KIPS Trans. Softw. and Data Eng. Vol.10, No.9 pp.349~358
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.9.349

Expanded Object Localization Learning Data Generation Using CAM and Selective Search and Its Retraining to Improve WSOL Performance

Sooyeon Go† · Yeongwoo Choi††

ABSTRACT

Recently, a method of finding the attention area or localization area for an object of an image using CAM (Class Activation Map)[1] has been variously carried out as a study of WSOL (Weakly Supervised Object Localization). The attention area extraction from the object heat map using CAM has a disadvantage in that it cannot find the entire area of the object by focusing mainly on the part where the features are most concentrated in the object. To improve this, using CAM and Selective Search[6] together, we first expand the attention area in the heat map, and a Gaussian smoothing is applied to the extended area to generate retraining data. Finally we train the data to expand the attention area of the objects. The proposed method requires retraining only once, and the search time to find an localization area is greatly reduced since the selective search is not needed in this stage. Through the experiment, the attention area was expanded from the existing CAM heat maps, and in the calculation of IOU (Intersection of Union) with the ground truth for the bounding box of the expanded attention area, about 58% was improved compared to the existing CAM.

Keywords : WSOL(Weakly Supervised Object Localization), CAM(Class Activation Map), Selective Search, Localization

CAM과 Selective Search를 이용한 확장된 객체 지역화 학습데이터 생성 및 이의 재학습을 통한 WSOL 성능 개선

고 수 연† · 최 영 우††

요 약

최근 CAM[1]을 이용해서 이미지의 객체에 대한 주의 영역 또는 지역화(Localization) 영역을 찾는 방법이 WSOL의 연구로서 다양하게 수행되고 있다. CAM을 이용한 객체의 히트(Heat) 맵에서 주의 영역 추출은 객체의 특징이 가장 많이 모여 있는 영역만을 주로 집중해서 객체의 전체적인 영역을 찾지 못하는 단점이 있다. 여기서는 이를 개선하기 위해서 먼저 CAM과 Selective Search[6]를 함께 이용하여 CAM 히트맵의 주의 영역을 확장하고, 확장된 영역에 가우시안 스무딩을 적용하여 재학습 데이터를 만든 후, 이를 학습하여 객체의 주의 영역이 확장되는 방법을 제안한다. 제안 방법은 단 한 번의 재학습만이 필요하며, 학습 후 지역화를 수행할 때는 Selective Search를 실행하지 않기 때문에 처리 시간이 대폭 줄어든다. 실험에서 기존 CAM의 히트맵들과 비교했을 때 핵심 특징 영역으로부터 주의 영역이 확장되고, 확장된 주의 영역 바운딩 박스에 대한 Ground Truth와의 IOU 계산에서 기존 CAM보다 약 58%가 개선되었다.

키워드 : WSOL(Weakly Supervised Object Localization), CAM(Class Activation Map), 선택적 탐색, 주의 영역

1. 서 론

이미지에서의 객체 검출을 위한 학습 방법으로서 찾고자 하는 물체의 위치를 정확히 알려주고 이미지를 학습하는 지도 학습(supervised learning) 방식과 달리, 위치 정보를 알려주지 않고 이미지와 찾고자 하는 객체의 라벨 또는 클래스만을 알려주고 물체의 위치를 찾도록 모델을 학습시키는 것을 Weakly Supervised Object Localization(WSOL)이라 한다. WSOL을 이용하여 객체를 정확히 추출할 수 있다면 대량의 이미지 데이터를 객체 단위로 레이블링하는 데 크게 도움이 되며, 객체 인식에도 중요하게 사용될 것이다.

이미지의 객체에 대한 라벨만을 알려줌으로서(supervised)

※ 이 논문은 한국연구재단 기초연구과제에 의하여 연구되었음(No. NRF-2017R1D1A1B04035633).
† 비 회 원 : 숙명여자대학교 컴퓨터과학과 석사과정
†† 정 회 원 : 숙명여자대학교 컴퓨터과학과 교수
Manuscript Received : February 9, 2021
First Revision : May 3, 2021
Accepted : May 31, 2021
* Corresponding Author : Yeongwoo Choi(ywchoi@sookmyung.ac.kr)

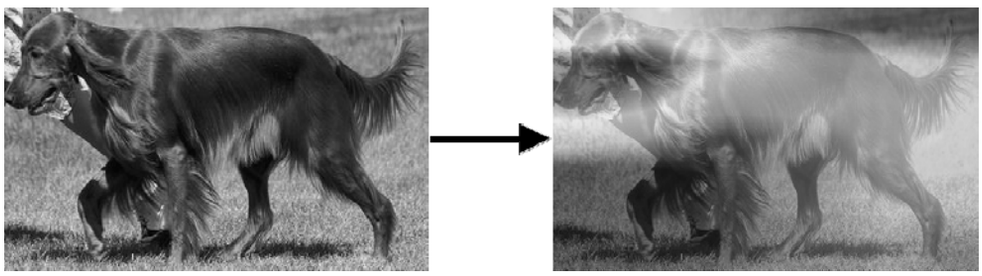


Fig. 1. Image of Labeled Dog and its CAM Heatmap

객체의 위치를 정확하게 찾아주는 딥러닝 모델을 찾는 것은 매우 어려운 일이다. 최근에 이에 대한 연구로서 CNN(Convolutional Neural Network)의 마지막 레이어(layer)를 Global Average Pooling(GAP) 레이어로 변경하여, Convolution 특징맵의 비중을 집계한 후 이미지 내의 물체의 위치를 주의(attention)하는 Class Activation Map(CAM)[1]을 제안하여 WSOL 연구에 많이 활용하고 있다. CAM[1]은 클래스 별로 특징맵의 비중을 반영한 최종 지역화(localization) 맵을 생성하여 객체의 위치를 파악하는데 중요한 역할을 하는 장점이 있다. 그러나 CAM은 주어진 레이블을 기반으로 특징맵의 비중을 과도하게 의존하여 지역화맵을 생성하는 것과 이미지 내의 레이블된 객체의 부분 영역만을 찾는 등의 단점이 있다. 한 예로서 Fig. 1은 이미지에서 “개”를 찾고자 할 때 CAM의 히트맵에서 개의 머리 부분이 특히 강조되어 주의된 결과를 보여주며, 몸통과 다리 부분은 비교적 적게 주의되어 최종 지역화맵에서 개를 전체적으로 찾지 못하는 경우가 빈번하게 발생한다.

