논문 2021-16-19

신경망 근사에 의한 다중 레이어의 클래스 활성화 맵을 이용한 블랙박스 모델의 시각적 설명 기법

(Visual Explanation of Black-box Models Using Layer-wise Class Activation Maps from Approximating Neural Networks)

강 준 규[1], 전 민 경[1], 이 현 석, 김 성 찬*
(JuneGyu Kang, MinGyeong Jeon, HyeonSeok Lee, Sungchan Kim)

Abstract : In this paper, we propose a novel visualization technique to explain the predictions of deep neural networks. We use knowledge distillation (KD) to identify the interior of a black-box model for which we know only inputs and outputs. The information of the black box model will be transferred to a white box model that we aim to create through the KD. The white box model will learn the representation of the black-box model. Second, the white-box model generates attention maps for each of its layers using Grad-CAM. Then we combine the attention maps of different layers using the pixel-wise summation to generate a final saliency map that contains information from all layers of the model. The experiments show that the proposed technique found important layers and explained which part of the input is important. Saliency maps generated by the proposed technique performed better than those of Grad-CAM in deletion game.

Keywords : Visual Explanation, Knowledge distillation, Attention, Grad-CAM

Ⅰ. 서 론

인공지능에서의 블랙박스 (Black-Box)는 출력을 인간과 유사하게 혹은 원하는 대로 도출할 수 있지만 무엇을 근거로 그런 결과가 나왔는지 알 수 없는 경우나, 입력과 출력만 알고 모델이 어떤 구조로 이루어져 있는지 알지 못하는 경우를 의미한다 [1]. 본 논문에서는 블랙박스 모델을 해석하기 위한 문제에 접근한다. 예측 정확도가 높아도 모델의 내부를 모르는 경우, 입력에 대해 모델의 출력이 올바르더라도 그 결과를 신뢰하기 어렵다. 따라서 예측에 대한 근거를 설명하는 것은 모델의 결과를 신뢰할 수 있는 근거가 된다 [2, 3].

모델의 예측에 대한 근거를 보여주는 가장 대표적인 방법은 입력의 어느 부분이 예측에 중요한 영향을 끼쳤는지를 시각화하는 것이다 [4-8]. 예측 결과에 대한 중요한 부분을 시각화함으로써, 모델이 입력의 어느 부분을 중요하게 봤는지 이해할 수 있고, 결과를 신뢰할 수 있는 근거가 된다. 기존 연구에서는 모델의 시각적 설명을 만들어내기 위해 모델의 다양한 내부 정보를 활용한다. 특히 gradient, 특징 맵 (feature map) 등이 많이 사용된다. 하지만 모델의 내부 정보는 항상 얻을 수 있는 것이 아니며, 특히 얻을 수 있는 정보가 입력과 출력이 제한된 블랙박스 모델의 경우 기존 시각적 설명 기법을 사용하기 어렵다. 본 논문에서는 기존 연구의 한계를 극복하기 위해 어텐션 (attention)을 이용해 모델의 예측 근거를 설명할 수 있는 시각적 설명 기법을 제안한다. 입력과 출력만 아는 블랙박스 모델의 내부에 접근하기 위해 지식 증류 기법 (Knowledge Distillation, KD) [9]을 사용하여 설명하려는 블랙박스 모델 (teacher network)의 훈련 정보를 화이트 박스 모델 (student network)로 전달한다. KD를 적용하면 모델의 블랙박스 유무, 모델의 종류에 한정되지 않는다. 간단한 student network가 복잡한 teacher network보다 더 좋은 성능을 가지게 하는 기존 KD 기법과 다르게 제안하는 기법의 student network는 teacher network와 동일한 결과를 내도록 훈련된다.

어텐션 맵을 생성하는 부분은 student network를 통해서 이루어진다. Student network에 CBAM [10]을 삽입해 teacher network가 분류 시에 중요하게 본 특징 (feature)을 시각화한 어텐션 맵을 생성한다. 또 생성된 어텐션 맵과 전체 레이어로 확장한 Grad-CAM을 합쳐 더 정교한 돌출 맵을 생성한다.

본 논문은 다음과 같이 구성된다. 2장에서는 논문에서 소개하는 기법과 관련된 배경 정보 및 관련 연구를 소개한다. 3장에서는 제안하는 기법에 관해 설명한다. 4장에서는 실험을 통해 제안하는 기법의 성능을 입증하고 5장에서 결론을 내린다.

[1]Co-first author
*Corresponding Author (s.k@jbnu.ac.kr)
Received: Jun. 22, 2021, Revised: Jul. 9, 2021, Accepted: Jul. 22, 2021.
J.G. Kang: Jeonbuk National University (M.S. Student)
M.G. Jeon: DEEPNOID (M.S.)
H.S. Lee: Jeonbuk National University (Ph.D. Student)
S. Kim: Jeonbuk National University (Prof.)
※ 이 논문은 정부 (과학기술통신부)의 재원으로 한국연구재단 (No. 2019R1F1A1061941) 및 전북대학교 인공지능응용기술연구센터의 지원을 받아 수행된 연구임.

ⓒ IEMEK J. Embed. Sys. Appl. 2021 Aug. 16(4) 145-151
ISSN : 1975-5066
http://dx.doi.org/10.14372/IEMEK.2021.16.4.145

Ⅱ. 배경 정보 및 관련 연구

1. 지식 증류 기법

지식 증류 기법 (Knowledge Distillation, KD)은 간단한 모델도 복잡한 모델과 동일한 성능을 내기 위해 연구되었다. KD는 제한된 컴퓨팅 자원으로 복잡한 모델의 일반화 (generalization) 능력을 얻기 위한 기법이다.

