

CBM-KAN: KAN 기반 최종 분류기 도입을 통한 CBM 설명 가능성 확장

송재호¹ · 오상원² · 오승민² · 김광기³ · 한민수⁴ · 김진술^{5*}

¹전남대학교 지능전자컴퓨터공학과 석사과정

²전남대학교 지능전자컴퓨터공학과 박사과정

³나사렛대학교 IT인공지능학부 교수

⁴아스타나IT대학교 컴퓨터 및 데이터 과학과 교수

⁵전남대학교 지능전자컴퓨터공학과 교수

CBM-KAN: Extending Concept Bottleneck Model Interpretability by Introducing a Kolmogorov-Arnold Network-Based Final Classifier

Jaeho Song¹ · Sangwon Oh² · Seungmin Oh² · Kwangki Kim³ · Minsoo Hahn⁴ · Jinsul Kim^{5*}

¹Master's Course, Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea

²Ph.D. Course, Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea

³Professor, Department of IT Artificial Intelligence, Nazarene University, Cheonan 31172, Korea

⁴Professor, Department of Computational and Data Science, Astana IT University, Astana 010000, Kazakhstan

⁵Professor, Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju 61186, Korea

[요약]

설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI)의 핵심 방법론인 CBM(Concept Bottleneck Models)은 중간 개념을 통해 높은 설명력을 제공하지만, 최종 분류기가 블랙박스 MLP(Multilayer Perceptron)로 구성되어 설명의 연쇄가 단절되는 근본적인 한계를 가진다. 이 문제를 해결하기 위해, 내재적으로 설명 가능한 KAN(Kolmogorov-Arnold Network)을 활용한 CBM-KAN 아키텍처를 제안한다. 제안하는 모델은 Top-1 정확도 80.53%를 달성하여 기존 모델의 성능(79.75%)과 동등한 성능을 보였다. 또한, KAN의 활성화 함수를 직접 시각화함으로써, 기존 사후 설명 방식으로는 불가능했던, 개념 신뢰도 값과 최종 분류 사이의 명시적인 함수 관계를 파악할 수 있었다. 이를 통해 제안 모델이 성능 저하 없이 최종 분류 단계의 설명 가능성을 확보할 수 있는 효과적인 방법론임을 입증하였다.

[Abstract]

Concept bottleneck models (CBM), a core methodology of eXplainable Artificial Intelligence (XAI), provide high interpretability through intermediate concepts; however, because the final classifier is a black-box multilayer perceptron (MLP), it has a fundamental limitation in that the chain of explanation is severed. To address this problem, we propose a CBM-KAN architecture that employs an intrinsically interpretable Kolmogorov-Arnold network (KAN). The proposed model achieved a Top-1 accuracy of 80.53%, exhibiting a performance equivalent to that of the existing model (79.75%). Furthermore, by directly visualizing KAN's activation functions, we were able to identify an explicit functional relationship between concept confidence values and the final classification, which has been impossible with conventional post-hoc explanation methods. These results demonstrate that the proposed model is an effective approach for securing interpretability at the final classification stage without performance degradation.

색인어 : 개념 병목 모델, 컴퓨터 비전, 딥러닝, KAN, 설명 가능한 인공지능

Keyword : Concept Bottleneck Models, Computer Vision, Deep Learning, KAN, XAI

<http://dx.doi.org/10.9728/dcs.2025.26.11.3217>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 30 August 2025; **Revised** 25 September 2025

Accepted 02 October 2025

***Corresponding Author, Jinsul Kim**

Tel: +82-62-530-1808

E-mail: jsworld@jnu.ac.kr

I. 서론

의료 진단, 금융 심사, 자율주행 등 고신뢰 영역에서 인공지능 모델의 실제 배치는 단순한 예측 정확도를 넘어 의사결정 과정의 투명성을 요구한다. 특히 잘못된 결정이 생명이나 재산에 직접적 피해를 초래할 수 있는 상황에서는 모델이 어떤 근거로 특정 결론에 도달했는지를 사람이 이해하고 검증할 수 있어야 한다. 이는 설명 가능한 인공지능(eXplainable Artificial Intelligence, XAI)연구의 핵심 동기이며, 단순한 사후 설명을 넘어 모델 내부 추론 과정 자체를 설명 가능하도록 설계하는 방법론에 관한 관심이 높아지고 있다[1].

이러한 요구사항을 해결하기 위한 방법론 중 하나인 CBM(Concept Bottleneck Models)은 입력 X 에서 사람이 이해 가능한 개념(Concept) 집합 C 를 추론하고, 이를 통해 최종 클래스 라벨(label) Y 를 예측하는 2단 구조($X \xrightarrow{g(x)} C \xrightarrow{f(c)} Y$)를 취한다[2]. 이처럼 CBM은 최종 예측 이전에 이해 가능한 중간 단계인 개념을 통하여 높은 설명 가능성과 개입 가능성을 가지는 XAI 설계로 주목받고 있다. 최근에는 사후 설명 기법을 통해 예측 근거의 시각화·정량화 방법과 개념 주석 부담을 낮추기 위한 연구가 활발히 진행되고 있다.

