딥러닝 설명을 위한 슈퍼픽셀 제외·포함 다중스케일 접근법

(Superpixel Exclusion-Inclusion Multiscale Approach for Explanations of Deep Learning)

서다솜*, 오강한*, 오일석**, 유태웅**
(Dasom Seo, KangHan Oh, Il-Seok Oh, Tae-Woong Yoo)

요 약

딥러닝이 보편화되면서 예측 결과를 설명하는 연구가 중요해졌다. 최근 슈퍼픽셀에 기반한 다중스케일 결합 기법이 제안되었는데, 물체의 모양을 유지함으로써 시각적 공감이라는 장점을 제공한다. 이 기법은 예측 차이라는 원리에 기반을 두고 있으며, 슈퍼픽셀을 가리고 얻은 예측 결과와 원래 예측 결과의 차이를 보고 돌출맵을 구성한다. 본 논문은 슈퍼픽셀을 가리는 제외 연산뿐 아니라 슈퍼픽셀만 보여주는 포함 연산까지 사용하는 새로운 기법을 제안한다. 실험 결과 제안한 방법은 IoU에서 3.3%의 성능 향상을 보인다.

■ 중심어 : 딥러닝; 설명모델; 슈퍼픽셀; 시각화

Abstract

As deep learning has become popular, researches which can help explaining the prediction results also become important. Superpixel based multi-scale combining technique, which provides the advantage of visual pleasing by maintaining the shape of the object, has been recently proposed. Based on the principle of prediction difference, this technique computes the saliency map from the difference between the predicted result excluding the superpixel and the original predicted result. In this paper, we propose a new technique of both excluding and including super pixels. Experimental results show 3.3% improvement in IoU evaluation.

■ keywords : deep learning; explanation model; superpixel; visualization

Ⅰ. 서 론

최근 심층신경망은 영상 분류, 음성 인식, 로봇틱스, 자율주행 등과 같은 다양한 분야에서 인간과 동일하거나 더 우수한 성과를 나타낸다[1, 2, 3, 4, 5]. 딥러닝이 다양한 분야에 사용되고 있음에도 불구하고 딥러닝은 예측 결과에 대하여 어떻게 그런 결정을 하였는지 만족스럽게 설명하지 못하고 있다[6]. 유럽연합은 이러한 문제가 잠재적인 해가 될 수 있음을 인식하고, 기계학습 알고리즘 사용이 일상에 미칠 잠재적 영향을 기술하고 알고리즘의 의사 결정에 대한 설명을 요구할 수 있는 "설명의 권리(right to explanation)"를 다루는 규정 GDPR(General Data Protection Regulation)을 채택하였다[7].

블랙박스 알고리즘을 사용하는 딥러닝의 예측 결과에 대한 설명(explanation)은 알고리즘의 공정성을 보장하고 훈련 데이터의 잠재적인 문제점을 파악하고 알고리즘이 예상대로 수행되는지 검증하는데 매우 중요하다[6, 8]. 특히, 의학 또는 자율자동차 등 모델의 신뢰성이 보장되어야 하는 응용에서는 설명이 필수적이다[2]. 딥러닝의 결정을 설명하기 위해 다양한 설명 기법들이 연구되고 있으며, 사용자의 신뢰를 높이고 사용자가 더 나은 모델을 선택할 수 있도록 한다[9, 10, 11].

딥러닝의 예측에 대한 설명은 네트워크 모델의 데이터 처리를 설명하거나 네트워크 모델의 데이터 표현을 설명하는데 중점을 둔다. 데이터 처리에 대한 설명은 "특정 입력이 특정 출력으로 이어지는 이유는 무엇입니까?"라는 질문에 대한 답이며 프로그램 실행 단계를 추적하는 것과 유사하다[6]. 개발 초기에는 대부분 네트워크 모델 내의 데이터 표현, 즉 데이터 구조 기반의 학습 모델 자체를 해석하였으나 최근에는 개별 인스턴스에 대한 예측을 설명하는데 더 많이 주목하고 있다[6, 8, 11, 12].

딥러닝을 설명하기 위해서는 부류 분별력, 즉 주어진 부류를 결정하는데 가장 기여하는 특징을 식별할 수 있어야한다. 즉 데

* 준회원, 전북대학교 컴퓨터공학부
** 정회원, 전북대학교 컴퓨터공학부

접수일자 : 2019년 05월 29일
수정일자 : 2019년 06월 13일
게재확정일 : 2019년 06월 13일
교신저자 : 유태웅 e-mail : twyoooo@gmail.com

이터의 어떤 부분이 실제로 네트워크 출력에 영향을 미치는지 시각적으로 나타내는 돌출맵(saliency map)과 같은 방법이 필요하다[6, 12].

