 어텐션 유넷을 이해하기 위해 알아야 할 사항은 추가된 어텐션 게이트의 작동 원리밖에 없다고 할 수 있다.

어텐션 게이트의 구조는 Fig. 17과 같다. 먼저 x는 인코더 층의 특징 지도(Fig. 16에서 x)를 의미하며, g는 x보다 한 단계 더 인코딩(encoding) 된 한 층 아래의 특징 지도(Fig. 16에서 g)를 의미한

대한영상의학회지

Fig. 16. The architecture of Attention U-Net. Attention gate selects features by using the contextual information which is extracted from coarser scales.
Conv = convolutional layer, Nclass = number of output class, N_f = number of filters, ReLU = rectified linear unit

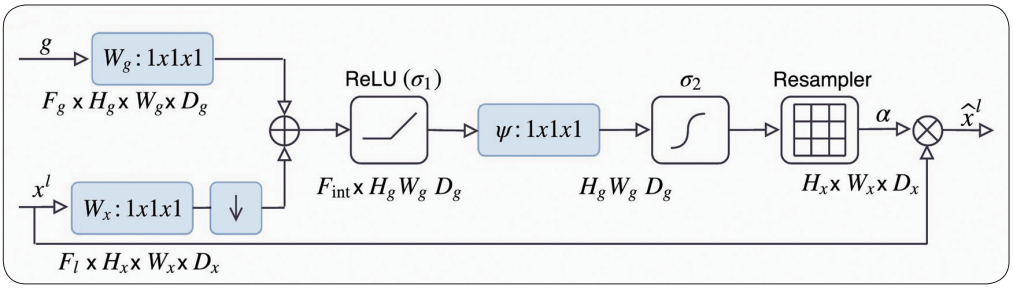


다. Fig. 17을 살펴보면 x와 g를 통해 어텐션 계수(attention coefficient) α를 계산한 후 이를 다시 x와 요소별 곱셈(elementwise multiplication) 함으로써 각 특징별 가중치가 반영된 특징 지도 $\hat{x}$를 창출한다. 즉, 어텐션 게이트의 역할은 인코더의 특징 지도 x를 디코더에 스킵 커넥션 해주기 전에, 핵심 영역을 더 활성화해주는 격자별 가중치(α)를 곱해줌으로써 네트워크로 하여금 그 부분에 더 집중하도록 하는 것이며, α를 계산하는데 필요한 정보를 얻기 위해 영상의 상황 정보(context)가 더 많이 내장되어 있는 한 층 아래의 특징 지도 g와 함께 시그모이드(sigmoid), 정류 선형 유닛(rectified linear unit)과 같은 비선형적인 처리를 진행함을 파악할 수 있다.

어텐션 계수의 효과는 Fig. 18를 통해 확인 가능하다. Fig. 18의 영상들은 원본 CT 영상에 학습 과정에서 생성된 어텐션 계수들을 중첩시켜 놓은 그림으로, 가로축으로 진행할수록 학습 에포크(epoch)의 증가에 따른 어텐션 계수의 변화를 관찰하고 있다. 학습이 진행됨에 따라 목적과 연관이 낮은 영역의 활성도는 줄어들고, 췌장, 신장(kidney), 비장(spleen) 등 중요 영역의 활성도가 높아지는 것을 확인할 수 있다.

어텐션 기법의 도입은 네트워크 학습에 요구되는 연산량 측면에서도 효율적이다. 목적과 연관이 큰 특징에 더 큰 가중치를 부여하며 핵심 영역 위주로 활성화되도록 네트워크 스스로 학습하기 때문에, 위에서 소개했던 하드 어텐션 기법처럼 순환 신경망이나 핵심 영역을 우선 추출하기 위한 선행 네트워크(Fig. 15의 전역 모듈) 등이 불필요해지기 때문이다. 즉, 소프트 어텐션을 통해 하드 어텐션의 역할을 효율적으로 대체할 수 있다. 어텐션 기법의 효과는 성능에서도 나타나는데, Schlemper 등(40)의 연구에 의하면 어텐션 유넷은 파라미터가 그보다 약 1.5배 이상 더 많은 일반 유넷보다도 다이스 계수(Dice coefficient) 및 재현율(recall) 등의 정량적 수치에서 더 높은 객체

