Distance Method를 바탕으로 한 eXplainable Artificial Intelligence(XAI) 기술의 정량 비교 동향

ㅁ **정재훈** / 한국항공대학교

요 약

본 기고문에서는 LIME 및 SHAP과 같은 Pixel-based XAI(eXplainable AI) 기법에 대해 검토하고, 이러한 방법을 정량적으로 평가하는 방법에 대해 서술한다. 현재 많은 XAI 기술들이 제시되었지만, 이를 정량적으로 평가하는 방법은 매우 어려운 문제이다. 이에 XAI 알고리즘 중 어떤 것이 가장 효과적이면서도 정확한지 파악하기 위하여 제시된 기술들을 소개한다. 본 기고문에서는 도장 기반의 distance method를 사용하여 이미지 데이터인 MNIST 데이터셋에서 도장(stamp)의 존재 여부를 예측하는 딥러닝 모델에 XAI 알고리즘으로 적용하여 생성된 설명을 통계 적기법으로 평가한다. 특히, LIME과 SHAP 알고리즘을 distance method로 평가하여, 각 알고리즘이 제공하는 설명의 성능을 비교했다. 결론적으로, Felzenszwalb 방법을 사용한 LIME 알고리즘이 다른 LIME과 SHAP 알고리즘보다 더 효과적인 설명을 제공하는 것으로 보인다.

I. 서 론

컴퓨터 하드웨어의 발달로 이전에는 사용하기 어려웠던 기술들이 현대에 와서 사용 가능해지기 시작했다. 특히, 인공신경망의 발전은 이미지 인식, 자연 언어 처리, 음성 인식 분야에 있어서 굉장히 눈부시다. 한편 인공신경망의 적용 분야가 넓어지고 성능이 좋아짐에 따라, 인공신경망은 점점 복잡해지고 사람들이 모델을 해석하는 것이 어려워졌다[1]. 따라서 성능이 좋은 복잡한 인공신경망을 해석하기 위해 XAI(eXplainable AI) 알고리즘들이 나타났다[2]. XAI란 해석하기 어려운 인공신경망을 사람에게 설명하는 방법이다[3]. 예를 들면, MNIST 데이터를 분류하는 모델이 주어졌을 때, 이 모델이 이미지의 어떤 부분을 보고 판단을 내리는지 알 수 있다면 이 모델은 해석이 가능하다고 할 수 있을 것이다. 만약 XAI 알고리즘이 표시한 이미지의 특징이 사람이 생각하는 특징과 다르다면 원 모델은 사람이 의도한 것과는 다른 방향으로 학습됨을 나타내고 이는 좋지 못한 결과를 초래할 수 있다. 최악의 경우 학습된 모델이 원하는 물체를 인식하는 것이 아니라 이미지의 배경에 공통적으로 나타난 특징을 찾을 수도 있다. 이러

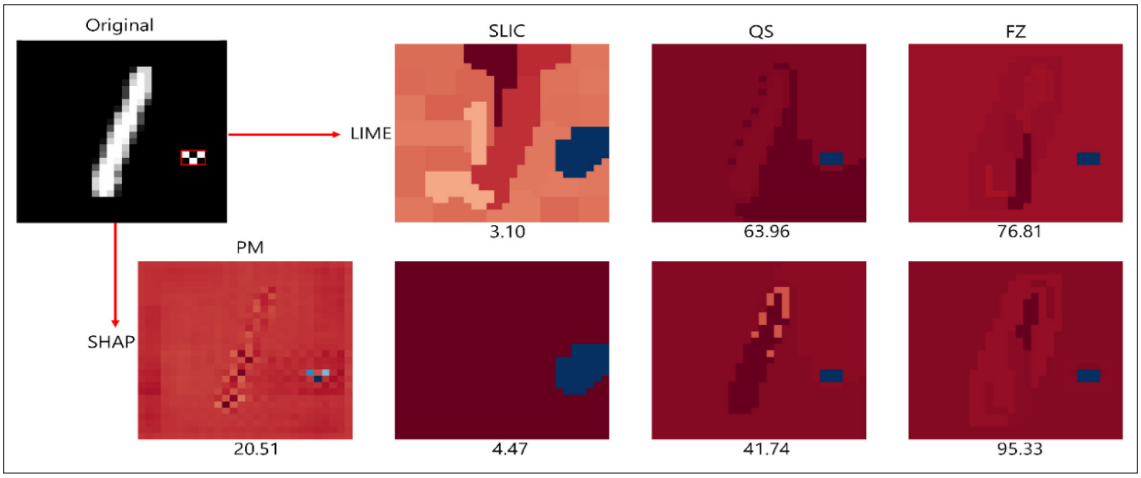
한 경우 XAI 알고리즘을 통해 모델에 입력된 이미지 중 중요 부위를 표시하여 모델이 적절히 학습되었는지 알 수 있게 해준다. 그런데 XAI 알고리즘은 여러 가지가 존재하며 각 알고리즘은 각자 다른 근거로 학습된 모델의 예상값에 대한 해석을 시도한다. 다시 말해서 서로 다른 XAI 알고리즘은 서로 다른 설명을 제시한다. 이렇게 제시된 여러 설명 중 어느 것이 더 나은 설명인지에 따라 어떤 XAI 알고리즘을 사용하는 것이 올바른지 알 수 있다. 따라서 어떤 알고리즘이 모델에 대한 가장 좋은 설명을 제시하는지 아는 것은 중요한 문제이다. 그러나 인간이 어떻게 이미지를 해석하는지는 아직 알려지지 않았으며 이는 곧 어떤 XAI 설명이 더 나은지 수치적으로 해석하기 매우 어렵다는 뜻이다.