KD는 많은 파라미터 (parameter)가 사용되는 복잡한 신경망 (teacher network, *TN*)에서 일반화 성능을 향상시킬 수 있는 지식 (knowledge)들을 분리해내는 방법이다. KD는 teacher network에서 학습한 지식 (dark knowledge)을 student network (teacher보다 간단한 신경망 모델)로 전달한다.

KD는 teacher network의 최종 결과인 *softmax* 레이어의 출력값이 teacher network가 지니고 있는 지식을 함축해서 담고 있다고 가정한다. 예를 들어, 이미지 분류 모델에는 타겟이 고양이일 때, 모델의 *softmax* 출력값은 고양이 90%, 개 1%, 양 0.01%, 자동차 0.001% 등 다른 클래스의 정보도 가지고 있다. 따라서 *softmax* 출력값은 학습에 도움이 될 수 있는 추가적인 지식을 담고 있다고 본다. 따라서 teacher network가 학습한 지식인 *softmax* 출력값을 student network의 학습에 활용하면 teacher network와 유사한 성능을 낼 수 있다.

$$P_i = \frac{\exp(\frac{z_i}{T})}{\Sigma_j \exp(\frac{z_j}{T})}. \quad (1)$$

위 예시 그림 1에서 양과 차에 대한 확률처럼 일부 클래스들에 대한 확률은 거의 0에 가까워서 학습 시 정보가 잘 전달되지 않을 수 있다. 따라서 수식 (1)의 *softmax* 함수에 확률 분포를 좀 더 부드럽게 만드는 온도 (temperature) 파라미터 T를 추가한다. T가 높을수록 기존보다 더 부드러운 확률 분포(probability distribution)를 얻을 수 있다.

어텐션은 기계 번역 (neural machine translation)에서 입력 문장이 길어질수록 모델의 성능이 나빠지는 문제를 해결

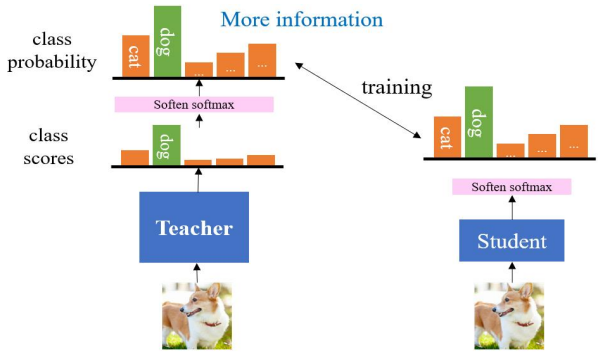


그림 1. Knowledge distillation의 개요
Fig. 1. Summary of knowledge distillation

하기 위해 도입되었다. 모델이 중요한 부분만 집중 (attention)하게 만들자는 것이 핵심 아이디어다. 기존 기계 번역 기법은 출력 단어를 예측할 때 입력 문장의 전체를 동일한 비율로 참고한다. 어텐션 메커니즘은 기존 기계 번역 기법처럼 전체 입력 문장을 참고하지만 출력 시점에서 예측해야 할 단어와 연관이 있는 부분에만 집중한다.

어텐션은 self-attention [11]의 등장으로 CNN에도 영향을 끼치게 되었다. CNN에서는 어텐션 가중치 (weights)를 이용해 각 픽셀 x_j간의 상관관계를 구하고 각 픽셀이 얼마나 주목받는지를 나타내는 어텐션 맵을 생성한다. 최근에는 Residual attention Networks [12], Squeeze-and-Excitation [13], Non-local Neural Network [14] 등 일반적인 모델의 구성요소로 주목 받고 있다.

2. 관련 연구

본 절에서는 시각적 설명 기법을 위한 기존 연구 기법들을 소개한다. 화이트 박스 모델의 대표적인 설명 기법인 Grad-CAM과 블랙박스 모델을 설명하기 위해 prediction difference 기법을 이용하는 세 가지 기법을 소개한다.

대표적인 시각화 기법인 CAM은 CNN 기반 모델에서 클래스를 결정할 때 시각적인 설명을 제공한다. 성능은 좋지만 FC 레이어 대신 GAP 레이어를 사용해야 하고 파인 튜닝 (fine-tuning)이나 재학습 (re-training) 같은 과정을 거쳐야 한다. 이런 문제들 때문에 다양한 목적을 수행하는 CNN에 CAM을 적용하기 어렵다. 이를 극복하기 위해 Grad-CAM이 연구되었다. Grad-CAM은 모델을 재학습할 필요 없이, 기존 모델이 가진 *softmax* 레이어의 입력과 합성곱 레이어 (convolution layer)의 특징 맵 (feature map)만으로 CAM을 얻을 수 있다. Grad-CAM은 gradient를 이용해 중요도를 구하므로, 마지막 합성곱 레이어에만 한정되지 않는다.