Seo et al.[12]는 예측 차이 원리에 기반을 두고 슈퍼픽셀을 사용한 다중스케일 결합 기법을 제안하였다. 이 기법은 슈퍼픽셀을 가리고 얻은 예측 결과와 원래 예측 결과의 차이를 보고 돌출맵을 구성한다. 이 방법은 입력 영상에 변화를 주었을 때 결과에 어떠한 변화가 발생하는지 측정하는 예측 차이(prediction difference) 기법을 사용한다. 입력 영상에 변화를 주기위하여 슈퍼픽셀(superpixel) 기반의 영역 분할과 다중스케일 기법을 사용한다. 슈퍼픽셀 기반의 영역 분할 기법과 다중스케일을 사용하는 이유는 기존 픽셀단위 영역 분할 처리방법보다 계산속도가 빠르고 영상 내 대상 객체의 크기에 관계없이 효율적으로 영역단위 분할을 수행하기 때문이다. 분할된 영역의 중요도 및 영향력 측정을 위하여 분할된 영역(슈퍼픽셀)들을 제외하는 연산을 사용한다. 입력 영상으로부터 대상 인스턴스를 제외하며 다중스케일로부터 생성된 맵을 최종 하나의 돌출맵으로 융합함으로써, 보다 분별적이고 시각적으로 공감이 가는 맵을 획득한다.

본 논문은 기존 알고리즘 [12]가 사용한 제외 연산뿐만 아니라 포함 연산까지 사용하여 성능 개선을 모색한다. 이 방법은 기존 알고리즘에 포함 연산만을 추가하였으므로 기존 알고리즘의 특성을 그대로 받는다. 따라서 제안한 알고리즘은 모델자유성(model-agnostic)과 시각적 공감성을 보장한다. 실험 결과 제안한 방법은 IoU에서 46.2%의 성능을 보이는데, 이는 기존 알고리즘에 비해 3.3% 향상에 해당한다.

II. 관련 연구

딥러닝은 영상인식, 자연어처리, 음성인식, 영상생성 등에서 획기적인 성능 향상을 가져왔다. 이러한 혁신에 힘입어 인공지능 제품이 시장에 속속 등장하고 있다. 하지만 딥러닝의 가장 취약한 점인 설명이 불가능하다는 사실에 따라 현장 응용에 한계를 노출하고 있다. NIPS(Neural Information Processing Systems) 학술대회의 위성 심포지움으로 WHI(Workshop on Human Interpretability in Machine Learning)은 이런 필요성에 따라 열리는 국제 학술대회이다. Lipton[13]과 Doran et al.[14]은 설명가능성을 개념적으로 정립한 논문이다.

현재 영상 분석에 가장 널리 사용하는 CNN(Convolution Neural Network) 모델을 설명하고 해석하는 방법이 많이 제안되었다. Erhan et al.[8]는 영상 공간에서 그레이디언트를 사용하여 최적화를 수행함으로써 관심있는 뉴런 활동을 최대화하는 입력 영상을 찾아 딥 모델을 시각화했다. Yosinski et al.은 그레이디언트를 사용하여 출력 단위에 대한 높은 활성화 또는 낮은 활성화를 유발하는 영상을 찾는다[15]. 돌출맵(saliency map)은 영상 각 영역의 시각적 중요도를 맵의 형태로 나타낸다. 돌출맵에서의 시각적 중요도는 시각적인 중요성이 높은 전경에서는 대체로 높은 값을 가지며, 중요성이 낮은 배경에서는 낮은 값을 가지는 형태이다. CNN에 예제 영상과 목표 부류를 입력하면 CNN은 부류를 예측(prediction)할 때 예제 영상의 화소 또는 영역이 예측에 미치는 중요도를 측정하여 돌출맵을 계산한다. 중요도를 측정하는 방법으로 그레이디언트 기반의 SA(Sensitive Analysis)[10], 디컨볼루션(deconvolution)[16], LRP(Layer-wise Relevance Propagation)[17], 그레이디언트 기반 접근법의 한계를 해결한 DeepLIFT 기법[18] 등이 있다. Zhou et al.은 전역 평균 풀링(global average pooling)을 이용하는 CAM 기법[19]을 제안했으며, Selvaraju et al.는 CAM 기법을 Grad-CAM(Gradient-weighted Class Activation Map)으로 확장하여 더 넓은 범위의 CNN에 적용하였다[20].

