심층신경망과 해석 가능성: 최신 연구 동향 및 시사점 고찰

조상균*

현대 인공지능 분야의 대표주자라 할 수 있는 심층신경망 학습(Deep learning)은 다양한 분야에 적용되어 그 위력을 입증하고 있다. 초기 심층신경망 연구의 주요 관심사가 효과적인 학습과 성능 향상에 방점이 찍혀있었다면, 최근의 심층신경망 연구는 심층신경망의 실제 적용이라는 관점에서 파생되는 다양한 측면의 문제들을 포괄하고 있다. 심층신경망에 해석 가능성을 부여하고자 하는 연구 흐름 역시 이러한 배경에서 등장했다고 볼 수 있다. 본 고에서는 심층신경망에 해석성을 이식하는 작업이 왜 필요한지 살펴보고 이와 관련된 최신 연구 동향을 논의하고자 한다.

Ⅰ. 서 론

최근 몇 년간 이어진 심층신경망의 눈부신 발전은 인공지능 분야뿐 아니라 인접한 여러 분야에도 파급효과를 미쳤다. 불과 몇 년 전까지만 해도 난공불락의 문제로 여겨지던 자연어처리(Natural language processing: NLP), 신약 후보 물질 디자인 등이 그것이며 지금도 끊임없이 새로운 응용 연구가 등장하며 발전하고 있다. 심층신경망은 기존의 고전 통계 기반 모형들보다 더 높은 성능, 즉 더 높은 정확성을 보인다. 하지만 뛰어난 성능에도 불구하고 심층신경망을 곧바로 일상의 문제에 적용하기 어려운 경우가 많다. 전통적인 심층신경망은 복잡한 내부 구조로 인하여 알고리즘의 투명성은 물론

* KAIST 경영대학 박사과정, trillionnt@kaist.ac.kr

도출되는 결과에 대한 직관적 해석조차 제공하지 못하기 때문이다. 이러한 문제는 “블랙박스 문제(Black-box problem)”으로 불리며 심층신경망의 치명적인 약점으로 인식되어 왔다. 본 고에서는 심층신경망의 해석가능성과 연관된 개념과 용어들을 살펴보고, 해석성을 심층신경망에 이식하고자 하는 최근의 연구동향을 소개한다. 이를 통하여 향후 심층신경망 연구가 풀어야 할 숙제에 대해서 고찰해보고자 한다.

Ⅱ. 심층신경망에서의 해석가능성

1. 해석가능성 이식의 필요성

심층신경망이 다른 방법론들을 압도할만한 높은 성능을 보인다면 해석성을 포기할 수도 있는 것이 아닐까? 굳이 해석성이라는 특성을 부여하려는 이유는 무엇일까? 몇 가지 예시를 통해서 그 필요성을 살펴보자. 먼저, 심층신경망을 연구하는 연구진이 자율주행 시스템을 구축한 상황을 생각해보자. 연구진은 최선을 다해서 심층신경망을 학습시키고 오류를 방지하고자 했지만 모든 일이 완벽할 수는 없기에 크고 작은 오류가 발생하여 사고로 이어졌다. 그런데, 전통적인 심층신경망에서 오류의 원인을 파악하기가 쉽지 않다. 원인을 파악하기 어렵기 때문에 그것을 교정하는 것은 더 어렵다. 연구자들이 할 수 있는 최선은 추가적인 데이터를 더 넣어주면서 모형이 미처 확인하지 못한 사각을 줄이는 것이다.

두 번째 예시에서는 사진을 심층신경망 넣어주면 개와 고양이로 분류하는 모형을 가정해 볼 것이다. 연구진은 개와 고양이 사진을 충분히 넣어서 학습시켰으며 만족스러운 정확도를 획득하였다. 만약 이 모형에 개나 고양이 사진이 아닌 돌고래 사진을 넣으면 어떻게 될까? 우리가 해결하고자 하는 문제의 범주를 벗어나 버린 것이지만 컴퓨터는 이것을 인식할 수 없고 개 혹은 고양이라는 결론을 출력할 것이다. 딥러닝이 일상에서 안정적으로 활용되기 위해서는 이런 분포 외 데이터셋(Out Of Distribution: OOD)이 입력될 경우에도 적절하게 대응할 수 있어야 한다(Liang, Li, and Srikant 2018).