Local Interpretable Model-agnostic Explanations (LIME)은 시각적 설명을 위해 결정 경계 (Decision Boundary)를 이용한다. LIME에서는 복잡한 비선형 모형을 전반적으로 해석하는 것이 불가능한 경우가 많지만, 작은 영역에서는 해석이 가능한 형태를 보일 것이라 가정한다. 따라서 모델 자체가 어떤 방식으로 동작하는지 이해하는 대신, 해석하고자 하는 값 근방에서 모델과 비슷하게 행동하는 선형 함수를 이용해 근사하고, 이를 통해 어떤 변수가 중요한 역할을 하는지 알 수 있다. 따라서 이 선형 함수의 예측값이 크게 변하는 변수가 있다면 이 변수가 예측에 결정적인 역할을 한 근거로 볼 수 있다. LIME은 이미지를 슈퍼 픽셀 단위로 나눠서 화소 일부를 회색으로 가려 입력과 유사한 샘플을 생성해 변수를 찾는다.

Randomized Input Sampling for Explanation of Black-box Models (RISE)은 수천 장의 무작위 마스크를 통해 클래스 별로 이미지에서 어떤 픽셀이 해당 클래스로 예측되는데 중요한지를 나타내는 돌출 맵을 만드는 방법을 제안했다. 먼저 입력 이미지에 무작위 마스크를 씌운 이미지

들에 대한 모델의 예측을 구한다. 예측된 클래스별 점수를 가중치로 이용해 해당 마스크들을 합산한 결과물로 돌출 맵을 생성한다.

MASK는 meta-predictor를 이용하여 블랙박스 모델을 설명한다. 설명을 특정 입력에 대한 블랙박스 f의 반응을 예측하는 규칙 Q로 놓고, 예측 오차를 이용해 설명의 신뢰성을 측정한다. 모델을 meta-predictor로 정의하면 신뢰성을 예측 정확도로 측정할 수 있다. 또, 자동으로 많은 규칙들 중에서 f에 적용되는 설명 규칙 Q를 발견할 수 있다. 블랙박스 $f(x)$에 대한 설명을 정의하기 위해 입력 x에 특정 변형을 가한다. 입력 이미지의 픽셀에 상수 값이나 노이즈, 흐리게 하는 세 가지 변형을 통해 새로운 이미지를 생성하고 이 이미지를 통해 $f(x_0)$이 변하는 최소한의 영역인 마스크를 생성한다.

Ⅲ. 어텐션과 확장된 Grad-CAM을 이용한 블랙 박스 심층 신경망 모델의 시각적 설명 기법

1. 블랙박스 심층 신경망 모델의 내부 접근법

입력과 출력만 알고 있는 블랙박스 모델의 예측 근거를 설명하려면 모델의 내부를 파악해야 한다. 블랙박스 모델의 내부를 파악하기 위해, 본 논문에서는 2.1장의 Knowledge Distillation (KD) 기법을 적용한다. 이미지 분류 task를 수행하는 블랙박스 모델이 teacher network (*TN*), 화이트박스 모델이 student network (*SN*)가 된다. 훈련 시 블랙박스 모델의 $softmax$ 출력값을 화이트박스 모델로 전달하고, 화이트박스 모델은 전달받은 정보를 통해 블랙박스 모델과 동일한 예측을 할 수 있도록 훈련된다.

기존 KD 기법은 *TN*보다 간단한 *SN*가 *TN*만큼 우수한 성능을 내도록 수식 (1)처럼 $softmax$ 함수에 파라미터 T를 추가했지만, 본 논문에서는 블랙박스 모델과 동일한 예측을 하는 모델을 구현하는 것이므로 $softmax$ 값이 변형 없이 전달된다. 따라서 T는 1이 되고, 이는 수식 (2)처럼 기존 $softmax$ 함수와 동일해진다.

$$P_i = \frac{\exp(\frac{z_i}{1})}{\Sigma_j \exp(\frac{z_j}{1})} = \frac{\exp(z_i)}{\Sigma_j \exp(z_j)}, \tag{2}$$

$$\begin{aligned} L_{KD} &= D_{KL}(softmax(O_t), softmax(O_s)) \\ &= D_{KL}(\frac{\exp(O_t)}{\Sigma_j \exp(O_j)}, \frac{\exp(O_s)}{\Sigma_j \exp(O_j)}). \end{aligned} \tag{3}$$

입력에 대한 *TN*의 예측 점수가 O_t이고 *SN*의 예측 점수는 O_s일 때, KD 모델의 loss 함수는 수식 (3)과 같다. *SN*의 O_s가 수식 (3)의 Kullback-Leibler divergence loss 함수를 통해 *TN*의 O_t의 확률 분포를 학습하도록 한다. 이를 통해 *SN*가 *TN*처럼 예측할 수 있도록 훈련된다. *SN*는 수식 (4)처럼 Cross Entropy loss 함수를 이용해 입력의 타겟 클래스 c를 맞출 수 있도록 훈련된다. 여기서 $softmax(O_s)$는 *SN* 모델에서 나온 예측값에 $softmax$를 거쳐 실제 모델에서 나오는 출력값이고, c는 그 값의 정답 (클래스)이다. 따라서 KD 모델의 최종 loss 함수는 수식 (5)와 같다.

$$CE(softmax(O_s), c), \tag{4}$$

$$\begin{aligned} L_{KD} &= D_{KL}(softmax(O_t), softmax(O_s)) \\ &+ CE(softmax(O_s), c). \end{aligned} \tag{5}$$

2. 시각적 설명 기법

3.1장의 방법론을 통해 *SN*가 *TN*인 블랙박스 모델과 동일한 예측을 하도록 훈련되었으므로 *SN*에서 생성된 어텐션 맵을 통해 블랙박스 모델의 예측을 시각적으로 설명할 수 있다. 따라서 *SN*에서 어텐션 맵을 생성하기 위해 그림 2의 Attention 부분처럼 CBAM을 *SN*의 ResBlock에 삽입한다.

