Multi-scale CAM을 이용한 X-ray 이물질 분류 신경망 성능 향상에 대한 연구

이성주 조남익

서울대학교 전기정보공학부 뉴미디어통신공동연구소

thomas11809@snu.ac.kr nicho@snu.ac.kr

A Study on the Performance Improvement of X-ray Foreign Matter Classification Neural Networks Using Multi-scale CAM

Lee, Sung Ju Cho, Nam Ik

Department of ECE, INMC, Seoul National University

요약

X-ray 영상 검사·검출 문제에 기존 딥러닝 모델을 사용하려는 시도들이 존재해왔고, 합성곱 신경망의 강력한 표현력 덕분에 대체로 준수한 성능이 보장되었다. 그러나 문제의 특성에 따라 기대한 만큼의 분류 및 검출 성능이 나오지 않는 경우가 존재한다. 이는 1) 검출 대상의 스케일이 다양하거나, 2) X-ray 영상은 흑백 영상으로 미세한 특징을 학습하기 어렵거나, 3) 지도학습을 하기에는 학습 데이터의 양이 부족하기 때문인 것이 주요 원인들이다. 본 논문에서는 다양한 스케일의 특징맵을 추출하여 종합적으로 학습하는 신경망을 통해, '생선살 X-ray 영상' 데이터셋에서 '생선 가시' 이물질 class가 모델 내에서 어떻게 학습되는지를 살펴본다. 그리고 X-ray 영상의 경우, 이물질 class를 크기별로 새롭게 labeling하여 성능 개선이 일어날 수 있음을 보인다. 또한 Multi-scale CAM을 통해 class에 따른 활성화 정도를 시각화하여 모델을 직관적으로 분석할 수 있음을 보일 것이다.

1. 서론

영상 분류 및 객체 검출의 성능을 평가하는 데 있어서, 실험실 상황의 정제된 테스트 이미지에서가 아닌 현실 상황에서 여러 상황을 고려하지 않고 취득된 실제 데이터에 대하여 해당 기법이 얼마나 잘 작동하는가가 가장 중요한 문제들 중 하나이다. 따라서 알고리즘의 성능 평가는 주로 현실 도메인의 영상에서 이루어졌고, 딥러닝의 부흥에도 이와 같은 경향은 변하지 않았다. 모델 학습을 위해 ImageNet[1]을 필두로 한 여러 현실 도메인의 데이터셋들이 등장했고, 이들의 공통점은 학습 및 테스트 영상들이 24-bits의 비트심도(Bit Depth)를 가지는 sRGB 컬러 영상이라는 것이었다.

의료 및 산업용 X-ray 검사·검출 문제를 푸는 기계학습 방법들에 대해서도, 현실 도메인 영상을 위해 만들어진 기성 딥러닝 모델을 사용하려는 시도들이 수행되어왔다. 그러사 X-ray 데이터는 주로 8~16-bits의 비트심도를 갖는 1채널 흑백 영상이기 때문에, 현실 컬러 영상과는 차원 및 도메인에서 확연한 차이가 존재한다. 이러한 차이에도 불구하고 합성곱 신경망(Convolutional Neural Networks, CNNs) 고유의 강력한 표현력 덕분에 sRGB로 학습된 해당 모델들도 대체로 준수한 성능을 보여준다. 그러나 다루는 문제의 특성에 따라서 기대한 만큼의 분류 및 검출 성능이 나오지 않는 경우가 종종 발생한다. 이에 대한 원인은 크게 다음과 같이 정리할 수 있다.

1) 검출 대상의 스케일(scale)이 다양한 경우

2) 흑백 영상으로 미세한 특징(features)을 학습해야 하는 경우

3) 지도학습을 하기에 데이터의 양이 적은 경우

먼저 검출 대상의 스케일이 다양한 경우 일관적이지 않은 학습 데이터로 인해 딥러닝 모델은 모호한 기준으로 판단을 하게 되는 오류를 범한다. 이런 문제를 방지하기 위해선 스케일 변화에 따른 별도의 대응책을 모델 내에 구현하거나, 엄청난 양의 지도학습 데이터로 모델을 학습시켜야 한다.