Fig. 17. A schematic of the attention gate.
The input features (xl) in the encoder are scaled with attention coefficients (α), and in this process the gating signal (g) from the coarser scale is used. Adapted from Schlemper et al. Med Image Anal 2019;53:197-207 (40) (https://doi.org/10.1016/j.media.2019.01.012), licensed under CC BY 4.0.
ReLU = rectified linear unit



분할 성능을 보였으며, 학습 데이터의 개수가 극단적으로 적은 상황에서도 일반 유넷에 비해 여전히 좋은 성능을 보였다. 이는 딥러닝 모델에 단순히 파라미터가 많은 것보다는 핵심 영역을 파악하는 것의 중요성 내지 효율성을 시사하며, 필터 개수를 늘리는 것보다는 어텐션 게이트 설계에 추가적인 파라미터를 사용하는 것이 더 효과적임을 의미한다(40). 또한, 딥러닝 모델에 파라미터가 많다고 해서 각 파라미터들이 모두 의미 있는 역할을 하는 것이 아니라 불필요하거나 중복되는 역할을 하고 있을 수도 있다는 사실을 드러낸다.

Fig. 19는 본 연구진이 Schlemper 등(40)의 연구와 같이 직접 어텐션 유넷을 구현하여 일반 유넷과의 성능을 비교한 결과로서, 어텐션 기법의 효과를 확연하게 드러내고 있다. 사용한 데이터는 BraTS 2018 데이터(41-43)로서, 고등급 악성 뇌교종(high grade glioma) 데이터 210개를 학습 및 시험 데이터로 사용하여 조영증강 T1 강조영상(post-contrast T1-weighted)으로부터 가돌리늄에 의해 증강되는 종양 영역(gadolinium enhancing tumor)을 분할하는 것을 목표로 하였다. 그 결과, 일반 유넷에 고작 약 1.6%의 파라미터를 추가(유넷: 2348246개, 어텐션 유넷: 2386896개. 파라미터 38650개 증가) 함으로써 다이스 계수에서 약 3.2%의 효과적인 상승(유넷: 0.7867 ± 0.205, 어텐션 유넷: 0.8124 ± 0.111)을 이뤄낼 수 있었다. 또한, Fig. 19에 나타나는 것처럼 정상 영역과 두드러지게 구별되는 이상 영역(종양 영역)에 어텐션 계수들이 집중적으로 형성됨으로써, 상대적으로 덜 증강되어 일반 유넷은 정확히 분할하지 못한 부분(흰색 화살표, Fig. 19G, H, I)을 어텐션 유넷은 더 민감하게(sensitive) 분할해내는 것을 확인할 수 있다.

채널 어텐션(Channel-Wise Attention)

지금까지 소개한 어텐션 기법들이 Fig. 20A와 같이 주로 공간축에서의 집중 영역 선별을 목표로 했다면, Hu 등(44)은 공간축이 아닌 채널 방향으로 어텐션 기법을 적용하였다. 즉, 딥러닝 네트워크에 의해 생성되는 특징 지도 내에는 특징 추출에 사용한 필터 수만큼의 채널이 존재하는데, 더 중요한 정보를 포함하고 있는 채널에 큰 가중치를 부여하고 그렇지 않은 채널에는 작은 가중치를 부여하는 채널 재조정(recalibration) 과정을 통해 모델의 표현력(representational capacity)을 높이는 방법이다.

제안하는 방법은 Fig. 20B와 같이 매우 간단하다. 크게 압축(squeeze)과 자극(excitation) 단계

Fig. 18. Attention coefficients are gradually learned to focus on the discriminative regions as the number of training epochs increase (3, 6, 10, 60, 150). Adapted from Schlemper et al. Med Image Anal 2019;53:197-207 (40) (https://doi.org/10.1016/j.media.2019.01.012), licensed under CC BY 4.0.