CBAM은 어텐션 맵을 채널 (channel-wise)/공간별 (spatial-wise)로 나누어 계산한다. CBAM의 각 모듈은 단순한 pooling과 convolution으로 이루어져 있고, 기존 모델에 쉽게 삽입되어 end-to-end 훈련이 가능하다. CBAM은 모델에서 공간 pooling이 이루어지는 병목 구간 (bottleneck)에 위치한다. 병목 구간에서 정보량이 줄기 전에 CBAM을 추가하여, 어텐션으로 중요한 부분의 값을 키우고, 덜 중요한 부분의 값을 줄이도록 한다. 채널 어텐션 모듈을 공간 어텐션 모듈보다 먼저 적용하는 것이 더 좋은 성능을 보여준다. 채널 어텐션 모듈은 합성곱 레이어의 특징 (feature) F를 입력으로 받고, average pool과 max pool, 다층 퍼셉트론 (Multi-Layer Perceptrons, MLP)을 통해 어텐션을 계산한다. sigmoid로 정규화 (normalize)된 어텐션 맵 $M(F)$는 F에 요소별로 (element-wise) 곱해진다. 공간 어텐션도 대칭적으로 구성되어, 하나의 합성곱 레이어로 공간 어텐션을 계산한다. 제안하는 방법론에는 *SN*는 총 16개의 블록으로 이

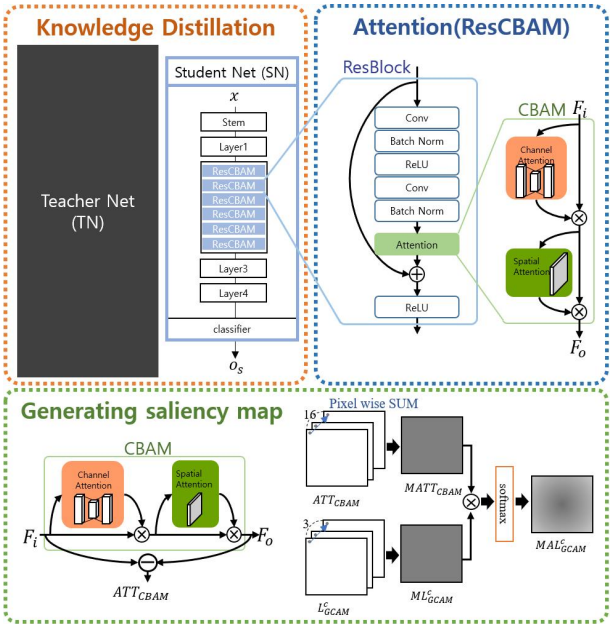


그림 2. 제안하는 모델의 구조
Fig. 2. Structure of the proposed model

루어져 있다. 모든 레이어의 ResBlock 대신 CBAM을 삽입한 ResCBAM (그림 2의 어텐션) 블록을 사용한다.

3. 레이어 별 어텐션 맵을 결합한 돌출맵 생성

제안하는 시각적 설명이 모델의 예측을 잘 설명할 수 있도록 하기 위해 SN에서 생성되는 어텐션 맵과 채널별로 확장한 2.2장의 Grad-CAM을 픽셀 단위로 더한다. 모델이 해당 클래스로 예측할 때 중요한 픽셀을 찾기 위해, 어텐션 맵에서 특정 클래스를 예측하는데 중요한 부분을 찾아야 한다. 따라서 특정 클래스로 예측하는데 중요한 픽셀을 나타내는 돌출 맵과 어텐션 맵을 합쳐 클래스를 예측하는데 중요한 픽셀을 강조한다. 돌출 맵을 생성할 때는 Grad-CAM 기법을 이용한다. 구하려는 Grad-CAM (L^c_{GCAM})은 타겟 클래스 c에 대해 k번째 특징 맵이 가지는 중요도 a^c_k에 마지막 합성곱 레이어의 k번째 특징 맵A^k를 곱하고 가중 합한 값을 ReLU를 거쳐 구한다.

$$L^c_{GCAM} = ReLU(\sum_k a^c_k A^k). \tag{6}$$

SN의 CBAM에서는 각 모듈에서 합성곱 레이어의 feature F_i와 어텐션을 생성하는 함수 $CBAM_{M_s}$를 이용해 어텐션을 순차적으로 계산한다.

$$\begin{aligned} F' &= CBAM_{M_c}(F_i)\otimes F_i \\ F_o &= CBAM_{M_s}(F')\otimes F_i \\ ATT_{CBAM} &= F_o - F_i. \end{aligned} \tag{7}$$

