딥러닝 해석을 위한 계층적 다중 뉴런 프레임워크 모델*

송치영[01], 남윤형[1], 송인혁[1], 김성태[1]
[1]경희대학교 컴퓨터공학과
cosmos88@khu.ac.kr, imyhnam@khu.ac.kr, thddlsgur0105@khu.ac.kr, st.kim@khu.ac.kr

A Hierarchical Interpretation Framework Based on Multi-Neuron Activation Patterns

ChiYoung Song[01], YoonHyung Nam[1], InHyuk Song[1], SeongTae Kim[2]
[1]Department of Computer Science and Engineering, Kyung Hee University

요 약

딥러닝 모델의 블랙박스 문제를 해결하기 위해, 본 연구는 뉴런 그룹 단위로 해석할 수 있는 확장 가능하고 계층적인 딥러닝 해석 프레임워크를 제안한다. 사전 라벨이 존재하지 않는 이미지들을 수집해 BLIP 모델[5]로 캡셔닝하고, Finch 알고리즘[4]으로 유사한 의미를 가지는 이미지들을 계층적으로 클러스터링하였다. 각 클러스터에 대표 캡션을 ChatGPT 로 생성하고, 분석 대상이 되는 모델(Resnet18)로부터 공통 활성 뉴런 그룹을 추출하여 해당 캡션을 라벨링 하였다. 상위 계층의 클러스터에서는 하위 뉴런 그룹 간 공통 뉴런을 통합하고, 보다 포괄적인 의미의 캡션을 추출해 라벨링한다. 또한 클러스터 간 의미적 차이를 뉴런 그룹 단위로 분석한다. 이를 통해, 사전 라벨 없이도 딥러닝 모델의 계층적 의미 구조를 해석할 수 있는 실질적 방법을 구현하였다.

1. 서 론

최근 딥러닝 기술은 컴퓨터 비전, 자연어 처리 등 다양한 산업 분야에 빠르게 응용되고 있으며, 그 성능 또한 지속적으로 향상되고 있다. 그러나 딥러닝 모델의 복잡성과 비선형성으로 인해, 내부의 결정 과정을 파악하기 어려운 '블랙박스(black-box)' 문제는 여전히 해결되지 않은 중요한 과제로 남아 있다.

이에 따라 모델의 내부 과정을 직관적으로 이해할 수 없다는 블랙박스 문제를 해결하기 위한 여러 연구들이 제안되어 왔다. 대표적으로는 뉴런이 특정 개념에 반응하는 정도를 정량화하거나[1], 생성 뉴런이 시각적 속성에 미치는 영향을 분석[2], 혹은 이미지의 활성화 영역을 바탕으로 자연어 설명을 붙이는 방식[3]이 있다. 하지만 이러한 방식들은 단일 뉴런 수준에서 의미를 도출하는 방식이기에 뉴런 간의 상호작용 및 딥러닝 모델의 계층적인 의미 파악은 어렵다. 또한 사전 정의된 클래스 레이블에 의존하고 있어, 레이블이 존재하지 않는 더 일반적인 이미지들은 사용하기 힘들다는 한계가 있다.

본 연구는 사전 라벨링 없이 임의로 수집된 이미지에 대한 레이어 별 분석을 통한 계층적 다중 뉴런 해석을 가능하게 하는 프레임워크를 제안한다. 사전 라벨 없이 수집된 이미지에 대해 BLIP[5] 기반의 이미지 캡셔닝과 Finch 클러스터링[4]을 결합하여, 의미 기반의 이미지 그룹을 구성한다. 각 그룹에 대해 공통적으로 활성화되는 다중 뉴런 집합을 추출하고, ChatGPT 를 활용하여 자동 라벨링을 수행함으로써, 뉴런 그룹에 대한 자연어 기반 의미 해석을 제공한다. 위 과정을 계층적으로 확장함으로써, 하위 뉴런 그룹 간 공통 뉴런을 상위 그룹으로 통합하고, 보다 포괄적인 설명을 생성하여 딥러닝 모델 내부의 의미 구조를 계층적으로 해석한다. 이를 통해 기존의 단일 뉴런 또는 사전 클래스 기반 해석과 달리, 임의의 데이터셋에 대한 확장 가능성과 뉴런 그룹 단위의 직관적인 해석이라는 장점을 지니며, 딥러닝 해석 연구의 새로운 방향을 제시한다.