이에 XAI 알고리즘을 평가하기 위해 정확성, 일관성, 민감성, 안정성 등의 기준이 제시됐다[4,5,6,7]. 대표적으로 안정성을 비교한 논문[7]에서는 안정성을 평가하기 위한 방법을 정의하고 Mean Decrease Accuracy(MDA)[8], LIME[9]과 SHAP[10]을 비교했다. 그러나 이미지 데이터에서 사람마다 분류 기준이 다르기 때문에, 이미지 데이터에 대해서 LIME과 SHAP을 정밀하게 비교한 연구는 찾기 어렵다. 최근에는 피부 이미지 데이터를 이용하여 LIME과 SHAP을 비교한 논문[11]이 있었지만, 정량적으로는 충분한 비교를 수행하지 못했다.

본 기고문에서는 이미지 데이터에서 LIME과 SHAP을 비교하기 위해 중점을 "사람이 명확하게 인지하는 무늬를 얼마나 정확히 찾는가"와 "다른 부분과 달리 무늬에만 얼마나 많은 중요도를 주는가"에 두고 정량적으로 LIME과 SHAP을 비교한다. 위 특징을 만족하는 비교 방법은 도장 기반의 distance method[12]이다. 이 방법은 우선 사람이 명확하게 인지할 수 있는 무늬인 도장을 만들고, 이 도장을 이미지에 찍은 후 XAI 알고리즘을 적용하여 중요도를 얻는다. 얻은 중요도를 정규화와 표준화를 이용한 Distance score 알고리즘을 이용하여 점수를 얻고 비교하는 방법이다. 〈그림 1〉에 과정을 간단히 소개하고 Distance method의 자세한 설명은 Ⅲ. 비교 방법에서 소개한다. 본 논문에서는 MNIST 데이터에 distance method를 이용해 LIME, kernel SHAP, permutation SHAP 알고리즘들을 비교한다.

II. LIME과 SHAP 개요

XAI 알고리즘들(LIME과 SHAP)에 대해 간단히 소개하고 또한 이 알고리즘들에 적용 가능한 이미지 분



<그림 1> 손 글씨 데이터 MNIST에 대해 모델을 학습시킨 후, XAI 알고리즘들(LIME, SHAP)을 데이터에 적용하여 픽셀 중요도를 얻는다. 그 후 distance method를 사용하여 각 설명에 대한 점수를 얻고 비교한다. 파란색은 높은 기여도를 의미하고 빨간색은 낮은 기여도를 의미한다.

할(segmentation) 방법도 간단히 서술한다. LIME은 Additive feature attribution methods[10]를 사용한 알고리즘 중 하나로 전체 데이터셋이 아닌 특정 데이터에 대한 예측에 대해 설명하는 알고리즘이다. Additive feature attribution method란 해석 가능한 설명 모델이 원 데이터 x가 주어지고 단순화시킨 데이터 x'이 주어졌을 때 $x=h(x')$가 성립하는 함수를 이용하는 방법이다. 해석 가능한 설명 모델 g는 원 모델 f가 주어졌을 때 $g(x')\approx f(h(x'))$로 표현 가능하다. 이때, LIME은 원 데이터에 x에 변형(perturbation)을 준 샘플인 z 데이터를 이용한다. 식으로 표현하면 $g(z')\approx f(h(z'))$로 표현된다. LIME은 해석 가능한 모델 중에서 g를 찾고 모델 복잡성($\Omega(g)$)과 거리에 대한 가중치(π_x)를 고려하면서 특정 데이터에 대해 해석 가능한 모델 g가 $f(h(z'))$에 가까워지도록 학습시킨 후 이 모델을 통해 원 데이터 x에 대한 예측을 설명한다.

