ContextAug: Combat Contextual-bias augmentation method for Multi-Label Classification

120210201 조석희, 120210393 김준태, 120210399 임희선,
120220182 공민석, 120220227 최규식

1. Introduction

시각적 맥락(Visual Context)은 사람과 컴퓨터 비전(computer vision) 영역 모두에서 장면 해석(scene interpretation) 및 객체 인식(object recognition)을 위한 시각 시스템의 중요한 보조 신호 역할을 담당한다 [1, 2, 3]. 일반적으로 시각적 맥락은 동시에 발생하는 객체 및 장면을 포함한 다양한 종류의 정보를 뜻하는데, 크게 객체와 배경(background)의 동시발생(co-occurrence)과, 객체와 객체의 동시발생으로 나눌 수 있다 [4]. 여기서 동시발생이란 객체나 배경이 한 이미지나 시각적 장면(visual scene)에 포함되어 두 관계 사이에 강한 의존성을 나타내는 경우를 의미한다. 전자는 자동차가 매우 높은 확률로 도로 위에 존재하는 경우를 예시로 들 수 있고, 후자는 책상이 매우 높은 확률로 의자와 함께 존재하는 경우를 예시로 들 수 있고, 이를 강한 맥락적 관계가 있다고 정의할 수 있다. 즉각적인 시각적 인지가 어렵거나 시각적 신호(visual signal)가 불완전할 때, 시각적 맥락은 시각적인 단서로서 인간의 인지 시스템에서 중요한 역할을 담당하게 된다. 이와 마찬가지로 이미지 분류(image classification) [5], 객체 탐지(object detection) [6, 7, 8, 9], 객체 세그맨테이션(object segmentation) [10] 등의 다양한 비전 태스크(vision task)에서도 시각적 맥락은 시각적으로 유사한 클래스(class)와 배경 간의 모호성을 해결하고 성능을 향상시키는데 도움이 될 수 있다 [11].

지난 연구는 시각적 맥락의 동시 출현 발생 확률이 높은 경우에 대한 긍정적인 해석을 통해, 주요 객체에 대한 성능 향상을 위주로 진행되었다. 하지만 비전 태스크에서 시각적 맥락에 대한 지나친 의존은 실제 환경에 적용 시 부정적인 영향을 줄 수 있다. 특히, 시각적 맥락에 지나치게 의존하여 성능을 높일 경우, 동시에 등장하지 않는 경우에 대한 특정 객체의 인식률을 개선할 수 없다는 단점이 있다. 객체 인식에 있어서 객체 자체에 대한 정보보다 시각적 맥락을 더 주요한 정보로 여기게 되는 주객전도가 나타날 수 있기 때문이다. 객체와 객체간의 동시발생을 예로 들면, 일반적으로 책상은 의자와 함께 존재하지만, 현실에서는 일반적인 시각적 맥락에서 벗어난 상황이 반드시 존재하기 때문에, 책상이 호랑이와 동시발생하는 비정형적인 경우에 대한 인식률이 저하될 수 있다. 사람은 이러한 특수한 시나리오를 인지하는 것이 어렵지 않은 반면, 기존의 모델들은 과거에 학습되었던 시각적 맥락에서 허상의 객체를 좇는 맥락적 편향(contextual bias)로 인해 객체 인식에 실패하는 경우가 다수 존재한다. 즉, 호랑이를 의자로 인식하는 틀린 판단을 하게 된다. 이처럼 시각적 맥락에 대한 무분별한 의존을 정제하지 않으면, 보편적인 맥락에서 벗어난 상황에서는 비전 모델의 성능이 오히려 저하될 수 있다.