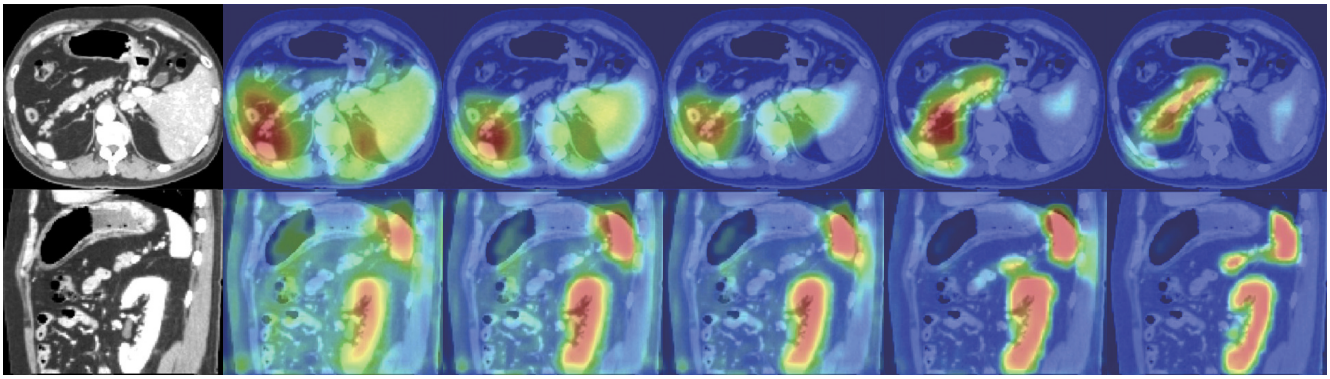
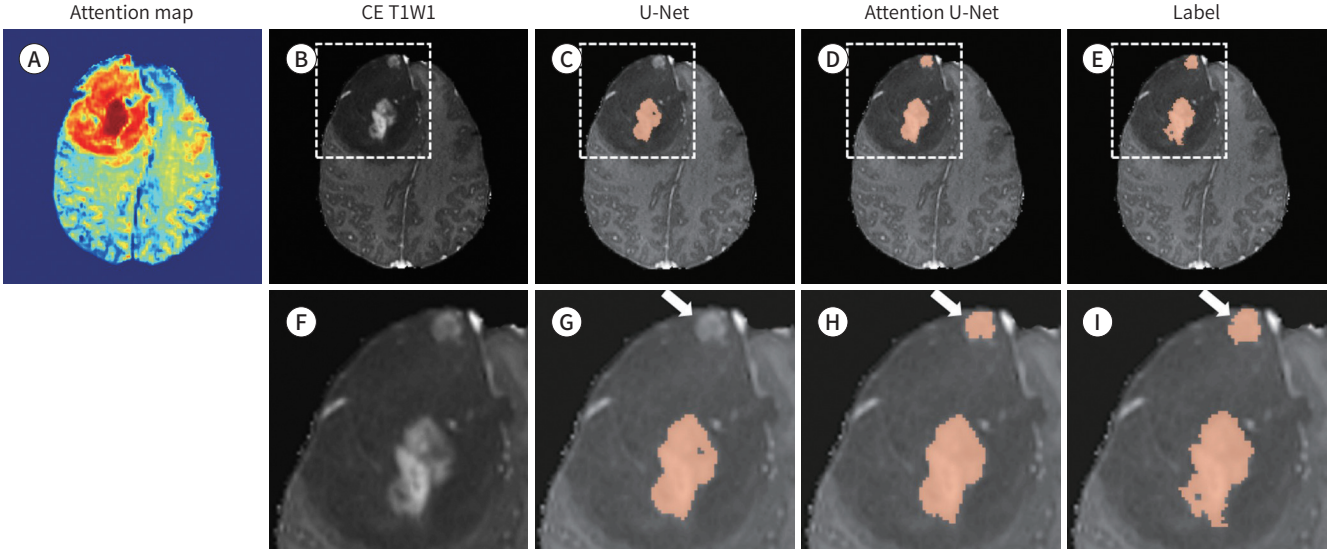


