특별논문 (Special Paper)
방송공학회논문지 제28권 제4호, 2023년 7월 (JBE Vol.28, No.4, July 2023)
https://doi.org/10.5909/JBE.2023.28.4.391
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

클래스별 평균 전경 맵을 이용한 약 지도학습 기반 객체 로컬라이제이션

박 세 진[a)], 이 소 은[a)], 강 병 근[a)‡]

Weakly Supervised Object Localization Using Class-Specific Foreground Maps

Sejin Park[a)], Soeun Lee[a)], and Byeongkeun Kang[a)‡]

요 약

최근 학습을 위한 레이블링 비용을 줄이기 위해 약 지도학습 기반의 인공지능 기술들이 많은 관심을 받고 있다. 특히, 약 지도학습 기반 객체 로컬라이제이션은 학습 단계에서 객체 위치정보 없이 학습하지만, 추론 단계에서 위치를 추정할 수 있어서 더욱 많은 관심을 받고 있다. 이는 웹에서 손쉽게 가져온 영상들로 학습할 수 있어 더욱 많은 관심을 받고 있다. 하지만 약 지도학습 기반으로 객체 로컬라이제이션을 학습하면, 영상 단위의 클래스 레이블만 사용하기에 객체의 특징적인 영역만 현지화하는 한계가 존재한다. 따라서, 본 논문에서는 클래스별 평균 전경 맵을 활용한 학습 방법을 제안하며 이를 활용하면 로컬라이제이션 정확도를 높일 수 있음을 보인다. 또한, 기존의 클래스 활성화 맵을 정규화하는 방법을 개선하여 추가로 정확도를 개선할 수 있음을 보인다. 실험은 약 지도학습 기반 로컬라이제이션에서 주로 사용되는 CUB200 데이터 세트와 ImageNet 데이터 세트를 사용하여 제안하는 방법의 효율성을 검증하였다.

Abstract

Recently, weakly supervised learning-based methods have received significant attention due to their ability to reduce labeling costs. In particular, weakly supervised object localization has become an important research topic as it aims to learn object localization without the need for precise location labels. Consequently, it can be trained using easily obtainable image data from online sources. However, since it is trained solely using image-level class labels, it has limitations in terms of localization accuracy, as it usually identify only the most distinctive regions of the object. To address this issue, we propose a training strategy that incorporates class-specific foreground maps to improve localization accuracy. Additionally, we present an enhanced method for normalizing class activation maps to further enhance localization accuracy. The effectiveness of the proposed method is validated using the publicly available CUB200 dataset and ImageNet dataset.

Keyword : Object localization, weakly supervised learning, class activation map

a) 서울과학기술대학교 전자공학과(Department of Electronic Engineering, Seoul National University of Science and Technology)
‡ Corresponding Author : 강병근(Byeongkeun Kang)
E-mail: byeongkeun.kang@seoultech.ac.kr
Tel: +82-2-970-6426
ORCID: https://orcid.org/0000-0003-2537-7720

※ This work was supported by Korea Evaluation Institute of Industrial Technology(KEIT) grant funded by the Korea government (MOTIE) (No. 20018635, Development of smart farm management system and growth monitoring).
· Manuscript May 31, 2023; Revised July 17, 2023; Accepted July 17, 2023.

Ⅰ. 서 론

영상 기반 객체 탐지, 객체 로컬라이제이션, 객체 분할 등 인공지능 기술을 활용한 영상 내 객체 인식 및 이해에 관한 연구 및 개발이 활발히 진행되고 있다. 이러한 연구들은 자율주행, 의료, 보안 등 다양한 산업에서 활발히 활용되고 있으며 좋은 성능을 보여주고 있다. 그러나 이러한 기법들은 대부분 지도학습에 의존하고 있어 학습을 위해서는 높은 라벨링 비용이 발생한다. 따라서, 이러한 비용을 줄이기 위해 약 지도학습, 비지도 학습 기술이 연구되고 있다.