두 번째 원인은 1채널 흑백 X-ray 영상에선 공간적인 특징(spatial feature)밖에 학습할 수 없다는 제한조건과 관련이 있다. 3채널 컬러 영상의 경우 채널 간 조합에 따른 색상 정보도 객체의 특징으로 학습되는데. 흑백 영상의 경우 해당 정보를 학습할 수 없으므로 미세한 특징 조합을 추출하는 것이 제한적이다. X-ray 검사·검출 문제의 특성상 전문가 수준으로(의료, 산업) 미세한 특징을 추출해야 하는 경우가 많기 때문에, 이와 같은 부족한 특징 조합이 모델 성능 저하의 주요 원인이 된다. 이 경우에도 많은 지도학습 데이터를 확보할 수 있다면 어느 정도 문제를 해결할 수 있다.

마지막 원인은 앞서 언급한 두 경우에 대한 근본적인 원인이면서 X-ray 검사·검출 모델이 겪는 고질적인 문제점이기도 하다. 학습을 위한 X-ray 데이터의 수는 현실 도메인 데이터의 수보다 현저하게 적다. 그리고 각 X-ray 영상들은 다루는 문제마다(task specific) 특징적인(distinctive) 도메인을 가지기 때문에 학습에 서로 큰 도움이 되지 못한다. 예를 들어 서로 다른 도메인에 존재하는 의료용 질병 검사 흉부 X-ray와 산업용 이물질 검사 X-ray는 서로의 모델에게 일반화 성능을 부여하지 않는다.

본 논문에서는 X-ray 영상에서 이물질을 검출하는 문제에 대해, 이물질의 다양한 스케일을 고려하여 합성곱 신경망 분류 네트워크의 성능을 향상시키는 방법에 대해서 소개한다. 또한 영상 분류 문제에 주로 사용되는 시각화 도구인 class activation maps[2]을 multi-scale로 이용하여[3], '생선살 X-ray 영상' 데이터셋에서 스케일에 따른 '생선 가시' 이물질의 class를 어떻게 학습하는지 살펴본다. 또한 이를 통해, 지도학습을 위한 X-ray 데이터를 얻기 어려운 경우 효율적인 모델 설계 및 학습 영상 자원 분배 등에 대해 Multi-scale CAM이 직관적인 도움을 줄 수 있음에 대한 실험과 분석을 한다.

2. 관련 연구

2.1. Class activation maps (CAMs)

Class activation maps[2](이하 CAM)은 합성곱 신경망이 데이터를 어떻게 학습했는지를 나타내는 시각적인 도구이다. CAM은 주로 영상 분류 문제에 적용되는데 해당 모델은 다음과 같이 설계된다.

합성곱 신경망이 깊어짐에 따라 입력 영상은 추상화 단계를 거쳐 합성곱 필터에 의해 추출된 특징을 가지는 특징맵(Feature maps)으로 변하게 된다. 일반적으로 분류 네트워크의 마지막 단에는 2D 특징맵(Feature maps)을 1차원 벡터로 펼치기(Flatten) 위한 완전연결 계층(Fully-Connected layer)이 도입된다. 그러나 CAM이 적용된 설계에서는 완전연결 계층 대신 Global Average Pooling(GAP)를 이용하여 Flatten 작업이 이루어지고, 이어서 추가적인 완전연결 계층을 통해 Class를 결정하는 Logits 벡터를 출력하게 된다. 그리고 이때 Class c에 대한 CAM은 다음과 같은 수식으로 표현될 수 있다.

$$M^c(x,y) = \sum_k w_k^c f_k(x,y)$$

w_k^c: 마지막 완전연결 계층에서 Class c에 대한 k번째 입력의 계수

$f_k(x,y)$: (GAP 이전) 마지막 2D 특징맵 중 k번째 채널

마지막 완전연결 계층은 2D 특징맵의 각 채널 $f_k(x,y)$가 Class 결정에 어떻게 영향을 미치는지에 대한 가중치 역할을 하게 된다. 영상 분류 신경망이 학습되어감에 따라, 입력 영상에 대한 각 Class의 결정요인이 활성화맵(Activation maps)으로 나타날 수 있다는 것이 CAM 기법의 주요 장점이다.

2.2. Layer relevance weights

일반적인 영상 분류 모델에서, 신경망의 깊이가 깊어짐에 따라 특징맵의 너비(width)와 높이(height)는 점점 축소된다. 이는 곧 특징맵의 해상도가 낮아지고 입력 영상에 대한 스케일은 커진다는 것을 의미한다. 다양한 스케일의 특징맵이 가진 정보들을 동시에 이용하려면 전체 네트워크의 중간 계층 특징맵들을 추출해야 한다. 그리고 중간 계층 특징맵들을 어떤 방식으로 재조합하는지에 따라 분류 모델의 성능이 결정될 것이다.