심층 신경망(Deep Neural Network)을 활용하는 비전 태스크는 학습(training)을 위해 이미지넷(ImageNet) [12], MS COCO [13]와 같이 주석(annotation)이 달린 대규모 이미지 데이터셋의 가용성에 의존하는데, 이러한 데이터셋은 제작자의 노력에도 불구하고 맥락적 편향을 지니고 있다 [4]. 자동차를 포함한 이미지를 예를 들어 생각해보면, 자동차 범주(category)의 이미지들의 높은 확률로 도로를 배경으로 하고 있으며, 버스, 트럭,

오토바이, 신호등과 같이 도로 위에서 보편적으로 보이는 객체들과 동시발생한다. 이는 데이터셋의 맥락적 편향을 유도하게 되고, 해당 데이터셋을 학습한 모델에 녹아들게 된다. 이러한 맥락적 편향을 학습한 모델은 크게 2가지 상황에서 문제가 발생한다. 1) 인지 대상인 객체가 맥락적 편향에 어긋난 배경과 함께 존재하는 상황, 2) 객체가 학습된 맥락과는 상이한 객체와 동시에 존재하는 상황 이다. 앞서 언급한 예시를 들어 설명하면, 전자는 자동차가 도로가 아닌 바다를 배경으로 하는 경우이고, 후자는 자동차가 돌고래 동시발생하는 경우라고 할 수 있으며, 맥락적 편향을 학습한 모델은 2가지 경우 모두에서 객체를 잘못 인지하는 오류를 범하게 된다.

우리의 목표는 이러한 오류를 고려하여 편향에서 자유로운 시각적 분류기를 학습시키고, 맥락에서 벗어난 시나리오에 대해 multi-label classification 태스크 성능을 향상시키는 것이다. 맥락적 편향을 다루는 기존 연구는 동시발생하는 객체끼리의 맥락을 제한하는 연구와 객체와 배경간의 맥락을 줄이는 연구로 나누어진다. [4]와 [11]는 객체와 객체 사이에서 발생하는 맥락적 편향에 집중한다. [11]에서는 이미지에서 객체 제거(object removal)를 제거를 동반한 데이터 증강(data augmentation)을 제안하는 반면, [4]에서는 CAM을 사용한 CAM-based method와 feature splitting method를 맥락적 편향의 해결책으로 제시한다. 한편, [14]에서는 ContraCAM (Contrastive class activation map)을 활용하여 동시발생하는 객체끼리의 편향을 방지하는 Object-aware random crop과 객체와 배경 사이의 편향을 막는 background mixup이라는 두 가지 데이터 증강 방법을 제안한다.

이러한 연구들에 영감을 받아 우리는 동시발생하는 객체끼리의 시각적 편향을 줄이는 새로운 데이터 증강 방법인 ContextAug (CA)를 제안한다. 아이디어의 핵심은 시각적 편향이 작은 이미지끼리 혼합(mixing)하는 방식의 데이터 증강을 통해, 편향을 상쇄시키는데에 있다. ContextAug는 크게 두 단계로 나뉜다. 첫번째로 객체 위치 정보가 식별되어 있는 데이터셋에서 레이블 범주(label category)끼리의 동시발생이 임계값 k보다 작은 범주 쌍(category pair)을 찾는 것이다. 두번째는 찾은 범주 쌍에 속해 있는 인스턴스(instance)를 랜덤으로 선택하고, 인스턴스끼리 혼합을 진행하여 동시발생을 임계값 k까지 증가시키는 것이다. 이 때, 혼합(mixing)은 목표 객체(target object)가 픽셀 단위로 세그멘테이션(segmentation)된 이미지끼리 이루어지며, 세그멘테이션된 부분만 랜덤한 크기로, 랜덤하게 회전시켜서, 다른 이미지의 랜덤한 위치에 붙이는 방식으로 진행된다.

ContextAug가 세그멘테이션된 이미지 데이터셋에만 적용가능하다는 점이 제한적으로 보일 수도 있지만, 이는 기존 이미지가 보유한 배경 편향(background bias)을 최소화하기 위함이다. 이미지를 혼합하는 과정에서 바운딩 박스(bounding box), 폴리곤(polygon)으로 주석 처리된 객체를 붙여넣게 되면, 픽셀 단위로 주석 처리된 세그멘테이션과는 다르게 기존 이미지의 배경 일부분이 추가된다. 그렇다고 세그멘테이션 되어 있지 않은 데이터셋에 직접 주석 처리를 하기에는 많은 비용과 시간이 소요된다. 한편 이미지 인스턴스를 선택하는 과정부터 객체를 이미지에 붙이는 과정까지 전부 랜덤한 방식을 채택한 이유는 맥락적 편향을 줄이는데 있어서, 또다른 편향인 사람의 편향이 더해지는 것을 억제하기 위해서이다.