Fig. 19. An example depicting the effectiveness of using an attention mechanism in U-Net to increase the accuracy of the segmentation result. It can be seen that U-Net with attention mechanism quite successfully performed segmentation close to the label, whereas the normal U-Net missed some areas (white arrows in the magnified image). This is attributed to the attention map, which can guide focus to more salient regions.
A. Learned attention coefficients, red areas depicting highly attended regions.
B. Original contrast-enhanced T1-weighted brain image.
C-E. U-Net, attention U-Net, and label segmentation results (orange areas) overlayed on original brain image.
F-I. Enlarged view of the white box area of B-E, respectively. The white arrows indicate regions that were missed on U-Net (G) but successfully segmented on Attention U-Net (H) close to the label segmentation (I).
CE = contrast-enhanced, T1WI = T1-weighted image



가 있는데, 압축 단계에서는 GAP를 통해 각 채널의 중요 전역 정보(global information)를 하나의 값으로 압축한다. 이후 자극 단계에서는 완전 연결층과 같은 비선형적인 신경망을 통해 채널 간 의존성(interdependencies between channels)을 계산하여 채널에 포함된 특징의 중요도에 비례하는 가중치를 생성한다. 이후, 생성된 가중치가 압축되기 전의 특징 지도에 곱해지며 채널별 가중치를 부여하게 된다. 이와 같은 압축 및 자극(squeeze-and-excitation) 구조 또한 소프트 어텐션 기법에 해당하기 때문에 기존의 여러 가지 딥러닝 네트워크에 유연하게 적용이 가능하며,

Schlemper 등(40)의 연구와 같이 파라미터 증가량에 비해서 모델의 성능 향상도가 매우 크다는 장점이 있다. 즉, 모델의 복잡도를 크게 증가시키지 않으면서도 뛰어난 성능 향상 효과를 도출할 수 있다. Rundo 등(45)은 유넷에 압축 및 자극 구조를 결합하여 향상된 성능 및 일반화 능력(generalization ability)을 보고하였으며, Guha Roy 등(46)은 뇌 MRI 및 전신 CT 영상에 공간 어텐션

Fig. 20. Comparison between spatial attention (A) and channel-wise attention (B).