그림 3과 같이 각 CBAM에서 생성된 어텐션 맵 ATT_{CBAM}은 CBAM의 출력 F_o와 입력 F_i의 차를 이용해 구해진다. 레이어별 어텐션 맵을 결합한 돌출 맵을 생성하는 순서는 다음과 같다. 학습이 끝난 SN에서 각 모듈 별로 어텐션 맵 ATT_{CBAM}을 생성한다. Grad-CAM L^c_{GCAM}은 모델의 레이어 4개 중 2, 3, 4번째 레이어의 마지막 블록에서 생성한다. 그 뒤 그림 4와 같이 16개의 ATT_{CBAM}을 픽셀별 합 (pixel-wise sum)을 통해 합쳐 $MATT_{CBAM}$을 생성한다. 생성된 세 개의 L^c_{GCAM}도 픽셀별 합을 통해 하나의 ML^c_{GCAM}을 생성한다. 생성된 $MATT_{CBAM}$과 ML^c_{GCAM}을 곱하고 $softmax$를 거쳐 논문에서 제안하는 MAL^c_{GCAM}을 생성한다.

$$\begin{aligned} MATT_{CBAM} &= (\sum_k ATT_{CBAM_k}) \\ ML^c_{GCAM} &= (\sum_k L^c_{GCAM_k}), \end{aligned} \tag{8}$$

$$MAL^c_{GCAM} = softmax(MATT_{CBAM} \times ML^c_{GCAM}). \tag{9}$$

Ⅳ. 실 험

1. 실험 방법

제안하는 기법을 검증하기 위해서 실험 자료로 ImageNet

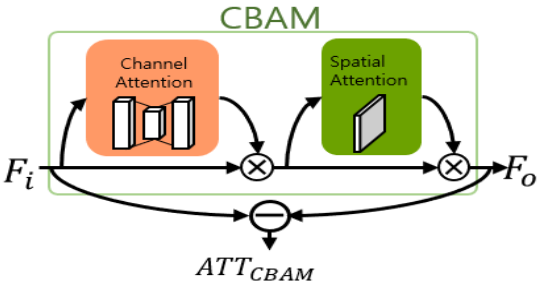


그림 3. CBAM을 이용한 어텐션 맵 생성
Fig. 3. Creating attention maps using CBAM

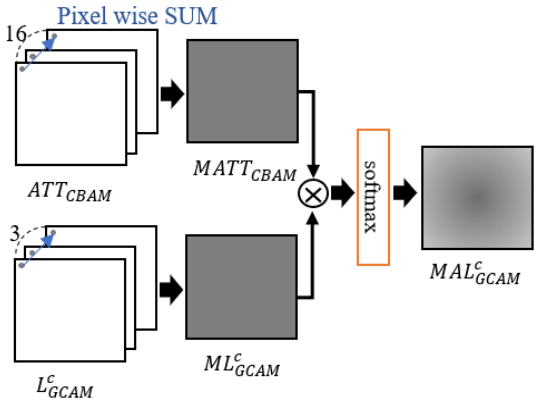


그림 4 ATT_{CBAM}과 L^c_{GCAM}을 MAL^c_{GCAM}으로 병합하는 방법
Fig. 4. Merging ATT_{CBAM} and L^c_{GCAM} into MAL^c_{GCAM}

[15], CUB-200-2011 [16], Stanford Cars [17], FGVC Aircraft [18] (Aircraft) 데이터 셋을 사용한다. Aircraft 데이터의 경우 분류 기준에 따라 클래스가 달라진다. 제안하는 모델 학습에 사용되는 데이터는 224×224 크기로 무작위 위치에서 크롭 한다. 학습에서 검증 단계에 사용되는 데이터는 256×256 크기로 조정하고 224×224 크기로 중앙 크롭 한다. 검증 실험에 사용되는 데이터는 224×224 크기로 변형한다. 또 데이터의 평균을 빼고 표준편차를 나눠서 정규화 한다.

TN로 PyTorch 패키지에서 제공하는 ImageNet 데이터로 사전 훈련된 ResNet-50 모델을 사용하고, SN로 ResBlock 대신 3.2장의 ResCBAM을 삽입한 ResNet-50 모델 사용한다. Fine-grained 데이터 (CUB-200, Stanford Cars, FGVC Aircraft)로 진행되는 실험에서 TN는 ImageNet 데이터로 사전 훈련되어 각 데이터에 맞게 FC 레이어를 변형해 fine-tuning한 ResNet-50 모델이고, SN는 ImageNet 데이터로 사전 훈련된 KD 모델의 SN의 가중치를 이용해 학습한다. 제안하는 모델은 SGD 최적화 기법을 사용하여 100 epoch 동안 훈련되었다. ImageNet 데이터로 훈련될 때는 학습률이 0.1, fine-grained 데이터는 0.01이다. momentum, weight_decay는 0.9, 0.0001이다. 학습률은 PyTorch에서 제공하는 MultiStep 스케줄러를 이용해 조정된다. ImageNet 데이터는 30 epoch 마다, fine-grained 데이터는 30, 50, 60, 70 epoch에서 학습률에 0.1씩 곱해진다.

