

딥러닝을 이용한 초음파 유방 종괴의 감별: 전향 연구 소개 및 초기 결과 보고

김 원 화 · 김 혜 정

칠곡경북대학교병원 영상의학과

목적: 초음파 유방 결절의 감별진단에서 convolutional neural network (CNN)방식의 적용은 아직 연구가 부족한 상황이다. 본 기관에서 1000개의 유방 결절을 이용한 후향 분석의 초기 결과를 공유하고, 이를 이용하여 대한유방영상의학회와 대한유방검진학회의 지원을 받고 있는 전향적 연구로 여러 기관에서 얻은 초음파 결절 영상을 이용하여 딥러닝을 적용해보고자 한다.

대상 및 방법: 2011년부터 2018년까지 조직검사 및 수술로 확진된 500개의 악성 결절과 2년 이상의 추적검사에서 변화를 보이지 않은 500개의 양성 결절, 모두 1000개의 초음파 결절을 대상으로 하였다. 전처리한 영상은 머신러닝에서 발생할 수 있는 과대적합 (overfitting)을 방지하기 위하여 1000개의 결절 중 800개를 무작위로 선택하여 학습 모델을 만들고 (training and validation), 이 모델을 나머지 200개 결절을 대상으로 하여 5-fold cross-validation test하였다. 학습모델의 진단성적은 area under the curve (AUC), 정확도, 민감도, 특이도로 계산하였다. 본 연구에서 학습된 CNN 모델은 최근 각광을 받고 있는 Global Averaging Pooling (GAP)기법을 사용하여 분류 학습을 하면서 위치 선정을 할 수 있는 정보를 동시에 제공하였다.

결과: 5-fold cross-validation test를 한 결과 set 1, 2, 3, 4, 5에 대한 AUC는 0.926-0.976, 민감도는 81%에서 94%, 특이도는 85%에서 92%, 정확도는 86%에서 93%이며, 모두 80%로 이상의 좋은 성적을 보였다. 또한 GAP기법을 이용한 attention heatmap을 적용하고 확인하였을 때 모델에서 제시하는 시그널 (적색)이 대부분의 악성 및 양성 병변과 잘 부합하는 것을 확인하였다.

결론: 본 연구의 초기결과에서는 분할기법이 없이도 CNN 진단 정확도가 좋은 것을 확인하였으며 이러한 연구결과를 다기관 연구로 얻은 외부 데이터를 통해 다시 학습하고 검증하고자 한다.

Index words: Breast; Ultrasound; Diagnosis; Convolutional neural network

서 론

통신저자: 김혜정
(41404) 대구광역시 북구 호국로 807, 경북대학교 의과대학
칠곡경북대학교병원 영상의학과
Tel. (053) 200-3379, Fax. (053) 200-3349
E-mail: mamrad@knu.ac.kr

유방 초음파 (US)는 유방 종괴 감별에 유용한 검사법으로 잘 알려져 있으며 초음파에서 보이는 종괴의 형태, 경계, 주변에코, 내부에코, 후방에코를 포함한 여러 가지 특징을 주관적으로 분석하여 종괴의 양성 및 악성 여부를 감

별한다 (1). Stavros 등이 1995년도 발표한 750개의 초음파 종괴를 대상으로 한 전향연구에 따라 전형적인 양성 종괴의 기준에는 균질하고 강한 고에코를 보이거나, 평행한 방향과 완전한 얇은 에코성 피막을 보이면서 난원형이거나 네 개 미만의 완전한 분엽형인 경우가 포함된다. 또한 악성의 기준에는 침상형, 각침형, 미세분엽형 변연, 고에코 달무리, 평행하지 않은 방향, 후방음향그림자 등의 소견이 포함된다 (2). 이러한 주관적 평가 시스템을 좀 더 표준화하기 위해 미국영상의학학회에서 제시한 breast imaging reporting and data system (BI-RADS)을 이용하고 있다. BI-RADS에서는 상기 소견을 종합적으로 판단 후 악성이 의심되는 정도에 따라 범주 1에서 5까지 나누어 판정을 한다 (3).