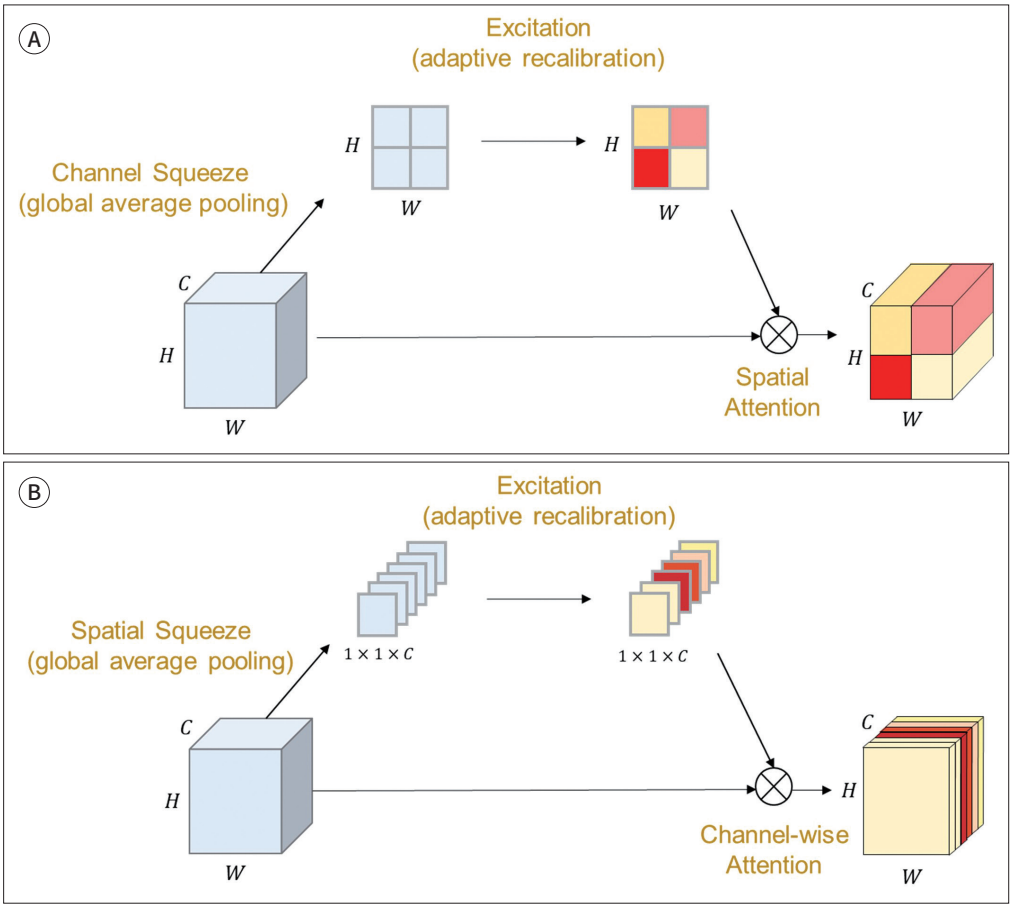


Table 2. List of Studies that Applied Attention Mechanisms to Medical Image Analysis

Attention Mechanism	
Post-hoc attention	
CAM	Bien et al. (32), Yune et al. (52), Rajpurkar et al. (53), Lee et al. (54)
Grad-CAM	Cheng et al. (33), Oh et al. (55), Li et al. (56)
Saliency Map	Pasa et al. (35), Cohen et al. (57)
LRP	Böhle et al. (36)
Trainable attention	
Hard attention	Ypsilantis & Montana (38), Guan et al. (39), Pesce et al. (49), Shaikh et al. (58)
Soft attention	Schlemper et al. (40), Rundo et al. (45), Guha Roy et al. (46), Zhou et al. (47), Li et al. (48), Zhang et al. (50)

CAM = Class Activation Mapping, Grad-CAM = Gradient-weighted CAM, LRP = Layer-wise Relevance Propagation

과 채널 어텐션을 모두 결합하여 향상된 분할 결과를 선보였다.

소프트 어텐션 기법은 사후 분석용 어텐션처럼 모델 해석의 용도로 사용(최종적으로 형성된 어텐션 지도를 관찰하여 모델이 집중한 영역을 파악할 수 있다) 될 수 있을 뿐 아니라, 하드 어텐션과 비교했을 때 학습이 훨씬 용이한 덕에 의료 영상 관련 연구에 활발하게 이용되고 있다. 특히 MRI, CT, X-ray와 같은 다양한 의료 영상에서 뇌, 흉부, 복부, 무릎 등에 존재하는 여러 기관 및 병변 등의 분류 및 분할 정확도를 높이기 위해 주로 연구되고 있으며(46-49), 피부 사진에서의 피부 병변 분류(50), 수술 영상에서의 외과 기계 분할(51) 등의 분야에서도 연구에 적용되고 있다. Table 2는 각 어텐션 기법 별로 의료 영상 분석에의 적용 사례를 정리한 것으로서, 본문에서 자세히 언급한 사례 외에도 다양한 목적을 위해 어텐션 기법이 활발하게 사용되고 있음을 확인할 수 있다(46-58).