첫 번째와 두 번째 예시가 기술적 측면에서의 문제점과 밀접한 관련이 있었다면 마지막 예시는 사회적, 윤리적 문제와 결부되어 있다. 이 예시는 가상의 예가 아닌 실제

2016년 미국에서 밝혀진 사례로서 인공지능의 편견(bias) 방지와 공정성(fairness) 확보라는 숙제를 남겼다. 한 독립 언론의 탐사보도를 통해 법원의 범죄 분석 알고리즘 콤파스(COMPAS)가 흑인을 차별한다는 사실이 밝혀졌다(Mehrabi et al. 2019). 물론, 이러한 차별이 사전에 인간에 의해서 설계된 것이 아니었을 뿐만 아니라 이 알고리즘은 인종을 변수로서 포함하지도 않았다. 유사한 예로 같은 해에 러시아 과학자들이 진행한 AI 미인대회에서는 선정된 44명의 미인 중 단 1명 만이 유색인종이었다고 한다(Levin 2016).

앞서 살펴본 세 가지 예시는 각각 독립적인 심층신경망의 과제라고 할 수 있지만, 한편으로는 심층신경망 학습에 왜 해석성이 필요한지 공통적으로 입증하는 사례라고 할 수 있을 것이다.

2. 해석가능성과 설명가능성

심층신경망에 해석성을 부여하는 연구는 현재진행형이며 연관 개념들도 연구자들 사이에 완전한 합의에 도달하지는 못한 상황이라고 할 수 있다. 연구 맥락에 따라 용어가 혼용되기도 하고 상이한 의미로 사용되기도 한다. 여기서는 많은 후속 연구자들에게 지지받아 빈번하게 피인용되고 있는 논문을 바탕으로 심층신경망의 해석가능성(Interpretability)과 설명가능성(Explainability)의 정의를 알아본다.

해석가능성과 설명가능성이라는 단어는 종종 동의어처럼 사용되기도 하지만 대체로 해석가능성이 좀더 광범위한 개념이라고 볼 수 있다. Miller는 해석성이란 어떤 결정의 근거를 사람이 이해할 수 있는 정도라고 정의하였으며(Miller 2019), Kim은 해석이란 사람에게 설명을 제공하는 프로세스라고 정의하였다(Doshi-Velez and Kim 2017). 설명가능성은 이러한 해석가능성에서 더 나아가 사회적(social), 인지적(cognitive)인 요인까지 고려하는 특성을 말한다. 예를 들어 한 사람이 다른 사람에게 어떠한 대상을 설명하는 행위는 화자와 청자의 사회적, 인지적 배경에 영향을 받는다. 똑같은 말을 전달하더라도 이러한 문맥에 따라 전혀 다른 의미로 해석될 수도 있다. 인공지능이 설명가능한 특성을 갖도록 만드는 것은 해석가능한 특성을 갖도록 하는 것보다 더 어려울 수 밖에 없는 이유다. 결국, 우선적인 과제는 심층신경망에 해석가능성을 부여하는 작

업이라고 볼 수 있다.

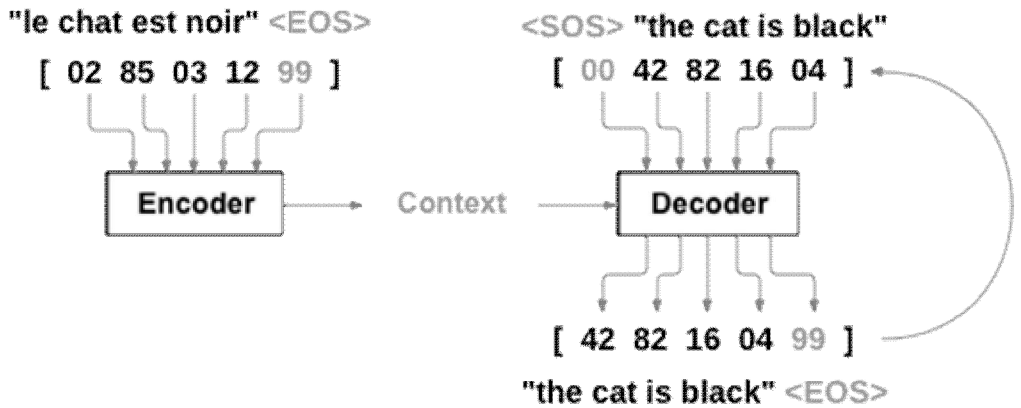
Ⅲ. 베이지안 추론 기반 해석가능한 심층신경망

본 고에서 논의할 주요 내용은 크게 세 가지 주제로 압축될 수 있다. i) 어텐션(attention) 구조 ii) 베이지안 심층신경망(Bayesian Deep neural networks) iii) 불확실성(uncertainty) 정량화가 이에 해당된다.