그럼에도 CBM이 최종 의사결정 단계까지 투명성을 자동으로 보장되지는 않는다. 실제 구현에서는 $f(c)$ 는 일반적으로 선형 분류기 혹은 MLP(Multilayer Perceptron)로 구성된다. 하지만 선형 모델은 개념 간 복잡한 상호작용을 포착하기 어렵고, MLP는 이를 유연하게 학습하지만, 내부 작동은 블랙박스에 가깝다. 그 결과 ‘주어진 개념 값이 최종 예측 라벨에 어떤 함수적 관계로 작용하는가?’에 대한 직접적인 답을 내리기 어렵다. 사후 설명은 학습 이후의 근사에 머물러 한계가 있다. 이러한 방식은 각 개념의 개별적 중요도는 확인할 수 있지만, 개념의 신뢰도 값 변화가 최종 결정에 어떤 함수적 관계로 영향을 미치는지 그 명확한 메커니즘을 보여주지는 못한다. 따라서 중간 표현이 설명 가능하더라도 마지막 단계가 불투명하면 설명의 연쇄가 단절된다.

기존 CBM의 한계를 보완하려는 여러 시도에도 불구하고, 최종 클래스 분류 함수 자체를 근본적으로 설명 가능한 모델로 대체하려는 연구는 미미하였다. 이에 본 논문은 CBM의 최종 클래스 분류 함수 $f(c)$ 에 KAN(Kolmogorov-Arnold Network)을 도입하여 개념과 라벨 간의 관계를 함수 형태로 직접 관찰할 수 있도록 한다[3]. 기존의 MLP 기반의 최종 분류 단계와는 다르게 사후 설명 없이 모델의 학습과 동시에 설명을 제공하며, CBM의 설명 가능성을 최종 분류 단계까지 확장하여, 고신뢰 영역에서의 AI 모델 수용성과 신뢰도를 한 단계 높일 수 있다.

이 연구의 핵심 기여는 CBM의 최종 분류기를 KAN 기반 분류기로 대체하여 기존 정확도는 유지하면서 최종 단계의 내재적 설명성을 확보한 데 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 CBM 계열 모

델과 사후 설명, KAN 관련 연구 동향을 정리하고, 제3장에서는 제안하는 CBM-KAN 아키텍처와 학습·손실 구성을 기술한다. 제4장에서는 실험 설정과 결과를 분석하며, 제5장에서는 결론과 향후 과제를 논의한다.

II. 관련 연구: CBM, 사후 설명 기법, KAN

2-1 Concept Bottleneck Models 및 주요 변형

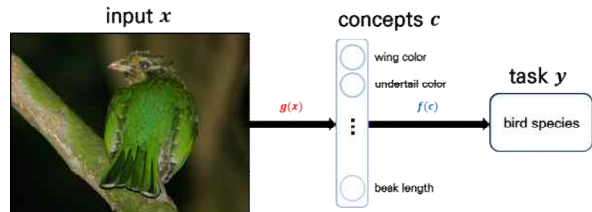


그림 1. Concept Bottleneck Model의 기본 구조
Fig. 1. The basic architecture of a Concept Bottleneck Model

CBM은 Koh et al.에 의해 처음 제안된 설명 가능한 딥러닝 아키텍처로, 입력에서 최종 분류까지의 과정을 두 단계로 분리한다. 그림 1과 같이 입력 이미지 X 에서 사람이 이해 가능한 개념 집합 C 를 추론하는 $g(x)$ 단계, 개념 C 를 통해 최종 클래스 라벨 Y 를 예측하는 $f(c)$ 단계로 이루어져 있다.

CBM의 핵심 기여점은 중간 bottleneck 표현을 의미론적으로 이해 가능한 개념들로 구성한다는 점이다. 그림 1에서 확인할 수 있듯, 새 분류 작업에서 ‘날개 색’, ‘꼬리 색’, ‘부리 길이’ 등의 개념들이 중간 표현으로 사용된다. 예를 들어, 새 사진(X)이 입력되면 모델은 ‘부리의 모양이 뾰족하다.’, ‘머리 색이 붉다.’와 같은 개념 집합 C 를 먼저 예측하고, 이러한 개념들을 통해 ‘홍관조’(Y)라는 최종 분류를 한다. 이는 기존의 End-to-End 모델의 고차원 잠재 표현과 달리, 각 표현이 명확한 의미를 가지며 사람이 직관적으로 이해할 수 있다.

개입(Intervention)은 CBM의 또 다른 핵심 특징이다. 테스트 시점에서 예측된 개념 값을 수정하고 이에 따른 최종 분류의 변화를 관찰함으로써, 각 개념이 최종 결정에 미치는 인과적 영향을 확인할 수 있다. 이러한 특성은 모델의 추론 과정을 검증하고 편향을 발견하는데 유용하다. 또한, 모델이 실제로는 ‘초록색 날개’를 가지는 새를 입력받았으나 ‘날개 색이 빨강다’라고 중간 개념을 잘못 예측하여 최종 분류를 잘못 내렸을 경우, 외부에서 이를 수정하여 모델이 더 정확한 분류를 하도록 유도할 수 있다.

