**An XAI Approach to Deep Learning Models in the Detection of Ductal Carcinoma in Situ**

<https://arxiv.org/pdf/2106.14186.pdf>

1. **AIMS AND OBJECTIVES**

|  |
| --- |
| 이 연구는 **assistive model이 질병을 진단하는 데 도움을 줄 수 있다**는 것을 증명한다. **(인공지능 및 medical field)** 또한 **Deep Taylor LRP methodology**를 통해 모델의 classification이 어떻게 이루어지는지를 이해한다.  이를 위해 다음의 목표를 달성한다.   * 인공지능을 이용하여 질병에 대한 전문가의 일반적인 태도에 대한 정보를 수집 및 분석한다. * **현존하는 CNN 모델의 hidden layer의 기능을 연구**하며, 이때 mammogram이 DCIS의 focus를 포함하고 있는지 예측하는 방법에 중점을 둔다. * 특히 Deep Taylor Decomposition methodology에 대해 XAI method를 적용하여 **이미지가 인공신경망의 서로 다른 hidden layer를 통해 어떻게 process되는지**에 대한 justification을 생성한다. * Breast cancer와 XAI에 대해, 모델의 성능을 다른 유사한 모델과 비교한다.   **[선택한 CNN 모델]**  여기서는 **end2end-all-conv open-source system**을 사용하여 모델을 훈련시키고, benign과 malignant image를 구분하게 한다.   * 이 모델은 **CBIS-DDSM dataset**을 사용한다. * 이미지를 CNN에 입력시킨다. * Dataset을 **60-40 K-fold cross validation** 시킨다.   **[XAI의 사용]**  본 프로젝트에서는 **Deep Taylor Decomposition Layer-wise Relevance Propagation (LRP) method**를 사용한다.   * **Deep Taylor** 기반 접근은 **ReLU neuron을 decompose**하는 데 사용된다. * 이 알고리즘을 이용할 때, source image의 heatmap의 품질은 다음과 같이 측정한다.   + 먼저 **highest pixel**을 perturb한다.   + **Area Under Curve (AUC) score**를 계산한다.   + **Lower AUC**일 경우 heatmap에서 pixel relevance를 더 잘 나타낼 것이다. |

1. **LAYER-WISE RELEVANCE PROPAGATION**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Layer-wise Relevance Propagation은 **입력 이미지를 모든 input으로 backward**시키는 것이다. 이때 redistribution process는 다음과 같이 간단하다.   * 이 수식은 **뉴런의 활성화와 weight의 연결 간의 relevance를 weight**한다. 각각의 역할은 다음과 같다.  |  |  | | --- | --- | |  | 주어진 레이어의 뉴런 j에 대한 activation value | |  | 주어진 레이어의 뉴런 j와 (주어진 레이어 + 1)의 뉴런 k 간의 연결에 대한 weight | |  | 첫 번째 주어진 레이어에서의 각 neuron의 relevance score | |  | (주어진 레이어 + 1)에서의 각 neuron의 relevance score | |

1. **METHODOLOGIES**

|  |
| --- |
| **<CBIS-DDSM Dataset>**  CBIS-DDSM dataset은 Cancer Imaging Archive에 의해 curate되었으며, DDSM Breast Imaging Subset의 subset이다.  이미지 데이터는 각 참여자가 여러 개의 patient ID를 가지므로, DICOM metadata에 6,671명이 있는 것으로 보인다. 그러나, **실제 참여자 수는 1,566명**이다.  Final dataset은 163.6GB이며, **1,566명의 6,775건의 case로부터 얻은 mammogram**을 갖는다.   * 이 dataset은 Kaggle 및 Tensorflow Datasets Catalog 등 유명한 데이터셋 사이트에서 찾을 수 있다.   **<Resnet-50 CNN>**  CNN은 **서로 다른 레이어를 통해 process된 입력이 있을 때 가장 잘 작동**한다.   * 이때 각각의 레이어는 단일한 수학적 함수의 연산을 수행한다. * Max Pooling 레이어는 서로 다른 레이어의 중간에서 사용되며, **translational invariance를 증가**시키고 **feature map size를 감소**시킨다.   Resnet-50 block은 2x2 max pooling layer 대신 **첫 번째 convolutional layer에서 2의 stride**를 가지며, 이는 몇 개의 convolutional layer의 stack 이전에 있는 블록의 시작 지점에서 **feature map size를 줄이기 위한** 것이다. |