이러한 단점을 보완하기 위해서 다양한 연구가 진행되고 있으며[2-5], 그 중 특히 ACoL[2]은 기존의 CAM과 같이 이미지들을 학습시킨 후 지역화맵을 기반으로 그 전 단계의 특징맵 중에서 강하게 반응한 상위 특징맵에서 주의 영역의 픽셀값을 0으로 하고 즉 검은색으로 가리고 다시 학습시킨다. 이는 처음 강하게 주의된 영역 주변이 학습되어 주의되기를 바라는 것이다. 이와 같이 강하게 주의된 영역을 지우는(erasing) 학습 과정을 반복하면서 반복된 결과를 모두 합쳐서 주의 영역을 전반적으로 확장해서 객체 전체 영역을 포함하도록 하였다.

이 결과 ACoL은 ILSVRC 데이터셋에서 검증 진리표(Ground Truth) 레이블과 비교하여 정확도를 확인했을 때 37.04%의 낮은 에러률을 보였다. 그러나 이 방법의 단점으로서 여러번의 반복된 재학습이 필요하여 학습 시간이 오래 걸리고, 기존의 CAM 모델에 비하여 객체 분류 정확도가 낮은 단점이 있다.

본 논문에서는 단일객체 이미지에 대해서 CAM 및 ACoL의 단점을 보완하는 확장된 주의 영역을 제공하고 이를 통해서 객체 분류의 정확성은 유지하면서 객체 검출의 정확성 즉 객체의 지역화 영역의 정확성을 높이는 방법을 제안한다. 제안하는 방법에서는 이미지의 특징을 분석하여 객체를 검출하는 Selective Search(SS)[6] 방법을 활용한다. SS는 이미지의 저수준(low level) 특징 분석으로부터 시작하는 상향식 방식으로서 딥러닝 방법을 사용하기 이전에 이미지의 객체 검출에 많이 활용된 방법이다. 이를 이용해 CAM의 결과로 얻어진 핵심 주의 영역과 SS에서 찾은 후보 영역 중에서 CAM 주의 영역과 겹침 정도가 큰 영역을 포함시켜 먼저 주의 영역을 확장한다. 확장된 바운딩 박스(bounding box) 영역에 가우시안 스무딩 필터를 적용하여 강하게 주의된 영역을 흐리게 스무딩한 후 이를 재학습 데이터로 사용하여 CNN 기반의 CAM을 다시 훈련시킨다. CAM이 집중하는, 즉 이미지에서 특징이 모여 있는 영역을 흐리게 만들어서 CAM이 핵심 주의 영역 이외의 주변 영역을 포함해서 주의하도록 하고자 한 것이다. 재학습된 새로운 CAM 모델이 테스트 입력 이미지의 객체 지역화에 사용된다. 여기서 전역적 탐색(exhaustive search)을 기반으로 한 SS는 모델의 재학습에 필요한 학습데이터를 생성하는데만 사용되고 최종 객체 지역화에는 재학습된 CAM만 사용되어 CAM 모델의 처리 시간만으로 확장된 즉 더 정확한 객체 영역을 지역화할 수 있다.

이 논문은 2절에서 관련 연구, 3절은 제안 방법을 설명하고, 4절에서 제안한 방법에 적용한 데이터와 결과를 보여준다. 끝으로 결론과 향후 연구를 5절에서 언급한다.

2. 관련 연구

WSOL(Weakly Supervised Object Localization)은 [7-9] 이미지와 객체의 라벨만이 주어졌을 때 이미지 안에 있는 객체의 위치를 예측하는 것을 목표로 한다. 보통 클래스 분류 네트워크로부터 지역화(localization)맵을 찾는 것이 일반적인 방법이며, 지역화맵은 목표 객체의 정확한 바운딩 박스를 얻기 위해서 이용된다. CAM[1]은 학습된 클래스 별로 지역화맵을 생성하는데 가장 널리 사용되는 방법이지만 CAM을 통해서 이미지 내의 객체 위치를 찾을 때 객체의 전체가 아닌 일부분에 집중하여 주의하는 단점이 있다. 이러한 CAM의 지역화맵에 표현된 주의 영역의 일부 또는 전부를 지우며(erase) 지운 영역의 재학습을 통해 객체 영역의 전반적

인 영역을 포함시키려는 방법들이 제안되고 있다[2-5]. 이 방법들을 이용하면 기존 CAM의 핵심 주의 영역이 가려져서 네트워크가 학습되는 과정에서 핵심 영역 주변의 객체에 포함되는 보다 넓은 영역이 주의되는 경향이 있다.

Hide and Seek(HaS)[3] 방법은 입력한 이미지 내의 일부를 임의로 지우고, 지운 이미지들로 재학습한다. HaS의 결과로 객체에 대한 주의 영역은 확장되지만 재훈련 결과와 처음의 CAM 결과를 합쳐야 하는 과정이 추가로 필요하다. 또한 이미지 내의 영역 일부를 랜덤하게 선택하여 지우기 때문에 학습이 제대로 되지 않을 가능성이 있으며, 이 경우에는 다시 랜덤하게 영역을 선택하고 재훈련하는 과정을 필요로 한다.

CutMix[4]는 학습하는 동안 지워진 영역을 전혀 다른 이미지의 일부분으로 채우며, 이는 일반적으로 학습 이미지의 수를 늘리기 위해서 사용하지만 이미지 내 객체의 주의 영역을 확장 시켜주는 효과도 있다.

Self-Enhancement Map(SEM)[5]은 CAM의 학습 결과로 만들어진 주의 영역 내에서 K개의 상위 시드(Seed)를 뽑아, 각각의 시드들과 이미지 내에서 유사도가 높은 픽셀들로 이루어진 맵을 만들어 이를 더한다. 이는 픽셀간의 유사도가 높은 부분들은 모두 하나의 객체로 보는 방법으로, 이미지 내 객체의 주의 영역을 확장 시켜주는 효과가 있다. 그러나 하나의 객체가 아니면서 유사도가 높은 부분들 역시 동일한 객체로 보는 경우가 많기 때문에 객체 추출의 정확도가 높지 않다.