이러한 장점에도 불구하고 CBM은 개념-분류 함수 $f(c)$ 의 불투명성이라는 근본적인 한계를 가진다. 대부분의 구현에서 $f(c)$ 는 선형 분류기 혹은 MLP로 구성되는데, 선형 모델은 개념 간 복잡한 상호작용을 모델링하지 못하고, MLP는 높은 표현력을 가지지만 내부 동작의 해석이 어렵기 때문이다.

CBM의 후속 연구들은 이러한 한계 속에서 다양한 방향으

로 연구가 진행되었다. 대표적인 예시로, Post-hoc CBM은 사전 학습된 모델로부터 개념을 추출하는 방법을 제안하였다 [4]. 이러한 사후 분석 기법들과 더불어 CBM의 가장 큰 제약 중 하나인 개념 주석 데이터 구축의 어려움을 해결하기 위한 방향으로 주로 발전하였다. GPT(Generative Pre-trained Transformer)등을 활용해 개념을 생성·정제하는 Label-Free CBM[5], 비전-언어 모델(Vision-Language Model)을 활용하여 개념 학습을 보조하는 VLG-CBM[6], 대규모 언어 모델(Large Language Models, LLM)을 병목 구조에 통합한 접근 [7], 연속 학습 시나리오[8], Incremental/Residual 구조[9] 등이 제안되었다. 그러나 이들 연구 모두 개념-분류 함수의 설명 가능성 문제는 근본적으로 해결하지 못하였다.

2-2 사후 설명 기법(Post-hoc Explanations)과 SHAP

이미 학습이 완료된 모델을 블랙박스(black-box)로 간주하고, 모델의 예측 결과를 설명하기 위해 사용되는 기법을 사후 설명(post-hoc explanations)이라고 한다. 이미지 분야에서는 CAM(Class Activation Mapping)계열의 연구들이 CNN(Convolutional Neural Network)의 마지막 레이어(layer)의 활성화 맵을 활용하여 예측에 근거가 되는 이미지상의 중요 영역을 시각화한다[10]-[13]. 또 다른 대표적인 기법으로는 게임 이론(game theory)에 기반한 SHAP(SHapley Additive exPlanations)이다[14]. SHAP은 특정 예측값에 대한 각 입력 특성(feature)의 기여도를 정량적으로 계산하여, 어떤 모델에도 적용할 수 있는 model-agnostic 방식으로 모델의 결정을 투명하게 설명한다.

이러한 SHAP의 특성은 CBM의 설명 가능성을 보완하는데 사용될 수 있다. 앞서 지적한 CBM의 불투명한 개념-분류 함수 $f(c)$ 에 SHAP을 적용할 수 있다. 이때, 입력 특성은 CBM의 개념들이 되며, SHAP 분석을 통해 각 개념이 최종 분류 과정에서 어떤 개념이 중요하게 작용했는지에 대한 사후적으로 추정할 수 있다.

하지만 SHAP을 포함한 모든 사후 설명 기법은 이미 학습 완료된 블랙박스 모델의 입출력 관계를 외부에서 근사적으로 추정한다는 근본적인 한계를 가진다. 즉, 각 개념의 중요도는 대략적으로 알 수 있지만, 개념 값 변화가 최종 분류에 어떤 함수적 관계로 작용하는지에 대한 내부 매커니즘을 직접 제시하지는 못한다. 따라서, 이러한 사후 분석의 추정적 한계를 넘어 결정 과정 자체의 투명성을 확보하려면, 모델 내부 구조에 설명 가능성을 직접 부여하는 접근이 필요하다.

2-3 Kolmogorov-Arnold Networks

KAN은 다변수 함수를 간단한 1차원 함수들의 합과 중첩으로 분해할 수 있다는 KAT(Kolmogorov-Arnold representation Theorem)에 이론적 기반을 둔 신경망 아키텍처이다. KAT에 따르면, 수학적으로 모든 다변수 함수 f 는

수식 (1)과 같이 표현될 수 있다.

$$f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right) \quad (1)$$

이 정리는 복잡한 고차원의 함수도 내부 함수 ϕ 와 외부함수 Φ 라는 두 종류의 1차원 함수들로 분해 가능함을 시사한다. 여기서 n 은 입력 차원, x_p 는 p 번째 입력값이며, q 은 외부 함수 Φ 인덱스 값이다.