이미지에 사용되는 LIME은 픽셀별로 존재하는 원 이미지를 분할 방법을 이용해 슈퍼 픽셀(super-pixels)별로 존재하게 만든다. 다음으로 슈퍼 픽셀의 존재 여부에 따라 부분집합을 만들어 변형 데이터를 생성한다. 따라서 분할 방법에 따라 LIME이 제공하는 해석이 달라질 수 있다. LIME에서 제공하는 분할 방법으로는 medoid shift 알고리즘에서 커널 함수의 크기가 작아질 경우 발생하는 local mode에 빠지는 현상을 해결하고 계산을 단순화하여 속도를 높인 quick shift[13] 알고리즘을 이용한 LIME(QS), K-means 클러스터링을 슈퍼 픽셀을 생성하는데 적용하는 방법인 SLIC(Simple Linear Iterative Clustering)[14] 방법을 적용한 LIME(SLIC)과 그래프 이론에 기반하여 그리드와 그리드 사이의 모서리로 그래프를 만들고, 모서리에 가중치를 부여한 후 가중치를 주위와 비교하여 비슷한 크기의 가중치를 가지는 부분으로 영역을 분할하는 Felzenszwalb[15] 방법을 적용한 LIME(FZ) 등이 있다.

SHAP(Shapley Additive exPlanation)은 게임 이론을 기반으로 한 shapely value[16]를 통해 모델을 설명하는 방법이다. LIME과 마찬가지로 Additive feature attribution methods를 이용한다. 실제로 shapely value를 구하는 것은 어렵기 때문에 조건부 기대값을 이용해 근사하는 SHAP value로 shapely value를 대체한다. 또 LIME을 SHAP value를 통합하여 shapely value를 근사하는 kernel SHAP이 존재한다. 본 논문에서는 딥러닝 모델을 이용하기 때문에 사용 가능한 SHAP 알고리즘이 제한된다. 여기서 비교하는 SHAP 알고리즘은 Permutation SHAP(SHAP(PM)), Kernel SHAP으로 아래에 간략히 설명한다.

SHAP(PM)은 모든 순열에 대해 shapely value를 구하는 것은 어렵기 때문에 반복할 개수만큼 임의의 순열들을 고르고 이 순열들에 대해 전 방향, 후 방향 과정에 대해 기여도를 구하여 shapely value를 추정하는 방법이다. 이 방법은 반복할 횟수가 증가할수록 더 정확한 shapely value를 얻을 수 있지만 시간이 오래 걸린다는 단점이 있다.

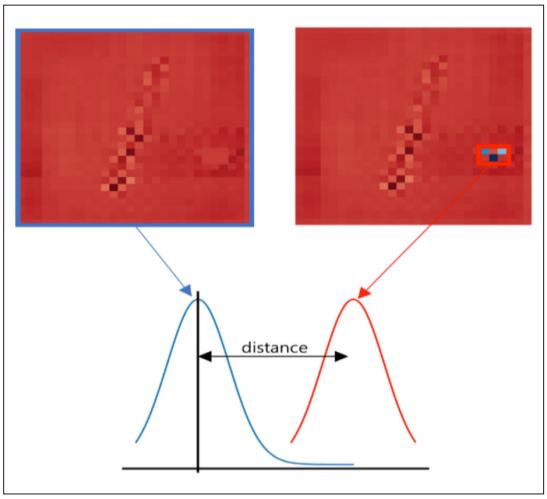
Kernel SHAP은 LIME과 SHAP value를 연계하여 더 빠른 시간 내에 shapely value를 근사하는 방법이다. Kernel SHAP 또한 이미지에 대해 분할 방법을 이용하여 슈퍼 픽셀을 만들어 기여도를 제공한다. 따라서 LIME과 동등한 조건에서 비교하기 위해 quick shift 분할을 사용한 SHAP(QS), SLIC 분할을 사용한 SHAP(SLIC), Felzenswalb 분할을 이용한 SHAP(FZ)을 이용하여 비교한다.