ContextAug의 특이점은 크게 두 가지가 있다. 첫번째로 기존 연구들이 편향이 큰 데이터에 집중하여 그와 관련된 방법론을 제안한 반면에, 우리는 편향이 작은 데이터들을 결합하여 기존의 맥락을 희석시키고, 이를 통해 맥락적 변동(contextual variation)에 구애받지 않는 방법론을 제시하였다는 점에서 차이가 있다. 두번째로 ContextAug는 Mixup [15], CutMix [16]과 같이 널리 쓰이는 데이터 혼합(data mixing) 방식의 데이터 증강과는 다르게 온전한 객체만을 추출하여 다른 이미지에 붙히는 방식으로 진행된다. 이는 Mixup, CutMix가 인식

성능 자체에 초점을 맞추고 있는 것과 달리 ContextAug는 기존 이미지의 배경적 편향을 최소화하여 데이터셋의 맥락적 편향을 줄이는 데에 그 목적이 있기 때문이다.

ContextAug에 대한 검증(validation)은 정량적 평가와 정성적 평가로 나눠 진행한다. Pascal VOC 2012, MS COCO 데이터셋을 CNN 기반의 Resnet backbone에 ContextAug를 적용하여 학습시키고, 베이스라인(baseline)과의 mAP를 비교하여 정량적으로 판단한다. 또 ContextAug의 하이퍼파라미터 K에 따른 효과를 검증하기 위해 K값 별 mAP를 비교한다. 정성적 평가는 ContextAug 사용 전후의 이미지를 각각 추출하고, CAM(Class Activation Map)을 사용하여 이미지 내 클래스 위치 정보(localization information)를 비교한다.

요약하면, 이 논문의 주요 기여는 다음과 같다. 첫번째로 우리는 비전 모델의 맥락적 편향을 줄이는 새로운 데이터 증강 방법인 ContextAug를 제안한다. 두번째로 ContextAug를 비전 모델에 적용하여 맥락적 편향에 보다 강건하도록 하고, 제안한 방법이 여러 데이터셋에 대한 multi-label classification 태스크 성능 향상에 부분적으로 효과적임을 입증한다. 마지막으로, 기존의 연구가 맥락적 편향이 큰 데이터에 집중한 것에 비해, 우리는 편향이 작은 데이터끼리의 혼합을 통하여 편향을 줄이는 새로운 관점으로 접근한다.

2. Related Work

Multi-label classification에 주로 사용되는 MS-COCO, PascalVoc 데이터셋은 두 객체간의 동시 발생이 많이 나타나는 이미지들을 포함하고 있다. 기존 연구는 해당 문제에 대한 인식 정확도를 개선하기 위해 시각적 맥락의 긍정적인 영향에 집중하였다.[?] 하지만, 시각적 맥락(Context)을 맥락적 편향(Contextual-bias)로 여기는 부정적인 영향에 대한 연구 또한 최근 이루어져 왔다[4, 11, 14].

2.1 Object-Object contextual bias

[4]의 저자는 맥락적 편향을 제거하기 위해 CAM을 활용하여 맥락이 강한 두 객체에 대한 feature representation을 de-correlation 시키는 방법을 제안한다. 특히 CAM 결과를 weak annotation으로 사용하여 loss function에 추가하는 방법(Ours-CAM)과, 네트워크 아키텍처를 두 부분으로 나누어 데이터셋 상 동시 발생 확률이 높은 두 객체에 대해서는 한 부분을 freezing 시켜 gradient 역전파가 되지 않도록 제한하는 방법(Ours-feature-split)을 제안하였다. [11]의 저자는 classification과 segmentation 문제에 대하여 맥락적 편향의 부정적인 영향을 개선하기 위해 inpainting 네트워크를 이용한 object removal 기반의 데이터 증강 방법을 제안하여 각 객체를 독립적으로 인식하는 효과를 입증한다.