하지만, 초음파는 시행자에 따라 영상의 질과 해석이 달라질 수 있어 영상학과 의사들간의 진단율이 서로 다르다는 제한점이 있다 (4). 영상의 질은 자동 유방 초음파를 개발하고 보급함으로써 객관성 및 표준화의 측면에서 개선될 수 있을 것으로 생각되며 (5), 영상의 해석의 효율 및 표준화를 위해서는 크게 두 가지 방법이 알려져 있다. 첫째는 초음파에서 발견된 유방 종괴의 특징을 정량적으로 분석 및 측정하는 방법이다. 히스토그램 및 텍스처 분석을 포함한 정량적 방식을 적용한 몇몇 연구에서 좋은 성적을 보여주고는 있으나, 정량적 방식을 통한 유방 종괴의 감별은 대부분 수동 분할 (segmentation) 방식을 채택하고 있어 분할에 따르는 시간과 노력이 상당히 소모되며 주관적으로 개입할 수 있다는 점에서 한계가 있다 (6, 7). 둘째는 최근에 각광을 받고 있는 딥러닝 (deep learning, DL) 방식을 이용하여 유방 종괴 감별진단에 도움을 받는 방법이다. 특히 의료영상에서는 여러 가지 딥러닝모델 중 컨볼루션 신경망 (convolution neural network, CNN)을 이용

한 여러 가지 시도들이 이루어 지고 있다 (8, 9). 몇몇 연구자들은 갑상선 결절의 감별진단에서 CNN 방식을 이용한 자동 분류방식이 영상학과 의사들의 분류 성적과 비슷하거나 약간 낮은 성적을 보여준다고 보고하고 있다 (10, 11). 이에, 유방 결절의 감별진단에서 CNN 방식의 적용은 아직 연구가 부족하여 본 기관에서 획득한 1000개의 유방 결절을 이용하여 얻은 후향 분석의 초기 결과를 공유하고, 이를 이용하여 대한유방영상의학회와 대한유방검진학회의 지원을 받고 있는 다기관 연구과제로 여러 기관에서 얻은 초음파 결절 영상을 이용하여 딥러닝을 적용해보고자 한다.

대상 및 방법

본 연구 중 후향적 데이터를 이용한 연구는 임상 윤리심의 위원회에서 승인을 받았으며 다기관 데이터 수집 및 분석을 이용한 연구는 진행 중이다. 2011년부터 2018년까지 조직검사 및 수술로 확진된 500개의 악성 결절과 2년 이상의 추적검사에서 변화를 보이지 않은 500개의 양성 결절, 모두 1000개의 초음파 결절을 대상으로 하였다. 결절은 학습 신뢰성을 확보하기 위해서 초음파에서 직경 0.5 cm이상의 결절로 선별하였다. 초음파 영상은 PHILIPS 장비로 얻었으며 분석에 영향을 줄 수 있는 바디마크 및 주석 및 텍스트들을 포함한 비관심영역은 소프트웨어를 이용하여 자동 크롭핑하여 전처리하였다 (그림 1). 이렇게 전처리한 영상은 머신러닝에서 발생할 수 있는 과대적합 (overfitting)을 방지하기 위하여 1000개의 결절 중 800개를 무작위로 선택하여 학습 모델을 만들고 (training and validation), 이 모델을 나머지 200개 결절을 대상으로 하여 테스트를 하는 과정을 5번 반복하는