앞서 살펴본 기법들은 신경망의 내부 정보를 사용하는데 이런 이유로 내부 구조를 필요에 따라 변경한다거나 특정 구조를 가진 신경망에만 적용 가능한 한계를 안고 있다. 예측 차이(prediction difference) 접근방법은 이런 한계로부터 자유로우며 결과적으로 모델자유성(model-agnostic)이라는 중요한 성질을 만족한다. 예측 차이 기법은 특징 벡터의 예측 값과 특징 i가 없는 특징 벡터 예측 값 사이의 차를 계산하는 방법이다. 예측 차이 기법을 사용하기 위해서는 원본 특징 벡터로부터 특징 i를 제외하는 방법을 고안해야한다. Robnik-Šikonja et al.은 주변 확률에 기반하여 특징을 제외하는 방법을 제안하였다[11]. Zintgraf et al.은 [11] 방법을 기반으로 영상 특성에 대한 사전 지식을 이용하는 방법을 제안하였다[23]. 주어진 픽셀 값은 인접 픽셀에 크게 의존한다는 사실을 고려하여 조건부 확률을 공식화하였다. Dong et al.은 모델에 예측 차이 최대화 연산을 삽입함으로써 비디오 캡션 CNN-RNN 모델의 해석 가능성을 향상시켰다[22]. Seo et al.[12]는 슈퍼픽셀 기반 영역 분할을 수행하고 각 영역을 제외시켜 가며 생성된 예측 값을 원본 영상의 분류 예측 값과의 차이를 계산하여 하나의 돌출맵으로 융합시킨다. 각각의 영역을 제외(Ex)함으로써 해당 영역의 중요도 및 영향력을 측정한다.

III. 알고리즘

제안하는 알고리즘은 예측 차이라는 원리를 따른다. 예측 차이는 모든 특징을 가지고 예측한 결과 $f(\mathbf{x}_{\setminus i})$와 i 번째 특징을 제외하고 예측한 결과 $f(\mathbf{x}_{\setminus i})$ 사이의 차이인 $f(\mathbf{x})-f(\mathbf{x}_{\setminus i})$ 값을 측정한다. 이 값을 i 번째 특징의 중요도로 간주하며, 이 값을 돌출맵에 기록한다. 여기에서 특징 벡터 $\mathbf{x}$는 d 차원 벡터

$\mathbf{x}=(x_1, x_2, \cdots, x_d)$이다. $f(\mathbf{x})$는 특정 클래스 y에 대한 예측값이므로 x가 y에 속할 확률로 해석할 수 있다. 따라서 알고리즘 기술에서는 $f(\mathbf{x})$를 $p(y|\mathbf{x})$로 대체한다.

1. 제외(Ex) 다중스케일 예측 차이

Zintgraf et al.[23]은 CNN으로 영상을 처리하기 위해 Robnik-Šikonja et al.이 제안한 일반적인 예측 차이 알고리즘[11]을 조건부 샘플링과 다변량 분석을 이용하여 픽셀 단위의 예측 차이를 계산하는 알고리즘으로 확장하였다. 이 알고리즘으로 생성된 돌출맵은 그림 6의 (d) 열에 나타나듯이 돌출 값이 물체 영역 내에 분산되고 배경 영역으로 삽입되는 문제를 나타낸다. 또한 픽셀 단위의 계산을 수행하므로 아주 느린 문제를 안고 있다[23]. Seo et al.은 이런 문제를 해결하는 기법을 제안했는데, 핵심 아이디어는 슈퍼픽셀을 사용한 돌출맵을 다중스케일로 구하고 이들을 하나로 결합하는 것이다[12]. Algorithm 1은 [12]의 알고리즘을 설명한다.