약 지도학습은 학습을 위해 필요한 모든 레이블이 아닌 그중 일부 정보만을 가지고 학습하는 것을 목표로 한다. 따라서 지도학습보다 레이블을 생성하는 비용이 낮다. 예를 들어, 클래스 정보만을 가지고 객체를 로컬라이제이션(localization)을 하거나, 세그멘테이션(segmentation)하는 것이다. 따라서 약 지도학습 기술의 목표는 상대적으로 적은 레이블 비용으로 지도학습의 결과와 비슷한 결과를 얻는 것이다. 구체적으로 약 지도학습 기반의 객체 로컬라이제이션이란, 클래스 정보만으로 객체의 종류와 위치를 정확하게 예측하는 것이다. 하지만 이러한 약 지도학습 기반의 객체 로컬라이제이션 방법들은 특정 영역에만 집중하는 한계가 존재한다. 또한, 클래스별로 가지고 있는 특정한 영역이나 모양을 반영하지 못한다. 따라서 본 논문에서는 클래스별 전경 맵을 활용한 학습법을 제안한다. 이러한 학습법을 통해 클래스별로 가지고 있는 특징을 클래스 활성화 맵을 만드는 과정에서 반영할 수 있도록 하였다. 또한, 클래스 활성화 맵을 이용하여 로컬라이제이션하는 과정에서 이용되고 있는 정규화 방법을 개선하였다. 결론적으로 이러한 두 가지 방법을 통해 약 지도학습 기반의 객체 로컬라이제이션의 성능을 높여 방법이 유효함을 증명한다. 2장에서는 약 지도학습 기반의 객체 로컬라이제이션에 대한 기존 연구들을 설명하고, 3장에서는 전경 맵을 활용한 학습법과 개선된 정규화 방법을 설명한다. 4장에서는 이러한 방법들을 활용한 결과를 기존의 연구들과 비교하며 양적 결과와 질적 결과를 모두 보여준다.

Ⅱ. 관련 연구

약 지도학습 기반의 객체 로컬라이제이션은 객체의 위치 정보 없이 클래스 정보만으로 학습한다. 따라서, 라벨링에 대한 부담이 적어 많은 연구가 활발히 이루어지고 있다. 특히 약 지도학습 기반의 객체 로컬라이제이션의 핵심이라고 할 수 있는 클래스 활성화 맵[1]의 등장으로 많은 연구가 시작되었다. 클래스 활성화 맵이란 설명 가능한 인공지능에서도 자주 쓰이는 기술로, 네트워크가 사진의 어디를 보고 예측을 했는지 알 수 있게끔 시각화한 맵이다. 따라서 약 지도학습 기반의 객체 로컬라이제이션에서는 객체의 위치 정보 레이블의 부재를 클래스 활성화 맵을 활용하여 해결하고 있다. 하지만 클래스 활성화 맵을 이용함에도 한계점이 존재한다. 구체적으로 클래스 정보만으로 객체의 클래스와 로컬라이제이션을 동시에 학습하기에, 네트워크는 객체를 구분하기 위해 객체의 특정 부분에 더욱 집중하게 되고, 클래스 활성화 맵 역시 특정 부분만 집중하게 되는 현상을 보인다. 따라서, 이러한 한계점을 해결하기 위해 다양한 방법들이 제안되었다. HaS[2]는 클래스 활성화 맵이 특정 영역에만 집중하는 문제를 해결하기 위해 무작위 제거 학습법을 제안했다. 클래스 활성화 맵을 패치로 나누어 무작위로 제거하며 학습을 진행하였다. 반면 SPOL[3]은 네트워크의 얇은 층에서 출력되는 클래스 활성화 맵에 집중하였다. 기존 클래스 활성화 맵은 네트워크의 마지막 컨볼루션 층에서 출력하기에 얇은 층에서 뽑힌 클래스 활성화 맵보다 특정 영역에 집중하는 경향을 보인다. 따라서 그들은 얇은 층에서 출력된 맵과 마지막 층에서 출력된 맵을 곱하여 보완된 클래스 활성화 맵을 이용하여 문제를 해결하였다. 한편, NM[4]은 클래스 활성화 맵을 이용하기 위해 정규화를 거치는 과정에 집중하였다. 다른 논문들의 방법을 사용할 경우, 발생하는 클래스 활성화 맵의 분포를 분석해서 기존의 정규화 방법을 개선한다. 특히, 클래스 활성화 맵 안의 낮은 값들을 무시하여 상대적으로 높은 값들에 집중하는 방법을 제안한다. I2C[5]에서는 배치 단위로 클래스 활성화 맵에서 정답 레이블에 해당하는 맵을 추출한 후, 무작위로 특징 벡터를 몇 개 선택한다. 그리고 같은 레이블의 영상끼리 이러한 특징 벡터의 값을 손실함수를 통해 가까워지게 하였으며 이러한 특징 벡터를 시드 벡터라고 부른다. 또한,