2.2 Object-Scene contextual bias

객체와 배경 간의 맥락적 편향을 제거하기 위한 시도도 진행되고 있다. Background Challenge[17]를 통해 객체가 배경 문맥에 의존적이지 않기 위한 다양한 모델이 제안되고 있다. 특히, [14]의 연구는 두 가지 데이터 증강을 통해 객체와 배경, 객체와 객체 두 편향을 모두 개선하는 시도한다. CAM을 이용한 contrastive learning(ContraCAM)을 제안하여 객체 바운딩박스를 이용한 object-aware random crop과 background mixup을 통해 각 객체에 대하여 다양한 배경정보를 결합함으로써 객체와 배경 간의 맥락적 편향을 제거하고, ContraCAM의 결과를 이용하여 두 객체 간의 맥락적 편향을 제거한다.

3. Method

우리는 동시발생 편향 문제를 해결하기 위하여 데이터셋의 동시 발생 횟수를 탐색하여 행렬(matrix)로 나타내고, 이 행렬을 기반으로 모든 객체의 동시 발생 횟수가 임계값 k만큼 나타낼 수 있도록 데이터를 증강한다. 이 때 논문에서 제안하는 새로운 맥락적 데이터 증강 ContextAug를 진행하며 세부 내용은 다음과 같다.

3.1 동시발생 행렬(Co-occurrence Matrix)

동시 발생 행렬은 데이터셋의 주석을 통해 생성하며 과정은 다음과 같다. 먼저 데이터셋이 가진 모든 클래스에 대한 행렬을 만들어 초기화 한다. 그 후 각 사진 별로 해당하는 클래스를 살펴 등장 횟수를 표기하며 모든 클래스의 등장 횟수를 완성한다. 자세한 알고리즘은 아래와 같다.

Algorithm 1 Get co-occurence matrix

Input: Dataset (e.g. COCO, PASCALVOC)

```
co_occur_matrix[len(class)][len(class)]=0;
for image in Dataset do
    annotations=get_annotation(image);
    for i in annotations do
        for j in annotations do
            co_occur_matrix[i][j]+=1;
        end for
    end for
end for
```

Output: co_occur_matrix

3.2 맥락적 데이터 증강 (Contextual data augmentation)

우리는 데이터셋으로 계산한 동시 발생 행렬을 기반으로 모든 객체의 동시 발생 횟수가 최소 임계값 k가 될 수 있도록 동시 발생 횟수가 낮은 객체의 사진들 간의 합성을 진행하며 과정은 아래와 같다.

먼저 현재 배치 내 사진의 레이블 중 하나를 무작위로 선택하고, 동시 발생 행렬을 기반으로 k개 이하의 객체를 선택하고, 객체가 등장하는 이미지를 미리 저장해 둔 객체 세그멘테이션 큐에서 무작위 추출한다. 무작위 추출된 사진에서 해당 객체를 지정 영역 (e.g. 범위 상자, 다각형 범위, 세그멘트)에 따라 잘라 목표 사진에 무작위의 위치와 크기 각도로 합성한다.

위의 과정을 학습하면서 반복적으로 수행하게 되고 k개 이하의 객체를 추출하기 때문에 모든 객체가 k번 이상 학습된다. 이런 과정을 통해 학습 모델이 최소 k번 이상 원하는 객체를 학습할 수 있게 하여 모든 객체간의 동시 발생 횟수를 골고루 분배할 수 있도록 하여 결과적으로 맥락적 편향을 제거할 수 있도록 하였다.