[Algorithm 1] Regional Multi-scale Prediction Difference by using Ex

Input: trained classifier f, image $\mathbf{x}$, target class c, multi-scale num. r

Output: saliency map $\mathbf{s}$

1: Run f to get $p(y=c|\mathbf{x})$
2: Estimate $N(\mu, \sigma^2)$ from the boundary segmented with scale 2^8
3: **for** $j=1$ to r **do** // for each scale
4: Perform a segmentation with the scale 2^j // 슈퍼픽셀 분할
5: **for** each region i **do** // 각각의 슈퍼픽셀에 제외 연산 적용
6: Simulate the exclusion of i using $N(\mu, \sigma^2)$, i.e., $\mathbf{x}_{\setminus i}$
7: Run f to get $p(y=c|\mathbf{x}_{\setminus i})$
8: **for** each pixel q in the region i **do**
9: $s_q^j = max\ (0, g(p(y=c|\mathbf{x}), p(y=c|\mathbf{x}_{\setminus i})))$
10: **end for**
11: **end for**
12: **end for**
13: $\mathbf{s} = \frac{1}{r}\sum_{j=1}^{r} s^j$

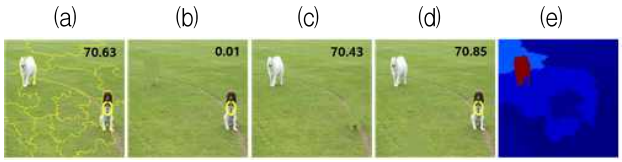
Algorithm 1의 1 행은 모든 특징을 가지고, 즉 원래 영상 x를 가지고 예측을 수행하고 그 결과를 $p(y=c|\mathbf{x})$라 표기한다. 2 행은 2^8=256개의 슈퍼픽셀로 분할한 영상에서 영상 경계에 위치한 슈퍼픽셀만 가지고 화소들의 색상 분포를 계산한다. μ는 화소들의 (R,G,B) 평균이며 σ는 표준편차이다. 경계에 위치한 슈퍼픽셀로 부터 계산한 가우시언 분포를 가지고 슈퍼픽셀을 지우는 이 기법을 경계 프라이어(boundary prior)라 부른다. 그림 1은 경계 프라이어 기법이 사용하는 경계 영역을 보여준다. 경계 프라이어를 사용하는 이유는 영상의 경계 부분에는 물체보다는 배경이 자리 잡을 확률이 크기 때문이다.

3~12 행은 다중스케일($2^1 \sim 2^r$)로 영역을 분할하고 슈퍼픽셀에 제외 연산을 적용한다.

그림 1. 그레이 영역을 정규분포 계산에 사용

6 행은 영역 제외(exclusion) 연산을 시뮬레이션한다. 2 행에서 준비해둔 정규분포를 이용하여 색상을 생성하고 이 색상을 슈퍼픽셀 i에 덧씌우는 연산을 수행한다. 7 행은 이렇게 영역 i를 가린 영상을 가지고 심층신경망의 전방 계산을 실행하여 $p(y=c|\mathbf{x}_{\setminus i})$를 구한다. 이 연산은 영역 i를 제외하는 효과를 거둔다. 8~10 행은 제외된 영역에 속한 모든 화소에 대해 예측 차이를 저장한다. 이때 예측 차이는 g라는 함수로 구하는데 g로 여러 가지 형태가 가능한데[11], 본 논문은 가장 단순한 형태인 $g(a,b)=a-b$를 사용하였다.

3~12 행을 마친 후에는 2^1, 2^2, ..., 2^r 스케일에서 구한 다중스케일 돌출맵을 하나로 결합한다. 이 일은 13 행이 수행한다.



(a) 입력영상(16레벨) (b) 사모예드 제외 (c) 다른 개 제외
(d) 잔디 삭제 (e) 돌출맵

그림 2. Samoyed 클래스 예제 영상(우측-상단은 예측 값)

그림 2는 목표 클래스(class)가 사모예드(Samoyed)인, 즉 c=Samoyed인 경우를 예시한다. 스케일이 16인 예제이다. 그림 2(a)는 슈퍼픽셀로 분할될 영상을 보여준다. $p(y=c|\mathbf{x})$=0.7063이다. 그림 2(b)는 Samoyed에 해당하는 영역이 제외된 상황으로서 $p(y=c|\mathbf{x}_{\setminus i})$=0.0001이 되었다. 하지만 Samoyed가 아닌 영역을 제외한 그림 2(c)와 (d)에서는 $p(y=c|\mathbf{x}_{\setminus i})$가 $p(y=c|\mathbf{x})$와 거의 같은 사실을 확인할 수 있다.