1. 심층신경망과 어텐션

심층신경망 모형에서 해석성을 부여할 수 있는 대표적인 방안 중 하나가 어텐션 구조를 심층신경망에 삽입하는 방법이다. 어텐션은 자연어처리 분야에서 순차적(Sequential) 정보를 적절하게 반영하기 위한 목적으로 처음 등장했으며, '집중'이라는 단어 뜻 그대로 어떤 설명변수(input feature)가 얼마나 중요한지 정량화하여 나타낸다(Bahdanau, Cho, and Bengio 2015; Sutskever, Vinyals, and Le 2014). 예를 들어 [그림 1]은 불어 문장을 영어 문장으로 번역하는 문제를 표현한다. 출력되는 영어

그림 1 번역을 위한 심층신경망 구조(Encoder-Decoder)

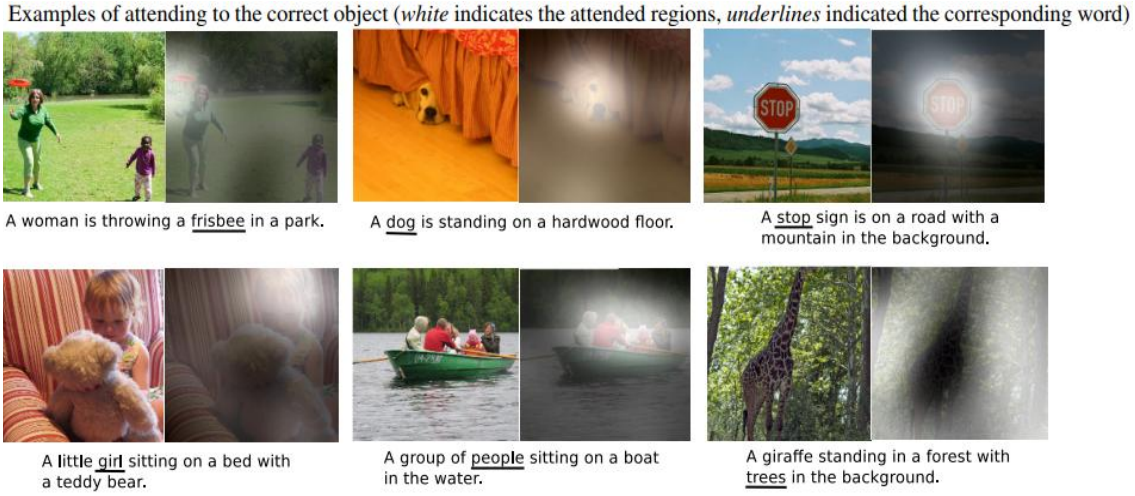


자료: https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

단어 "*black*"은 어떤 요인들의 영향을 받아 선택되었을까? 이미 앞서 출력된 "*cat*", "*is*"의 영향도 받았을 것이고 동시에 불어 원문 단어들 "*le*", "*chat*"과 같은 단어들에도 영향을 받을 것이다. 어텐션은 출력되는 단어를 선정하는 과정에서 앞서 등장한 단어가 얼마나 중요하게 작용하는지 나타내는 일종의 가중치라 볼 수 있다.

어텐션이 자연어 처리 분야에서 처음 도입된 후, 다른 분야에서도 이를 활용하기 시작했다. 특히 이미지처리 분야에서는 어텐션을 시각화하여 보여줌으로써 어텐션이 어떤 역할을 수행하는지 좀더 직관적으로 표현하였다(Xu et al. 2015). [그림 2]는 사진 이미지를 입력받아 이를 묘사하는 캡션이 생성되는 것을 보여준다. 여기서 어텐션은 밑줄친 각 단어가 선정되는데 사진의 어떤 입력(=픽셀)이 중요하게 고려되는지를 알려준다. 입력 사진 바로 오른쪽 사진에서 밝음의 정도를 이용하여(픽셀의 밝기가 밝을수록 그 픽셀이 중요하게 작용) 어텐션을 표현하였다.