결론

의료환경에서 딥러닝은 점차 확대적으로 사용되고 있으며 그 변화 속도 또한 매우 빨라지고 있으나, 아직까지 임상에서 안전하게 사용되기에는 이른 단계이다. 어텐션 기법은 모델의 판단 근거를 시각화함으로써 딥러닝을 임상에 적용하는 데 있어 하나의 안전장치 역할을 할 것으로 기대된다. 또한, 의료서비스 불균형 해소를 위해 의료 취약 지역에 인공지능 기반의 자동 진단 시스템을 도입하는 방법이 제시되고 있는데, 어텐션 기법은 이 과정에서도 의료 교육의 수준이 높지 않은 의료인을 보조하고 진단의 안전성을 높이는데 활용될 수 있을 것이다.

학습 가능한 어텐션의 경우 목적과 연관된 핵심 영역에 더 집중하도록 학습되는데, 이 과정에서 학습 데이터와 분포가 다른 데이터 및 적대적 공격에 대한 모델의 일반화 능력 향상을 기대할 수 있다(24, 59). 이는 특히 촬영 변수가 다양한 MRI나 방사선량을 조절하며 촬영하는 CT 등 촬영 장치나 그 목적에 따라 영상의 화질 및 대조도가 변화하는 의료 영상을 분석함에 있어 딥러닝의 범용성을 높이는 데 핵심적으로 기여할 것으로 생각된다.

더 나아가, 의학적으로 아직 명확히 연구되지 않은 새로운 진단법의 발견을 위해서도 어텐션 기법이 효과적으로 활용될 수 있을 것으로 보인다. Yune 등(52)은 손 방사선 사진(hand radiograph)을 통해 성별을 구별하는 딥러닝 모델을 구축함으로써 영상의학과 전문의 대비 훨씬 높은 정확도(영상의학과 전문의: 58%, 딥러닝 모델: 95.9%)뿐 아니라 구체적으로 영상의 어느 부분을 통해 남녀의 구별이 가능한지를 시각화하였다. Li 등(56)은 흉부 CT 영상에 딥러닝을 적용하여 신종 코로나바이러스(COVID-19) 환자의 영상을 비슷한 영상적 특성을 보유한 지역 획득성 폐렴(community acquired pneumonia) 및 폐렴이 아닌 폐 질환(non-pneumonia lung disease) 환자의 영상으로부터 구별하는 연구를 진행하였으며 히트맵 시각화를 통해 신속한 신종 코로나바이러스 확진에 딥러닝 및 어텐션 기법이 기여할 가능성을 제시하였다.

본 종설에서는 데이터에 기반하여 스스로 특징을 추출하는 딥러닝의 판단 근거를 어텐션 기법을 통해 시각화함으로써, 인간이 의료 영상을 분석하고 진단하는 데 있어 딥러닝 및 어텐션 기법이 효과적으로 활용될 수 있는 측면에 대하여 소개하였다. 이러한 기술을 활용하여 인공지능으로부터 새로운 의학적 지식을 발견하는 시대가 가까운 미래에 펼쳐질 것으로 기대된다.

Author Contributions

Conceptualization, S.H., L.J., H.D.; formal analysis, S.H., L.J.; supervision, H.D.; validation, all authors; visualization, S.H., L.J.; writing—original draft, S.H., L.J.; and writing—review & editing, all authors.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2019R1A2B5B01070488).

This research was results of a study on the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

This work has been supported by Y-BASE R&E Institute a Brain Korea 21, Yonsei University.