2. 관련 연구

기존의 뉴런 해석 연구는 주로 단일 뉴런에 초점을 맞추어, 해당 뉴런을 가장 크게 활성화시키는 대표 이미지를 기반으로 설명을 생성하였다.

대표적으로, Network Dissection 은 CNN 의 중간 계층 뉴런이 '사전 정의된 시맨틱 개념(예: texture, object, part)'에 얼마나 민감하게 반응하는지를 정량적으로 평가하여 뉴런 단위 해석 가능성을 평가하였다. 이 방법은 픽셀 수준의 수작업 annotation 을 필요로 한다는 단점이 있다 [1]. GAN Dissection 은 생성 모델의 내부 뉴런이 특정 시각적 속성을 조절하는 메커니즘을 분석함으로써, 뉴런이 이미지 생성에 어떤 영향을 미치는지 규명하였다 [2]. MILAN 은 단일 뉴런을 크게 활성화시키는 이미지를 수집한 뒤, 해당 뉴런의 기능을 자연어로 설명하는 문장을 자동 생성하는 방식으로 뉴런 해석의 언어적 직관성을 높였다 [3].

단일 뉴런 기반 딥러닝 해석은 명백한 한계점을 가진다. 하나의 뉴런이 모델의 출력에 미치는 영향은 제한적이기에, 설명이 제공된 해당 뉴런이 결정적으로 예측에 어떻게 기여했는지 파악하기 어렵다. 또한 하나의 개념을 표현하는 데는 여러 뉴런이 협력하는데, 단일 뉴런 분석은 이러한 다중 뉴런의 협동적 표현을 전혀 고려하지 않기 때문에 개념 단위

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 2025 년도 SW 중심대학사업의 결과로 수행되었음 (2023-0-00042)

해석에는 부적합하다. 마지막으로, 딥러닝 네트워크가 커질수록 단일 뉴런 분석은 비효율적이다. 한 개의 뉴런마다 설명이 붙기 때문에 분석 비용이 매우 커지며 전체적인 통찰을 얻기 어렵게 된다.

다중 뉴런 기반의 접근 방식들은 이러한 한계를 극복하고자 시도되었다. Net2Vec 은 개별 뉴런이 아닌 뉴런 집합이 특정 개념을 어떻게 표현하는지를 정량적으로 분석하는 프레임워크로, 필터 기반의 활성화 맵과 시맨틱 개념 사이의 관계를 회귀 모델을 통해 학습함으로써, 개념이 여러 뉴런에 걸쳐 분산 표현된다는 사실을 입증하였다 [6]. 이는 단일 뉴런 중심 해석과 달리, 실제 모델이 개념을 다중 뉴런 협력 구조로 인식한다는 점을 정량적으로 보여 주며, 보다 실제적인 해석 가능성을 제시한다. 그러나 Net2Vec 은 사전에 정의된 개념 집합에 의존하기 때문에, 새로운 도메인이나 레이블이 없는 데이터셋에 직접 적용하기 어렵고, 계층적인 의미 구조나 뉴런 간 상호 작용을 반영하기에는 한계가 존재한다.

3. 본론

3.1. 이미지 데이터셋 자동 분류

본 연구는 Unsplash 이미지 25,000 장을 사용하였다. 라벨 없는 이미지 컬렉션을 기반으로 하여, 사전 정의된 클래스 없이도 의미 기반 해석이 가능한 프레임워크를 구축하고자 하였으며, 이를 위해서는 선행적으로 이미지들을 자동으로 의미적으로 분류하는 과정이 필요하다. 이러한 분류는 이후 뉴런 그룹 해석의 기반이 되는 의미 단위 분석을 가능하게 하며, 연구의 일반성과 확장성을 확보한다.

사전 라벨 없는 이미지에 대한 분류 과정을 위해, BLIP 모델 [5]을 활용하여 각 이미지에 대한 자연어 설명을 생성하였다. BLIP 은 사전 학습된 Vision-Language Transformer 기반 모델로, 이미지와 텍스트 간 정렬 능력을 통해 시각적 내용을 자연어로 표현할 수 있다. 생성된 설명 문장은 후속 단계에서 의미 기반 클러스터링의 입력으로 활용된다.