그림 2 **이미지 세분화 문제와 어텐션 시각화**



자료: Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International Conference on Machine Learning (ICML 2015).

어텐션이 입력 변수의 중요도를 나타내는 지표로 사용될 수 있다면, 통상적인 회귀(regression) 문제에서도 이를 활용할 수 있을 것이다. Choi et al.은 전자의료기록(electronic health records: EHR)을 활용하여 심부전 위험도를 예측하는 심층신경망

구조(REverse Time AttentIoN model: RETAIN)를 제안하였다(Choi et al. 2016). 이 구조에서는 어텐션을 변수축과 시간축 각각에 대해서 설정하였으며, 이를 통하여 심부전을 일으키는데 주요하게 작용하는 위험인자(risk factor)와 그 시점을 식별하였다. 이 성과 이후 많은 후속 연구가 동일한 심층신경망 구조를 활용하였으나 연산된 어텐션이 얼마나 안정적이며 신뢰성 있는 값인지에 대한 의문은 해소하지 못했다. 이 문제는 베이지안 심층신경망 기법을 통하여 해결할 수 있었는데, 이 장의 마지막 섹션에서 이를 다시 논의하겠다.

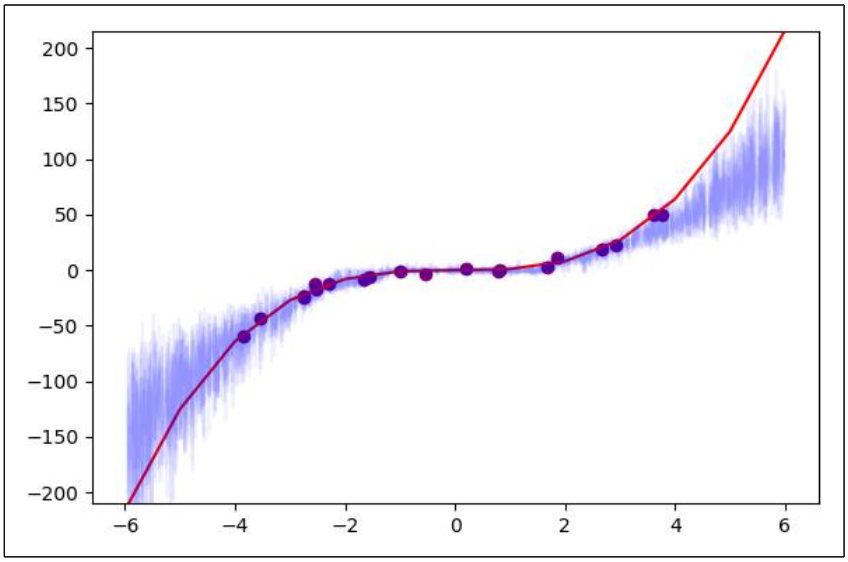
2. 베이지안 심층신경망의 불확실성 정량화

통계학에는 빈도주의(frequentist) 관점과 베이지안 관점의 시각이 있다. 베이지안 관점에서 사건과 확률을 바라본다는 것은 관측된 정보 외에 추가적으로 선험적인 지식이나 경험적인 정보를 추가적으로 부여하겠다는 것이다. 이러한 맥락에서 추정하고자 하는 모수(parameter)는 더 이상 점추정의 대상이 아니라 분포를 갖는 확률변수로 간주된다. 그렇다면 심층신경망의 수많은 네트워크 가중치들을 어떻게 베이지안 관점에서 계산할 수 있을까? Gal et al.은 심층신경망에서 과적합(overfitting) 방지 방법으로 사용되던 dropout과 정규화(regularization)를 동시에 적용할 경우, 베이지안 관점에서 출력변수의 확률을 계산하는 것과 수학적으로 동등하다는 것을 증명하였다(Gal and Ghahramani 2016). 즉, 네트워크 가중치들이 기댓값이 0인 정규분포를 따른다고 가정했을 때 출력변수가 관측될 확률은 가우시안 프로세스(Gaussian Process: GP)가 되며, 이를 다시 정리하면 최우도추정(Maximum likelihood estimation: MLE)에 dropout과 정규화를 적용한 꼴과 동일해진다는 것이다.