각 배치에서의 시드 벡터와 미니 배치에서의 시드 벡터의 크기를 가깝게 만들어 클래스별 전경 맵의 일부분을 따라갈 수 있게 하였다. 그리고 최근 BAS[6]는 클래스 활성화 맵을 전경 맵과 배경 맵으로 구분하여 학습에 이용했다. BAS[6]는 기본 네트워크와 생성기, 클래스 분류기를 기본으로 하여 학습하는 방법으로, 배경 맵과 특징 맵을 합성하여 배경에 대해서 억제한다. 동시에 전경 맵으로 클래스 활성화 맵을 지역 억제한다. 따라서, 두 개의 억제 과정의 균형을 바탕으로 객체에 대해 정확하게 로컬라이제이션 할 수 있는 방법이다. 특히 배경 영역의 활성 크기를 억제함으로써 네트워크가 객체 전체의 영역에 대해 정확하게 로컬라이제이션하게 만들어 주었다. 따라서 본 논문에서는 MobileNetV1[7]을 Backbone으로 하는 방법 중에서 CUB200[8] 데이터 세트에 대해 가장 성능이 높은 BAS[6]를 초기 학습에 이용하여 연구하였으며 이를 ImageNet[12] 데이터 세트에 대한 실험에도 이용하였다. 또한, 기존에 클래스별 특징을 클래스 활성화 맵에 반영하지 못하는 점을 해결하기 위해, BAS[6]로부터 영감을 받아 클래스별 평균 전경 맵을 활용한 학습법을 제안하였으며, 기존의 정규화 방법에서 더 개선된 정규화 방법을 제안한다.

Ⅲ. 제안 방법

1. 초기 학습

본 논문에서는 BAS[6]와 같이 본 논문의 모델 구조 역시 기본 네트워크, 생성기, 그리고 클래스 분류기로 구성하였다. 여기서 생성기는 클래스 활성화 맵을 이용해 전경 및 배경 맵을 생성하는 모듈을 뜻한다. 먼저 본 논문에서는 학습 단계를 초기 학습 단계, 후기 학습 단계로 두 단계로 구분한다. 초기 학습 단계에서는 BAS[6]의 방법을 활용하여 학습하여 기초 모델을 만든다. 따라서, 본 논문에서 제안하는 방법은 후기 학습 단계와 학습 후 정규화하는 방법에서 제안된다. 먼저, 초기 학습 단계에서 영상이 기본 네트워크에 들어가서 특징 맵이 생성된다. 그리고 해당 특징 맵은 생성기와 분류기에 각각 입력된다. 먼저 생성기에서 클래스 활성화 맵을 만든다. 이 클래스 활성화 맵은 모든 클래스에 대한 정보를 가지고 있고, 여기서 모든 클래스에 대한 정보란, 모든 클래스에 대해 클래스마다 특정 전경 맵을 가지고 있다는 것을 뜻한다. 따라서, 정답 레이블에 해당하는 맵을 추출하여 이를 전경 맵이라 칭한다. 생성기는 마지막에 시그모이드 층을 포함하고 있어, 0과 1 사이의 값을 출력하게 된다. 이러한 성질을 이용해, 전경 맵의 각 특징을 (1-특징) 값으로 치환하여, 배경 맵을 만들게 된다. 그림 1에서 특징 맵이 생성기에 들어가 전경 맵과 배경 맵을 만드는 과정을 확인할 수 있다. 그리고 출력된 전경 맵은 지역 억제를 위한 손실함수와 전경을 이용한 추가 교차 엔트로피 손실함수에서 이용된다. 배경 맵은 분류기에서 출력되는 특징 맵과 연산된다. 그리고 분류기에선 특징 맵이 컨볼루션 층을 통과하며, 그 과정에서 생성기에서 만든 배경 맵과 합성하여 연산이 진행된다. 따라서 분류를 위한 교차 엔트로피 손실함수와 배경을 억제하는 손실함수에 대한 연산이 진행된다. 손실함수에 대한 자세한 수식은 3.3절에서 설명하고 있다.

2. 클래스별 평균 전경 맵을 활용한 후기 학습

초기 학습이 끝난 후, 후기 학습이 시작된다. 만약 초기 학습만을 진행하게 될 경우, 클래스별 전경 맵의 특징에 대

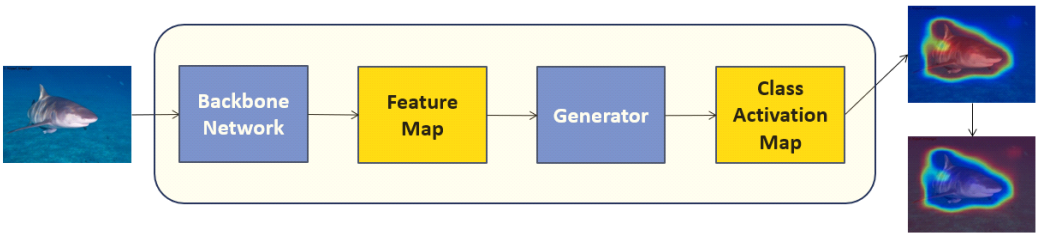

그림 1. 전경 맵과 배경 맵의 생성과정
Fig. 1. Process of generating foreground and background maps