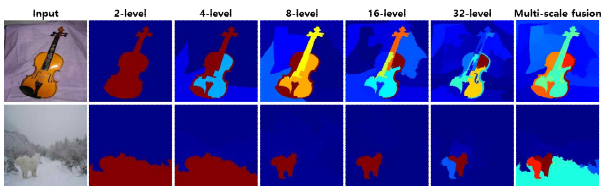


그림 3. 다중스케일 레벨에 따른 violin과 Samoyed 클래스 예제 영상

그림 3은 바이올린과 사모예드 부류에 속하는 영상을 가지고 다중스케일의 장점을 예시한다. 바이올린 영상의 경우 물체가 크기 때문에 스케일 2에서 가장 좋은 결과를 보인다. 반면에 물체가 작은 사모예드 영상의 경우에는 스케일 8 또는 16에서 좋은 결과를 보인다. 하지만 물체의 스케일을 미리 알 수 없기 때문에 단일스케일을 사용하는 경우 운에 따라 성공이 달라지는 한계가 있다. Algorithm 1은 다중스케일을 사용하여 이런 문제를 해결하였다.

2. 제외-포함(Ex-In) 다중스케일 예측 차이

Seo et al.[12]이 제안한 Algorithm 1은 제외(Ex) 연산만 사용한다. 본 논문은 제외뿐만 아니라 포함 연산까지 사용하는 알고리즘을 제안한다. 그림 4는 본 논문이 제안하는 아이디어를 설명한다. 슈퍼픽셀로 분할하고 경계에 속하는 슈퍼픽셀을 가지고 가우시언을 구하는 과정은 Algorithm 1과 같다. 하지만 다중스케일을 적용하는 단계에서는 개별 영역을 제외하고 돌출맵을 구하는 과정과 개별 영역만 포함하여 돌출맵을 구하는 과정을 병행한다. 이들 과정에서 구한 두 돌출맵을 결합하여 최종 결과를 출력한다. Algorithm 1은 그림 4에서 Inclusion process를 제외한 것으로 볼 수 있다.

[Algorithm 2] Regional Multi-scale Prediction Difference by using Ex-In
Input: trained classifier f, image $\mathbf{x}$, target class c, multi-scale number r
Output: saliency map $\mathbf{s}$

Run f to get $p(y=c|\mathbf{x})$
Estimate $N(\mu,\sigma^2)$ from the image $\mathbf{x}$ segmented with scale 2^8
for $j=1$ to r **do** // for each scale
  Perform a segmentation with the scale 2^j // 슈퍼픽셀 분할
  for each region i **do** // Ex 연산
    Simulate the exclusion of i using $N(\mu,\sigma^2)$, i.e., $\mathbf{x}_{\setminus i}$
    Run f to get $p(y=c|\mathbf{x}_{\setminus i})$
    for each pixel q in the region i **do**
      $(es)_q^j = max\ (0, g(p(y=c|\mathbf{x}), p(y=c|\mathbf{x}_{\setminus i})))$
    end for
  end for
  for each region i **do** // In 연산
    Simulate the Inclusion of i using $N(\mu,\sigma^2)$, i.e., $\mathbf{x}_{/i}$
    Run f to get $p(y=c|\mathbf{x}_{/i})$
    for each pixel q in the region i **do**
      $(is)_q^j = max\ (0, g(p(y=c|\mathbf{x}), p(y=c|\mathbf{x}_{/i})))$
    end for
  end for
end for
$\mathbf{s} = mean(\frac{1}{r}\sum_{j=1}^{r}(es)^j + \frac{1}{r}\sum_{j=1}^{r}(is)^j)$

Algorithm 2는 제외-포함(Ex-In) 연산을 사용하여 Algorithm 1을 확장한다. 영역 i를 제외한 예측을 $p(y=c|\mathbf{x}_{\setminus i})$로 표기한다. 영역 i를 포함한 예측, 즉 영역 i 만 남기고 나머지 영역을 모두 제거하고 예측한 결과를 $p(y=c|\mathbf{x}_{/i})$로 표기한다.

5~11 행은 제외 연산, 즉 Ex 연산을 적용한다. 12~18 행은 포함 연산, 즉 In 연산을 적용한다. 20 행은 제외 연산으로 구한 돌출맵 es와 포함 연산으로 구한 돌출맵 is를 하나의 돌출맵으로 결합한다. 그림 4의 오른쪽 부분은 es 맵과 is 맵, 그리고 이들을 결합한 최종 돌출맵을 보여준다.