본 연구에 가장 근접한 방법인 ACoL[2]은 CAM으로의 학습이 끝난 후 지역화맵에 핵심 주의 영역으로 간주되는 영역을 픽셀값 0(검은색)으로 하여 지운다. 이후 이렇게 만든 학습데이터들을 재훈련시켜 이미지 내의 객체에 대한 주의 영역을 확장시킨다. 그러나 ACoL 방법을 통해서 재훈련시킨 새로운 CAM 모델의 지역화맵이 처음 CAM의 주의 영역을 포함하지 않는 경우가 많이 발생하는 단점이 있으며, 이를 해결하기 위해서 단계별로 재학습한 결과 영역들을 모두 합치는 과정을 별도로 수행한다. 또한 ACoL은 이미지 내의 객체의 특징이 집중된 영역을 지운 후 학습하기 때문에 객체의 클래스를 분류하는데 정확도가 낮아지는 단점도 있다.

Selective Search[6]는 객체 인식이나 검출을 위해 이미지로부터 가능한 후보 영역을 모두 찾아내는 알고리즘이다. 이 방법은 완전 탐색(exhaustive search) 방식과 분할(segmentation) 방식을 결합하여 후보 영역을 잘 찾는 장점이 있다. 완전 탐색 방식은 이미지 내의 특징과 상관없이 후보가 될 수 있는 모든 영역을 전부 조사하는 방식이지만, 분할 방식은 완전 탐색 방식과는 다르게 입력 수준의 저수준 특징을 고려하여 분리하는 방식이다. 색, 질감, 모양, 영역의 크기 및 서로 간의 거리 등 다양한 기준에 따라 분할이 가능하지만 모든 이미지에 대해서 동일하게 적용할 수 있는 분할 방식을 찾는 것은 불가능하다. 따라서 SS는 분할 방식을 활용하여 후보 영역을 찾기 위한 시드를 설정하고 그 시드에 대하여 완전 탐색 방식을 적용하여 객체를 인지하는 후보를 찾는다. SS 방법은 이미지의 모든 영역을 검색하기 때문에 계산 비용이 크지만, 이미지 데이터의 특징과 유사도를 함께 고려해서 객체를 찾기 때문에 객체를 찾는데 적한한 방법 중 하나이다.

본 논문에서는 CAM의 장점과 SS의 장점을 함께 이용하여 보다 정확하고 포괄적으로 이미지 내의 객체 영역을 찾는 방법을 제안한다. 제안하는 방법은 CAM의 결과로 만들어진 핵심 주의 영역을 가리는 대신에 가우시안의 필터의 스무딩 효과를 이용해서 영역내의 특징의 세기를 약화시킨 후 다시 학습에 참여시키는 방법이다. 이를 통해 객체의 핵심 주의 영역에 대한 특성을 어느 정도 남겨둠으로써 처음 CAM이 주목한 주의 영역으로부터의 확장된 주의 영역이 만들어지도록 한다. 또한 이 과정에서 처음 CAM의 핵심 주의 영역에 Selective Search 방법을 포함시켜 블러링 영역을 확장시켜 적용하여 CAM의 재학습 데이터로 사용한다. 또한 핵심 주의 영역이 흐려져서 객체의 클래스 분류 정확도가 오리지널 CAM에 비해서 낮아질 수 있기 때문에 이를 보완하기 위해서 클래스 분류 예측은 처음의 CAM 결과를 사용하며, 핵심 주의 영역을 찾는 것을 재학습을 통한 새로운 CAM 모델을 사용한다.

3. 제안 방법

제안하는 방법의 전체 흐름은 Fig. 2와 같이 2단계로 구성된다. 1단계에서는 클래스 기반의 하향식(top-down) 방식인 CAM으로 Fig. 2(b)와 같이 히트맵을 찾고, 여기서 강하게 반응하는 영역을 핵심 주의 영역으로 바운딩 박스를 Fig. 2(c)와 같이 결정한다. 또한 이미지 특징 기반의 상향식(bottem- up) 방식인 SS(Selective Search)를 이용해서 객체의 후보 영역을 Fig. 2(d)와 같이 찾는다. CAM의 바운딩 박스 영역을 중심으로 그 영역과 일정부분 이상 겹치는 SS의 영역들을 결합하여 Fig. 2(e)를 거쳐 Fig. 2(f)의 결합된 영역을 만든다. 여기까지 수행한 후 Fig. 2(f)의 확장된 주의 영역에 가우시안 스무딩 필터를 적용하여 Fig. 2(g)와 같이 해당 영역을 흐리게(Blur) 만들며, 이는 CAM을 이용한 재학습 과정에서 처음 단계에서의 핵심 주의 영역에 대한 집중을 낮추기 위함이다.

제안한 방법의 2단계에서는 1단계에서 얻어진 결과 이미지들을 학습데이터로 이용하여 CAM을 재학습시켜 새로운 모델을 만든다. 재학습된 새로운 CAM 모델에 Fig. 2(a)의 입력 이미지를 넣어주면 Fig. 2(h)와 같은 결과를 얻게 되며 처음 만든 CAM 모델로 추출한 주의 영역보다 핵심 주의 영역이 확장된 것을 알 수 있다. 여기서 가우시안 스무딩 필터로 영역을 흐리게 만든 이유는 CAM이 일반적으로 이미지 특징이 많이 몰려 있는 영역에 대해서만 집중적으로