흉부 병리(Chest Pathologies)에 대한 분류(Classification) 그리고 위치파악(Localization) 문제를 해결하기 위해, *S Sedai et al* [3]은 중간 계층 특징맵을 고루 이용할 수 있게 Layer relevance weights를 도입했다. 저자는 원하는 스케일의 수만큼 계층마다 중간 특징맵을 각각 추출했고, 이를 통해 얻은 스케일별 Logits들을 Convex Combination으로 재조합한다. 스케일별 Logits의 계수에 해당하는 가중치를 학습 가능한 파라미터로 두면, 예측하는 클래스의 종류에 따라서 계층 간의 가중치가 각각 부여되고 이를 Layer relevance weights라고 한다. 이후 재조합된 최종 Logits 벡터는 활성화 함수를 거쳐 클래스별 예측 확률로 출력된다.

$$p_c = \sigma\left(\sum_{b=1}^{B} h_b^c \times l_b^c\right)$$

p_c : Class c 에 대한 예측 확률

σ : 활성화 함수 (Sigmoid)

l_b^c : 스케일이 b인 Logits 벡터의 c 번째 성분 (Class c 예측)

h_b^c : 스케일이 b인 Logits 벡터의 가중치 (Layer relevance)

Convex Combination: $\forall h_b^c \in [0,1] \;\; s.t. \;\; \sum_{b=1}^{B} h_b^c = 1$

분류 네트워크와 함께 학습된 Layer relevance weights는 클래스별 맞춤형 스케일 가중치를 제공한다. 저자에 따르면 흉부 질병은 종류에 따라 다양한 크기로 나타나는데, 이때 중간 계층 특징맵에 대한 가중치가 개별적으로 학습되면서 분류 성능이 좋아졌다고 한다.

2.3. Multi-scale attention map

S Sedai et al [3]의 연구에서, 추출된 중간 계층 특징맵들은 각 스케일별로 GAP와 완전연결 계층을 통해 Logits 벡터가 된다. 이때 스케일별 CAM과 학습된 Layer relevance weights로 선형 결합을 이루면 다양한 스케일의 CAM 정보를 담은 Multi-scale attention map(이하 Multi-scale CAM)이 생성된다. 스케일별 재조합 가중치가 클래스에 따라 다르기 때문에, 해당 기법은 관심 있는 클래스에 대한 예측 근거가 다음과 같은 스케일 가중치를 반영한 CAM으로 나타난다.

$$S_c = \sum_{b=1}^{B} h_b^c \; R(M_b^c)$$

M_b^c : 스케일이 b인 Class c에 대한 CAM

R : 스케일별 CAM의 너비, 높이를 통일하기 위한 Resize

3. 실험 설계

3.1. 실험 모델 구조

본 논문에서는 영상 분류 모델에 사용한 백본(Backbone) 네트워크로 DenseNet-121[4]을 사용했다. 합성곱 신경망의 경량화를 목표로 개발된 DenseNet-121은 4개의 Dense Block과 사이사이에 존재하는 Transition Layer들, 그리고 마지막 단에서 분류를 위한 GAP와 완전연결 계층으로 이루어졌다.