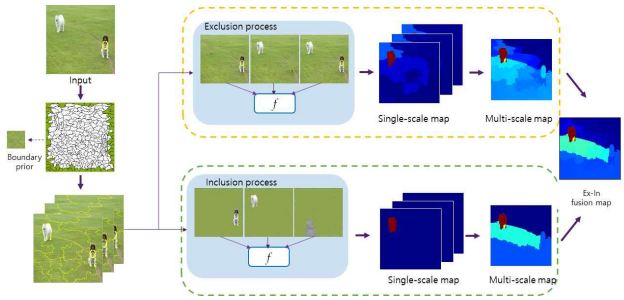
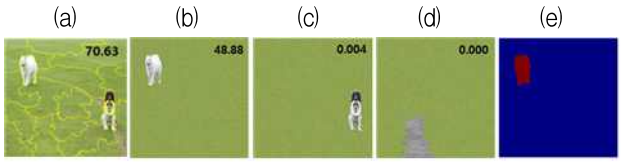


그림 4. 대상 객체 제외-포함(Ex-In) 방법 프로세스

그림 5는 In 연산의 효과를 예시한다. 그림 2와 마찬가지로 목표 클래스가 샤모예드인, 즉 c=Samoyed인 경우를 예시한다. 스케일이 16인 예제이다. 그림 5(a)는 슈퍼픽셀로 분할될 영상을 보여준다. $p(y=c|\mathbf{x})$=0.7063이다. 그림 5(b)는 Samoyed만 포함된 상황으로서 $p(y=c|\mathbf{x}_{/i})$=0.4888이 되었다. 하지만 Samoyed가 아닌 영역만을 포함한 그림 5(c)와 (d)에서는 $p(y=c|\mathbf{x}_{/i})$가 거의 0이라는 사실을 확인할 수 있다.



(a) 입력영상(16레벨) (b) 사모예드 포함 (c) 다른 개 포함
(d) 잔디 포함 (e) 돌출맵
그림 5. Samoyed 클래스 예제 영상(우측-상단은 예측 값)

IV. 실험 결과

실험은 ILSVRC 2012 데이터셋(ImageNet)과 GoogLeNet [27]을 가지고 수행하였다. NVIDIA GTX 1080 GPU가 장착된 Intel Core i5-4670 CPU(3.40 GHz)에서 수행하였다.

[12]의 특성 물려받기: 그림 6은 5 가지 기존 방법(DC[25], SA[21], LRP[17], PD[29], Grad-CAM[19])과 [12]의 결과, 그리고 본 논문이 제안한 제외-포함 연산을 사용한 결과를 비교한다. 기존 방법을 이용한 결과 영상은 공개 소프트웨어 (https://github.com/lmzintgraf/DeepVis-PredDiff) 및 공개

사이트(https://lrpserver.hhi.fraunhofer.de/image-classification)를 이용하였다.

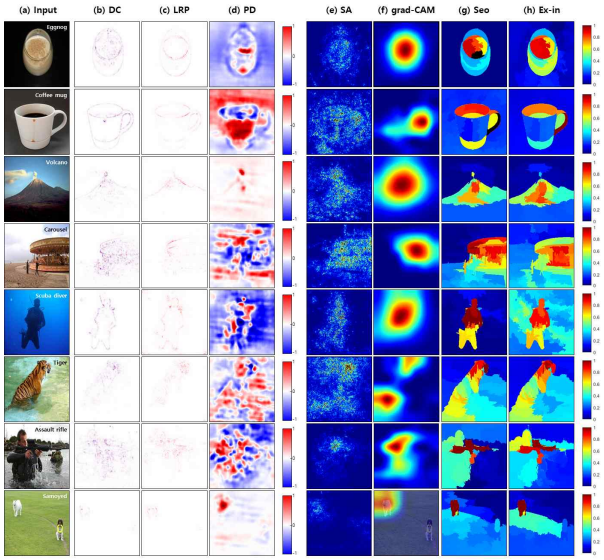


그림 6. 기존 방법과 제안된 방법의 시각적 비교

DC, LRP 및 SA의 돌출맵은 점 구름과 비슷해 보이고 PD와 Grad-CAM의 돌출맵은 각각 얼룩과 윤곽선 같아 보인다. 부류-분별력(class-discriminability)의 관점에서, DC 및 LRP 방법은 명백한 한계를 드러낸다. 이 방법들은 객체의 에지를 강조하는 경향이 있고 목표 부류에 대한 올바른 영역을 히트하지 않는다. 예를 들어, "에그 노그"및 "커피 머그" 부류에 속하는 영상은 각각 내부 내용물과 용기 표면을 표시해야한다. 그러나 DC 및 LRP 방법은 이들 영상에서 용기의 에지만 찾아 에그 노그와 커피 머그를 구별하지 못한다. Grad-CAM 방법은 이 두 부류에 대해 각각 용기 내부 내용과 손잡이를 올바르게 히트한다.