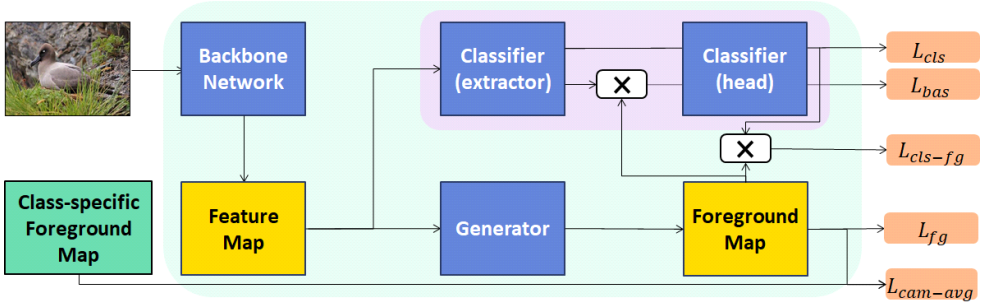

그림 2. 제안된 모델의 구조
Fig. 2. Framework of proposed method

해서는 학습되지 않는다. 실제 환경에서는 클래스별로 전경과 배경이 달라 이를 학습하는 것은 중요한 문제이다. 따라서 이러한 클래스별 특징을 이용한 방법도 있다. 예를 들어, I2C[5]에서는 배치 단위로, 전경 맵에서 시드 벡터를 무작위로 선택해, 그 값들을 서로 닮게 하였다. 하지만 이러한 방법을 적용할 경우, 전경 맵에서 시드 벡터를 무작위로 선택하기에 객체의 전혀 다른 부분에서 시드 벡터를 받아오는 한계점이 있어 만약 전경 맵이 활성화되는 범위가 넓어지더라도 이상적이지 않다. 또한, 배치와 미니 배치 단위에서 적용하였기에, 모델의 학습이 배치 사이즈에 민감할 수 있으며 하이퍼 파라미터에 따라 성능의 결과가 달라질 수 있다. 따라서 우리는 클래스별 평균 전경 맵을 활용한 학습을 제안한다. 후기 학습에서는 학습을 시작하기 전, 모든 영상에서 정답 클래스를 뜻하는 전경 맵을 먼저 출력한다. 출력된 모든 전경 맵을 토대로, 클래스별 평균을 내어 학습 데이터 세트의 클래스 수만큼 평균 전경 맵을 가지게 된다. 후기 학습에서는 이러한 클래스별 평균 전경 맵을 활용한다. 초기 학습과 똑같이 모든 연산이 진행되지만, 후기 학습에서는 각 영상의 전경 맵이 미리 생성해 둔 클래스별 평균 전경 맵과 가까워지게 학습을 진행된다. 그림 2에서 출력된 전경 맵과 클래스별 평균 전경 맵을 이용해 학습하는 것을 확인할 수 있다. 단순히, 몇 개의 시드 벡터를 뽑는 것이 아닌 각 영상의 전경 맵의 모든 특징값과 클래스별 평균 전경 맵의 특징값들이 닮게 학습된다. 이러한 방법을 토대로 모델은 클래스별 전경 맵의 분포를 일부 학습하고, 모델이 전경 맵을 만드는 과정에서 클래스별 전경 맵의 특징을 일부 반영하게 하는 효과를 보인다. 후기 학습 손실함수 역시 3.3절에서 설명하고 있다.

3. 손실 함수

이 절에서는 초기 학습과 후기 학습에서 사용되는 손실함수에 대해 정리한다. 우선 초기 학습 단계의 손실함수이다. 영상 X는 기본 네트워크 $E(\bullet)$를 통해 특징 맵을 출력한다.

$$F = E(X) \tag{1}$$

그리고 앞에서 설명한 것과 같이 이 특징 맵은 생성기 $G(\bullet)$와 분류기 $C(\bullet)$에 들어가게 되며, 생성기에서는 전경 맵을 출력하게 된다.

$$M_{fg} = G(F, Y) \tag{2}$$

그리고 여기서 생성된 전경 맵은 지역 억제 손실함수[6]를 통해 계산된다.

$$L_{fg} = \frac{1}{N}\sum_{i}\sum_{j} M_{fg}(i,j) \tag{3}$$

여기서, N은 전경 맵의 총 특징 맵의 요소 수를 뜻한다. 또한, 전경 맵과 전경 맵을 이용한 배경 맵, 그리고 기본 네트워크에서 출력된 특징 맵은 분류기에 들어가서 계산된다. 기본적으로 기본 네트워크에서 출력된 특징 맵이 분류기에 들어가 교차 엔트로피 손실함수가 계산되며, 전경 맵과 분류기의 출력인 특징 맵을 원소별 곱을 통해 연산하여, 전경 맵을 고려한 교차 엔트로피 손실함수[6]가 계산된다.