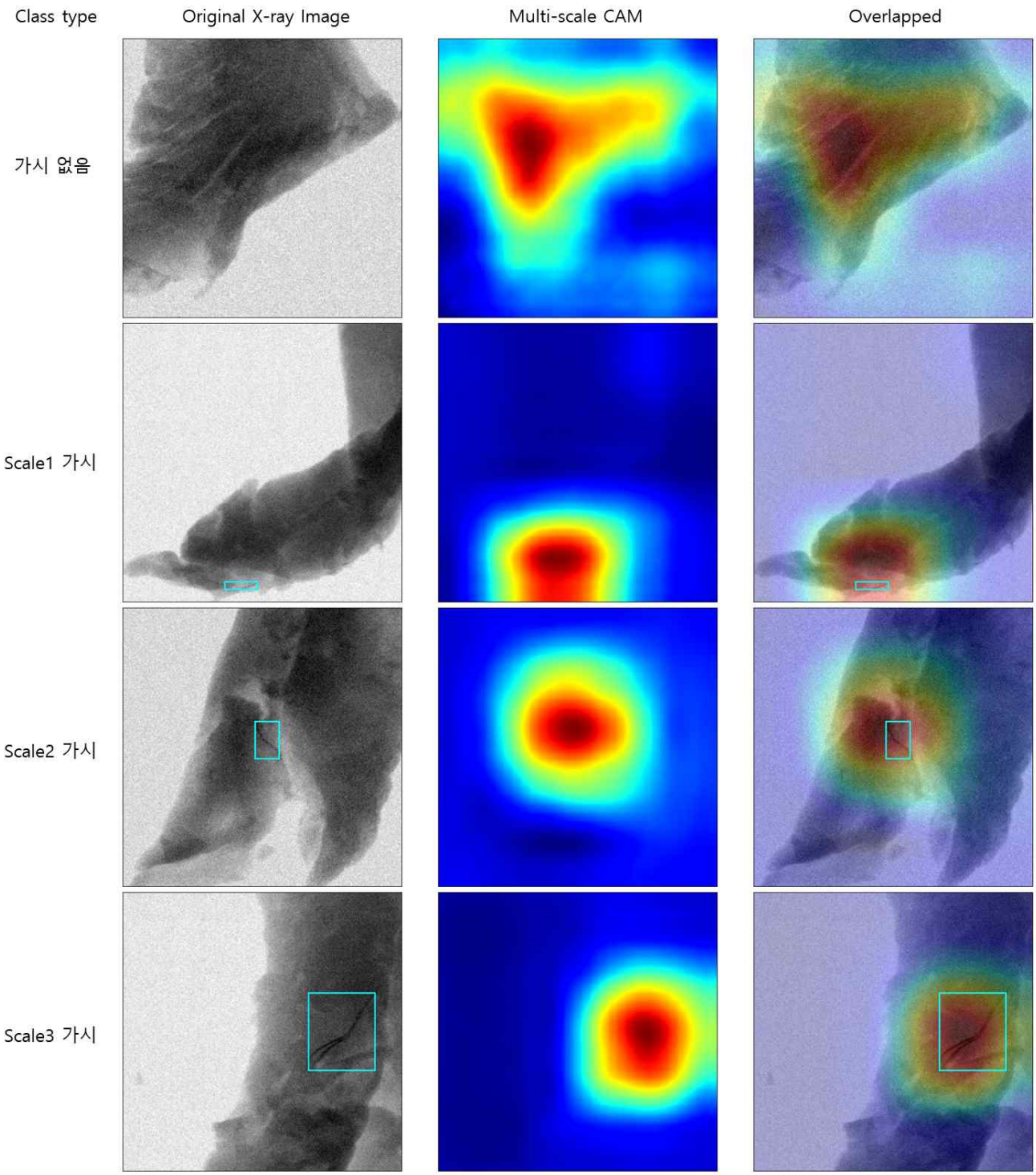

그림 1. Multi-scale CAM 시연 결과: 4개 클래스로 이루어진 '생선살 X-ray 영상'에 대한 Multi-scale CAM 시연 결과이다. 빨강에 가까울수록 해당 영역에 Class를 예측하기 위한 활성화가 많이 일어난 것을 의미하고, 파랑에 가까우면 낮은 활성화 정도를 나타낸다.

Layer relevance weights을 도입하고 궁극적으로 Multi-scale CAM을 이용하기 위해 2,3,4번째 Dense Block의 출력에 해당하는 특징맵 정보를 추출했다. 스케일별로 추출된 특징맵은 각각 GAP와 완전연결 계층을 통해 Logits을 만들고, 이어서 Convex Combination을 통해 최종 Logits 벡터가 생성된다. 학습 시 Layer relevance weights의 조건을 만족시키기 위해 Regularization 수식을 Loss에 추가했다.

$$Loss = CE\,Loss + \lambda \times \sum_{c=1}^{C}(1 - \sum_{b=1}^{B} h_b^c)$$

3.2. 실험 데이터

실험에 사용된 데이터는 '생선 가시' 이물질이 포함된 '생선살 X-ray 영상'으로, 총 371개의 256 × 256 크기의 패치로 구성되어있다. 처음 데이터셋의 Class는 생선 가시 유무에 따라 2개로 나뉘었는데, 이후 가시의 크기를 고려한 분류를 위해 단일 '가시 Class'를 'Scale1,2,3 가시'로 새롭게 labeling을 했다. 실험은 먼저 Class가 2개인 처음의 데이터셋으로 학습 및 평가되었고 이후 Class를 4개로 개정한 데이터셋에 대해서도 동일 과정이 진행되었다. Train:Validation:Test Set의 비율은 7:1:2로 나누었다.