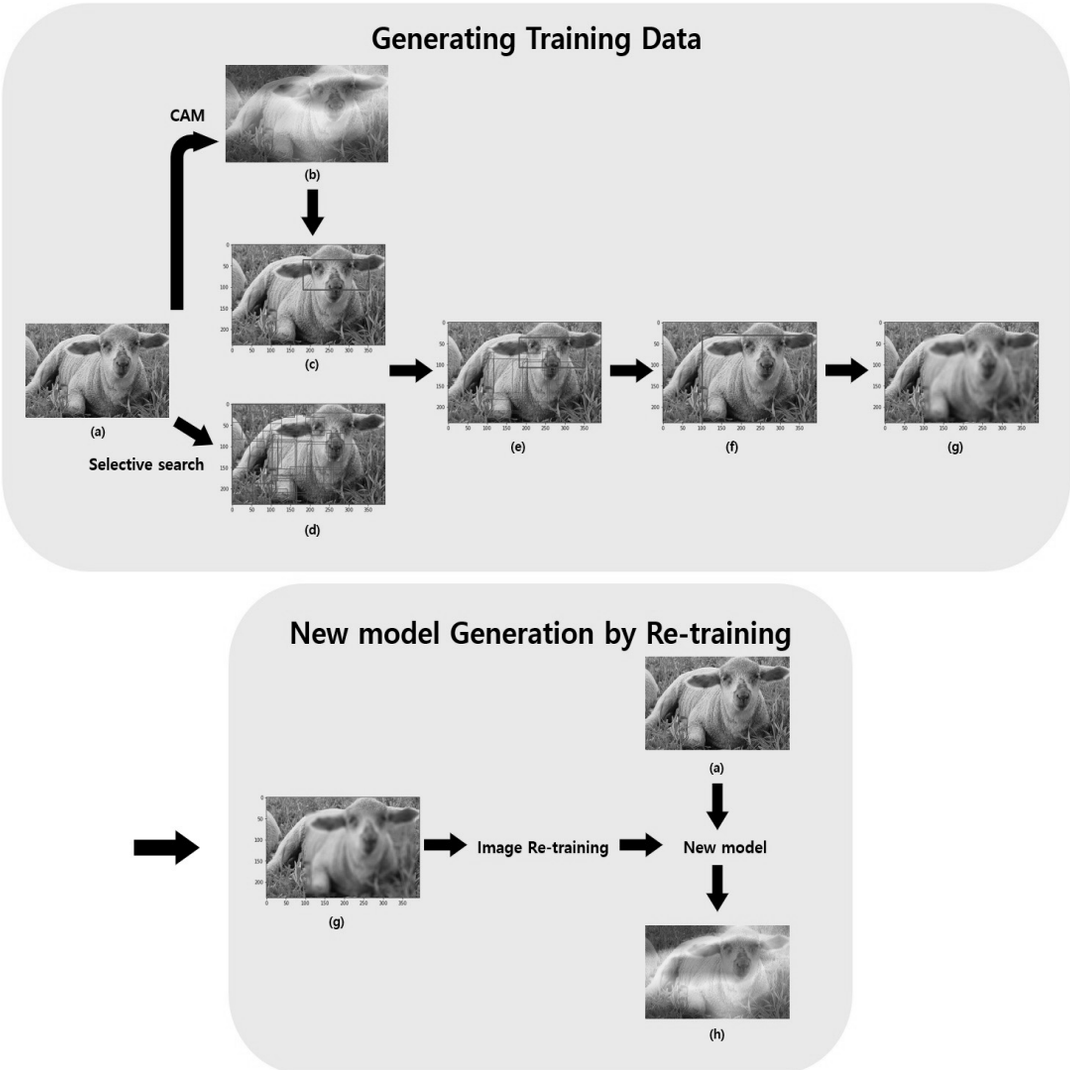


Fig. 2. Overview of the Proposed Method

반응하는 경향이 있어서 이 부분을 완화시킨 새로운 학습데이터를 만들기 위함이다. 각 과정에 대한 자세한 설명은 다음과 같다.

3.1 CAM을 이용한 주의 영역 추출

이미지와 이미지 객체의 클래스 라벨만을 주어 CNN 모델을 학습시킨다. 여기서 사용한 CNN 모델은 특징 추출 단계의 레이어와 연결된 Fully-Connected(FC) 레이어를 제거하고 대신 Global Average Pooling(GAP) 레이어로 대체한 것이다[1]. 이와 같이 GAP 레이어를 사용하면 이미지 내의 객체의 위치 정보를 파악할 수 있는 장점이 있다. 학습된 CNN 모델에 이미지를 입력하면 CNN의 마지막 특징맵들에 가중치가 반영된 각 클래스별 CAM 히트맵(heat map)을 만들 수 있다. Fig. 3은 이에 대한 예로서 이미지 내부에 있는 객체에 대한 분류 라벨과 그 객체의 위치에 따른 히트맵 정보를 얻게 된다. Fig. 3의 CAM은 주어진 이미지 객체인 “양(Sheep)”을 인식하는데 필요한 영역의 중요도를 히트맵으로 보여주며, 여기서 빨간색과 노란색에 해당하는 부분들이 바로 전 단계의 다양한 특징맵에서 강하게 반응한 중요 특징들이 반영된 것을 보여준다.

3.2 Selective Search를 이용한 영역 추출

동일한 이미지에 Selective Search(SS) 방법을 적용한다. 이 방법의 초기 단계에서는 [10]에서 제안한 그래프 기반의 이미지 영역 분할(Segmentation) 방식을 적용하여 작은 객체 후보 영역을 많이 만들어 낸다. 다음으로 Greedy 알고리즘을 적용하여 작은 영역을 반복적으로 통합해 간다. 이 과정은 우선 각 영역의 유사도를 계산한 후 유사도가 가장 높은 영역을 통합해서 점점 큰 영역으로 만들어가며 설정한 멈춤 조건에 도달할 때까지 반복된다. 이 논문에서 사용한 멈춤 조

Fig. 3. CAM Algorithm

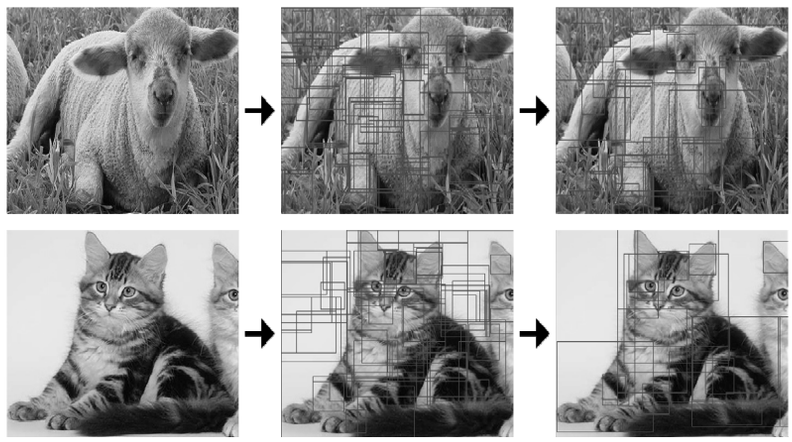


Fig. 4. Extracted Bounding Boxes by SS

건은 새롭게 합쳐진 바운딩 박스의 크기가 이미지 전체 크기의 1/4를 넘어서는 시점으로, 이보다 바운딩 박스가 더 커진다면 바운딩 박스의 생성이 멈추게 된다. 끝으로 통합된 영역들을 바탕으로 후보 영역이 만들어진다.

각 영역의 통합에 사용한 유사도는 영역의 색, 질감, 크기 및 다른 영역의 포함 정도를 함께 사용한다. 유사도에 따라 통합한 결과 영역에 바운딩 박스를 씌우면 SS를 이용한 결과 영역을 만들어진다. Fig. 4는 SS를 통해서 영역이 통합되는 과정과 예를 보여준다.