KAN은 이러한 KAT를 학습 가능한 신경망 형태로 재해석한 것이며, 그림 2에서 볼 수 있듯, 기존의 MLP가 노드(node)에 고정된 활성화 함수를 사용하는 것과 달리, KAN은 엣지(edge)에 학습 가능한 1차원 활성화 함수를 배치하여 KAT를 모방하며, 각 입력 변수가 출력에 미치는 영향을 개별적으로 학습한다. 이러한 제안 방법은 모델의 예측 정확도 성능의 저하 없이 최종 예측 단계의 설명 가능성을 확보한다.

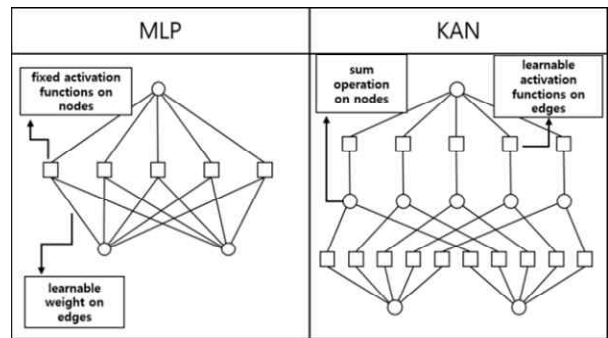


그림 2. MLP와 KAN의 구조적 비교
Fig. 2. Structural comparison of MLP and KAN

KAN의 하나의 레이어는 KAT의 내부 합 $\sum \phi_{q,p}(x_p)$ 에 대응하도록 설계되며, 입력 벡터 (x_1, \dots, x_n) 가 주어졌을 때, 다음 레이어의 j 번째 출력값은 수식 (2)와 같이 계산된다.

$$x_{l+1,j} = \sum_{i=1}^{n_l} (\phi_{i,j,i}(x_{l,i})) \quad (2)$$

수식 (2)에서 l 은 레이어 인덱스이며 n_l 은 노드 수이다. i 는 l 번째 레이어의 노드 인덱스이고 j 는 $l+1$ 번째 레이어의 노드 인덱스를 뜻한다. ϕ 는 이전 레이어와 이후 레이어를 연결하는 엣지에 위치한, 학습 가능한 1차원 함수이다. KAN에서는 이러한 1차원 함수를 유한한 수의 기저 함수(basis function)의 선형 결합으로 근사하는데 주로 B-spline이 사용되며 수식 (3)과 같이 구성된다.

$$\phi(x) = w_b b(x) + w_s \sum_i c_i B_i(x) \quad (3)$$

수식 (3)에서 $b(x)$ 는 잔차 기본 함수이며 보편적으로는 SiLU(Sigmoid Linear Unit)를 사용한다. $B_i(x)$ 는 미리 정

의된 B-spline 기저 함수이다. w_b, w_s 는 각 함수에 대한 가중치이며, c_i 는 학습을 통해 최적화되는 계수이다. 학습이 완료되면, 각 계수 c_i 가 결정되어 1차원 함수 ϕ 의 최종적인 형태가 만들어진다.

결론적으로, KAN의 이러한 구조는 모델에 내재된 설명 가능성(Inherent Interpretability)을 부여한다. 각 엣지의 학습된 $\phi_{i,j}$ 를 직접 시각화함으로써, 특정 입력 개념(c_i)이 최종 클래스(y_i)에 어떤 명시적인 함수 관계로 기여하는지 직관적으로 파악할 수 있다. 이는 기존 CBM이 가지는 개념-분류 함수 $f(c)$ 의 불투명성 문제를 해결할 가능성을 보인다.

III. CBM-KAN 아키텍처와 학습 및 손실 구성

3-1 CBM-KAN의 전체 구조

본 논문에서는 CBM의 최종 분류 함수 $f(c)$ 가 가지는 설명 가능성의 한계를 극복하기 위해, KAN을 결합한 새로운 CBM-KAN 아키텍처를 제안한다. CBM의 장점인 중간 개념 표현의 설명 가능성은 그대로 유지하면서, 최종 클래스 분류 단계의 설명 가능성 또한 확보하는 것을 목표로 한다.

제안하는 모델 전체 데이터 흐름은 그림 3과 같다. 입력 이미지(X)는 기존 CBM과 동일한 방식인 ImageNet으로 사전 학습된 Inception-v3 백본 네트워크 $g(x)$ 를 통과하여, 사람이 이해할 수 있는 개념 집합 C 을 생성한다. 예를 들어, 조류 이미지에 대해 ‘부리의 모양이 뾰족하다.’와 같은 각 개념의 신뢰도를 예측하여 개념 집합 C 을 생성한다. 이후, 생성된 개념 벡터는 제안 모델의 핵심인 KAN 기반의 최종 분류 $f(c)$ 의 입력으로 전달된다. KAN 분류기는 각 개념의 신뢰도 값이 최종 클래스 분류에 미치는 함수적 관계를 학습하여 최종 클래스 라벨 Y 를 예측한다. 이는 각 개념의 중요도만 알려주었던 기존의 CBM 모델과 달리 개념 값의 변화가 예측에 ‘어떻게’ 영향을 미치는지 직접 확인할 수 있어 설명의 연쇄가 단절되는 기존의 문제점을 해결할 수 있다.

