프로토타입 생성 기반 딥 러닝 모델 설명 방법

박재훈O, 김광수*
O성균관대학교 소프트웨어학과
*성균관대학교 소프트웨어학과
e-mail: pk9403@skku.eduO, kim.kwangsu@skku.edu*

Interpretable Deep Learning Based On Prototype Generation

Jae-Hun ParkO, Kwang-Su Kim*
OCollege of Computing and Informatics, Sungkyunkwan University
*College of Computing and Informatics, Sungkyunkwan University

요 약

딥 러닝 모델은 블랙 박스 (Black Box) 모델로 예측에 대한 근거를 제시하지 못해 신뢰성이 떨어지는 단점이 존재한다. 이를 해결하기 위해 딥 러닝 모델에 설명력을 부여하는 설명 가능한 인공지능 (XAI) 분야 연구가 활발하게 이루어지고 있다. 본 논문에서는 모델 예측을 프로토타입을 통해 설명하는 딥 러닝 모델을 제시한다. 즉, "주어진 이미지는 티셔츠인데, 그 이유는 티셔츠를 대표하는 모양의 프로토타입과 닮았기 때문이다."의 형태로 딥 러닝 모델을 설명한다. 해당 모델은 Encoder, Prototype Layer, Classifier로 구성되어 있다. Encoder는 Feature를 추출하는 데 활용하고 Classifier를 통해 분류 작업을 수행한다. 모델이 제시하는 분류 결과를 설명하기 위해 Prototype Layer에서 가장 유사한 프로토타입을 찾아 설명을 제시한다. 실험 결과 프로토타입 생성 기반 설명 모델은 기존 이미지 분류 모델과 유사한 예측 정확도를 보였고, 예측에 대한 설명력까지 확보하였다.

▸ Keyword : 설명 가능한 인공지능 (XAI), 프로토타입 기반 설명 (Prototype Based Explanation), 딥 러닝 (Deep Learning)

I. Introduction

딥 러닝 모델은 컴퓨터 비전에서부터 자연어 처리까지 다양한 문제를 해결하면서 발전해왔다. 그러나, 모델이 깊어져 감에 따라 성능은 높아졌으나 비선형적인 특징으로 인해 사람이 모델을 직관적으로 이해하기 어렵다. 이런 특징 때문에 딥 러닝 모델은 블랙 박스 (Black Box) 모델로 불리는데, 모델의 예측 성능은 뛰어나지만, 어떠한 근거로 해당 결과가 나왔는지 설명하기 어렵다. 따라서, 딥 러닝을 해석하기 위해 설명 가능한 인공지능 (XAI)에 대한 연구가 활발히 이루어지고 있다.

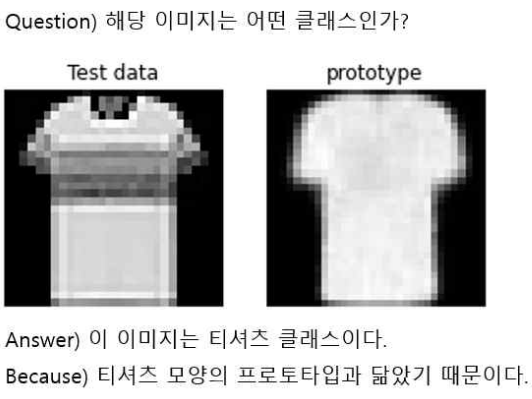

Fig. 1. Overview

본 논문에서는, 프로토타입 생성 기반 설명 가능한 딥 러닝 모델을 제시한다. 프로토타입 설명 모델은 그림 1처럼 어떤 이미지의 예측 (티셔츠)을 설명할 때 프로토타입 (티셔츠 대표 이미지)을 활용하는 모델이다. 프로토타입 설명 모델은 encoder, prototype layer, classifier로 구성되어 있다. Encoder를 통해 특징을 추출하고, classifier를 통해 분류 작업을 수행한다. Prototype layer를 통해 모델의 예측을 설명한다.
실험 결과 프로토타입 생성 기반 설명 가능한 딥 러닝 모델은 기존의 딥 러닝 모델과 비슷한 성능을 유지하면서 예측에 대한 설명력까지 확보할 수 있었다.

II. Preliminaries

1. Related works

딥 러닝 모델의 해석을 위해 다양한 시도가 이루어지고 있다. DeConvNet [1]은 학습 과정을 역계산하고 각 계층을 시각화하여 이미지 분류 근거를 확인하는 설명 기법이다. LIME [2]은 모델을 지역적으로 선형 근사시켜 설명력을 확보한 기법이다. SHAP [3]은 각 feature가 모델에 미치는 영향도를 계산함으로써 예측 결과에 대한 근거를 제시한 기법이다. TCAV [4]은 이미지 분류에서 사람이 직관적으로 이해할 수 있는 concept의 개념을 정의하고 이를 바탕으로 모델의 예측을 설명하는 기법이다.
본 논문에서 프로토타입을 기반으로 설명하는 기법에 대해서 다룬다. 기존 논문 [5]과의 차이점은 학습 방법, prototype vector, 전이 학습 여부 등이 있다.