3.3 CAM과 SS 영역들의 결합을 통한 주의 영역 확장

CAM의 바운딩 박스를 기준으로 하여 SS의 바운딩 박스들과의 겹침 정도를 비교한다. 이 둘의 겹침 정도가 충분한 SS 바운딩 박스는 남겨두고 나머지 바운딩 박스들은 모두 제거한다. 예를 들어 Fig. 5에서 A+B는 CAM의 바운딩 박스이고, B+C는 SS의 바운딩 박스라고 할 때 겹치는 역인 B의 크기가 SS의 바운딩 박스인 B+C 크기의 70% 이상이 된다면 충분히 겹치는 바운딩 박스로 간주하여 남겨둔다. 바운딩 박스를 합쳐 새로운 확장된 영역으로 만든다.

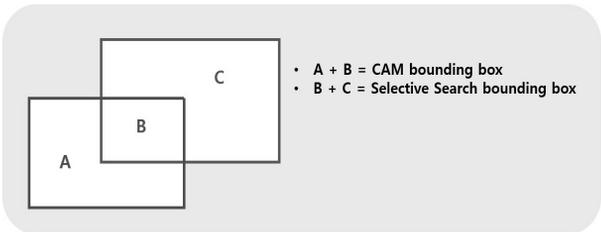


Fig. 5. Combining Boxes of CAM and SS

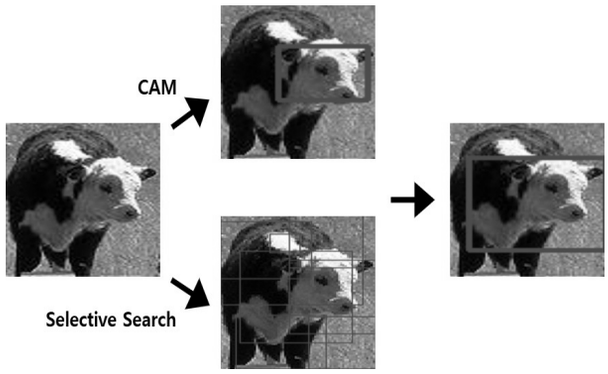


Fig. 6. Extended CAM Bounding Box Considering with SS Bounding Boxes

Fig. 6은 CAM에서 추출한 핵심 주의 영역과 겹치는 SS의 바운딩 박스 영역들을 결합하여 만들어진 확장된 주의 영역의 예를 보여준다.

3.4 주의 영역의 블러링 통한 재학습 데이터 생성

위에서 구한 확장된 주의 영역에 9x9 크기의 가우시안 스무딩 필터를 적용하여 영역 내부를 흐리게 만든다. ACoL에서는 CAM으로 추출한 주의 영역만을 픽셀값 0으로 한 검은색 마스크를 사용해서 특징이 많이 모여 있는 영역을 지우고 재학습을

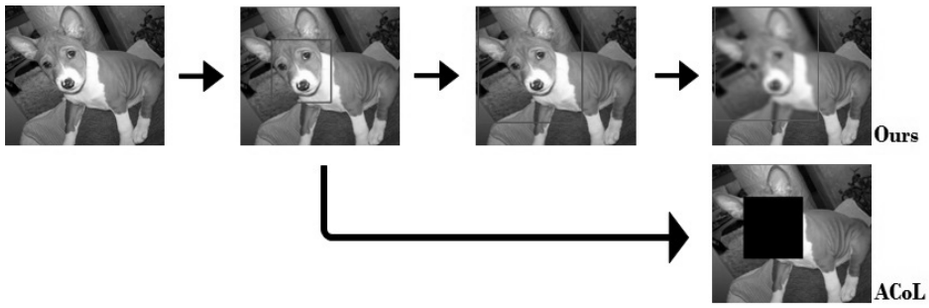


Fig. 7. Generation of Our Re-training Blurred Data and Comparison with ACoL

반복적으로 수행했지만, 제안한 방법에서는 SS를 통해 확장한 주의영역에 객체의 특징이 어느 정도 살아있도록 블러링하고 단 한번의 재학습을 시키는 차이점이 있다. Fig. 7은 제안하는 방법으로 블러링시킨 재학습 데이터와 ACoL 방법으로 지운 재학습 데이터를 보여준다. 실험 결과에서 보겠지만 제안한 방법의 주의 영역에 대한 IOU 개선 정도가 검은색으로 지운 마스크를 사용한 방법보다 뛰어남을 확인할 수 있다.

3.5 재학습을 통한 새로운 CAM 모델 생성

1단계에서 사용한 학습한 이미지들에 대해서 확장된 주의 영역을 블러링한 후 이를 다시 CNN 기반의 CAM 모델에 적용하여 새로운 학습을 수행한다.학습된 결과로 만든 새로운 CAM 모델을 CAM2라 하고, 처음의 CAM 모델은 CAM1이라 할 때, CAM2가 만들어 낸 주의 영역이 CAM1이 제시한 주의 영역보다 전반적으로 확장되어 IOU 값이 개선된 것을 실험에서 확인할 수 있다. 또한 재학습을 통해서 만들어진 CAM2가 새로운 주의 영역을 찾는 과정에서 계산량이 많은 SS 알고리즘을 사용할 필요가 없는데 이는 주의 영역을 확장시켜 재학습 데이터를 만들 때만 SS가 사용되기 때문이다. Fig. 2(h)를 다시 보면 새로운 지역화맵은 기존 CAM의 지역화맵과는 달리 이미지 내부의 객체를 전반적으로 포함하고 있는 것을 알 수 있고, 이는 다양한 데이터로 실험한 결과에서도 확인할 수 있다.