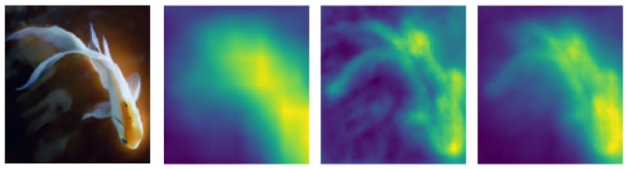
첫 번째 실험에서는 *TN*의 정보가 *SN*로 잘 전달되었는지 평가한다. 이를 검증하기 위해 *TN*와 *SN*의 예측 결과 간의 KD 유사도 (KD similarity)를 측정한다. 본 논문의 실험에서 사용한 *TN*와 *SN*는 입력에 대해 각 클래스의 신뢰도 (confidence)를 출력하게 된다. 여기서 가장 신뢰도가 높은 클래스가 입력에 대한 분류 예측 결과로 사용된다. *TN*와 *SN*의 분류 예측 결과가 같다면 *TN*의 정보가 *SN*로 잘 전달되었다고 평가할 수 있을 것이다. 본 실험에서는 *TN*와 *SN*의 분류 예측 결과만을 비교하는 방법을 Sim 1이라 칭한다. 이에 더해 상대적으로 신뢰도가 낮은 클래스에 대한 정보가 잘 전달되었는지를 Sim N을 통해 측정한다. Sim N이란 각 모델의 출력에 대해 신뢰도가 높은 순서대로 N개 클래스의 순서와 분류가 맞는지를 평가하는 방법이다. 예를 들어 *TN*과 *SN*이 개, 고양이 그리고 물고기를 분류하는 모델이라고 가정해보자. 같은 입력에 대해서 *TN*는 고양이, 물고기, 개의 순서대로 신뢰도가 높았고 *SN*는 고양이, 개, 물고기의 순서대로 신뢰도가 높았다면 Sim 1은 1, Sim 2는 0이 될 것이다.

두 번째 실험에서는 제안하는 시각적 설명 기법이 기존의 시각화 기법보다 효과적인지 확인하기 위해 2.2장의 RISE에서 제안한 AUC (Area Under Curve) 기반의 삽입 (insertion) 게임과 삭제 (deletion) 게임을 통해 정량적으로 평가한다. 삭제 게임은 생성한 어텐션 맵에서 픽셀 점수가 높은 순서대로 픽셀들을 지우면서 새로운 이미지를 생성하고, 해당 이미지가 모델에서 예측된 확률을 그래프로 나타낸다. 이미지의 모든 픽셀이 지워질 때까지 반복한다. 그 후 그래프의 AUC를 계산한다. 어텐션 맵이 잘 생성된다면 예측에 중요한 영역부터 지워지게 되므로 예측 성능이 빠르게 떨어질 것이다. 따라서 AUC가 작을수록 어텐션 맵의 품질이 좋다.

삽입 게임은 이미지를 모두 블러 처리해 흐리게 만든 후 어텐션 점수가 높은 순서대로 해당 픽셀의 블러를 없앤 이미지를 생성하고, 해당 이미지가 모델에서 예측된 신뢰도를 그래프로 나타낸다. 이미지의 모든 픽셀이 복원될 때까지 반복한다. 그 후 그래프의 AUC를 계산한다. 중요한 픽셀이 삽입될수록 모델의 예측 점수가 높아질 것이므로 삽입 게임 그래프의 AUC가 높을수록 어텐션 맵의 품질이 좋다.

2. 실험 결과 및 분석

그림 5. (b)는 *SN*에서 생성된 L^c_{GCAM}을 합친 결과, 그림 5. (c)는 모든 레이어의 ATT_{CBAM}을 합친 결과다. 특정 클래스를 예측하는데 중요한 부분인 그림 5. (b)와 모델이 이미지에서 집중하는 영역인 그림 5. (c)가 서로 다른 양상을 보임을 확인할 수 있다. 따라서 이미지에서 모델이 집중하는 부분 그림 5. (c) 중 특정 클래스로 예측하는데 중요한 부분 그림 5. (b)를 뽑아내면 모델의 특정 클래스로 예측하는데 중요한 부분을 시각적으로 설명하기 쉽다. 또 둘을 합침으로써 배경이나 다른 물체 같은 불필요한 부분이 포함된 것을 제거할 수 있다. 이를 통해 최종 어텐션 맵인 그림 5. (d)의 MAL^c_{GCAM}을 생성한다.



(a) input (b) Merged L^c_{GCAM} (c) Merged ATT_{CBAM} (d) MAL^c_{GCAM}

그림 5. 제안하는 방법론에서 생성된 돌출 맵, (a) 입력 이미지, (b) 생성된 Grad-CAM을 합친 결과, (c) 모든 레이어의 어텐션 맵을 합친 결과, (d) 합친 어텐션 맵과 Grad-CAM을 합친 결과
Fig. 5. Saliency maps by the proposed approaches: (a) Input image, (b) combining saliency maps of layers using Grad-CAM only, (c) combining the attention maps of layers, and (d) combining the attention maps and the results from Grad-CAM

표 1. ImageNet 데이터에서 *TN*와 *SN*의 KD 유사도
Table 1. KD similarity of *TN* and *SN* on ImageNet dataset

	Sim 1	Sim 2	Sim 3	Sim 4	Sim 5
Similarity	0.807	0.701	0.633	0.574	0.534

표 2. 각 기법 별 삭제 게임 결과
Table 2. Deletion game results for each of four models

Data type	Grad-CAM	LIME	RISE	Ours
ImageNet	0.1253	0.1217	0.1077	0.1236
CUB-200-2011	0.0805	0.1287	0.0588	0.0627
Stanford Cars	0.0796	0.1345	0.0658	0.0698
Aircraft Variant	0.0740	0.1508	0.0569	0.0628
Aircraft family	0.1049	0.1935	0.0762	0.0878
Aircraft manufacturer	0.1735	0.3009	0.1388	0.1627