3-2 KAN 기반 분류기 구현

제안 모델의 핵심인 KAN 기반 분류기는 입력을 통해서 개념을 예측하는 $g(x)$ 의 출력을 입력으로 받아 최종 클래스를 분류하는 단일 KAN 레이어 구조를 갖는다. KAN이 학습 가능한 1차원 함수 ϕ 를 구현함에 있어서, KAN 논문의 원저자가 제안하는 B-spline 대신하여 가우시안 RBF(Radial Basis Function)를 기저 함수로 사용하였다. B-spline은 높은 표현력을 가지나 다수의 기저 함수로 인하여 느린 학습 속도를 보인다. 이를 보완하기 위해 RBF를 활용하는 방법[15], 사인(sinusoidal) 기반 함수를 사용하는 방법[16], 위 두 가지 방법을 혼합 적용하는 방법[17] 등을 통해 B-spline 함수를 효율적으로 근사하려는 연구가 주목받고 있으며, 본 논문

에서도 이러한 접근 방식들 중 RBF를 채택하였다. RBF 기반 활성화 함수는 수식 (4)와 같다. 수식 (4)의 μ 은 RBF의 중심, σ 은 표준편차를 의미한다.

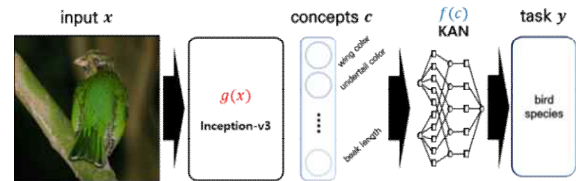


그림 3. 제안하는 CBM-KAN 모델 아키텍처
Fig. 3. The proposed CBM-KAN architecture

$$\phi(x) = \sum_i c_i \exp\left(-\frac{(x - \mu_i)^2}{2\sigma^2}\right) \tag{4}$$

3-3 모델 학습 및 손실 함수

제안하는 CBM-KAN은 개념을 예측하는 $g(x)$ 와 KAN 기반 최종 분류 함수 $f(c)$ 를 End-to-End 방식으로 함께 학습시키는 Joint Training 방식을 사용한다. 이를 위해 전체 손실 함수는 최종 클래스 분류의 정확도를 위한 분류 손실과 중간 개념 예측의 정확도를 위한 개념 손실의 가중합으로 구성된다. 전체 손실 함수는 수식 (5)와 같다.

$$L_{total} = L_{label} + \lambda L_{concept} \tag{5}$$

수식 (5)에서의 L_{label} 은 분류 문제에서 표준적으로 사용되는 Cross Entropy 손실 함수이며, $L_{concept}$ 은 각 개념의 유무를 예측하는 다중 이진 분류 문제이므로 Binary Cross Entropy 손실 함수를 사용한다. λ 는 두 손실 간의 중요도를 조절하는 하이퍼파라미터이다.

IV. 실험 및 결과

4-1 실험 설계

1) 데이터셋(Dataset)

본 논문에서는 CBM 계열에 있어서 표준 데이터셋이라고 할 수 있는 조류 이미지 분류 데이터셋인 CUB-200-2011 데이터셋을 사용하였다[18]. 해당 데이터셋은 200종의 조류에 대한 총 11,788개의 이미지로 구성되어 있으며, 각 이미지에는 312개의 이진(binary) 개념 주석이 함께 제공된다. 다만, 모든 개념이 충분한 학습 샘플을 가지는 것은 아니므로, CBM 논문과 동일하게 데이터셋에서 최소 10개 이상의 샘플을 가지는 개념만을 선별하여 실험을 진행하였다. 이 과정을 통해 최종적으로 111개의 개념이 실험에서 사용되었다. 데이터 분할은 CUB-200-2011 데이터셋에서 제공하

는 공식적인 학습(train) 및 테스트(test) 분할을 사용하였다.

2) 비교 모델(Baseline Model)

본 논문에서 제안 모델의 효과를 검증하기 위해 CBM의 표준 구조인 CBM-MLP를 핵심 비교 모델로 사용하였다. 이 모델은 최종 클래스 분류 함수 $f(c)$ 가 MLP로 구성되어 있으며, 제안 모델과 동일한 방식인 Joint 방식으로 학습되었다.

3) 구현 상세(Implementation Details)

• 백본 네트워크(Backbone Network)

CBM-MLP와 CBM-KAN 모델의 개념 추출 함수 $g(x)$ 는 공정하게 비교하기 위하여, CBM 논문에서 진행한 실험과 동일하게 ImageNet으로 사전 학습된 Inception-v3 모델을 백본으로 사용하였다.

• KAN 예측기(KAN Predictor)

제안 모델의 $f(c)$ 는 111개의 개념을 입력받아 200개의 클래스를 예측하는 단일 KAN 레이어로 구성되었으며, 활성화 함수의 기저 함수로는 가우시안 RBF를 사용하였다.