REFERENCES

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436-444
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410
3. Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol* 2019;20:728-740
4. Bello GA, Dawes TJW, Duan J, Biffi C, De Marvao A, Howard LSGE, et al. Deep learning cardiac motion analysis for human survival prediction. *Nat Mach Intell* 2019;1:95-104
5. Eo T, Jun Y, Kim T, Jang J, Lee HJ, Hwang D. KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. *Magn Reson Med* 2018;80:2188-2201
6. Jun Y, Eo T, Shin H, Kim T, Lee HJ, Hwang D. Parallel imaging in time-of-flight magnetic resonance angiography using deep multistream convolutional neural networks. *Magn Reson Med* 2019;81:3840-3853
7. Eo T, Shin H, Jun Y, Kim T, Hwang D. Accelerating Cartesian MRI by domain-transform manifold learning in phase-encoding direction. *Med Image Anal* 2020;63:101689
8. Han Y, Ye JC. Framing U-Net via deep convolutional framelets: application to sparse-view CT. *IEEE Trans Med Imaging* 2018;37:1418-1429
9. Kim S, Jang H, Jang J, Lee YH, Hwang D. Deep-learned short tau inversion recovery imaging using multicontrast MR images. *Magn Reson Med* 2020 [in press] doi: https://doi.org/10.1002/mrm.28327
10. Jun Y, Eo T, Kim T, Shin H, Hwang D, Bae SH, et al. Deep-learned 3D black-blood imaging using automatic labelling technique and 3D convolutional neural networks for detecting metastatic brain tumors. *Sci Rep* 2018;8:9450
11. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019;293:38-46
12. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342-1350
13. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793-795
14. Kang D, Park JE, Kim YH, Kim JH, Oh JY, Kim J, et al. Diffusion radiomics as a diagnostic model for atypical manifestation of primary central nervous system lymphoma: development and multicenter external validation. *Neuro Oncol* 2018;20:1251-1261
15. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep* 2017;7:10353
16. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 2019;25:954-961
17. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. *Radiology* 2019;293:607-617

18. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A* 2018;115:11591-11596
19. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-118
20. Rauschecker AM, Rudie JD, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology* 2020;295:626-637
21. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci Rep* 2019;9:6381
22. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA: IEEE; 2017:2097-2106
23. Kim H, Jung DC, Choi BW. Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks. *J Korean Soc Radiol* 2019;80:259-273
24. Jetley S, Lord NA, Lee N, Torr PH. Learn to pay attention. *ArXiv Preprint* 2018;arXiv:1804.02391
25. Wen PY, Macdonald DR, Reardon DA, Cloughesy TF, Sorensen AG, Galanis E, et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J Clin Oncol* 2010;28:1963-1972
26. Theiler R, Stucki G, Schütz R, Hofer H, Seifert B, Tyndall A, et al. Parametric and non-parametric measures in the assessment of knee and hip osteoarthritis: interobserver reliability and correlation with radiology. *Osteoarthritis Cartilage* 1996;4:35-42
27. Mnih V, Heess N, Graves A. Recurrent models of visual attention. Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014); 2014 Dec 8-13; Montreal, Canada: NIPS; 2014:2204-2212
28. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27-30; Las Vegas, NV, USA: IEEE; 2016:2921-2929
29. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22-29; Venice, Italy: IEEE; 2017:618-626
30. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *ArXiv Preprint* 2013;arXiv:1312.6034
31. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* 2015;10:e0130140
32. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med* 2018;15:e1002699
33. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *Eur Radiol* 2019;29:5469-5477
34. Kim B, Seo J, Jeon S, Koo J, Choe J, Jeon T. Why are saliency maps noisy? Cause of and solution to noisy saliency maps. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27-28; Seoul, Korea (South): IEEE; 2019:4149-4157
35. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Sci Rep* 2019;9:6268
36. Böhle M, Eitel F, Weygandt M, Ritter K. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Front Aging Neurosci* 2019;11:194
37. Radiopaedia. Rad_doc, rID 47997, "Childhood Pneumonia". Available at. https://radiopaedia.org/cases/childhood-pneumonia-1?lang=us. Published Sep 13, 2016. Accessed Aug 1, 2020
38. Ypsilantis PP, Montana G. Learning what to look in chest X-rays with a recurrent visual attention model. *ArXiv Preprint* 2017;arXiv:1701.06452
39. Guan Q, Huang Y, Zhong Z, Zheng Z, Zheng L, Yang Y. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognit Lett* 2020;131:38-45
40. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, et al. Attention gated networks: learning to leverage salient regions in medical images. *Med Image Anal* 2019;53:197-207
41. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor im-