<표 1> LIME과 SHAP의 distance score 결과

Segmentation	Scores	
	Median	Mean
SHAP(PM)	10.50	11.37±4.75
SHAP(FZ)	8.28	14.16±21.74
SHAP(QS)	1.80	7.01±17.43
SHAP(SLIC)	3.42	3.44±1.10
LIME(FZ)	15.41	28.18±27.89
LIME(QS)	4.52	14.37±24.81
LIME(SLIC)	2.37	2.49±0.85

Ⅲ. 비교 방법

$I(\in R^{n\times m})$를 XAI가 만든 설명이라 할 때, I는 각 픽셀의 중요도를 의미한다. Distance method는 도장 영역(O)



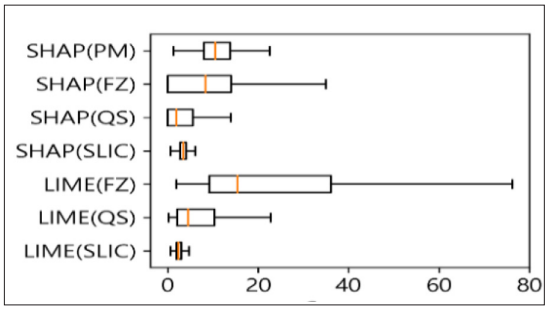
<그림 2> Distance method 개요

과 나머지 영역의 픽셀 중요도(R)를 두 개의 집합으로 분리한다. 〈그림 2〉에서, 이 과정을 파란 이미지(R, 배경)와 빨간 이미지(O, 도장)로 분리하여 나타낸다. 이를 바탕으로 distance score를 z-score를 이용해 구한다.

$$z_i^O = K \cdot \frac{O_i - \mu}{\sigma} \quad (1)$$

$$d = \frac{\sum_i z_i^O}{|O|} \quad (2)$$

여기서, $\mu = \frac{\sum_i R_i}{|R|}$, $\sigma^2 = \frac{\sum_i (R_i - \mu)^2}{|R|}$로 정의하고 이것은 나머지 영역(R)의 평균과 표준편차이다. 이 둘을 이용하여 도장 영역(O)의 픽셀 중요도가 평균으로부터 떨어진 정도를 구하고 이들을 평균한 것이 distance score이다. 이때, distance method를 소개한 논문에서 K는 도장 영역 픽셀이 흰색: +1 검은색: -1이다. 그 이유는 이 논문이 비교한 XAI 알고리즘이 양수 값만 가지는 것과 음수 값도 가지는 것이 존재했기 때문이다. 한편, 본 논문에서 비교하는 알고리즘은 음수, 양수를 모두 나타내기 때문에 K가 존재할 필요가 없다. 따라서, K를 제거한 distance method를 이용하여 비교한다. 한편, distance score의 값이 양수이면 분류를 하는 데 긍정적인 영향을 주었다는 의미이고, 음수



<그림 3> 3 LIME과 SHAP의 distance scores

이면 부정적인 영향을 주었다는 뜻이다. 또한 score의 범위는 [-100,100]으로 제한한다.