$$L_{cls} = -\sum_{i=1}^{K} y_i \log \frac{e^{\hat{y_i}}}{\sum_{j}^{K} e^{\hat{y_j}}} \quad (4)$$

$$L_{cls-\mathrm{fg}} = -\sum_{i=1}^{K} y_i \log \frac{e^{\hat{y_i}^{fg}}}{\sum_{j}^{K} e^{\hat{y_j}^{fg}}} \quad (5)$$

여기서, y는 정답 클래스를 의미하며, y^{fg}는 전경 맵과 특징 맵을 원소별 곱을 진행한 것을 의미한다. 수식 4는 영상을 넣어 출력한 교차 엔트로피 손실함수를 뜻하며, 수식 5는 전경 맵이 곱해진 후, 출력된 교차 엔트로피 손실함수를 의미한다.

또한, 배경 맵의 활성화를 줄여주는 손실함수가 있다. 분류기에 들어갔던 전경 맵을 앞서 설명한 것과 같이 배경 맵으로 변환해 준 후, 분류기의 중간 특징 맵과 원소별 곱을 한 후에 다시 분류기의 남은 컨볼루션 층을 통과한다. 최종적으로, 배경 맵과 곱해져 들어간 후의 최종 결과와 곱해지지 않은 결과의 비율로 손실함수를 계산한다. 따라서 배경 억제 손실함수[6]가 배경 맵을 억제하고, 지역 억제 손실함수가 전경 맵을 억제하게 되어 이 균형을 유지하며 학습한다. 이러한 균형을 통해 객체의 위치정보 레이블 없이 객체에 대한 로컬라이제이션을 수행하게 된다. 해당 손실 함수는 다음과 같다.

$$L_{bas} = \frac{s^{bg}}{s+\varepsilon} \quad (6)$$

여기서, ε는 매우 작은 수를 의미한다. 그리고 s^{bg}는 분류기에 배경 맵과 분류기의 중간 특징 맵이 요소별 곱이 적용된 후에 다시 분류기로 입력되어 나온 출력 값을 의미한다. 그리고 s는 배경 맵과 연산 되지 않은 결과 스코어를 의미하며, 최종적으로 s^{bg}와 s는 정답 클래스의 스코어만 추출하여 계산된다. 최종적으로 초기 학습을 위한 손실함수는 다음과 같으며, 자세한 설명은 BAS[6]에서 확인할 수 있으며, α, β, γ는 하이퍼 파라미터를 의미한다.

$$L_{first} = L_{cls} + \alpha L_{cls-\mathrm{fg}} + \beta L_{fg} + \gamma L_{bas} \quad (7)$$

다음은, 제안된 후기 학습의 손실함수 계산이다. 미리 추출해둔 클래스별 평균 전경 맵을 활용해 L1 손실함수를 이용해 계산된다.

$$L_{cam-\mathrm{avg}} = \frac{1}{B}\sum_{i=1}^{B} \| M_{fg}^{i} - M_{cls-\mathrm{avg}}^{i} \|_1 \quad (8)$$

여기서 B는 배치 사이즈를 의미하며, $M_{cls-\mathrm{avg}}$가 클래스별 평균 전경 맵을 의미한다. 이 손실함수를 통해서 클래스별 전경 맵의 특징을 학습하게 된다.

4. 개선된 정규화 방법

현재, 대부분 정규화 방법으로 최소-최대 정규화 방법, 혹은 최대 정규화 방법을 사용하고 있다. 최근, NM[4]에서는 약 지도학습 기반의 객체 로컬라이제이션을 수행하는 과정에서 정규화 방법이 굉장히 중요함을 증명했다. 실험적으로 클래스 활성화 맵을 정규화하는 과정에서 음수 값부터 일정 비율의 작은 값들은 정규화하는 것이 좋지 않은 방법임을 보였고, 따라서 해당 부분을 제외하고 정규화하였다.

하지만 본 논문에서는 시그모이드 함수를 포함하는 생성기를 사용하기에 본래부터 음수값이 존재하지 않는다. 또한, NM[4] 이전에 발표된 PaS[9]에서는 예외적으로 큰 값들이 상대적으로 작은 값을 무시하게 만들기 때문에, 예외적으로 큰 값들을 제거하는 방법을 제안하였다. 하지만 이 논문 역시, 음수가 존재하지 않는 전경 맵을 고려하지 않았다. 따라서 본 논문에서는 예외적으로 큰 값을 일부 제외할 수 있으며, 음수를 고려하지 않는 정규화 방법을 사용하였다.

$$\overline{M_{fg}} = H(M_{fg}) \quad (9)$$

$$H(m) = \frac{m}{Pct_p(\max(m))} \quad (10)$$

여기서, $\overline{M_{fg}}$은 정규화된 전경 맵, $H(\bullet)$는 정규화 함수를 뜻하며, max는 전경 맵의 최댓값, Pct_p는 p퍼센트의 값을 의미한다.