3.6 CAM1 예측 결과를 이용한 CAM2 주의 영역 추출

CAM2 모델은 학습데이터의 핵심 특징 영역을 흐리게 한 후 이를 학습데이터로 사용하였기 때문에 CAM1에 비해 객체에 대한 분류 정확성이 떨어진다. 즉 CAM1의 객체 분류 정확도가 더 높기 때문에 이 분류된 클래스에 해당되는 CAM2 결과의 확장된 히트맵에서 영역을 추출하도록 한다. 이는 CAM2의 해당 클래스의 분류 확률이 가장 높지 않을 수 있기 때문이다. Fig. 8의 예와 같이 테스트 과정에서 주어진 이미지의 객체 클래스는 'Sheep'이다. CAM1에서는 'Sheep'이 1순위로 분류되었지만, CAM2에서의 1순위는 'Cow'로 분류되었다. CAM2에서 'Sheep'은 2순위이지만 분류 정확성이

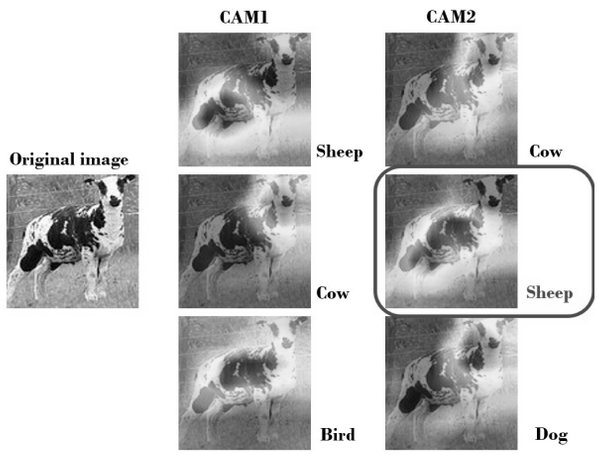


Fig. 8. Selection of CAM2 Heatmap(attention area) by Considering CAM1 Prediction Results

높은 CAM1의 결과를 인정하여 CAM2의 'Sheep'에 해당하는 히트맵에서 확장된 주의 영역을 찾는다. 이를 반영한 IoU 개선 정도를 실험에서 확인할 수 있다.

4. 실험 및 결과

4.1 실험 환경

학습 및 테스트 데이터 셋은 Visual Object Classes Challenge 2012 (VOC 2012)의 이미지를 사용하였다. VOC 데이터는 총 20개의 클래스로 구성되어 있으며, 학습, 검증, 테스트 셋으로 각각 나누어져 있다. 여기서 학습은 20개의 각 클래스 별로 400개의 이미지를 사용했으며, 검증(validation) 및 테스트 이미지는 각 100개를 사용하였다.

VOC 데이터 셋은 이미지 검출(detection)[11,12] 및 분할(segmentation)[13]을 위한 바운딩 박스가 제공되지만, 본 연구는 WSOL에 대한 연구로서 학습 시에 이미지의 클래스 레이블만을 사용한다. 또한 학습데이터 수를 증가시키기 위해서 랜덤으로 이미지를 선택하여 수평으로 뒤집어(flip) 사용하였다. 여기서 가우시안 스무딩 필터의 크기는 9*9로 설정하였다.

CAM을 구축하기 위한 CNN 모델은 ILSVRC으로[14] 사전 학습된 Resnet50 모델을[15] 사용했으며, Cross Entropy 손실 함수와 Adam 알고리즘을 사용하여 CNN을 최적화시켰다. 전체 Epoch은 25, 초기 학습률은 0.001로 설정했으며 Epoch 7마다 학습률이 1/10씩 줄어들도록 하였다.

4.2 IOU(Intersection Of Union) 비교

Table 1은 기존 CAM1의 검증 진리표(Ground Truth)와의 IOU 결과와 제안한 방법으로 만든 CAM2의 검증 진리표와의 IOU 결과를 비교한 것이다. IOU는 검증 진리표 영역과 예측 바운딩 박스 사이의 (Area of overlap)/(Area of Union)으로 측정하며, IOU가 0.5 정도가 되면 객체의 검증 진리표 영역과 예측 바운딩 박스의 겹치는 정도가 검증 진리표 영역을 기준으로 약 67%에 해당되는 수치이다. Table 1에서 20개의 모든 클래스에 대해서 제안한 방법의 IOU가 향상된 것을 확인할 수 있다. CAM1의 평균 IOU는 0.2735이며, 제안한 방법인 CAM2의 평균 IOU는 0.4315로서 약 57% 개선되었다.

여기서 CAM2는 이미지 객체 영역의 핵심 부분을 흐리게 한 후 학습하였기 때문에 20개의 클래스에 대한 테스트 데이터의 객체 분류 정확성은 81%로서 CAM1의 92%보다 11% 낮은데, 이는 객체의 중요한 특징이 속한 부분이 흐려진 상태에서 학습하여 얻어진 결과이기 때문이다. 따라서 3.6에서 제시한 분류 정확성을 유지하면서 주의 영역을 확장할 수 있는 방법으로서 CAM1의 분류 결과에 대응하는 CAM2의 해당 클래스의 히트맵에서 확장된 주의 영역을 찾도록 하였다. 이를 이용한 결과가 Table 1의 (CAM1 classification + CAM2 attention area extraction)으로서 전체 IOU 평균은 0.4348로서 CAM2 보다 1.86% 개선되었고, CAM1보다는 약 58.86% 개선되었다.

Table 1에서 (CAM1 classification + CAM2 attention area extraction)의 IOU 결과가 CAM2에 비해서 낮아진 클래스들도 있음을 볼 수 있다. 이는 전체 분류 정확도는 CAM1이 CAM2보다 높지만, 클래스 5, 6, 7, 8, 9, 13, 14, 17에서 CAM2의 정확도가 평균 90.8%로 CAM1의 평균 88.1%보다 높기 때문이다. 또한 클래스 11, 12에서는 CAM1의 분류 정확도가 CAM2보다 높지만, CAM2에서 클래스 11 dog을 클래스 7 cat으로, 클래스 12 horse는 클래스 9 cow로 잘못 예측하는 경우가 다수 발견되었다. 이 때 dog과 cat, horse와 cow 클래스 사이에 세부적인 요소들은 다르지만 자전거, 새, 병, 기차 등의 다른 클래스들과 비교할 때 전체적인 모습이 비슷하고, 클래스 7, 9가 클래스 11, 12보다 더 큰 주시 영역으로 물체를 주시하는 경향으로 인해서 IOU 스코어가 더 높게 나온 것으로 분석된다.

Fig. 9는 다양한 객체에 대해서 기존 CAM의 객체 추출 영역과 제안한 방법에서 SS를 함께 사용하여 추출한 바운딩 박스 영역을 비교한 것이다. SS를 이용했을 때 기존 CAM의 바운딩 박스보다 포괄적으로 객체를 포함하는 것을 확인할 수 있다.