**2. METHODOLOGIES (CONT.)**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **<Model>**  이 모델에서는 **bottleneck design**을 이용하며, **3개의 convolutional layer의 반복되는 unit**을 포함한다. 필터 사이즈는 각각 **1x1, 3x3, 1x1**이다.  한편, Resnet block의 중요한 feature는 **각 유닛의 양끝이 직접 연결**되어 있다는 것이다. 이 feature는 directly carry over되므로, 따라서 **각 유닛은 residual information을 학습하는 데 집중**할 수 있다.  **Resnet patch-classifier에서 Batch normalization**은 모든 convolutional layer에서 사용되며, 이것은 수렴 속도를 빠르게 하기 위한 것으로 알려져 있다.   * Resnet-50 block은 **의 패턴**으로 표현할 수 있는데, 여기서 L, M, N, K는 각각 다음을 나타낸다.  |  |  | | --- | --- | | L, M, N | 3개의 convolutional layer의 depth | | K | Unit의 개수 |   **Resnet-50 모델이 이 논문의 연구를 위한 CBIS-DDSM dataset에서의 실험에 사용**되었으며, 50개의 layer는 CNN의 출력 이전에 사용되었다.  Patch classifier로부터 전체 image classifier를 만들기 전에, 먼저 **heatmap을 flatten시키고 이것을 이미지의 classification output (fully connected layers)에 연결**한다.   * 이때 max-pooling layer가 heatmap 이후에 사용되어 모델의 translation invariance를 증가시킨다. * 또한, heatmap과 output 간의 shortcut을 만들어서 학습을 보다 쉽게 한다. * **heatmap에서의 결과는 patch classifier의 출력으로부터 직접 도출**되는데, 이때 다음과 같이 softmax 함수를 사용한다.  |  |  | | --- | --- | | Train / test / valid | **85%** train, **15%** test, validation for **10%** of training set | | Background patches and ROI | 각 이미지에 대해서 **100개의 background patch**가 생성되며, 이때 ROI (region of interest) 간의 중복이 없다.  각 ROI에 대해서, **ROI에 대해 최소 50%가 중복되도록 100개의 patch**가 생성된다. | | Input data | Residual net으로의 입력 : **3x256x256** (red, green, blue의 3가지 채널) | | optimizer | **Keras의 Nadam 알고리즘** | | Batch size | 32 | |

**2. METHODOLOGIES (CONT.)**

|  |
| --- |
| **<Layer-Wise Relevance Propagation>**  **LRP는 CNN의 예측을 설명하기 위한 가장 성공적인 방법**들 중에 하나이며, Resnet-50 patch-classifier에 대해서는 특히 그렇다.   * LRP의 목적은 인공신경망의 inner layer의 예측에 대한 classification explanation을 제공하는 것이다. (궁극적으로 최종 output) * LRP에 의한 설명은 **원본 이미지의 pixel들의 map**으로 표현되며, 이때 **최종 판단에 기여한 pixel이 표시**된다.   LRP의 가장 핵심적인 장점은 **인공신경망의 학습 과정과 상호작용하지 않으며**, 따라서 dataset을 통해 **이미 학습된 신경망에 쉽게 적용이 가능**하다는 것이다.   * 다른 연구에 따르면 LRP는 **환자가 큰 risk 또는 다른 건강 문제가 있는지 verify하는 데 mammogram을 사용**할 수 있게 한다.   LRP는 **network weight과 neural activation을 network의 반대쪽을 따라 backpropagate하여 최종적으로 input layer에 도달**하게 하는 것이다.   * 따라서 **CNN에서 어떤 픽셀이 classification에 기여했는지 시각화**할 수 있으며, 최종적으로 final diagnosis에 어떤 픽셀이 기여했는지 알 수 있게 한다. * 따라서 의료 전문가가 propagation에 대해 학습하면 그들은 모델이 어떻게 작동하는지 이해할 수 있다.   LRP에서는 각 픽셀의 contribution magnitude 또는 중간 뉴런의 relevance value를 **R로 표시**한다. |

1. **RESULTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LRP는 **VGG-16 architecture를 back-propagate하는 데 사용되지 않으므로** Resnet-50을 사용하는 것으로 한다.  각 Resnet-50 block의 AUC는 다음과 같다. (2개의 Resnet-50 block 사용)   |  |  | | --- | --- | | [512 – 512 – 1024] x 2 | AUC = 0.86 | | [256 – 256 – 256] x 1  [128 – 128 – 128] x 1 | AUC를 크게 감소시키지 않음 |  * 이 결과는 **Resnet-50 block의 depth는 전체 image classifier의 성능과 상대적으로 큰 관련이 없다**는 것을 보여준다. |

**3. RESULTS (CONT.)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Deep Taylor method는 입력 feature들의 기여를 **Taylor decomposition을 이용한 Taylor expansion의 first order로 identify**한다.   * 따라서 각각의 뉴런의 속성을 estimate할 수 있다.   각 setting에서의 saliency map의 역할은 다음과 같다.   |  |  | | --- | --- | | Classification setting | 각 픽셀이 **class prediction에 얼마나 기여**하는지 estimate한다. | | Regression setting | 각 픽셀이 **모델에 얼마나 많이 영향**을 주고 **예측 error (loss function)를 감소시키는 데 기여**하는지를 estimate한다. | |

**3. RESULTS (CONT.)**

|  |
| --- |
|  |
| **iNNvestigate library LRP**를 이용하여 classification에 기여한 area를 표시한다.  (class: normal / benign / malignant breast) |
|  |
| 위 그림은 benign breast cancer이며 추후 관찰이 필요하다고 분류한 경우이다.   * **Heatmap visualization은 이것을 benign breast cancer로 옳게 분류**하였다. |
|  |
| 위 그림에서는 heatmap visualization이 분류 가장 크게 기여한 픽셀을 표시한 것이다.   * **Purple area가 없으므로, normal non-cancerous breast mammogram으로 옳게 분류**한 것이다. |