최종적으로, 클래스별 평균 전경 맵을 활용한 후기 학습법과 개선된 정규화 방법을 통해 약 지도학습 기반의 객체

로컬라이제이션의 정확도를 높이는 방법을 제안하였다. 초기 학습과 후기 학습에 대한 전반적인 구조는 그림 2에서 확인할 수 있으며, 최종적인 손실함수는 수식 (11)과 같고, δ는 조절 가능한 하이퍼 파라미터를 의미한다.

$$L_{final} = L_{first} + \delta L_{cam-\text{avg}} \tag{11}$$

Ⅳ. 실험 및 결과 분석

본 장에서는 제안된 방법의 성능 검증을 위한 실험 방법 및 결과에 관해 설명한다. 본 논문에서는 MobileNetV1[7]을 Backbone으로 하여 학습하였다. 또한, 약 지도학습 기반의 객체 로컬라이제이션 논문들이 자주 사용하는 CUB200[8]와 ImageNet[12] 데이터 세트를 이용하여 기존의 제안된 방법들과의 성능을 비교하였다. CUB200[8] 데이터 세트는 200개의 클래스로 구성되어 있으며, 총 5,994장의 학습 데이터 세트, 5,794장의 테스트 데이터 세트로 구성되어 있다. ImageNet[12] 데이터 세트는 1,000개의 클래스로 구성되어 있으며, 약 130만 장의 학습 데이터 세트, 50,000장의 테스트 데이터 세트로 구성되어 있다. 그리고 성능 측정 방식은 기존 논문들의 기법을 사용하였으며[6], 표 1에서 제안된 방법의 성능을 확인할 수 있다. 기존 논문들에서는 Top-1 Loc, Top-5 Loc, GT-known의 기법으로 측정한다. Top-1 Loc는 분류의 예측 스코어가 가장 높은 클래스가 정답 클래스이면서, 로컬라이제이션한 결과와 정답 영역의 교차 영역 비율이 0.5 이상인 것을 기록한다. 여기서 교차 영역 비율이란, 두 개의 영역이 겹치는 부분의 영역을 두 영역의 전체 영역으로 나눈 비율을 나타낸다. 또한, Top-5 Loc는 분류의 예측 스코어가 높은 5개의 클래스 중에 정답 클래스가 있으며, 동일하게 로컬라이제이션한 결과와 정답 영역의 교차 영역 비율이 0.5 이상인 것을 기록한다. 마지막으로, GT-known은 정답 클래스에 대해서 로컬라이제이션한 결과와 정답 영역의 교차 영역 비율이 0.5 이상인 것을 기록한 것을 의미한다. 그림 3에서는 제안된 방법으로 학습한 모델의 로컬라이제이션 결과를 확인할 수 있다. 표 1을 살펴보면 제안된 방법이 가장 높은 성능이 보임을 확인할 수 있다. 또한, 그림 3을 살펴보면 제안된 방법으로 원래의 영상에서 전경 맵을 합성하여 객체에 대해 로컬라이제이션을 한 결과를 살펴볼 수 있으며, 객체에 대해 전체적인 영역을 찾는 것을 확인할 수 있다. 그림 4에서는 각 손실함수가 빠졌을 때의 결과를 시각적으로 확인할 수 있다. 이 결과를 통해 각 손실함수의 중요성을 알 수 있으며, 또한 제안된 방법이 로컬라이제이션을 정확하게 하는 모습을 확인할 수 있다. 특히, L_{fg}가 제외되었을 때는, 전경 맵이 제한 없이 무한히 커지는 모습을 확인할 수 있다. 그리고 시각화한 결과에서 빨간색 경계 상자는 정답을 초록색 경계 상자는 네트워크가 예측한 상자를 의미한다. CUB200[8] 데이터 세트는 초기 학습과 후기 학습 모두 100 epoch 동안 학습되었으며, ImageNet[12] 데이터 세트는 초기 학습과 후기 학습을 8 epoch, 4 epoch 동안 하였다. 학습률은 0.001로 설정하였다.