심층신경망에 베이지안 관점을 도입하게 되면서 우리는 출력변수의 불확실성을 계산할 수 있게 되었다. 다시 말해 모형이 출력하는 종속변수가 단순히 어떤 특정 값이라는 결론이 아니라 그 종속변수의 추정구간을 출력하게 된다. 불확실성이 커지면 커질수록 구간도 넓어진다. [그림 3]은 시뮬레이션을 통해 생성된 회귀 데이터를 베이지안 심층신경망으로 복구한 결과를 나타낸다. 붉은색 실선은 데이터를 생성시킨 실제 함수(ground truth)이고 보라색 점들은 실제 관측치이다. 파란색 음영으로 처리된 부분이

베이지안 심층신경망을 활용하여 예상된 예측치(구간)로서 관측치가 없었던 영역에서 불확실성이 증가하는 것을 확인할 수 있다.

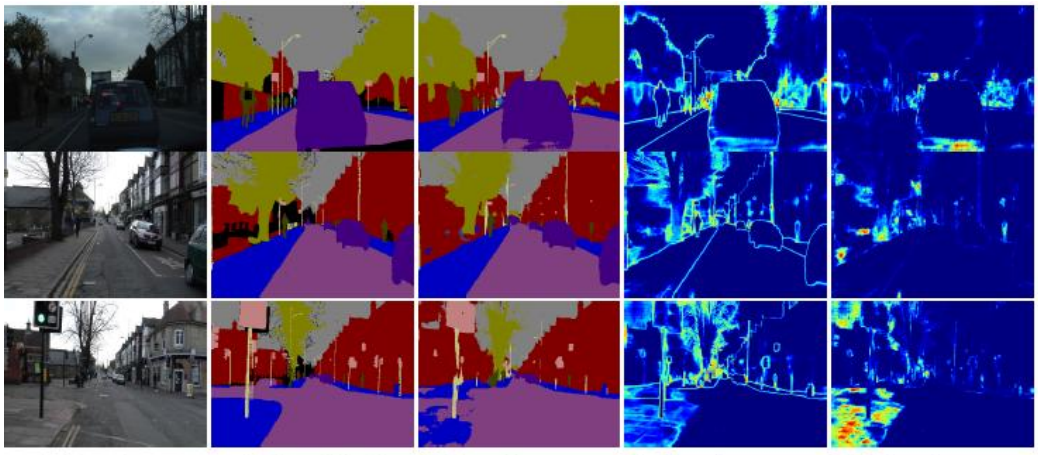
그림 3 **베이지안 심층신경망을 이용한 예측**



자료: Gal, Yarin, and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning.” International Conference on Machine Learning (ICML 2016).

Gal et al.은 후속연구에서 불확실성을 다시 크게 두 가지로 분류하였다. Epistemic 불확실성과 aleatoric 불확실성이 그것이다(Kendall and Gal 2017). Epistemic 불확실성은 모형의 불완전성에 의해 발생하는 것으로, 일반적으로 관측치가 많아지면 해소될 수 있는 불확실성이다. 반면, aleatoric 불확실성은 관측 오류(measurement error)와 같은 내재적 환경에서 기인하는 불확실성으로 감소시킬 수 없는 불확실성이다. [그림 4]는 이미지 분할 문제에 이러한 베이지안 방법론을 적용시킨 결과를 보여주고 있다. 물체의 경계면에서는 aleatoric 불확실성이 높은 값을 나타내고, 충분히 관측되지 못한 대상(해당 사진에서는 하단의 인도 노면)에서는 epistemic 불확실성이 높게 나타나는 것을 확인할 수 있다.

그림 4 Epistemic/Aleatoric 불확실성



자료: Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." Advances in Neural Information Processing Systems (NIPS 2017).

이와 같이 베이지안 심층신경망을 활용하면 출력 변수의 불확실성을 획득할 수 있지만, 이것만으로는 입력 변수의 영향력과 그 불확실성을 산출할 수 없다. 앞선 문단에서 독립적으로 소개했던 어텐션과 베이지안 심층신경망 기법 각각의 이점을 모두 포용할 수 있어야 비로소 입력 변수의 영향력과 불확실성을 정량화하고 산출하는 것이 가능할 것이다.