본 연구는 의미 기반 클러스터링을 위해 Finch(First Neighbor Clustering) 알고리즘 [4]을 채택하였다. Finch 는 자동으로 적절한 클러스터 수를 결정하며, 각 데이터 포인트의 가장 가까운 neighbor 정보를 기반으로 연결 그래프를 구성하고, 이를 반복 병합하여 계층적 클러스터 구조를 형성한다. 이러한 구조는 라벨이 없는 이미지 데이터셋에서 의미 기반의 그룹을 자동으로 도출하는 데 매우 적합하다. 본 연구에서는 BLIP 로부터 생성된 25,000 개의 캡션 임베딩을 Finch 알고리즘에 입력함으로써, 유사한 시각적 의미를 공유하는 이미지 클러스터를 계층적으로, 총 6 개의 level 로 구성하였다.

3.2. 공통 활성 뉴런 추출 및 계층적 라벨링

이미지 설명 문장들이 의미 기반으로 클러스터링 된 이후, 각 클러스터를 대표할 수 있는 '요약 문장(대표 캡션)'을 생성한다. 본 연구는 대표 캡션 선정에 있어서 기존 문장을 단순히 재사용 하는 방식 대신, 해당 클러스터에 포함된 모든 이미지 설명을 포괄하면서도 공통된 핵심 개념만을 부각시키는 새로운 문장을 생성하는 접근을 도입하였다. 클러스터에 속한 이미지들의 모든 설명 문장을 ChatGPT 를 통해 입력하고, 적절한 프롬프트를 전달하여 대표 캡션을 생성하였다. 이 과정을 통해 생성된 대표 문장들은 클러스터에 속한 설명 간의 공통 개념과 시각적 특성을 효과적으로 요약하였으며, 이후 단계에서 공통 활성 뉴런 그룹에 의미적 라벨을 부여하고 계층적으로 해석하는데 있어 직관적인 해석 단위로 기능하였다.

Level 1 클러스터는 의미적으로 밀접하게 유사한 이미지들로 구성된다. 이제 각 level 1 클러스터에 대해 의미적으로 연관된 공통 활성 뉴런 그룹을 식별하고, 이 뉴런 그룹에 클러스터의 대표 캡션을 라벨링함으로써 뉴런 해석 단위를 구축한다. 이 과정은 다음과 같이 구성된다.

1) **클러스터 내 이미지 입력** : 각 클러스터에 포함된 모든 이미지를 분석대상이 되는 딥러닝 모델에 입력하여, 뉴런 활성화 값을 수집한다.
2) **활성화 뉴런 정의 및 추출** : 한 이미지에 대해 활성화 값이 전체 뉴런 중 상위 10% 이내에 해당하는 뉴런을 '활성화 뉴런'으로 정의한다. 이후, 클러스터에 포함된 모든 이미지에 대해 이러한 활성화 뉴런을 추적하며, 모든 이미지 중 절반 이상에서 공통적으로 활성화된 뉴런만을 해당 클러스터의 대표 뉴런 그룹으로 선정한다.
3) **대표 캡션 기반 뉴런 그룹 라벨링** : 추출된 뉴런 그룹은 해당 클러스터를 설명하는 핵심적 의미 표현 단위로 간주하며, 해당 클러스터의 대표 캡션을 라벨링한다.

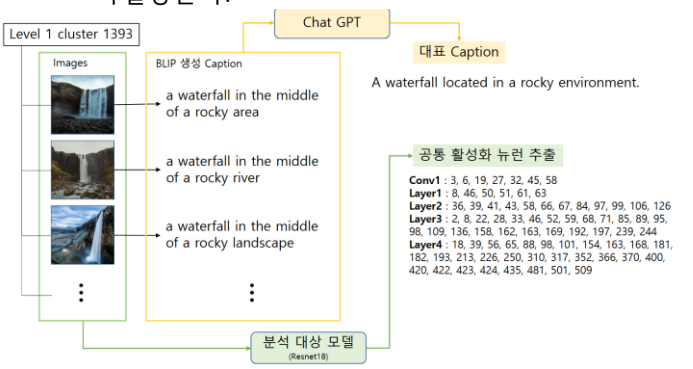

Figure 1: level 1 클러스터 1393 에 속한 이미지들에서 대표 캡션과 공통 활성화 뉴런을 추출하는 과정. 각 이미지를 Resnet18 에 넣어 활성화 뉴런을 수집하고, 절반 이상의 이미지에 대해서 중복적으로 나타난 활성화 뉴런을 '공통 활성화 뉴런'으로 추출하였다. 위의 공통 활성화 뉴런 그룹에는 'A waterfall located in a rocky environment.' 라는 설명 라벨이 붙는다.