표 3. 각 기법 별 삽입 게임 결과
Table 3. Insertion game results for each of four models

Data type	Grad-CAM	LIME	RISE	Ours
ImageNet	0.6785	0.6940	0.7335	0.6055
CUB-200-2011	0.6982	0.6531	0.7461	0.6415
Stanford Cars	0.7197	0.6447	0.7720	0.6785
Aircraft Variant	0.6742	0.5647	0.7248	0.6168
Aircraft family	0.7480	0.6532	0.8026	0.7261
Aircraft manufacturer	0.8011	0.7091	0.8475	0.7798

KD의 성능을 측정하기 위한 첫 번째 실험 결과는 표 1과 같다. 훈련이 잘 되었다면 동일한 입력 이미지에 대해 *TN*와 *SN*의 예측값이 유사할 것이다. ImageNet 데이터에서 입력 이미지에 대해 *TN*와 *SN*가 동일한 클래스로 예측한 것은 Sim 1에서 80.70%이다. 따라서 *TN*와 *SN*가 유사하게 동작함을 알 수 있다.

두 번째 실험은 생성된 어텐션 맵을 측정하기 위한 삭제와 삽입 게임 실험이다. 표 2와 표 3은 각 기법에서 생성된

어텐션 맵의 성능을 평가하기 위한 삭제와 삽입 게임의 결과다. 삭제 게임에서 RISE, 본 논문에서 제안하는 기법, LIME, Grad-CAM 순서로 AUC가 낮았다. 삽입 게임의 경우 RISE, LIME, Grad-CAM, 제안하는 기법 순서로 AUC가 높았다.

실험에서 Grad-CAM은 영역이 넓게 퍼져서 생성되므로 배경이 포함되는 경우가 많았고 이미지가 흔들려 물체가 또렷하지 않을 경우 제대로 생성되지 않았다. LIME의 경우 슈퍼 픽셀 단위로 쪼개지기 때문에 중요한 영역의 색이 중요하지 않은 부분 (배경)의 색과 유사할 경우 배경도 중요한 부분으로 함께 인식되는 한계가 있었다. 또 돌출 맵 생성 시 평균 9초로 가장 많은 시간이 소요된다. RISE는 가장 우수하지만 픽셀 단위로 변형을 반복하기 때문에 돌출 맵 생성 시간이 5초가 걸리고 계산량이 크다. 또 뒷모습이거나 훈련 데이터에서 주로 보이는 형태가 아닐 경우 결과가 좋지 않았다.

표의 IN, CUB, CARS, A-V, A-F, A-M은 각각 ImageNet, CUB-200-2011, Stanford Cars, Aircraft Variant, Aircraft family, Aircraft manufacturer의 약어다. 그림 6은 Cars 데이터에서 각 기법 별 생성된 돌출 맵의 예시다. 그림 6의 Chevrolet Silverado 1500 Extended Cab 2012 예시처럼 종의 특징에 해당하는 영역에 잘 집중하는 경우도 있지만 전반적으로 물체 전체를 중요하게 인식하는 경우가 많았다.

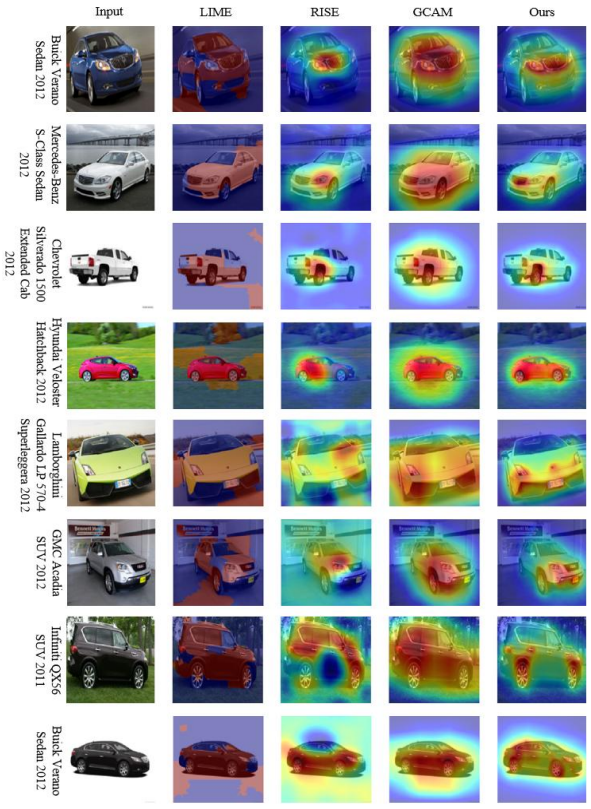


그림 6. Cars dataset에서의 각 기법별 돌출맵
Fig. 6. Feature maps on Cars dataset

V. 결론 및 향후 연구 방향

본 논문에서는 블랙박스 모델을 설명할 수 있는 시각적 설명 기법을 제안하였다. 지식 증류 기법을 적용함으로써 블랙박스 모델과 동일하게 예측하는 화이트 박스 모델을 생성하고 이 모델을 이용해 예측 근거를 시각화할 수 있는 돌출 맵을 생성한다. 블랙박스 모델과 동일하게 예측하는 화이트 박스 모델을 생성해 학습하기 때문에 설명하려는 모델의 종류와 관계없이 적용될 수 있다. 또 화이트 박스 모델의 후반부 레이어에서 생성된 일부 어텐션 맵만 사용하는 것이 아니라, 모델의 전체 레이어에서 생성된 어텐션 맵을 이용함으로써 모델의 예측 근거에 대한 정보를 더 많이 포함하고 있다. 특정 클래스로 예측한 근거를 설명하기 위해 돌출 맵 생성 시 Grad-CAM을 적용해 클래스에 해당하는 부분에 더 집중할 수 있도록 한다.