3. 어텐션의 불확실성 부여 및 신뢰구간 산출

Heo et al.은 Choi et al.이 사용한 것과 동일한 데이터셋을 활용하여 입력 변수의 어텐션 값과 함께 그 불확실성을 산출할 수 있는 베이지안 심층신경망 구조(Uncertainty Aware deep neural networks: UA)를 제안하였다(Heo et al. 2018). 이들은 베이지안 심층신경망 구조에 어텐션을 추가하고 몬테카를로 샘플링(Monte Carlo sampling)하여 네트워크에 반영하는 방식으로 두 방법론을 통합하였다.

지금까지의 선행연구를 종합 요약하자면 다음과 같다. 어텐션이 도입된 베이지안 심

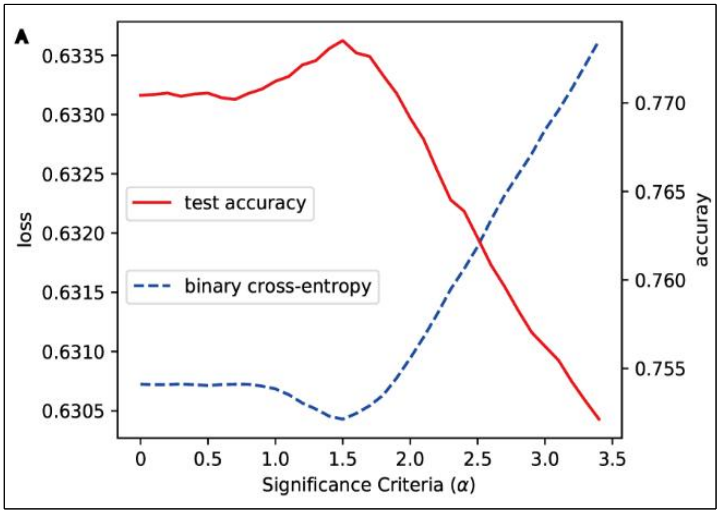
층신경망 구조를 활용하면 입력 변수가 얼마나 중요하며 또 그것이 얼마나 안정적인지 얻어낼 수 있다. 만일 입력 변수의 중요도라는 측면에서 “어텐션”을 고전통계 모형의 “계수(혹은 모수)”에 대응시킨다면, 같은 맥락에서 어텐션의 불확실성은 계수의 표준오차에 대응시킬 수 있지 않을까?

이에 착안하여 Jo et al.은 내생성(endogeneity)이 존재하는 상황에서 내생성을 교정하는 동시에 설명변수들의 영향력을 정량화할 수 있는 심층신경망 구조를 생성했다(Uncertainty Aware Deep Instrumental 2-stage network: UA Deep IV). 나아가 어텐션의 크기와 불확실성을 이용해서 어텐션의 신뢰구간을 산출하고 어텐션의 통계적 유의성의 개념을 제안하였다(Jo, Jun, and Park 2020). 이를 바탕으로 건강검진이 의료비 지출에 미치는 효과는 통계적으로 유의하지 않다는 것을 실증분석하였으며, 제안된 통계적 유의성이 심층신경망의 가지치기(pruning)에 유용하게 활용될 수 있다는 것을 입증하였다.

심층신경망의 가지치기는 본래 기존 네트워크의 성능은 유사하게 유지하면서도 네트워크의 크기를 압축하여 컴퓨터 연산 자원을 절약할 수 있도록 고안된 방법이다(KISDI AI Outlook(2020년 봄 Vol. 1) 참고). 예를 들어, 크기가 작은 가중치들을 네트워크에서 제거한다고 했을 때 약 90% 정도의 가중치를 제거하더라도 네트워크의 성능은 처음과 비슷한 수준으로 유지된다는 것이다(Han et al. 2015). 다만, 이러한 방식의 가지치기는 재훈련을 시키지 않는 이상 가지치기만으로 원래 성능보다 향상된 성능을 기대하기는 어렵다. 그런데, 어텐션의 신뢰구간을 기준으로 통계적으로 유의하지 않은 어텐션을 가지치기 했을 때, 원래 네트워크의 성능보다 향상된 성능의 정확도를 획득할 수 있다.