학습 및 검출 처리 시간을 고려해 볼 때 재학습 데이터 한 개를 만들기 위해서 가우시안 필터링과 Selective search 방법이 적용되어 평균 3.65초가 소요되며, 학습 및 검증에 사용되는 전체 10,000개의 데이터를 재학습 데이터로 만드는데 모두 36,500초(10.13시간)가 소요된다. 이는 Selective search 알고리즘 안에 Exhaustive search 방법이 사용으로

Table 1. Comparisons of IOU by CAM1,CAM2 and (CAM1 Classification+CAM2 Attention Area Extraction)

IOU	0. Aeroplane	1. Bicycle	2. Bird	3. Boat	4. Bottle
CAM1	0.26	0.31	0.31	0.29	0.26
CAM2	0.39	0.39	0.42	0.48	0.50
CAM1 recognition+CAM2 attention area extraction	0.47	0.41	0.43	0.68	0.69
IOU	5. Bus	6. Car	7. Cat	8. Chair	9. Cow
CAM1	0.30	0.23	0.25	0.29	0.28
CAM2	0.49	0.55	0.48	0.45	0.49
CAM1 recognition+CAM2 attention area extraction	0.39	0.44	0.33	0.33	0.48
IOU	10. Dining table	11. Dog	12. Horse	13. Motorbike	14. Person
CAM1	0.27	0.27	0.28	0.29	0.19
CAM2	0.40	0.37	0.39	0.37	0.38
CAM1 recognition+CAM2 attention area extraction	0.41	0.35	0.34	0.36	0.30
IOU	15. Potted plant	16. Sheep	17. Sofa	18. Train	19. Tv/monitor
CAM1	0.25	0.27	0.28	0.32	0.27
CAM2	0.49	0.50	0.44	0.35	0.30
CAM1 recognition+CAM2 attention area extraction	0.52	0.56	0.34	0.36	0.42

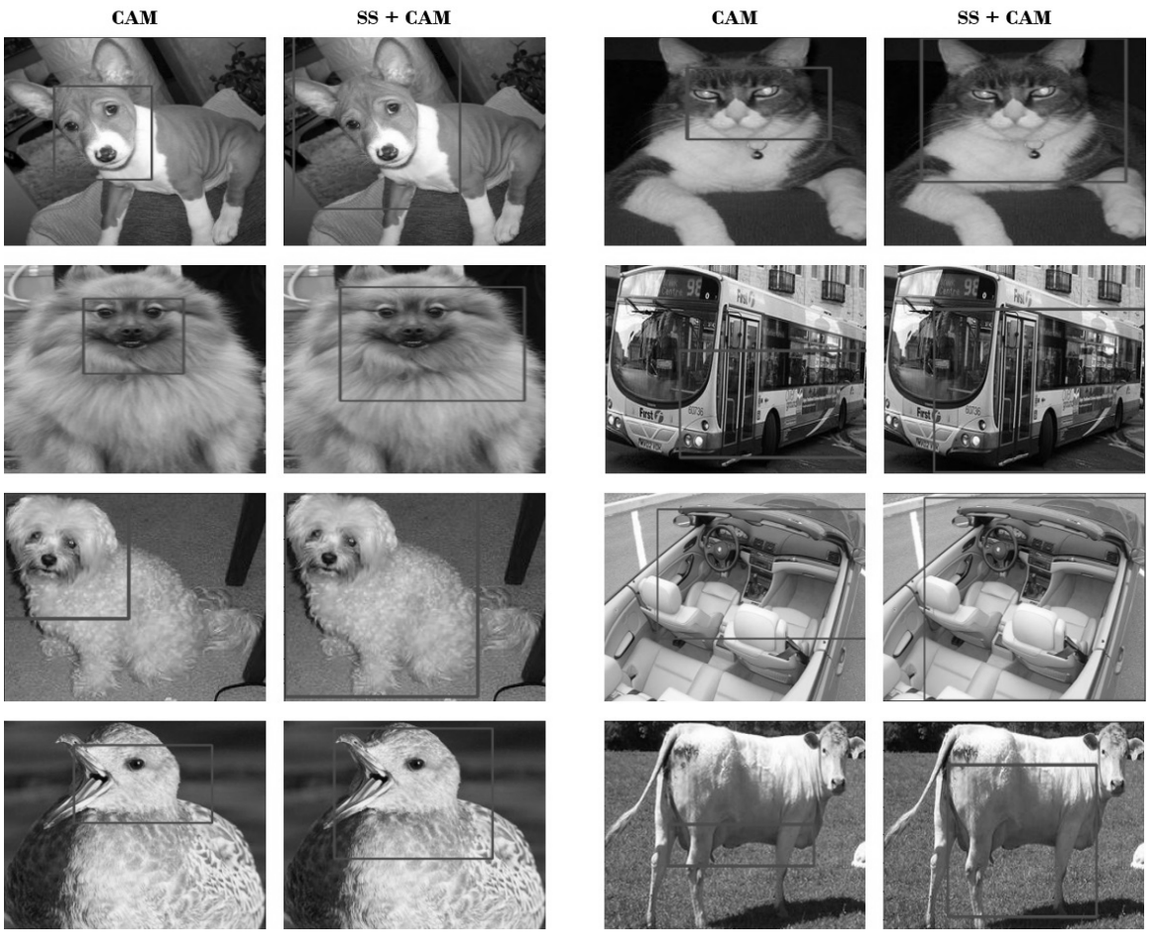


Fig. 9. Comparing Bounding Boxes between CAM and CAM+SS Methods

인해서 시간이 꽤 걸리지만, 재학습 데이터는 처음 한 번만 만들어서 모델을 학습하고 영역을 찾는 과정에서는 사용되지 않는다. 이는 주의 영역을 확장하는 대표적인 방법인 ACoL [2]에서는 한 개의 재학습 데이터를 만드는 데 2.83초가 소요되어 제안하는 방법보다 빠르지만, ACoL은 여러 번의 재학습이 필요한 방법으로서 전체 재학습 시간은 제안하는 방법보다 오래 걸리는 것으로 추정된다.

4.3 정성적 실험 결과

Fig. 10은 기존의 CAM을 이용하여 만든 히트맵과 제안한 방법의 CAM 히트맵을 비교한 것이다. 기존 CAM의 주의 영역은 객체의 특징이 많이 모여 있는 일부만을 가리키고 있다. 또한 한 번의 재훈련을 거친 ACoL 결과에 대한 히트맵은 ACoL 제안 의도와 같이 기존 CAM의 히트맵에서 가리키고 있는 객체의 특징 영역과 다른 영역을 가리키며 이 둘을 합칠 때 주의 영역이 확장되는 것을 알 수 있다. 그러나 제안한 방법의 CAM은 주의 영역이 대부분 객체로부터 시작하여 확장되며 보다 포괄적으로 객체를 포함하고 있는 것을 확인할 수 있다.