표 1. CUB200 데이터 세트[8]와 ImageNet 데이터 세트[12]에 대한 결과 비교
Table 1. Quantitative comparison using the CUB200 dataset[8] and ImageNet dataset[12]

Method	CUB-200[8]			ImageNet[12]		
	Top-1 Loc	Top-5 Loc	GT-known	Top-1 Loc	Top-5 Loc	GT-known
CAM[1]	48.07	59.20	63.30	43.35	54.55	58.97
HaS[2]	46.70	-	67.31	42.73	-	60.12
ADL[10]	47.74	-	-	43.01	-	-
PaS[9]	59.41	-	78.60	44.78	-	61.69
FAM[11]	65.67	-	85.71	46.24	-	62.05
BAS[6]	69.77	86.00	92.35	52.97	66.59	72.00
Ours	70.68	86.30	92.77	53.31	66.84	72.77

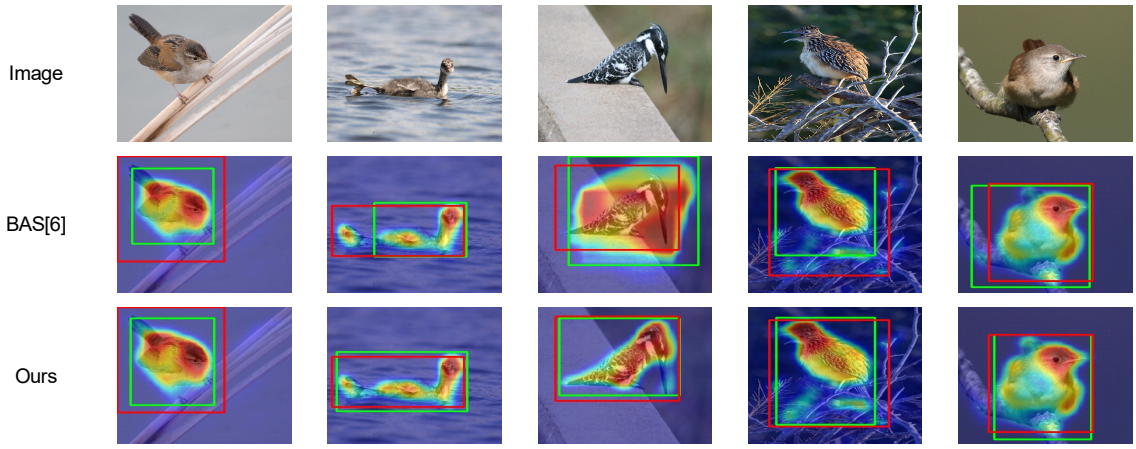

그림 3. CUB200 데이터 세트[8]에 대한 시각적 결과 비교
Fig. 3. Qualitative comparison using the CUB200 dataset[8]

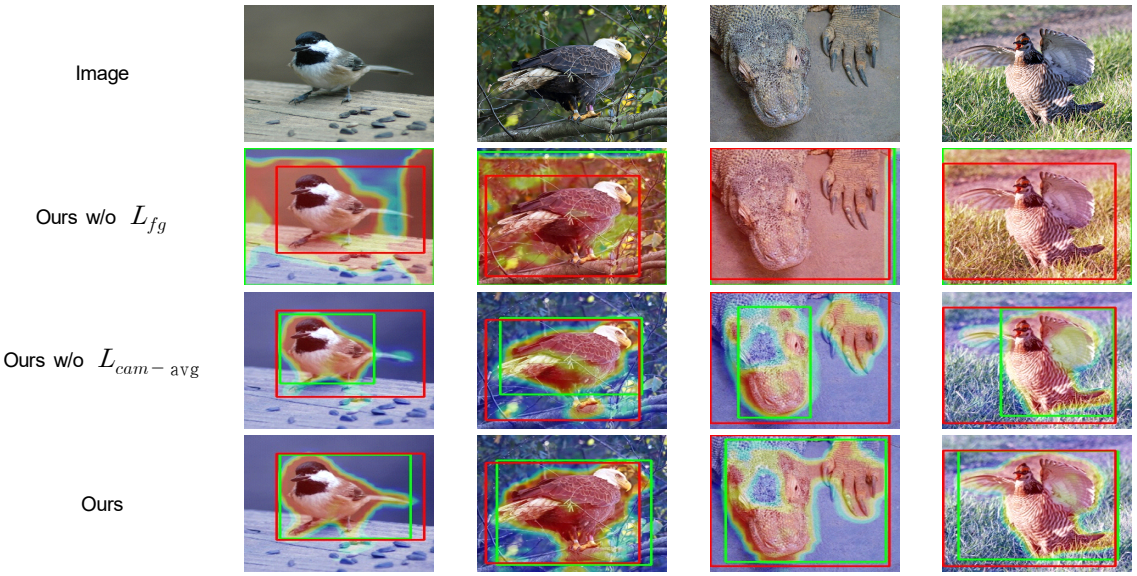

그림 4. 손실함수의 여부에 따른 ImageNet 데이터 세트[12]에 대한 시각적 결과 비교
Fig. 4. Ablation study on the loss terms using the ImageNet dataset[12]