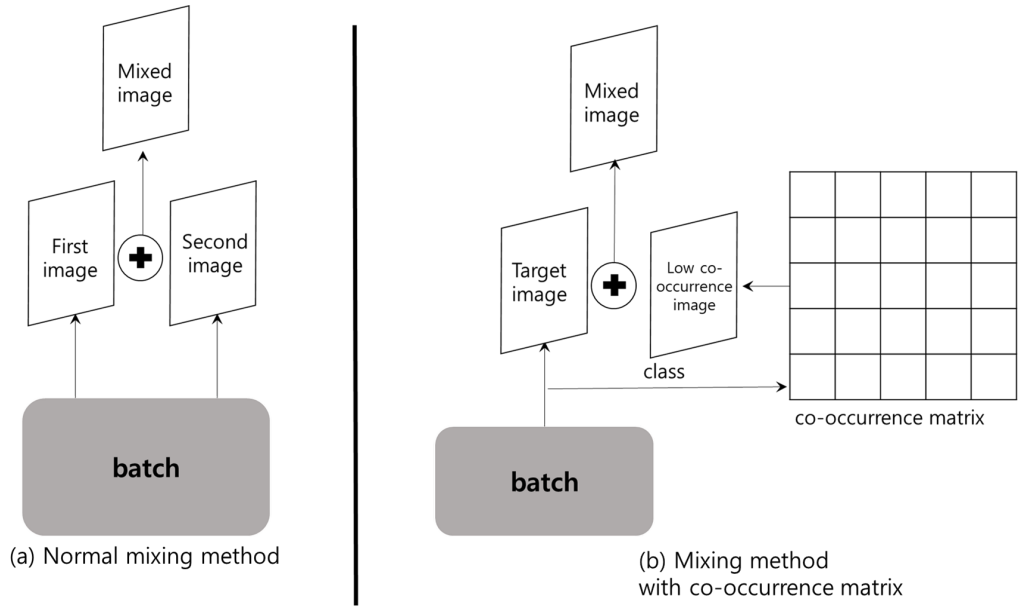
Algorithm 2 Context Augmentation

Input: co_occur_matrix,data,label,k

```
class_label=random_choice(label);
candidate=random_choice(where(co_occur_matrix[class_label]>k));
random_image=random_choice(get_image_by_class(candidate));
instance=get_instance(random_image,candidate)
result=mix_image(instance,data)
```

Output: result

[그림 1] 생성된 augmentation 샘플 이미지, 레이블 예시



[그림 2] 동시 발생 행렬을 사용한 이미지 증강 구조

맥락적 증강 방법은 동시 발생 행렬을 기반으로 동시 발생이 적은 클래스에 대한 데이터 증강을 진행한다. 때문에 기존 mixing method와 같이 배치 내부에서 증강을 진행[16]할 수 없고 동시 발생 행렬을 미리 계산하고 이를 통해 배치 외부의 데이터를 활용하는 방식을 사용한다. 동시 발생 행렬을 기반으로 증강 하는 방식을 통해 클래스를 지정해서 증강 함에도 매번 다른 객체를 붙이기 때문에 iteration 마다 다른 이미지를 증강하게 하여 적절한 randomness가 부여된다. 또한 동시 발생 행렬에서 증강에 이용된 각 클래스 페어에 대한 최대값 K를 지정함에 따라 randomness가 과도하게 커지는 것을 방지할 수 있으며, 동시 발생이 적은 클래스에 더욱 집중하는 regularization을 부여할 수 있다.

결과적으로 우리는 Context Aug 방법을 통해 맥락적 편향이 적은 데이터를 제공하여 맥락에 대한 의존을 낮추고 클래스의 특징을 확실하게 학습하고자 하였다. 이를 통한 결과를 증명하기 위한 실험은 4장에서 서술한다.

4. Experiments

이 장에서는 multi-label classification에 제안한 방법들을 적용하여 co-occurrence가 높은 객체 쌍에 대하여 맥락적 편향을 낮추기 위한 다양한 실험을 소개한다. 먼저, 우리의 데이터 증강 기법이 나타내는 효과를 테스트하기 위해 20개 클래스를 포함하고 있는 Pascal-VOC 2012 데이터셋으로 간이 테스트를 진행하고, 이후 CNN backbone(ResNet)에 대하여 MS-COCO 데이터셋으로 최종 성능을 검증한다. 정량적 지표로는 mAP를 사용하며, 증강을 적용할 때 각 클래스 마다 미치는 영향을 확인하기 위해 class 당 AP를 사용하고, 정성적 평가를 위해 CAM을 활용한다.

4.1 Validation on Pascal VOC

우리는 다양한 방식의 객체 stitching 방법을 테스트하기 위해 Pascal VOC 2012 데이터셋에 대해 실험을 진행했다. 바운딩박스의 객체 stitching 방법에 비해 배경정보를 포함하지 않는 세그멘테이션 단위의 객체 stitching을 적용할 때 유의미한 결과를 가지며, 이는 동시 발생

확률이 낮은 두 객체에 대하여 성능 개선의 가능성이 있다는 것을 의미한다. [표1]의 결과와 같이 기존의 높은 AP를 보이는 클래스에 대한 성능을 유지하면서, 하위 5개 클래스에 대한 성능이 개선됨을 확인할 수 있다.