• 학습 파라미터(Training Parameters)

모든 모델은 SGD(Stochastic Gradient Descent)를 사용하여 학습되었으며, 초기 학습률(learning rate)은 0.001, 가중치 감쇠(weight decay)는 0.0004로 설정하였다. 학습은 100에폭 진행하였으며, 배치 사이즈는 128을 사용하였다. 전체 손실 함수인 수식 (5)에서의 λ 는 0.01로 설정하고 실험을 진행하였다.

4-2 실험 결과

1) 분류 성능

본 논문에서는 분류 성능을 정확도(Accuracy, Acc)로 평가하고 Top-1/Top-5 지표를 보고한다. 개념 예측 성능은 개념 정확도(Concept Acc)로 측정한다. CUB-200-2011 데이터셋에서 모델별 최종 분류 및 개념 예측 성능은 표 1과 같다. 제안하는 CBM-KAN 모델은 Top-1 정확도 80.53%, Top-5 정확도 96.01%를 기록하며, 기존의 CBM-MLP의 성능(Top-1 79.75%, Top-5 95.63%)과 비교했을 때 소폭 향상된 분류 성능을 보였다. 개념 정확도의 경우, 두 모델 모두 약 91% 수준의 높은 정확도를 보여 제안하는 모델이 기존 모델의 개념 학습 능력 및 최종 클래스 분류 능력을 성공적으로 유지하였음을 확인하였다.

표 1. 모델 성능 비교

Table 1. Model performance comparison

Model	Top-1 Acc(%)	Top-5 Acc(%)	Concept Acc(%)
CBM-MLP	79.75	95.63	91.06
CBM-KAN	80.53	96.01	90.78

2) 설명 가능성 분석

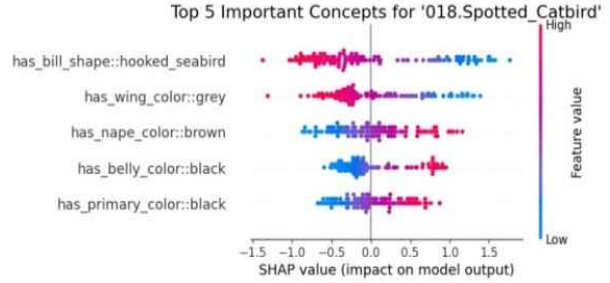


그림 4. CBM-MLP에 대한 SHAP 개념 중요도

Fig. 4. SHAP concept importance analysis for the CBM-MLP

CBM-MLP의 설명 가능성을 분석하기 위해, 앞서 설명한 게임 이론에 기반한 사후 설명 기법인 SHAP을 사용하여 각 개념의 중요도를 측정하였다. SHAP은 특정 예측 및 분류에 대한 각 개념의 기여도를 공정하게 배분하여 어떤 개념이 예측에 긍정적 또는 부정적인 영향을 미쳤는지 정량적으로 보여준다.

그림 4는 특정 클래스 예측에 대한 SHAP 요약 플롯(summary plot)으로, 예측에 가장 큰 영향을 미친 상위 5가지의 개념들을 보여준다. 세로축은 각 개념이 나열되어 있으며, 하나의 점은 개별 테스트 샘플을 의미하며, 점의 가로축 위치는 해당 샘플에 대한 SHAP값(기여도)을 나타낸다. 점의 색상은 개념의 값(빨간색은 높음, 파란색은 낮음)을 의미한다.

이처럼 SHAP 분석을 통해 어떤 개념이 중요한지와 그 영향의 방향성까지 파악할 수 있었지만, 이는 모델의 입출력 관계를 외부에서 추정한 결과일 뿐, 개념의 신뢰도 값 변화에 따라 분류가 '어떻게' 변하는지에 대한 명시적인 함수 관계를 설명하지 못하는 한계가 있다.

반면, 제안하는 CBM-KAN 모델은 각 개념과 최종 분류 사이의 명시적인 함수 관계를 직접 시각화할 수 있다. 그림 5는 KAN 자체 중요도 분석을 통해서 상위 4가지 주요 개념에 대해 KAN이 학습한 1D 활성화 함수를 시각화한 것이다.