V. 결 론

본 논문에서는 약 지도학습 기반의 객체 로컬라이제이션을 연구하였다. 특히, 클래스별 평균 전경 맵을 활용한 후기 학습 과정을 통해서 클래스별 전경 맵의 특징을 학습할 수 있게 하였다. 또한, 정규화 방법을 개선하여 양수로만 구성되어 있는 전경 맵에서 활용할 수 있는 방법을 제안하였으며, 두 가지의 방법을 통해 다른 방법들에 비해 보다 정확한 로컬라이제이션 성능을 얻을 수 있었다. 결론적으로, 이러한 과정들을 통해 객체의 위치정보 없이 객체를 로컬라이제이션하는 것을 수행하였으며, 객체의 특정 영역만을 잡는 점과 클래스별 전경 맵 특징을 따라가지 못하는 한계점들을 해결하였다.

참 고 문 헌 (References)

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, “Learning Deep Features for Discriminative Localization,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2921-2929, 2016.
doi: https://doi.org/10.1109/CVPR.2016.319

[2] K. Singh and Y. J. Lee, “Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-Supervised Object and Action Localization,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 3544-3553, 2017.
doi: https://doi.org/10.1109/ICCV.2017.381

[3] J. Wei, Q. Wang, Z. Li, S. Wang, S. K. Zhou and S. Cui, “Shallow Feature Matters for Weakly Supervised Object Localization,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 5989-5997, 2021.
doi: https://doi.org/10.1109/CVPR46437.2021.00593

[4] J. Kim, J. Choe, S. Yun and N. Kwak, “Normalization Matters in Weakly Supervised Object Localization,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 3407-3416, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00341

[5] X. Zhang, Y. Wei, and Y. Yang. “Inter-Image Communication for Weakly Supervised Localization”. Proceedings of the European Conference on Computer Vision, Glasgow, UK, August 23–28, pp. 271-287, 2020.
doi: https://doi.org/10.1007/978-3-030-58529-7_17

[6] P. Wu, W. Zhai and Y. Cao, “Background Activation Suppression for Weakly Supervised Object Localization,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 14228-14237, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01385

[7] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “ Efficient Convolutional Neural Networks for Mobile Vision Applications”, arXiv, 2017.
doi: https://doi.org/10.48550/arXiv.1704.04861

[8] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, “The Caltech-UCSD Birds-200-2011 Dataset”, Technical Report CNS-TR-2011-001. California Institute of Technology, 2011.
url: http://www.vision.caltech.edu/datasets/cub_200_2011/

[9] W. Bae, J. Noh, and G. Kim, “Rethinking Class Activation Mapping for Weakly Supervised Object Localization”, Proceedings of the European Conference on Computer Vision Glasgow, UK, pp. 618-634, 2020. doi: https://doi.org/10.1007/978-3-030-58555-6_37

[10] J. Choe and H. Shim, “Attention-Based Dropout Layer for Weakly Supervised Object Localization,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 2214-2223, 2019.
doi: https://doi.org/10.1109/CVPR.2019.00232

[11] M. Meng, T. Zhang, Q. Tian, Y. Zhang and F. Wu, “Foreground Activation Maps for Weakly Supervised Object Localization,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 3365-3375, 2021.
doi: https://doi.org/10.1109/ICCV48922.2021.00337

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge.” International Journal of Computer Vision (IJCV) Vol. 115, No. 3, pp.211-252, 2015.
doi: https://doi.org/10.1007/s11263-015-0816-y

저 자 소 개

박 세 진

- 2022년 2월 : 서울과학기술대학교 전자IT미디어공학과 공학사
- 2022년 2월 ~ 현재 : 서울과학기술대학교 전자공학과 석사과정
- ORCID : https://orcid.org/0009-0006-5524-0337
- 주관심분야 : 컴퓨터 비전, 영상 처리, 기계 학습

저 자 소 개

이 소 은

- 2020년 2월 ~ 현재 : 서울과학기술대학교 전자공학과 학사과정
- ORCID : https://orcid.org/0009-0004-0116-5197
- 주관심분야 : 컴퓨터 비전, 영상 처리, 기계 학습

강 병 근

- 2018년 12월 : 미국 University of California San Diego 전기컴퓨터공학과 공학박사
- 2018년 ~ 2019년 : 미국 Carnegie Mellon University Postdoctoral Fellow
- 2019년 ~ 현재 : 서울과학기술대학교 전자공학과 조교수
- ORCID : https://orcid.org/0000-0003-2537-7720
- 주관심분야 : 컴퓨터 비전, 영상 처리, 기계 학습