위 과정을 모든 level 1 클러스터에 반복 적용함으로써, 의미 기반 뉴런 그룹과 자연어 라벨이 자동으로 연결된다. level 1 에서 추출된 뉴런 그룹과 의미 라벨을 기반으로, level 2 이상의 상위 계층 클러스터에 대해 뉴런 그룹을 확장, 통합하고 클러스터 간의 개념적 공통점, 차이점 분석을 수행한다.

1) **상위 level 뉴런 그룹 생성 및 캡션 통합**
   각 상위 level n 의 클러스터에는 여러 개의 하위 level n-1 클러스터가 속해있다. 상위 level 클러스터의 뉴런 그룹과 캡션은 다음과 같이 도출된다.
   공통 뉴런 추출 : 해당 n level 클러스터에 포함된 모든 n-1 level 클러스터에서 각각 추출된 뉴런 그룹을 비교하여, 이들 간에 공통적으로 존재하는 뉴런들만을 상위 n level 클러스터의 뉴런 그룹으로 선택한다.
   대표 캡션 생성 : 해당 n level 클러스터에 포함된

모든 n-1 level 클러스터의 대표 캡션을 종합하여 ChatGPT 를 활용해 공통 개념을 일반화한 새로운 문장을 생성한다. 이 문장은 해당 n level 클러스터의 대표 캡션이 되며, 하위 개념들을 포괄하는 의미를 형성한다.

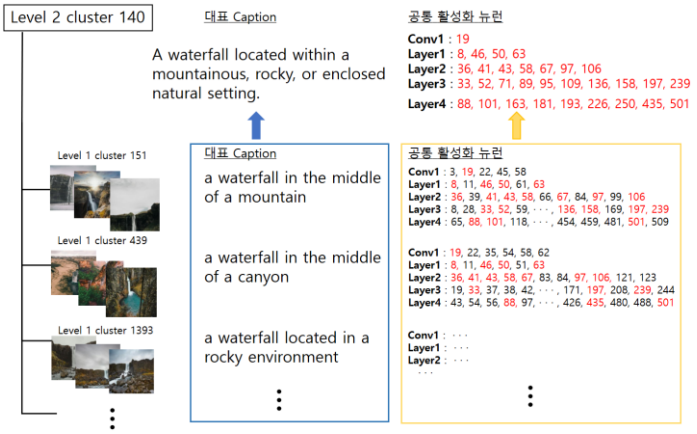

Figure 2: level 1 클러스터들의 대표 캡션을 일반화하여 level 2 클러스터의 캡션을 생성하였다. 모든 level 1 클러스터에서 동일하게 나타난 활성화 뉴런들이 level 2 클러스터의 공통 활성화 뉴런으로 추출된다. 빨간색으로 표시된 뉴런들이 중복되는 부분.

2) **상-하위 뉴런 그룹 간의 차이점 분석**

상위 level 뉴런 그룹이 하위 level 뉴런 그룹의 의미를 완전히 대변하지는 않기 때문에, 본 연구는 level n 의 뉴런 그룹과 각 하위 level n-1 의 뉴런 그룹 간 차이점도 분석하였다.

차이 뉴런 추출 : level n-1 클러스터의 뉴런 그룹 중, 상위 그룹(level n)에 포함되지 않는 뉴런을 차이 뉴런으로 정의한다.

차이 설명 생성 : 차이 뉴런들이 표현하는 개념을 설명하기 위해, 해당 level n-1 클러스터의 대표 캡션에는 존재하지만, 상위 대표 캡션에는 나타나지 않는 특징을 추출하여 차이 설명 문장을 생성한다. 이 과정은 ChatGPT 를 사용하여 진행된다.

설명 부착 : 생성된 차이 설명은 해당 차이 뉴런 집합에 의미적 주석으로 부착되며, 이는 모델 내부에서 특화된 의미가 어떤 뉴런들에 의해 나타나는지 분석할 수 있는 단서를 제공한다.