4. 실험 결과 및 분석

4.1. 분류 성능 결과

표 1. 분류 성능 결과

Class 수	2 개	4 개	
Model	Naive DenseNet	Naive DenseNet	Multi-CAM DenseNet
Accuracy	0.790	0.838	0.867

앞서 소개한 X-ray 영상 데이터셋을 학습하여 얻은 생선 가시 분류 성능은 **표 1**을 통해 확인할 수 있다. 우선 검출 대상의 크기에 따라 지도학습 데이터의 Class 수를 늘릴수록 분류 성능이 좋아지는 것을 볼 수 있다. 또한 기성 분류 네트워크를 그대로 사용하는 것보다 Multi-scale CAM을 목적으로 설계한 모델을 사용하는 것이 더 좋은 성능을 낸다는 것도 확인할 수 있다.

4.2. Multi-scale CAM 시연 결과

그림 1은 4개 Class 데이터셋에 대한 Multi-scale CAM을 시연한 결과이다. 가시가 없는 영상에서는 생선살 중앙부에 활성화가 많이 일어났고, 가시가 있는 영상에서는 가시 근처 부분에서 활성화가 많이 일어난 것을 볼 수 있다. 이는 상대적으로 적은 지도학습 데이터 양에도 불구하고, 스케일에 따라 학습된 분류 모델이 바람직하게 Class를 학습하는 것으로 볼 수 있다.

또한 가시의 크기, 모양에 따라서 Multi-scale CAM의 크기와 형태도 다양하게 나타나는 것을 볼 수 있는데, 이를 통해 스케일이 다른 중간 계층 특징맵들이 Convex Combination을 통해 재조합되었음을 암시한다고 할 수 있다.

4.3. Multi-scale CAM이 지닌 가능성 분석

Multi-CAM DenseNet 모델의 Class 예측이 실패하는 경우는 주로 'Scale1 가시'와 '가시 없음' 영상에서 오는 혼동 때문에 발생한다. 예컨대 사람 육안으로도 판단하기 어려울 만큼 작은 가시가 X-ray 영상 내에 있는 경우 모델이 '가시 없음'으로 예측하거나, 그 반대의 경우가 발생할 때 모델의 정확도 성능은 떨어지게 된다.

그런데 예측을 실패한 영상들에 대해서 Multi-scale CAM을 보면, 'Scale1 가시' Class로 판단하는 활성화 영역과 '가시 없음' Class로 판단하는 활성화 영역이 많이 겹치는 것을 볼 수 있다. 본 문제처럼 지도학습 데이터를 쉽게 얻기 어려운 경우, Multi-scale CAM을 이용한다면 모델의 성능 개선을 위한 효율적인 접근이 가능해진다. 예를 들어 앞서 말한 두 가지 Class 영상을 구분하기 위한 학습 영상을 추가적으로 취득할 때, 활성화에 의해서 겹칠만한 영역을 최대한 배제하여 데이터를 얻는다면 모델의 성능은 직관적으로 개선될 가능성이 클 것이다.

5. 결론

본 논문에서는 '생선살 X-ray 영상'에서 '생선 가시' 이물질을 검출하는 문제에 대해, 다양한 스케일을 고려한 합성곱 신경망 분류 네트워크의 성능을 향상시키는 방법에 대해서 알아보았다. 그리고 Class 예측을 위해 모델이 어떻게 학습하는지에 대해, Multi-scale CAM을 이용하여 활성화 영역을 시각화하여 살펴볼 수 있었다. 또한 지도학습을 위한 X-ray 데이터를 얻기 어려운 경우 효율적인 모델 설계 및 학습 영상 자원 분배 등에 대해, Multi-scale CAM이 지닌 가능성에 대해서 알아보았다.

감사의 글

이 논문은 2021년도 산업통상자원부와 한국산업기술진흥원의 "지역혁신클러스터육성사업(R&D, P0002072)"및 2021년도 BK21 플러스 창의정보기술 인재양성사업단에 의하여 지원되었음.

Reference

[1] A Krizhevsky, I Sutskever, GE Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in Neural Information Processing Systems(NIPS), 2012.

[2] B Zhou et al. "Learning deep features for discriminative localization." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[3] S Sedai et al. "Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in X-ray images." In International Workshop on Machine Learning in Medical Imaging (MLMI), 2018.

[4] G Huang et al. "Densely connected convolutional networks" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.