Grad-CAM 방법은 해당 물체를 등고선 형태로 표시해준다. 따라서 신경망이 영상의 어느 곳을 보고 분류를 수행했는지 대략적인 설명을 제공하지만 물체의 윤곽을 알 수 없는 한계가 있다. 이들과 달리 [12]는 에그 노그와 커피 머그를 잘 구별해주어 부류 분별력 측면에서 우수하다. 게다가 물체의 윤곽을 잘 드러내므로 시각적 공감을 제공한다고 평가할 수 있다. 본 논문이 제안한 방법의 결과는 맨 오른쪽 열에 있는데, [12]와 유사하다고 평가할 수 있다. 다시 말해 부류 분별력과 시각적 공감 측면에서 보면 [12]의 장점을 그대로 물려받는다고 볼 수 있다.

정량적 비교: 제외-포함 연산을 사용한 본 논문과 제외 연산만 사용한 [12]의 성능을 정량적으로 비교하기 위해 물체 검출(detection)에서 주로 사용하는 IoU(Intersection of Union) 점수를 이용한다. IoU는 정답 바운딩 박스와 돌출맵의 임계값별 바운딩 박스가 겹쳐지는 정도를 나타내는데, 그림 7은 IoU 계산 방법을 설명한다. IoU는 돌출맵을 임계값에 따라 이진화한 다음 계산한다.

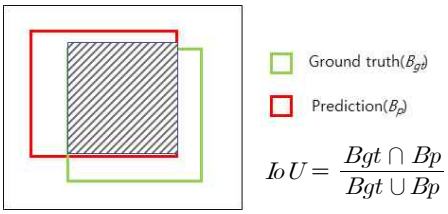


그림 7. IoU 계산방법

표 1은 기존 4개 기법, [12]의 기법, 그리고 제안한 기법의 평균 IoU, 최대 IoU를 보여준다. 제외(Ex) 연산만 사용한 [12]는 평균 IoU로 42.9를 보였는데, 제안한 기법은 3.3이 증가한 46.2를 보였다.

표 1 ResNet 모델을 사용한 평균 IoU(최대 IoU)

	IoU
Occlusion[23]	28.5(48.6)
LIME[14]	27.7(40.6)
SA[19]	15.3(53.8)
Grad-CAM[17]	28.9(51.4)
[12](Ex)	42.9(55.0)
Ours(Ex-In)	**46.2(55.5)**

그림 8은 임계값을 0에서 출발하여 0.1씩 증가시키면 IoU를 측정한 그래프이다. Grad-CAM은 0.1에서 가장 좋은 성능을 나타내는데, 이후로 급격히 점수가 하락하는 현상을 보인다. 제안한 방법은 40~50%까지 폭넓은 범위에서 안정적인 성능을 보여준다. 이 그래프에서 제안한 방법은 AUC(Area Under Curve)가 가장 넓다.

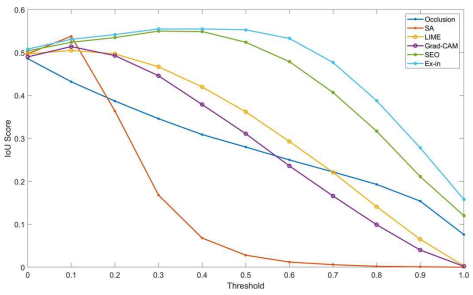


그림 8. 돌출맵의 임계값에 따른 IoU

V. 결 론

슈퍼픽셀에 기반한 다중스케일 결합 기법은 물체의 모양을 유지함으로써 시각적 공감이라는 장점을 제공한다. 이 기법은

예측 차이라는 원리에 기반을 두고 있으며, 입력 영상에 변화를 주기위하여 슈퍼픽셀을 가리는 제외 연산을 사용한다. 본 논문에서는 제외 연산뿐 아니라 슈퍼픽셀만 보여주는 포함 연산까지 사용하는 새로운 기법을 제안한다. 제안한 방법은 슈퍼픽셀의 제외·포함 연산을 사용하여 보다 부류 분별적이며 시각적 공감이 가는 돌출맵을 생성한다. 시각적 공감이 가는 설명 및 해석이 가능한 시스템은 사용자들로부터 높은 신뢰를 얻을 것으로 기대된다.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 521, pp. 436-444, May 2015.