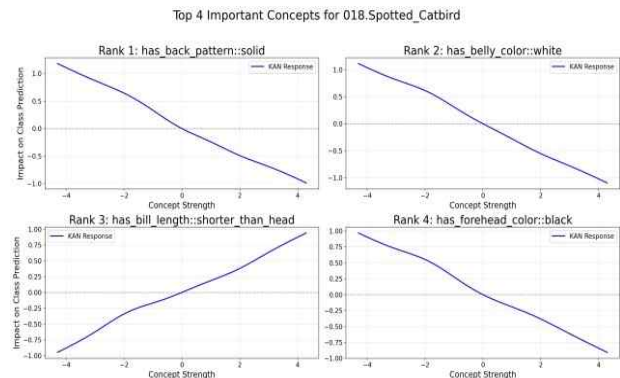


그림 5. 주요 개념(Top 4)에 대해 학습된 KAN 활성화 함수

Fig. 5. Learned KAN activation functions for top 4 concepts

그래프는 개념의 신뢰도(가로축)가 증가함에 따라 최종 분류에 미치는 영향(세로축)을 보여준다. 4가지 그래프 모두 선형적인 모습을 보여주며, Rank 3을 제외한 그래프들은 각 개념의 신뢰도 값이 증가함에 따라 해당 클래스로의 분류에 부정적인 영향을 준다. 반면, Rank 3의 그래프는 개념의 신뢰도 값이 증가함에 따라 해당 클래스로의 분류에 긍정적인 영향을 주는 것을 확인할 수 있다. 따라서 그림 5는 개념 신뢰도 변화가 최종 분류에 미치는 방향·강도 등을 직관적으로 보여주며, 이는 MLP 기반 CBM의 블랙박스형 최종 분류 단계와 달리 사후 설명 없이 연구자가 개념-라벨 함수 형태를 그래프로 직접 검증할 수 있게 한다는 점에서 근본적인 차이가 있다.

V. 결론

본 논문은 고신뢰 인공지능 시스템에서 요구되는 예측 과정의 투명성을 확보하기 위해 CBM의 설명 가능성을 최종 클래스 분류 단계까지 확장하는 방법을 제안하였다. 기존 CBM은 중간 개념을 통해 설명력을 제공하지만, 개념에서 최종 클래스 예측으로 이어지는 함수가 MLP와 같은 블랙박스로 구성되어 설명의 완결성이 약하다는 한계가 있다. 이에 최종 클래스 분류 단계를 KAN을 결합한 CBM-KAN 구조를 설계하여 개념과 클래스 간의 관계를 직관적으로 파악할 수 있음을 보였다. CUB-200-2011 데이터셋에서 제안하는 모델인 CBM-KAN은 Top-1 Acc 80.53%, Top-5 Acc 96.01%를 기록하여, 기존 모델 CBM-MLP의 성능 대비 소폭 개선을 보였다. 개념 정확도는 두 모델 모두 약 91%로 유사하여, 제안 모델이 기존 모델의 개념 예측 및 클래스 분류 성능을 보존하며 설명 가능성을 확보했음을 확인하였으며, 사후 설명 중요도 나열을 넘어 개념이 ‘어떻게’ 클래스 예측에 기여하는지를 드러낼 수 있었다.

본 논문의 의의는 기존 CBM의 성능을 유지하면서 End-to-End 설명 가능성을 강화하고, 간단하고 일반적인 전략을 제시했다는 점에 있다. 한편, 본 실험은 단일 데이터셋에서만 평가되었으며, 동일 실험 설정에서 CBM-KAN의 학습 시간이 기존 MLP 기반 CBM 대비 약 5% 증가하는 한계가 확인되었다. 향후 다양한 도메인의 데이터로 일반화 성능 검증, KAN 최적화를 통한 모델 경량화 및 추론 시간 개선을 진행할 예정이다.

감사의 글

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.RS-2021-II212068, 인공지능 혁신 허브 연구 개발)과 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행

된 연구임(No.RS-2022-II220215, 다시점 자유도를 제공하는 OTT 플레이어 지능화 기술개발).

참고문헌

- [1] Q. Wan, C. Gao, R. Wang, and X. Chen, “A Survey on Interpretability in Visual Recognition,” arXiv:2507.11099, July 2025. <https://doi.org/10.48550/arXiv.2507.11099>
- [2] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept Bottleneck Models,” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Virtual Event, pp. 5338-5348, July 13-18, 2020. <https://doi.org/10.48550/arXiv.2007.04612>
- [3] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, ... and M. Tegmark, “KAN: Kolmogorov-Arnold Networks,” arXiv:2404.19756, 2024. <https://doi.org/10.48550/arXiv.2404.19756>
- [4] M. Yuksekgonul, M. Wang, and J. Zou, “Post-hoc Concept Bottleneck Models,” arXiv:2205.15480, May 2023. <https://doi.org/10.48550/arXiv.2205.15480>
- [5] T. Oikarinen, S. Das, L. M. Nguyen, and T.-W. Weng, “Label-Free Concept Bottleneck Models,” arXiv:2304.06129, May 2023. <https://doi.org/10.48550/arXiv.2304.06129>
- [6] D. Srivastava, G. Yan, and T.-W. Weng, “VLG-CBM: Training Concept Bottleneck Models with Vision-Language Guidance,” *Advances in Neural Information Processing Systems*, 37, 2024. <https://doi.org/10.48550/arXiv.2408.01432>
- [7] C.-E. Sun, T. Oikarinen, B. Ustun, and T.-W. Weng, “Concept Bottleneck Large Language Models,” arXiv:2412.07992, December 2024. <https://doi.org/10.48550/arXiv.2412.07992>
- [8] L. Yu, H. Han, Z. Tao, H. Yao, and C. Xu, “Language Guided Concept Bottleneck Models for Interpretable Continual Learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, USA, pp. 14976-14986, June 2025. <https://doi.org/10.48550/arXiv.2503.23283>
- [9] C. Shang, S. Zhou, H. Zhang, X. Ni, Y. Yang, and Y. Wang, “Incremental Residual Concept Bottleneck Models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle: WA, pp. 11030-11040, June 2024. <https://doi.org/10.1109/CVPR52733.2024.01049>
- [10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A.