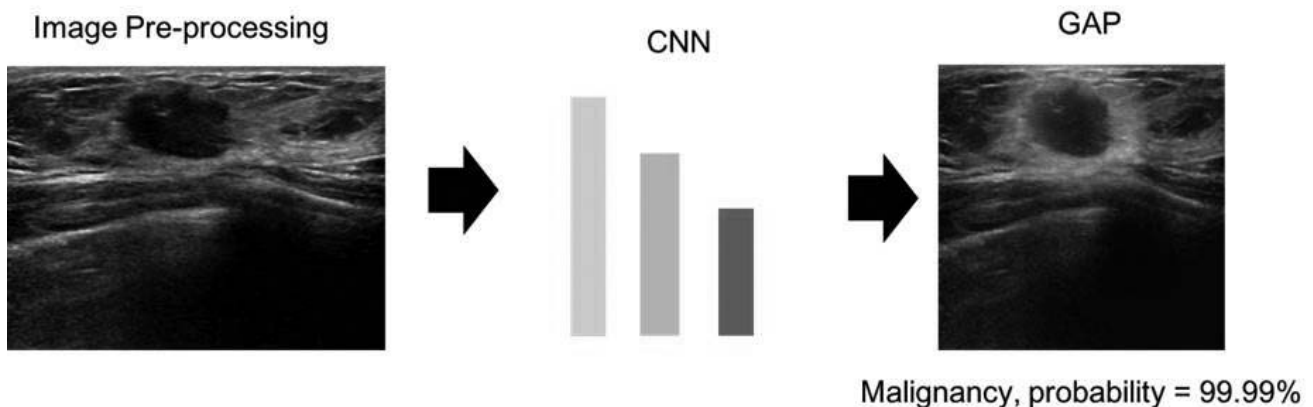


그림 1. CNN 학습의 workflow.

5-fold cross-validation을 사용하였다. 학습모델의 진단 성적은 area under the curve (AUC), 정확도, 민감도, 특이도로 계산하였다.

본 연구에서 학습된 CNN 모델은 최근 각광을 받고 있는 Global Averaging Pooling (GAP)기법을 사용하여 분류 (classification)을 하면서 위치 선정(localization)을 할 수 있는 정보를 동시에 보여주기 때문에 (그림 1) 별도로 영상의학과 의사가 수동 분할을 하지 않았다. 또한, 결절의 수가 충분하므로 미리 학습된 공개 모델을 사용하지 않고 엔드 투 엔드 (end to end) 기반의 CNN 모델을 개발하였다. 앞으로 진행될 다기관 연구에서는 7개 기관이 (칠곡경북대병원, 화순전남대병원, 동아대병원, 천안 단국대병원, 건양대병원, 원광대병원, 부산백병원) 참여하여 최종 2000개 정도의 결절에 대한 후향적 데이터를 획득하여 딥러닝 학습을 최종적으로 적용할 예정이다.

Table 1. 5-fold cross validation에 따른 CNN의 진단 성적 결과

	AUC	Accuracy	Sensitivity	Specificity
Set 1	0.942	86%	81%	91%
Set 2	0.975	93%	94%	92%
Set 3	0.926	86%	87%	85%
Set 4	0.976	91%	90%	91%
Set 5	0.949	89%	90%	87%

결 과

5-fold cross-validation test를 한 결과 set 1, 2, 3, 4, 5에 대한 진단 성적은 Table 1과 같다. AUC는 0.926-0.976, 민감도는 81%에서 94%, 특이도는 85%에서 92%, 정확도는 86%에서 93%이며, 모두 80%로 이상의 좋은 성적을 보였다. 또한 GAP기법을 이용한 attention heatmap을 적용하고 확인하였을 때 모델에서 제시하는 위치 시그널 (적색)이 대부분의 악성 병변과 잘 부합하는 것을 알 수 있었다 (그림 2). 또한, 양성 병변의 대부분도 크기가 비교적 작은 경우에도 CNN-GAP기법이 제시하는 위치가 양성 병변과 잘 부합하는 것을 알 수 있었다 (그림 3).

고 찰

딥러닝은 머신러닝의 일부로 인간의 인지 기능 시스템을 닮은 인공신경망 (artificial neural network, ANN)의 특별한 타입이다 (12). ANN은 1950년에 도입이 되었으나 몇몇 심각한 문제들로 인하여 (딥 알고리즘의 기술적 결함, 컴퓨터 성능 부적합, 빅데이터 부재) 적용에 한계가 있었다. 하지만, 그러한 문제가 해결되면서 (GPU 사용으로 인한 컴퓨터 성능 상승, 새로운 알고리즘 개발) 의료영상에서 인간의 역할과 비슷한 수행능력을 보이게 되었다. 특히 CNN 은 인간의 신경세포와 같이 연결 (시냅스)이 있