Ⅳ. 실험 및 결과

1. 데이터 및 실험 조건

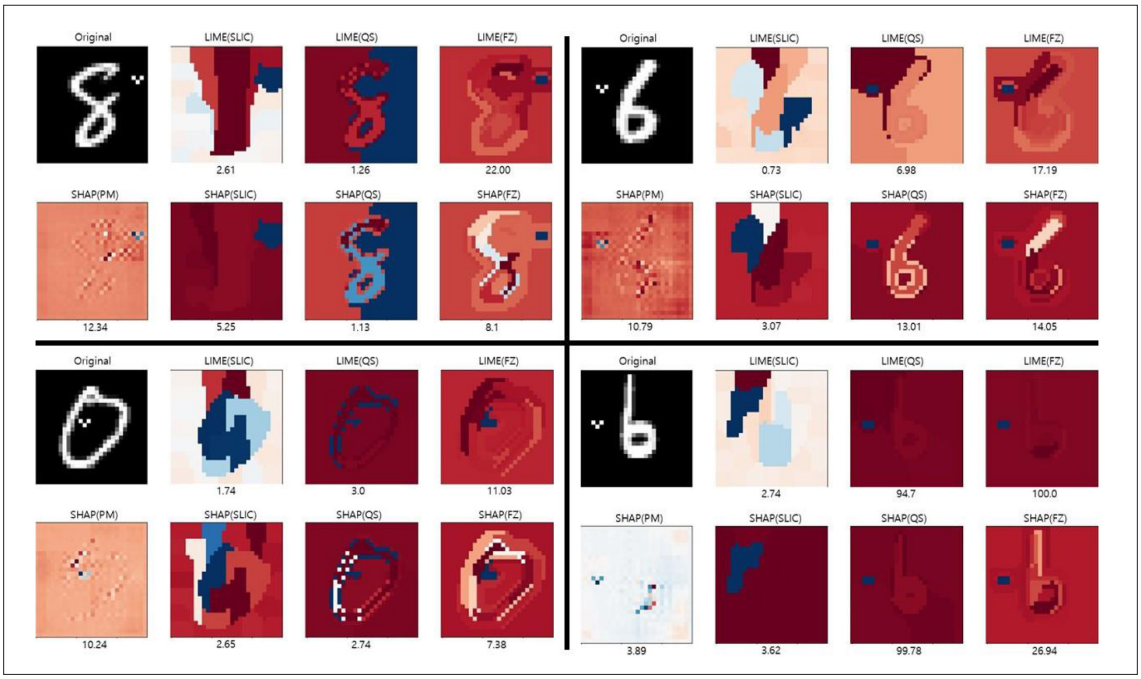
실험에 사용한 데이터는 MNIST 데이터에 학습, 테스트 데이터셋의 절반에 도장을 찍고 분할에 편리하도록 복사하여 3차원 데이터로 만들고, 이 데이터로 초기 LeNet[17]을 학습시켜 학습 정확도는 99.94%, 테스트 정확도는 99.91%를 얻었다. 분할에 사용한 초 매개변수는 〈표 2〉로 제시하고 결과는 〈표 1〉과 〈그림 3〉으로 제시한다.

<표 2> 분할에 사용한 초 매개변수

SLIC		Quick Shift		Felzenswalb	
N_seg	50	Kernel_size	4	Scale	10
Compact	1	Max_dist	5	Sigma	0.8
sigam	1	Ratio	0.6	Min_size	5
		sigma	0.8		

2. 정성적 비교

〈그림 4〉에 4개의 실험 결과 이미지와 distance scores를 나타냈다. Distance method는 도장과 나머지 영역의 중요도를 분리하여 distance score를 구하는 방법이기 때문에 파란색 영역이 도장과 완전히 일치하고 나머지 영역이



<그림 4> 4개의 이미지에 대한 XAI 알고리즘들의 distance scores

빨간색이면 좋은 점수를 가진다. 분할 알고리즘별로 살펴보면, SLIC은 도장을 포함한 더 넓은 영역을 분할하는 경향이 강하고, QS는 도장을 잘 분할할 때도 있지만 종종 적절하지 못한 분할을 하고, FZ는 가장 도장을 잘 분할하는 경향을 보인다. 따라서 FZ 분할이 나머지 분할보다 더 좋은 distance score를 가지게 된다. 한편, XAI 알고리즘으로 비교하면, LIME이 FZ 분할에서 SHAP보다 도장에만 집중하므로 더 높은 점수를 가지는 경향이 있고 SLIC 분할에서는 반대로 SHAP이 도장에 더 집중하는 경향을 보인다.

3. 정량적 비교

〈그림 3〉과 〈표 1〉에 LIME과 SHAP의 결과를 나타냈다. 〈그림 3〉은 극단치(outlier)를 제거하고 중앙값(median)을 노란색 선으로 표시하여 그림을 그렸다. 각 방법들의 distance score를 비교하면 LIME(FZ)〉SHAP(FZ)〉LIME(QS)〉SHAP(PM)〉SHAP(QS)〉SHAP(SLIC)〉LIME(SLIC) 순으로 순위를 가진다. 평균적으로 LIME이 SLIC 분할을 제외한 나머지 분할에서 SHAP보다 distance score가 높게 결과가 나왔다. 따라서, 사람이 인지하는 매우 작은 도장을 가장 잘 찾는 알고리즘은 LIME이고 특히, FZ 분할이 좋다는 결과를 얻는다.