[2] G. Montavon, W. Samek, and K.R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1-15. Feb 2018.

[3] 박선, 김종원, "오픈 소스 기반의 딥러닝을 이용한 적조생물 이미지 분류," *스마트미디어저널*, 제7권, 제2호, 34-39쪽, 2018년 6월

[4] 김서정, 이재수, 김형석, "딥러닝을 이용한 양파 밭의 잡초 검출 연구," *스마트미디어저널*, 제7권, 제3호, 16-21쪽, 2018년 9월

[5] 오정원, 김행곤, 김일태, "머신러닝 적용 과일 수확 시기 예측시스템 설계 및 구현," *스마트미디어저널*, 제8권, 제1호, 73-81쪽, 2019년 3월

[6] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," arXiv preprint arXiv:1806.00069v3, Feb 2019.

[7] B. Goodman, and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI MAGAZINE*, vol. 38, no. 4, pp. 50-57, Fall 2017.

[8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Technical Report*, University of Montreal, 1341. June 2009.

[9] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28(11), pp. 2660-2673. Aug 2016.

[10] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps," *In International Conference on Learning Representations Workshop*. 2013.

[11] M. Robnik-Šikonja, and I Kononenko, "Explaining classifications for individual instances," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20(5), pp. 589-600, May 2008.

[12] D.S. Seo, K.H. Oh, and I.S. Oh, "Regional Multi-scale Approach for Visually Pleasing Explanations of Deep Neural Networks," arXiv preprint arXiv:1807.11720v2, Aug 2018.

[13] Z.C. Lipton, "The mythos of model interpretability," *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. June 2016.

[14] D. Doran, S. Schulz, and T.R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," arXiv preprint arXiv:1710.00794v1, Oct 2017.

[15] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *In International Conference on Machine Learning*. 2015.

[16] M.D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," *In European Conference on Computer Vision*, pp. 818-833. Sep 2014.

[17] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS One*, 10(7), e0130140. July 2015.

[18] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *ICML'17 Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 3145-3153, Aug 2017.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *In IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921-2929, June 2016.

[20] R.S. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *In International Conference on Computer Vision*, pp. 618-626, 2017.

[21] S. Barratt, "InterpNET: Neural introspection for interpretable deep learning," *In Symposium on*

Interpretable Machine Learning, 2017.
[22] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 4306-4314, July 2017.
[23] L.M. Zintgraf, T.S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *In International Conference on Learning Representation*, 2017.
[24] M.T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," *In The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, Aug 2016.
[25] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *In IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, 2009.
[26] M.Y. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," *In IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097-2104, June 2011.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, and V. Vanhoucke, "Going deeper with convolutions," *In IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

저 자 소 개

서다솜(준회원)

2017년 전북대학교 전자공학부 학사 졸업.
2019년 전북대학교 컴퓨터공학부 석사 졸업.
현 재 국립농업과학원 석사후연구원

<주관심분야 : 컴퓨터비전, 기계학습, XAI, 자율주행>

오강한(준회원)

2010년 호남대학교 컴퓨터공학 학사 졸업.
2013년 전남대학교 전자컴퓨터공학과 석사 졸업.
2017년 전남대학교 전자컴퓨터공학 박사 졸업.
현 재 전북대학교 컴퓨터공학부 박사후연구원
<주관심분야 : 객체 검출, XAI, 뇌 영상 처리, 문서 영상 처리>

오일석(정회원)

1984년 서울대학교 컴퓨터공학과 학사 졸업.
1992년 한국과학기술원 전산학과 석사·박사 졸업.
현 재 전북대학교 컴퓨터공학부 교수
<주관심분야 : 기계학습, 컴퓨터비전>

유태웅(정회원)

1991년 전북대학교 수학과 학사 졸업.
1993년 전북대학교 전산통계학과 석사 졸업.
1998년 전북대학교 전산통계학과 박사 졸업.

<주관심분야 : 기계학습, 컴퓨터비전, 약물 상호작용>