III. The Proposed Scheme

1. Model

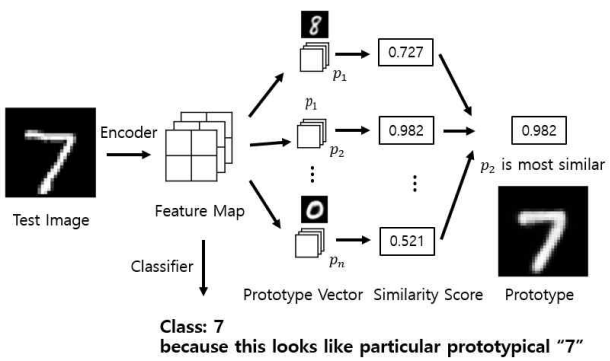

Fig. 2. Model Architecture

본 논문에서는 이미지 분류 딥 러닝 모델을 기반으로 모델을 구성하고 실험을 진행한다. 시계열 데이터에 대해서도 같은 모델 구조를 적용할 수 있다.
모델은 encoder, classifier, prototype layer로 이루어져 있다. Encoder는 특징을 추출하여 feature map을 생성한다. Classifier는 feature map의 특징을 파악하여 이미지의 클래스를 예측한다. Prototype layer는 feature map과 같은 차원이며, 프로토타입의 개수만큼 존재한다. 학습 과정에서 클러스터링을 진행하고, 클러스터 중심을 prototype vector로 설정한다. Prototype vector를 사람이 이해할 수 있는 형태로 변환하기 위해 decoder에 통과시켜 프로토타입을 시각화할 수 있다. Prototype vector와 feature map 간의 유사도를 비교하여 가장 유사하다고 판단되는 프로토타입을 해당 이미지의 설명으로 채택한다. 따라서 그림 2와 같이 테스트 이미지를 분류하고 설명하는 방식은 다음과 같다: 이 이미지는 7이라는 클래스로 예측되었는데, 그 이유는 7 모양의 프로토타입과 가장 유사하기 때문이다.

2. Training

2.1 Autoencoder

이미지의 특징을 추출하고 prototype vector를 시각화하기 위해 autoencoder 모델을 학습한다. Autoencoder 모델은 입력과 출력이 동일한 형태로, 차원을 축소 및 확대하면서 데이터의 특징을 학습하는 모델이다. 훈련 데이터로 학습이 완료되고 나면, encoder와 decoder를 분리한다. Encoder는 이미지 특징 추출에 활용하고, decoder는 프로토타입 시각화에 활용한다.

2.2 Prototype Generation

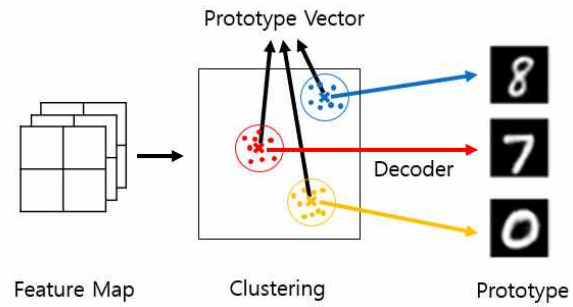

Fig. 3. Prototype Generation

그림 3과 같이 프로토타입 생성 및 시각화를 위해 2.1에서 훈련했던 encoder와 decoder를 활용한다. 훈련에 사용했던 학습 데이터들을 encoder에 통과시켜 feature map을 생성한다. 생성된 feature map들에

대해 클러스터링을 진행하고, 클러스터의 중심을 prototype vector로 설정한다. 클러스터의 개수는 프로토타입의 개수와 같다. Prototype vector를 학습된 decoder에 통과시켜 사람이 해석할 수 있는 이미지 형태의 프로토타입을 생성한다.

2.3 Classifier

이미지 분류를 수행하기 위해 classifier를 추가로 학습한다. 2.1에서 학습된 encoder 끝단에 classifier를 붙여 전이 학습 (Transfer Learning)을 수행한다. 전이 학습은 학습된 신경망을 활용하기 때문에 학습 속도가 빠르다는 장점이 있다. Encoder의 가중치는 학습되지 않도록 고정시키고, classifier 부분만 학습한다.

3. Experiment Setup

Encoder를 4개의 컨볼루션 레이어로 구성하였고 kernel size 3×3, zero padding, stride 2, relu activation을 적용하였다. 필터 크기는 32, 16, 16, 8로 설정하였고, adam optimizer, L2 loss로 200 epochs 만큼 학습을 진행했다. Decoder는 encoder의 역순으로 4개의 전치 컨볼루션 레이어를 구성했고, 마지막 부분에 이미지 형태로 출력하기 위해 필터 크기 1, sigmoid activation의 컨볼루션 레이어 1개를 추가했다. Classifier는 flatten layer와 2개의 dense layer로 구성했다.