하위 5개			상위 5개		
Class no.	Baseline	Ours	Class no.	Baseline	Ours
16	65.35	**65.46(+0.11)**	18	98.53	99.45(+0.92)
4	**67.45**	66.10(-1.35)	14	97.16	97.57(+0.41)
15	70.83	**72.41(+1.58)**	11	97.44	96.60(-0.84)
8	73.75	**74.97(+1.22)**	7	97.08	95.97(-1.11)
9	75	**75.66(+0.66)**	11	95.9	95.14(-0.76)

[표1] PascalVOC 2012 데이터셋에 ContextAug를 적용하여 Resnet-50을 학습시킨 결과의 하위 5개 class의 AP와 상위 5개 클래스의 AP(%).

4.2 Experiment setting

우리가 실험에 적용한 하이퍼파라미터를 소개한다. Learning rate 1e-3, multi label soft margin loss와 BCE, cosine, one-cycle scheduler를 사용해 실험에 적용하였다. Multi label soft margin loss는 max-entropy를 바탕으로 각 label에 대한 Binary Cross Entropy 값을 평균내는 loss 함수이다. Multi-label classification에서는 label 별 계산된 loss 값들이 개별적으로 쓰일 필요가 없기 때문에 BCE를 전체 class 개수로 나눈 평균값을 이용하는 multi label soft margin loss를 사용하였다.

$$loss(x, y) = -\frac{1}{C} * \sum_i y[i] * \log((1 + \exp(-x[i]))^{-1}) + (1 - y[i]) * \log\left(\frac{\exp(-x[i])}{(1 + \exp(-x[i]))}\right)$$

$$\text{where } i \in \{0, \cdots, \text{x.nElement}() - 1\}, y[i] \in \{0, 1\}$$

4.3 Models

Model	# Params	mAP
ResNet-50 (Baseline)	25.6M	**74.11**
ResNet-50 + Ours	25.6M	73.21(-0.9)
ResNet-101 (Baseline)	44.6M	76.3
ResNet-101 + Ours	44.6M	**76.58(+0.28)**

[표2] Resnet-50, Resnet-101 모델에 ContextAug를 적용하지 않았을 때와 적용했을 때 mAP(%).

[표2]에서는 두 크기의 backbone 모델에 적용한 결과를 비교한다. 실험에는 Imagenet에 사전훈련된 ResNet-50, ResNet-101을 사용한다. Backbone 모델들은 모두 ImageNet-1K pre-trained를 사용하여 MS-COCO 데이터셋에 Multi-label classification 으로

전이학습하였고, 제안한 ContextAug를 적용했을 때 ResNet-101에 대하여 성능이 다소 향상되는 결과를 확인할 수 있다. ResNet-50의 경우 Baseline에 비해 ContextAug를 적용했을 때 mAP 0.9% 감소하며, ResNet-101의 경우 mAP 0.28%, tresnet-l-448은 mAP 0.32% 향상된다.

4.4 Analysis upon different K values

Model	Baseline	K=10	K=20	K=30	K=40	K=50
Resnet-50	**74.11**	73.21(-0.9)	73.58(-0.53)	73.69(-0.42)	73.4(-0.71)	73.4(-0.71)
Resnet-101	76.31	75.78(-0.57)	76.10(-0.21)	76.46(+0.15)	**76.58(+0.27)**	76.38(+0.07)

[표3] MS-COCO 데이터셋에 ContextAug를 서로 다른 K값을 적용하여 Resnet-50, Resnet-101을 학습한 mAP(%).