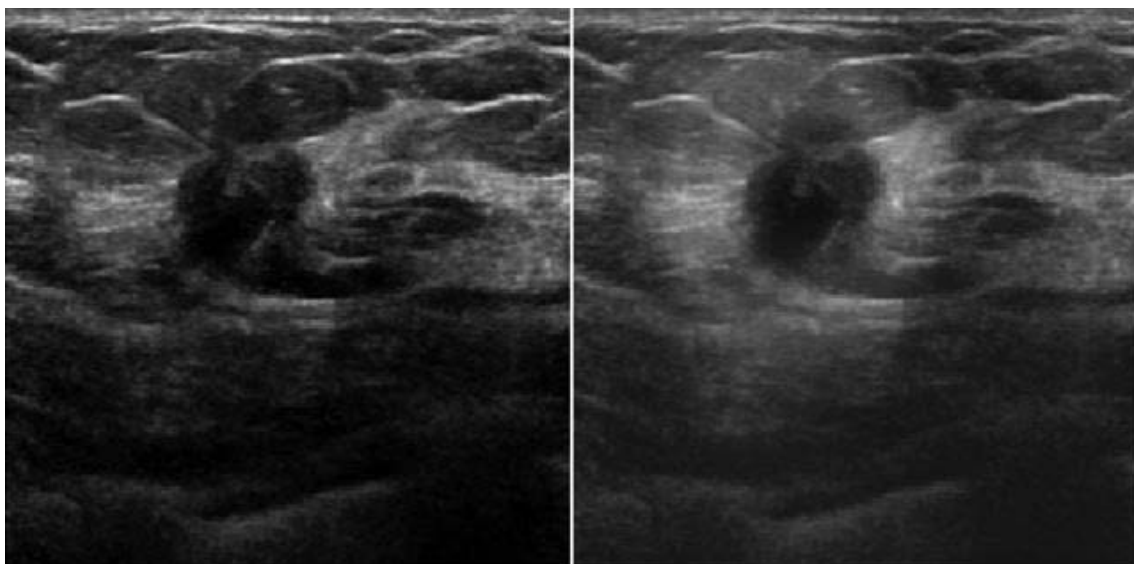


그림 2. 50세 여성의 악성 유방 종괴.

좌측이 pre-processing 영상이며 우측이 CNN을 이용한 activation heatmap을 overlapping한 영상. CNN이 종괴를 잘 인식한 것을 알 수 있으며 CNN 모델에서 악성 예측도는 99.25%였다.

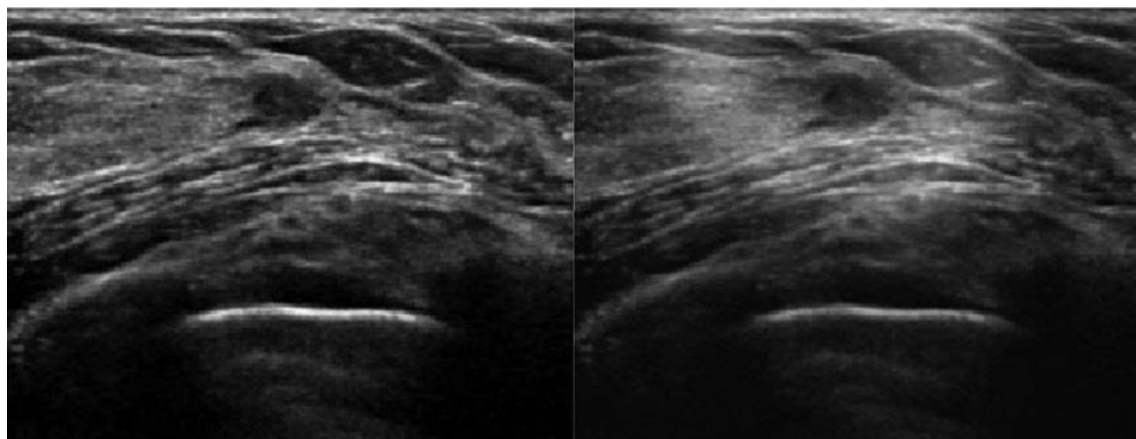


그림 3. 45세 여성의 양성 유방 종괴.