2개의 데이터셋에 대해 실험을 진행하였다. 손으로 쓴 숫자 이미지로 이루어진 MNIST 데이터셋과 옷 이미지로 이루어진 Fashion MNIST 데이터셋을 활용하였다. 두 데이터 셋 모두 60,000 개의 훈련 이미지와 10,000 개의 테스트 이미지로 구성되어 있으며, 이미지는 32×32×1 크기로 변형하였다.

4. Experiment Results

4.1 Autoencoder

Autoencoder 모델이 제대로 학습되었는지 확인하기 위해 그림 4와 같이 테스트 데이터를 활용하여 복원 결과를 확인해보았다. MNIST와 Fashion MNIST 데이터에 대해 복원이 잘 되는 것을 확인할 수 있었다.

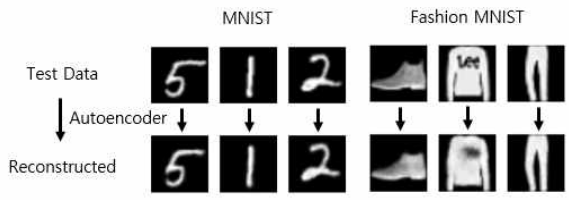

Fig. 4. Test Autoencoder Model

4.2 Prototype Generation

Prototype vector로부터 프로토타입을 잘 생성하는지 확인해보았다. 그림 5는 prototype vector를 decoder에 통과시켰을 때 생성되는 프로토타입을 나타낸 그림이다. MNIST와 Fashion MNIST 데이터 세트에 대하여 prototype vector가 특정 클래스의 클러스터를 대표하도록 클러스터링이 잘 되었고, 사람이 이해할 수 있는 형태로 시각화도 가능했다.

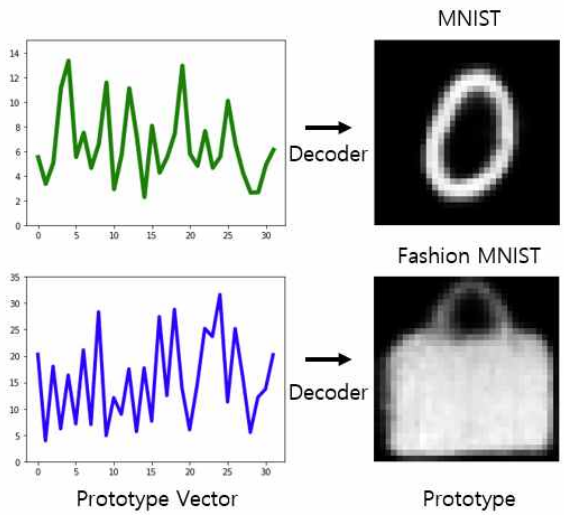

Fig. 5. Prototype Generation

그림 6은 생성된 프로토타입의 일부를 나타낸 그림이다. 프로토타입은 사람이 직관적으로 이해하기 쉬운 형태였고, 특정 클래스를 대표하는 이미지들로 구성되어 있었다. 클래스의 개수보다 프로토타입의 개수를 크게 설정했기 때문에, MNIST 데이터셋에서 7과 9가 여러 번 등장한 것처럼 같은 클래스 안에서 다양한 형태의 프로토타입이 등장하였다.

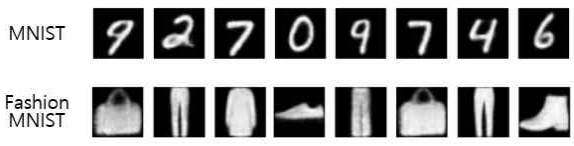

Fig. 6. Generated Prototypes

4.3 Model Accuracy

일반적인 이미지 분류 모델과 프로토타입 모델의 분류 성능을 비교하기 위해 프로토타입 모델과 같은 구조의 encoder와 classifier로 baseline 모델을 구성하고 훈련을 진행했다. 표 1에 따르면 프로토타입 모델은 기존 이미지 분류 모델과 비슷한 성능을 보였다.

Table 1. Classification Accuracy

	MNIST	Fashion MNIST
Baseline	0.9862	0.8967
Prototype	**0.9777**	**0.8747**

IV. Conclusions

본 논문에서는 프로토타입 생성 기반 딥 러닝 모델 설명 기법을 제시하였다. 프로토타입을 통해 모델의 예측을 설명할 수 있었으며, 예측 정확도는 설명력이 없는 모델과 유사한 수준을 보였다. 따라서, 본 논문에서 제시한 프로토타입 설명 모델은 예측과 설명 모두 가능한 모델이다.

Acknowledgement

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2021-0-02068, 인공지능 혁신 허브 연구 개발)과 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-2015-0-00742, 지능지역화혁신인재양성사업)

References

[1] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European conference on computer vision. Springer, Cham, 2014.

[2] 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. 2016.

[3] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[4] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.

[5] Li, Oscar, et al. "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.