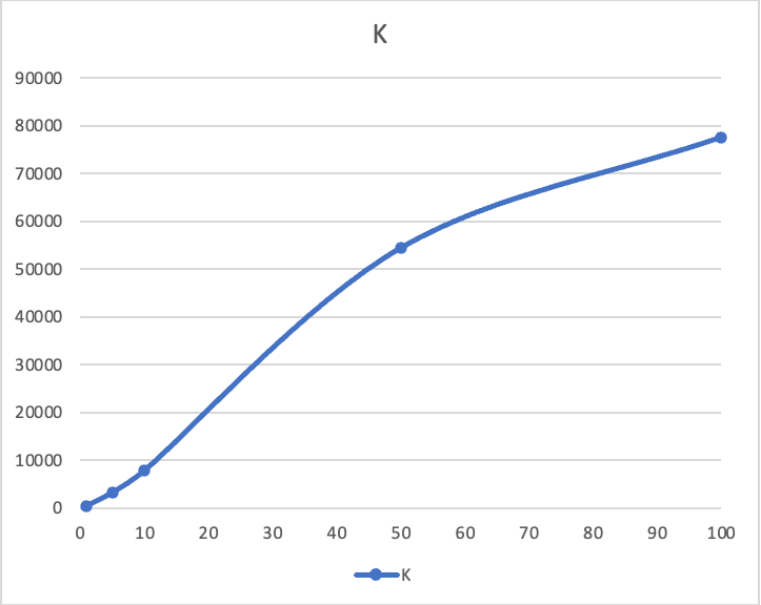
제안한 data augmentation 기법은 데이터셋의 co-occurrence matrix를 학습 도중에 확인하고, 레이블 중 하나의 객체를 랜덤하게 선정하여 matrix 상 최소 이미지 쌍에 대응되는 객체를 stitching 하는 방법이다. 이 때, K 값에 따라 각 객체마다 augmentation을 적용하는 최대값을 hyperparameter K를 통해 지정할 수 있으며 MS-COCO 데이터셋에 K 값별 ContextAug를 적용하였을 때 augmentation이 적용되는 이미지의 개수는 [표4]와 같다. 우리는 [표3]과 같이 K값을 5가지에 걸쳐 incremental하게 적용하여 실험을 진행하였다. 실험 결과 Resnet-50에서는 ContextAug가 유의미한 성능향상을 가져오지 않았고, Resnet-101에서는 K=40 으로 설정하였을 때 가장 높은 성능향상을 가져왔다.

5. Result

5.1 결과 분석

	k=1	k=5	k=10	k=50	k=100
#augmented	456	3279	7854	54466	77626

[표4] k 값별 ContextAug가 적용된 샘플의 개수(mean)