좌측이 pre-processing 영상이며 우측이 CNN을 이용한 activation heatmap을 overlapping한 영상. CNN이 종괴를 잘 인식한 것을 알 수 있으며 CNN 모델에서 양성 예측도는 100%였다.

는 다수의 층으로 구성되어 있고, 학습과정은 동물의 뇌피질의 구성과 흡사하여 가장자리-모양-구성-물체 (edge-shape-component-object)를 파악하여 영상을 성공적으로 분류할 수 있다. CNN의 아키텍처는 컨볼루션 계층 (convolutional layer)과 통합 계층 (pooling layer), 완전하게 연결된 계층 (fully connected layer)로 구성되어 있다. 그 중 컨볼루션은 가장 핵심적인 요소로 특수 필터 연산자로 주어진 영상 픽셀과 주변을 커널 (kernel)이라고 부르는 학습된 파라미터를 곱하는 것을 의미하며 이와 같이 의미있는 커널을 학습함으로써 이미지의 특징을 추출한다.

CNN을 이용하여 초음파 병변을 분류하는 연구는 최근에 증가하였으나 아직 걸음마 단계라고 할 수 있으며 주로 갑상선 결절에 대한 연구가 되어왔다. Ko 등이 2018년도에 발표한 연구에 따르면 589개의 갑상선 결절로 학습한 3개의 CNN 모델을 이용하여 갑상선 암을 진단하였을 때 0.835-0.850의 AUC값을 얻었으며 같이 시행한 2명의 영상의학과 의사는 AUC가 0.805-0.860으로 CNN 모델과 영상의학과 의사의 진단 성적이 통계학적으로 다르지 않았다고 한다 (11). 최근 lancet oncology에 실린 중국에서 시행한 다기관, 후향적 연구에 따르면, 10만개 이상의 영상으로 학습하여 16명의 영상의학과 의사들의 진단성과 비교해 본 결과 CNN모델이 비슷한 민감도와 우월한 특이도를 보였다고 한다 (13). 본 연구의 초기 결과도 진단 성적이 정확도에서 모두 80%가 넘는 기대 이상의 우월한 성적을 보여주었다. 특히 본 연구에서는 대부분 다른 딥러닝-의료영상 연구에서 시행하는 segmentation 혹은 boxing을 하지 않고 annotation만 cropping한 상

태에서 CNN model로 input한 경우도 상기 성적을 보였다는 것은 고무적이다. 그리고 본 연구에서 사용한 CNN 모델은 분류와 위치선정을 동시에 할 수 있는 weakly supervised learning이라고도 부르는 GAP기법으로, 이러한 방법으로 제시한 activation heatmap은 영상의학과 의사가 CNN 모델의 수행능력을 재검증 할 수 있어 인공지능을 통한 preliminary 판독 역할에 대한 기대에서 임상적 유용성이 높다고 할 수 있겠다. GAP기법을 이용한 CNN과 영상의학과 의사와의 커뮤니케이션을 통해 CNN 모델의 특성을 이해하고 모델의 진단 성능을 효율적으로 개선할 수 있을 것으로 기대된다.

인공지능 연구들이 의료영상에 많이 적용되었으나 연구 디자인에 대한 가이드라인은 아직 없는 상태이다. 예를 들어 학습에 필요한 영상의 수, 영상의 질적 수준에 대한 구체적인 방법론은 제시되지 않았다. 하지만 이러한 인공지능 연구를 다기관, 전향적으로 모든 데이터로 학습하고 외부기관의 데이터로 검증 시험 (external validation test) 하여 연구 디자인을 구성하는 것이 중요하다고 지적되고 있다 (14). 이에, 본 연구자들은 GAP 기법을 이용하여 CNN 모델의 특성을 이해하는 한편, 7개의 다기관 연구자들과 협업하여 유방 결절에 대한 딥러닝 적용 가능성을 앞으로 밝혀보고자 한다.

감사표시

본 연구는 2017년도 대한유방영상의학회 및 대한유방검진학회 다기관 연구지원사업의 지원을 받아 이루어진 것임.