[그림 5]는 통계적 유의성의 기준을 점점 더 엄격하게 만들면서 가지치기 대상을 늘려 손실함수와 정확도를 측정한 그래프이다. 기존 가지치기의 방법의 경우, 가지치기 대상을 늘리면 손실은 일정 수준에서 유지되다가 단조증가하고 정확도 역시 일정수준에서 유지되다가 단조 감소하는 모양이 나타날 것이다. 반면 통계적 유의성을 기준으로 가지치기를 수행하였을 때는 해당 그림에서 확인할 수 있는 것처럼 적정수준의 가지치기 수행 이후 오히려 정확도가 향상되는 것을 확인할 수 있다.

그림 5 **통계적 유의성 기반 가지치기 수행 시 정확도/손실함수**



자료: SangKyun Jo, Duk Bin Jun and Sungho Park. “Estimating the effect of General Health checkup using Uncertainty Aware Attention of Deep Instrument Variable 2-stage network.” 19th International Conference on Machine Learning and Applications(ICMLA 2020). (accepted)

해당 연구 사례는 심층신경망 분야와는 별개의 방법론으로 간주되던 계량경제학적 방법론을 차용하여 해석성을 높이고 심층신경망의 지상과제인 성능향상도 성취한 한 사례로 볼 수 있을 것이다.

IV. 결 론

본 고에서는 심층신경망에 해석가능성을 도입하는 것이 왜 필요한지 알아보고, 해석가능성을 심층신경망에 부여하기 위한 최근의 연구 동향을 살펴보았다. 입력 변수의 영향력을 정량화하기 위한 어텐션, 그 불확실성을 산출하기 위한 베이지안 심층신경망 등의 방법론이 제안되었으며 이를 기반으로 한 후속연구들이 이루어지고 있다. 한 가지 주지해야 할 사실은 심층신경망 분야 연구자들 모두가 동의하는 정석적인 해석성 부여 방법론은 아직 까지는 존재하지 않는다는 사실이다. 일례로 어텐션은 복잡한 연산 과정

의 부가적 산물이며 설명변수의 중요도를 정확하게 반영할 수 없다는 견해도 존재한다. 이러한 과도기적 연구에서는 다각적인 접근, 인접 분야와의 협력 등이 다음 단계로 나아가는 열쇠가 될 수도 있을 것이다. 앞으로 활발한 후속 연구가 이어져 심층신경망이라는 블랙박스를 적절히 통제하고 해석하게 될 날을 기대하며 이 글을 마친다.

참 고 문 헌

Bahdanau, Dzmitry, Kyung Hyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate." in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Choi, Edward, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. "RETAIN: An Interpretable Predictive Model for Healthcare Using Reverse Time Attention Mechanism." in *Advances in Neural Information Processing Systems*.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning." (Ml):1-13.

Gal, Yarin, and Zoubin Ghahramani. 2016. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." in *33rd International Conference on Machine Learning, ICML 2016*.

Han, Song, Jeff Pool, John Tran, and William J. Dally. 2015. "Learning Both Weights and Connections for Efficient Neural Networks." in *Advances in Neural Information Processing Systems*.

Heo, Jay, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. 2018. "Uncertainty-Aware Attention for Reliable Interpretation and Prediction." in *Advances in Neural Information Processing Systems*.

Jo, SangKyun, Duk Bin Jun, and Sungho Park. 2020. "Estimating the Effect of General Health Checkup Using Uncertainty Aware Attention of Deep Instrument Variable 2-Stage Network." in *International conference on machine learning and applications (ICMLA 2020)*. IEEE (accepted).

Kendall, Alex, and Yarin Gal. 2017. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in *Advances in Neural Information*

Processing Systems.

Levin, Sam. 2016. "A Beauty Contest Was Judged by AI and the Robots Didn't Like Dark Skin." *The Guardian*.

Liang, Shiyu, Yixuan Li, and R. Srikant. 2018. "Enhancing the Reliability of Out-of-Distribution Image Detection in Neural Networks." in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. "A Survey on Bias and Fairness in Machine Learning."

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence*.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. "Sequence to Sequence Learning with Neural Networks." in *Advances in Neural Information Processing Systems*.

Xu, Kelvin, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." in *32nd International Conference on Machine Learning, ICML 2015*.