age segmentation benchmark (BRATS). *IEEE Trans Med Imaging* 2014;34:1993-2024
42. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data* 2017;4:170117
43. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *ArXiv Preprint* 2018;arXiv:1811.02629
44. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018); 2018 Jun 18-22; Salt Lake City, UT, USA: IEEE; 2018: 7132-7141
45. Rundo L, Han C, Nagano Y, Zhang J, Hataya R, Militello C, et al. USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing* 2019;365:31-43
46. Guha Roy A, Siddiqui S, Pölsterl S, Navab N, Wachinger C. 'Squeeze & excite' guided few-shot segmentation of volumetric images. *Med Image Anal* 2020;59:101587
47. Zhou C, Ding C, Wang X, Lu Z, Tao D. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. IEEE Trans Image Process 2020 [in press] doi: https://doi.org/10.1109/TIP.2020.2973510
48. Li S, Dong M, Du G, Mu X. Attention dense-u-net for automatic breast mass segmentation in digital mammogram. *IEEE Access* 2019;7:59037-59047
49. Pesce E, Joseph Withey S, Ypsilantis PP, Bakewell R, Goh V, Montana G. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Med Image Anal* 2019;53:26-38
50. Zhang J, Xie Y, Xia Y, Shen C. Attention residual learning for skin lesion classification. *IEEE Trans Med Imaging* 2019;38:2092-2103
51. Ni ZL, Bian GB, Xie XL, Hou ZG, Zhou XH, Zhou YJ. RASNet: segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. *Annu Int Conf IEEE Eng Med Biol Soc* 2019: 5735-5738
52. Yune S, Lee H, Kim M, Tajmir SH, Gee MS, Do S. Beyond human perception: sexual dimorphism in hand and wrist radiographs is discernible by a deep learning model. *J Digit Imaging* 2019;32:665-671
53. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018;15:e1002686
54. Lee H, Yune S, Mansouri M, Kim M, Tajmir SH, Guerrier CE, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nat Biomed Eng* 2019;3:173-182
55. Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans Med Imaging* 2020;39:2688-2700
56. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT: evaluation of the diagnostic accuracy. *Radiology* 2020;296:E65-E71
57. Cohen JP, Dao L, Morrison P, Roth K, Bengio Y, Shen B, et al. Predicting covid-19 pneumonia severity on chest x-ray with deep learning. *ArXiv Preprint* 2020;arXiv:2005.11856
58. Shaikh M, Kollerathu VA, Krishnamurthi G. Recurrent attention mechanism networks for enhanced classification of biomedical images. Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019 Apr 8-11; Venice, Italy: IEEE; 2019:1260-1264
59. Wu L, Wang Y, Gao J, Li X. Where-and-when to look: deep siamese attention networks for video-based person re-identification. *IEEE T Multimedia* 2018;21:1412-1424

대한영상의학회지 2020;81(6):1305-1333

대한영상의학회지

어텐션 기법 및 의료 영상에의 적용에 관한 최신 동향

신형섭 · 이정룡 · 어태준 · 전요한 · 김세원 · 황도식*

딥러닝 기술은 빅데이터 및 컴퓨팅 파워를 기반으로 최근 영상의학 분야의 연구에서 괄목할 만한 성과를 이루어 내고 있다. 하지만 성능 향상을 위해 딥러닝 네트워크가 깊어질수록 그 내부의 계산 과정을 해석하기 어려워졌는데, 이는 환자의 생명과 직결되는 의료분야의 의사 결정 과정에서는 매우 심각한 문제이다. 이를 해결하기 위해 "설명 가능한 인공지능 기술"이 연구되고 있으며, 그중 하나로 개발된 것이 바로 어텐션(attention) 기법이다. 본 종설에서는 이미 학습이 완료된 네트워크를 분석하기 위한 Post-hoc attention과, 네트워크 성능의 추가적인 향상을 위한 Trainable attention 두 종류의 기법에 대해 각각의 방법 및 의료 영상 연구에 적용된 사례, 그리고 향후 전망 등에 대해 자세히 다루고자 한다.

연세대학교 공과대학 전기전자공학과

https://doi.org/10.3348/jksr.2020.0150