Ⅴ. 토의

비교를 통해서, FZ 분할이 가장 높은 distance score를 가지고 LIME과 SHAP 중에서는 전반적으로 LIME이 더 좋은 점수를 가지는 것을 확인했다. 한편, distance method에서 도장은 인간이 인식할 수 있고 신경망이 찾는 것이 가능한 가장 작은 특징이다[12]. 따라서, 작은 특징을 찾는 신경망을 설명하는 경우에, LIME(FZ)을 사용하는 것이 가장 좋은 설명을 얻는다. 한편, 알고리즘들의 특성을 보면 모두 클러스터링 알고리즘과 관련이 있다. 따라서,

분할 알고리즘을 이용하는 XAI 알고리즘들을 모두 클러스터링하는 기법이 발전한다면 이미지에서 더 좋은 설명을 얻을 것이다.

VI. 결론 및 향후연구

본 기고문에서는 MNIST 데이터셋에 신경망이 인식할 수 있는 최소 단위 특성을 가진 도장을 찍어서 존재 여부를 학습 후 이에 대한 XAI 알고리즘들의 설명을 distance method를 통해 비교했다. 전체적으로 LIME이 SHAP보다 우수한 점수를 갖는 것을 확인할 수 있었다. 또한 클러스터링 기법이 발전함에 따라 분할을 이용하는 설명들은 후에 더 좋은 점수를 얻을 수 있다는 것도 확인했다.

Distance method에도 문제점은 분명히 존재한다. 특히, 본 논문에서 말한 것처럼 XAI 알고리즘이 제공하는 해답의 의미에 따라 distance를 구하는 과정 중 -1을 곱하는 과정이 있지만 LIME과 SHAP에는 적절하지 않기 때문에 생략을 했다. 즉, 이 비교 방법은 적용하는 XAI 알고리즘에 따라 적절한 수정이 아직 필요해 보인다.

참 고 문 헌

[1] Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big data & society* 3.1 : 2053951715622512, 2016.
[2] Panigutti, Cecilia, et al. "FairLens: Auditing black-box clinical decision support systems." Information Processing & Management 58.5 (2021): 102657.
[3] Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." *arXiv preprint arXiv:2006.11371*, 2020.
[4] ElShawi, Radwa, et al. "Interpretability in healthcare: A comparative study of local machine learning interpretability techniques." *Computational Intelligence* 37.4 : 1633-1650, 2021.
[5] Amparore, Elvio, Alan Perotti, and Paolo Bajardi. "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods." *PeerJ Computer Science* 7 : e479, 2021.
[6] Doumard, Emmanuel, et al. "A comparative study of additive local explanation methods based on feature influences." *24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2022)*. Vol. 3130. No. paper 4. CEUR-WS. org, 2022.
[7] Man, Xin, and Ernest P. Chan. "The best way to select features? comparing mda, lime, and shap." *The Journal of Financial Data Science* 3.1 : 127-139, 2021.
[8] Breiman, Leo. "Random forests." *Machine learning* 45 : 5-32, 2001.
[9] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you? Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. p.1135-1144. 2016.
[10] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).
[11] Sun, Jia, Tapabrata Chakraborti, and J. Alison Noble. "A Comparative Study of Explainer Modules Applied to Automated Skin Lesion Classification." *XI-ML@ KI*. 2020.
[12] Jung, Jay Hoon, and Youngmin Kwon. "A Metric to Compare Pixel-Wise Interpretation Methods for Neural Networks." *IEEE Access* 8: 221433-221441, 2020.
[13] Vedaldi, Andrea, and Stefano Soatto. "Quick shift and kernel methods for mode seeking." *European conference on computer vision*. Springer, Berlin, Heidelberg, p. 705-718. 2008.
[14] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." *IEEE transactions on pattern analysis and machine intelligence* 34.11 : 2274-2282 ,2012.
[15] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient graph-based image segmentation." *International journal of computer vision* 59.2: 167-181, 2004.
[16] Shapley, L. "Quota solutions op n-person games1." *Edited by Emil Artin and Marston Morse* : 343, 1953.
[17] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1.4 : 541-551, 1989.

저 자 소 개

정 재 훈

- 2007년 : 뉴욕주립대학교 물리학 박사 수료
- 2021년 : 뉴욕주립대학교 컴퓨터 과학 박사
- 2021년 ~ 현재 : 한국항공대학교 인공지능학과 조교수
- 주관심분야 : XAI, 양자 컴퓨팅, 양자 인공지능