4. 실험결과

계층적 다중 뉴런 해석 프레임워크를 25,000 장의 이미지 데이터셋에 적용한 결과, 딥러닝 모델(Resnet18)의 뉴런 그룹을 계층적 의미 단위로 구조화 할 수 있었다. level1 부터 level6 까지 단계적으로 확장되는 뉴런 그룹과 이에 대응하는 자연어 기반의 설명 라벨이 형성되었다. 특히 상, 하위 뉴런 그룹 간의 차이점 분석을 통해 특화된 의미와 일반화된 의미의 분기점을 뉴런 단위에서 식별할 수 있었다.

5. 결론

본 연구는 기존 단일 뉴런 중심 해석의 한계를 극복하고, 딥러닝 모델 내부의 복잡한 의미구조를 계층적으로 해석할 수 있는 실질적 방법을 제시하였다. 이 프레임워크는 향후 XAI, 모델 편향 탐지, AI 안정성 평가 등 다양한 분야에 활용될 수 있는 확장 가능하고 실용적인 해석 도구로서 의의를 가진다. 특히 자율주행, 의료 영상 판독 등과 같이 모델의 예측 결과에 대한 신뢰성과 설명 가능성이 필수적인 분야에서는, 해석 가능한 뉴런 단위 설명이 실제 현장 적용을 위한 전제 조건이 된다. 본 연구에서 제안한 프레임워크는 위와 같은 분야에서 뉴런 그룹의 계층적 분석을 통해, 입력 데이터가 모델 내부에서 어떤 의미 단위로 처리되는지 언어적으로 설명함으로써, 모델의 의사결정 과정에 대한 통찰과 신뢰성을 확보하는 효과를 기대할 수 있을 것이다.

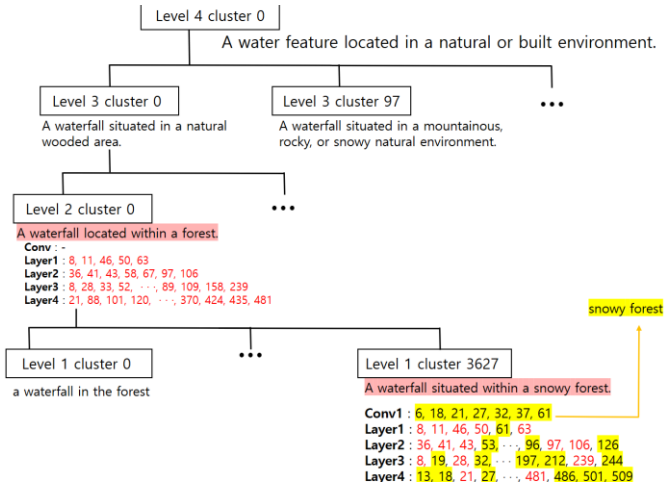

Figure 3: 계층적 라벨링 수행 결과의 일부 구성을 나타낸 도표이다. 상위 level 로 갈 수록 더 일반화된 캡션이 붙는 것을 확인할 수 있다. level 1 cluster 3627 의 활성화 뉴런 중, 상위 level 의 활성화 뉴런과 중복되지 않는 뉴런은 '차이 뉴런'으로, 노란색으로 하이라이트 표시하였다. 분홍색으로 하이라이트 표시된 두 대표 캡션의 차이점을 추출하여 'snowy forest'라는 '차이 설명'을 생성하였다. 이는 차이 뉴런에 라벨링 되었다.

참고 문헌

[1] Bau, D. et al., *Network Dissection: Quantifying Interpretability of Deep Visual Representations*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6541-6549, 2017.

[2] Bau, D. et al., *GAN Dissection: Visualizing and Understanding Generative Adversarial Networks*, International Conference on Learning Representations (ICLR), 2019.

[3] Hernandez, D. et al., *Natural Language Descriptions of Deep Visual Features*, Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 12465-12477, 2021.

[4] Sarfraz, F. et al., *Efficient Parameter-Free Clustering Using First Neighbor Relations*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8935-8944, 2019.

[5] Li, J. et al., *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*, Proceedings of the 39th International Conference on Machine Learning (ICML), vol. 162, pp. 12888-12900, 2022.

[6] Fong, R. C., Patrick, M., & Vedaldi, A. Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks. CVPR 2018, pp. 8730-8738, 2018.