참 고 문 헌

1. Harper AP, Kelly-Fry E, Noe JS, Bies JR, Jackson VP. Ultrasound in the evaluation of solid breast masses. *Radiology* 1983;146(3):731-736
2. Stavros AT, Thickman D, Rapp CL, Dennis MA, Parker SH, Sisney GA. Solid breast nodules: use of sonography to distinguish between benign and malignant lesions. *Radiology*. 1995;196(1):123-134.
3. Hong AS, Rosen EL, Soo MS, Baker JA. BI-RADS for sonography: positive and negative predictive values of sonographic features. *AJR Am J Roentgenol* 2005;184(4):1260-1265
4. Berg WA, Blume JD, Cormack JB, Mendelson EB. Operator dependence of physician-performed whole-breast US: lesion detection and characterization. *Radiology* 2006;241(2):355-365
5. van Zelst JCM, Mann RM. Automated Three-dimensional Breast US for Screening: Technique, Artifacts, and Lesion Characterization. *Radiographics* 2018;38(3):663-683
6. Lee SE, Han K, Kwak JY, Lee E, Kim EK. Radiomics of US texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. *Sci Rep* 2018;8(1):13546
7. Vidic I, Egnell L, Jerome NP, et al. Support vector machine for breast cancer classification using diffusion-weighted MRI histogram features: Preliminary study *J Magn Reson Imaging* 2018;47(5):1205-1216
8. Byra M, Galperin M, Ojeda-Fournier H, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. *Med Phys* 2019;46(2):746-755
9. Fujioka T, Kubota K, Mori M, et al. Distinction between benign and malignant breast masses at breast ultrasound using deep learning method with convolutional neural network. *Jpn J Radiol*. 2019.
10. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M. Thyroid Nodule Classification in Ultrasound Images by Fine-Tuning Deep Convolutional Neural Network. *J Digit Imaging* 2017;30(4):477-486
11. Ko SY, Lee JH, Yoon JH, et al. Deep convolutional neural network for the diagnosis of thyroid nodules on ultrasound. *Head Neck* 2019;41(4):885-891
12. Lee JG, Jun S, Cho YW, et al. Deep Learning in Medical Imaging: General Overview. *Korean J Radiol* 2017;18(4):570-584
13. Ha EJ, Baek JH, Na DG. Deep convolutional neural network models for the diagnosis of thyroid cancer. *Lancet Oncol*. 2019;20(3):e130
14. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology*. 2018;286(3):800-809

Differential Diagnosis of Breast Masses at US: Initial Experience with Deep Learning

Won Hwa Kim, MD, Hye Jung Kim, MD

*Department of Radiology, School of Medicine, Kyungpook National University,
Kyungpook National University Chilgok Hospital, Daegu, Korea*

Purpose: To design a deep convolution neural network (CNN) to diagnose breast nodules at ultrasound (US).

Materials and Methods: Between 2011 and 2018, 1000 nodules (500 malignant and 500 benign nodules) were diagnosed by biopsy or surgical excision or over 2-year follow-up. In order to prevent overfitting that may occur in machine learning, preprocessed images were divided by training and test sets; randomly selected 800 out of 1000 nodules for training set and 200 nodules for test set using 5-fold cross-validation. The diagnostic performance of the learning CNN model was calculated with accuracy, sensitivity, and specificity. The CNN model in this study provided information that can be used for localization while classifying using Global Averaging Pooling (GAP) technique.

Results: The 5-fold cross-validation test showed that the sensitivities for sets 1, 2, 3, 4 and 5 ranged from 81% to 94%, with specificity ranging from 85% to 92% and accuracy ranging from 86% to 93%. In addition, when the attention heatmap using GAP technique was applied and confirmed, the signal (red) presented in the model was found to be in good agreement with most malignant and benign lesions.

Conclusion: In the initial results of this study, it was confirmed that the accuracy of CNN diagnosis was good without the segmentation technique. The results of this study will be reexamined and verified through external data obtained from multicenter studies.

Index words: Breast; Ultrasound; Diagnosis; Convolutional neural network

Corresponding author: Hye Jung Kim, M.D.