Torralba, "Learning Deep Features for Discriminative Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas: NV, pp. 2921-2929, June 2016. <https://doi.org/10.1109/CVPR.2016.319>

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice: Italy, pp. 618-626, October, 2017. <https://doi.org/10.1109/ICCV.2017.74>

[12] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, ... and X. Hu, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, USA, pp. 111-119, June 2020. <https://doi.org/10.1109/CVPRW50498.2020.00020>

[13] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks," arXiv:1710.11063, 2017. <https://doi.org/10.48550/arXiv.1710.11063>

[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach: CA, pp. 4768-4777, 2017.

[15] Z. Li, "Kolmogorov-Arnold Networks are Radial Basis Function Networks," arXiv:2405.06721, 2024. <https://doi.org/10.48550/arXiv.2405.06721>

[16] E. Reinhardt, D. Ramakrishnan, and S. Gleyzer, "SineKAN: Kolmogorov-Arnold Networks Using Sinusoidal Activation Functions," *Frontiers in Artificial Intelligence*, Vol. 7, 1462952, January 2025. <https://doi.org/10.3389/frai.2024.1462952>

[17] H.-T. Ta, "BSRBF-KAN: A Combination of B-splines and Radial Basis Functions in Kolmogorov-Arnold Networks," in *Proceedings of the 13th International Symposium on Information and Communication Technology (SOICT)*, Danang: Vietnam, pp. 3-15, December 2024. https://doi.org/10.1007/978-981-96-4288-5_1

[18] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, The Caltech-UCSD Birds-200-2011 Dataset, California Institute of Technology, Pasadena, CA, Technical Report CNS-TR-2011-001, July 2011.



송재호(Jaeho Song)

2024년 : 전남대학교 물리학(이학사)

2024년~현 재: 전남대학교 지능전자컴퓨터공학과 석사과정
 ※ 관심분야 : 컴퓨터비전, 객체탐지, 딥러닝, XAI



오상원(Sangwon Oh)

2020년 : 전남대학교 컴퓨터정보통신공학과(공학사)
 2023년 : 전남대학교 ICT융합시스템공학과(공학석사)

2023년~현 재: 전남대학교 지능전자컴퓨터공학과 박사과정
 ※ 관심분야 : 인공지능, 딥러닝 생성 모델, 메타버스, 네트워크



오승민(Seungmin Oh)

2019년 : 한국나사렛대학교 디지털콘텐츠학과(공학사)
 2021년 : 전남대학교 ICT융합시스템학과(공학석사)

2021년~현 재: 전남대학교 지능전자컴퓨터공학과 박사과정
 ※ 관심분야 : 딥러닝, 인공지능



김광기(Kwangki Kim)

2002년 : 항공우주대학교 전자엔지니어링학(공학사)
 2004년 : 한국과학기술원 정보통신공학(공학석사)
 2011년 : 한국과학기술원 정보통신공학(공학박사)

2012년~2013년: 삼성전자 DMC연구소
 2013년~현 재: 나사렛대학교 IT인공지능학부 교수
 ※ 관심분야 : 신호처리, 3D 음향, 디지털콘텐츠 등



한민수(Minsoo Hahn)

1979년 : 서울 국립대학교 전자공학(공학사)

1981년 : 서울 국립대학교 전자공학(공학석사)

1985년 : 플로리다 대학(University of Florida) 전자공학(공학박사)

1982년~1985년: 대전 한국표준과학연구원 연구원

1990년~1997년: 한국전자통신연구원 연구원

1998년~현재: KAIST 전자공학과 및 아스타나IT대학교 컴퓨터 및 데이터과학과 교수

※ 관심분야 : 음성 및 오디오 코딩, 음성 합성, 음성 변환, 노이즈 감소, VoIP

김진술(Jinsul Kim)

2001년 : Computer Science from University of Utah, Salt Lake City, Utah, USA (공학사)

2005년 : 한국과학기술원 정보통신공학 (공학석사)

2008년 : 한국과학기술원 정보통신공학 (공학박사)

2005년~2008년: 한국전자통신연구원 IPTV 인프라 기술, 융·복합 방송/통신 분야 연구원

2009년~2012년: 나사렛대학교 멀티미디어학과 교수

2012년~현재: 전남대학교 지능전자컴퓨터공학과 교수

※ 관심분야 : QoS/QoE 예측/분석/관리, 모바일 미디어 처리/통신, 클라우드 컴퓨팅 디지털 미디어 및 네트워크 지능