KD 유사도 측정을 통해 화이트 박스 모델이 블랙박스 모델의 정보를 받아 유사하게 예측하도록 훈련된 것을 확인했다. 또 삽입, 삭제 게임을 통해 어떤 레이어에서 생성된 어텐션 맵과 Grad-CAM이 모델의 예측 근거를 더 잘 시각화하는지 확인할 수 있었다.

제안하는 시각적 설명 기법은 삽입과 삭제 게임에서 RISE 기법보다 더 좋은 결과를 내지는 못했지만 모델의 어떤 레이어가 예측 근거를 더 많이 포함하는지 확인했다. 또 어텐션을 합치는 과정, 합치는 레이어 등에서 다양한 변형과 조합이 가능하기 때문에 추후 연구를 통해 좀 더 우수한 성능을 내는 돌출 맵을 생성할 수 있을 것이다.

References

[1] Sample, Ian, "Computer says no: Why Mmaking AIs fair, Accountable and Transparent is Crucial". the Guardian. last modified, Vol. 5, No. 1, pp. 1-15, 2017.

[2] A. Holzinger, M. Plass, K. Holzinger, G.C. Crisan, C.M. Pintea, V Palade, "A Glass-box Interactive Machine Learning Approach for Solving NP-hard Problems with the Human-in-the-loop.", arXiv preprint arXiv:1708.01104, 2017

[3] L. Edwards, M. Veale. "Slave to the Algorithm: Why a Right to an Explanation is Probably not the Remedy you are Looking for." Duke L. & Tech, Vol. 16, No. 18, 2017.

[4] B. Zhou, A. Khosla, A. Lapedriza, "Learning Deep Features for Discriminative Localization." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 2921-2929, 2016.

[5] R.R. Selvaraju, M. Cogswell, A. Das, "Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization." Proceedings of the IEEE International Conference on Computer Vision., pp. 618-626, 2017.

[6] M.T. Ribeiro, S. Singh, C. 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. ACM, pp. 1135-1144, 2016.
[7] V. Petsiuk, A. Das, K. Saenko, "Rise: Randomized Input Sampling for Explanation of black-box models.", arXiv preprint arXiv:1806.07421, 2018
[8] C. Fong, A. Vedaldi, "Interpretable Explanations of Black Boxes by Meaningful Perturbation." Proceedings of the IEEE International Conference on Computer Vision., pp. 3429-3437, 2017
[9] G. Hinton, O. Vinyals, J. Dean, "Distilling the Knowledge in a Neural Network.", arXiv preprint arXiv:1503.02531, 2015.
[10] S.H. Woo, J.C. Park, J.Y. Lee, I.S. Kweon "Cbam: Convolutional Block Attention Module." Proceedings of the European Conference on Computer Vision (ECCV), pp. 3-19, 2018.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, "Attention is all you Need." Advances in Neural Information Processing Systems., pp. 5998-6008, 2017.
[12] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, "Residual Attention Network for Image Classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 3156-3164, 2017.
[13] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 7132-7141, 2018.
[14] X. Wang, R. Girshick, A. Gupta, K. He, "Non-local Neural Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition., pp. 7794-7803, 2018.
[15] J. Deng, W. Dong, R. Socher, .LJ. Li, K Li, L. Fei-Fei "Imagenet: A Large-scale Hierarchical Image Database." 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248-255, 2009.
[16] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, "The Caltech-ucsd Birds-200-2011 dataset.", 2011.
[17] J. Krause, M. Stark, J. Deng, L. Fei-Fei, "3d Object Representations for Fine-grained Categorization." Proceedings of the IEEE International Conference on Computer Vision Workshops., pp. 554-561, 2013.
[18] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, "Fine-grained Visual Classification of Aircraft.", arXiv preprint arXiv:1306.5151, 2013.

June Gyu Kang (강 준 규)

2020 Statistics from Jeonbuk National University (B.S.)
2020~Computer Science and Engineering from Jeonbuk University (M.S.)

Field of Interests:Medical Image Analysis, Artificial Intelligence
Email: chop4687@jbnu.ac.kr

Min Gyeong Jeon (전 민 경)

2017 Computer Science and Engineering from Jeonbuk National University (B.S.)
2020 Computer Science and Engineering from Jeonbuk National University (M.S.)

Career: Working in DEEPNOID
Field of Interests: Medical image analysis, Artificial Intelligence
Email: christy4526@gmail.com

Hyeon Seok Lee (이 현 석)

2018 Computer Science and Engineering from Jeonbuk National University (B.S.)
2020 Computer Science and Engineering from Jeonbuk National University (M.S.)
2020~Computer Science and Engineering from Jeonbuk National University (Ph.D)

Field of Interests: Explainable AI, Visual Tracking
Email: hslee0390@jbnu.ac.kr

Sungchan Kim (김 성 찬)

1998 Material Science and Engineering from Seoul National University (B.S.)
2000 Computer Science and Engineering from Seoul National University (M.S.)
2005 Computer Science and Engineering from Seoul National University (Ph.D)

Career:
2009~Division of Computer Science and Engineering, Jeonbuk National University (Prof.)
Field of Interests: Computer Vision, Artificial Intelligence
Email: s.k@jbnu.ac.kr