5. 결 론

이 연구에서는 기존 CAM이 이미지 내의 객체 영역을 전체적으로 찾는 데 어려움이 있는 단점을 보완하는 방법을 제안하였다. 기존의 방법들이 주의 영역을 확장하기 위해서 일부 영역을 랜덤하게 정하고 랜덤 이미지로 대체하거나 주의 영역만을 검은색으로 지우는 방법과는 달리 제안하는 방법은 SS를 이용한 영역 확장과 가우시안 블러링을 통해서 객체의 분류 정확성을 어느 정도 유지하면서 핵심 주의 영역을 확장하는 방법을 제안하였다. 제안한 방법을 통해서 기존의 CAM IOU보다 57%에서 58.86%의 개선된 IOU 결과를 얻었다. 또한 제안한 방법은 단 한번의 재훈련만으로도 초기 CAM의 주의 영역을 포함하면서 확장된 주의 영역을 얻기 때문에 다른 방법들 보다 우수한 장점이 있다.

향후 연구에서는 국내외의 다양한 데이터 셋으로 실험을 확장하여 성능 향상을 검증하고 IOU 성능을 지속적으로 향상시키는 연구를 진행하고자 한다. 또한 향후 연구에서는 객체의 영역 추출 성능 향상과 함께 객체 분류의 정확성도 유지하는 방법을 고려하고자 한다.

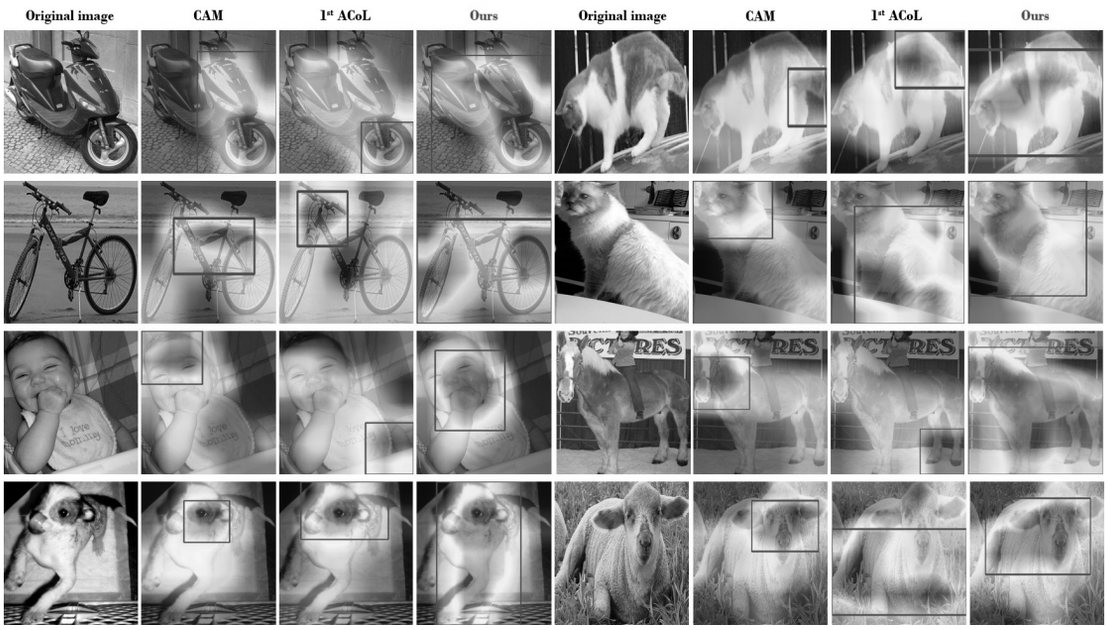


Fig 10. Comparing Attention Areas between CAM, ACoL and Our Method

References

[1] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Computer Vision and Pattern Recognition*, pp.2921-2929, 2016.

[2] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *IEEE Computer Vision and Pattern Recognition*, pp.1325-1334, 2018.

[3] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," arXiv preprint arXiv:1704.04232, 2017.

[4] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *International Conference on Computer Vision*, pp.6022-6031, 2019.

[5] X. Zhang, Y. Wei, Y. Yang and F. Wu. Rethinking Localization Map: Towards Accurate Object Perception with Self-Enhancement Maps. *Computer Vision and Pattern Recognition preprint*, 2020.

[6] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *International Computer of Computer Vision*, Vol.104, pp.154-171, 2013.

[7] L. Bazzani, A. Bergamo, D. Anguelov, and L. Torresani, "Self-taught object localization with deep networks," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2016.

[8] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. Manjunath, "Weakly supervised localization using deep feature maps," *European Conference on Computer Vision*, pp.714-731, Springer, 2016.

[9] D. Li, J. B. Huang, Y. Li, S. Wang, and M. H. Yang, "Weakly supervised object localization with progressive domain adaption," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, Vol.59, No.2, Sep. 2004.

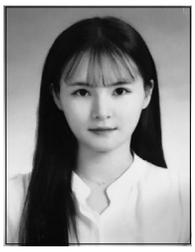
[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, Vol.39, pp.1137-1149, 2015.

[13] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," In *European Conference on Computer Vision*, pp.695-711, 2016.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Conference on Computer Vision and Pattern Recognition*, pp.248-255, 2009.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.



고 수 연

https://orcid.org/0000-0002-2373-8695
e-mail : sy1122@sookmyung.ac.kr
2020년 숙명여자대학교 컴퓨터과학과(학사)
2020년 ~ 현 재 숙명여자대학교 컴퓨터과학과 석사과정
관심분야 : 머신러닝, WSOL

최 영 우

https://orcid.org/0000-0003-0364-236X
e-mail : ywchoi@sookmyung.ac.kr
1985년 연세대학교 전자공학과(학사)
1986년 Univ. of Southern California 컴퓨터공학과(석사)
1994년 Univ. of Southern California 컴퓨터공학과(박사)
1994년 ~ 1997년 LG전자기술원 선임연구원
1997년 ~ 현 재 숙명여자대학교 컴퓨터과학과 교수
관심분야 : 시각정보처리, 머신러닝