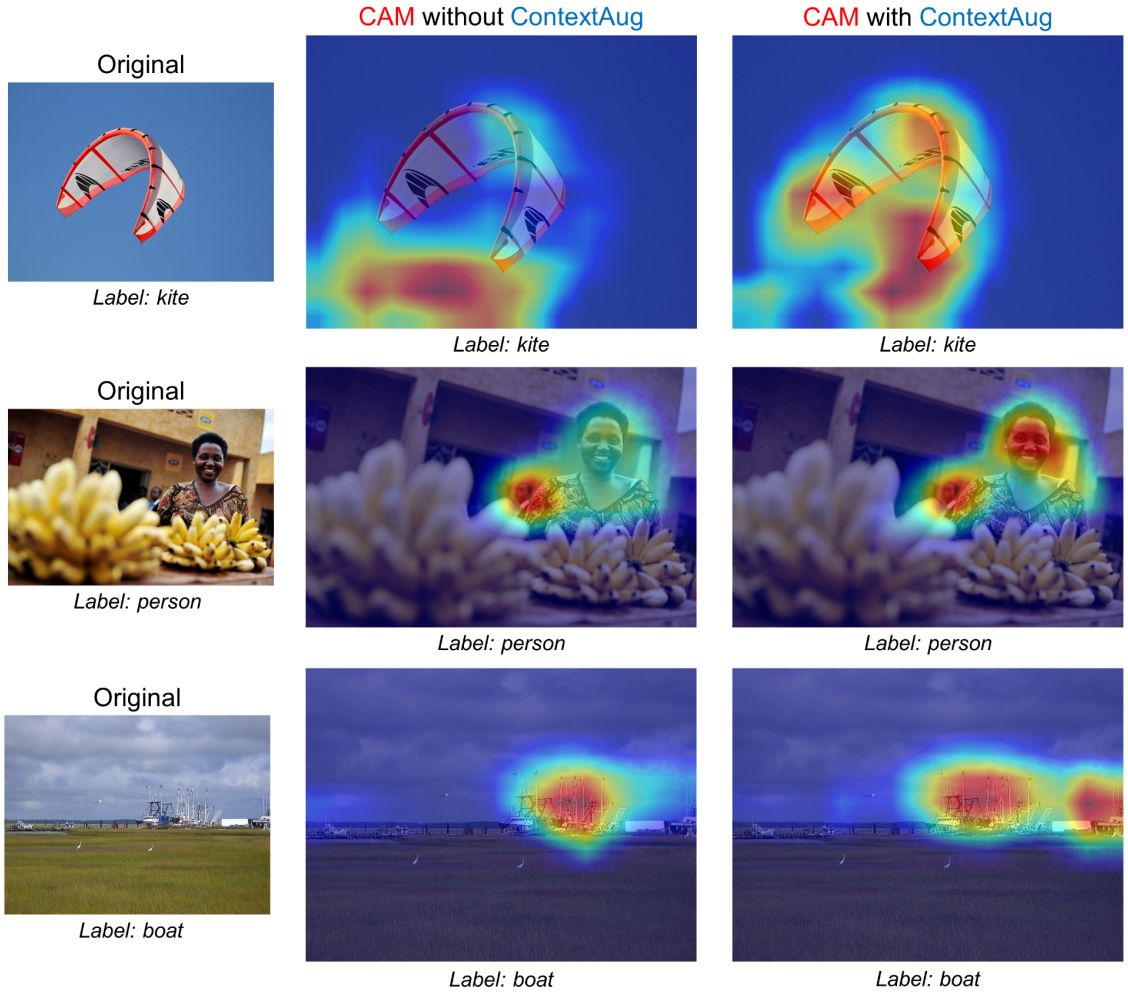
[그림3] K값 설정에 따른 증강데이터 개수 추이

5.1.1 정량적 분석

4.3에서 기술한 실험결과와 같이 ContextAug방법은 Resnet-50(25.6M params) 보다 Resnet-101(44.6M params)모델과 같이 비교적 크기가 큰 네트워크에서 성능향상을 가져온다. MS-COCO 데이터셋은 기본적으로 이미지당 평균 2.9개의 카테고리를 가지지만, ContextAug를 적용했을 때 이미지당 평균 카테고리 수는 최대 3.9가 된다. 즉 ContextAug는 모델로 하여금 더 까다로운 multi-label classification 문제를 풀도록 학습되기 때문에 크기가 큰 네트워크에서 성능 향상을 가져온 것으로 추정된다.

4.4에서는 하이퍼파라미터 K에 따른 효과를 검증하기 위해 K값을 서로 다르게 설정하여 얻은 결과를 기술한다. 적용한 K값 별 성능 차이는 [표3]에서 확인할 수 있다. [표4]와 [그림3]에서는 K값이 증가함에 따라 증강되는 샘플의 수를 표시하였고, [그림3]에서 K를 50 이상으로 설정할 때 증가율(기울기)가 감소하는 것을 확인할 수 있다. 즉 50 이상의 K에 대해 유의미한 성능 향상이 이루어지지 않는다. 이를 개선하기 위해 학습에 이용되는 이미지 한 장 당 하나의 객체만 stitching을 하는 것이 아닌 두개 이상을 stitching하는 방법을 시도하였으나, 두 개의 객체만 stitching하여도 이미지당 평균 카테고리의 개수는 최대 4.9개가 되어 오히려 성능이 떨어지는 결과를 가져온다.(ContextAug를 적용하지 않은 MS-COCO 데이터셋은 이미지당 평균 2.9개의 레이블을 갖게 되므로) 따라서 우리는 Resnet-101 모델에 image 한 장 당 하나의 객체만을 stitching하고 K 값을 40로 설정하였을 경우에만 유효한 결과를 가져온다는 최종 결론을 내린다.

5.1.2 정성적 분석



[그림4] 원본이미지(좌측)과 Resnet-101에 ContextAug를 적용하지 않고 학습한 모델의 gradCAM결과(중앙)와 ContextAug를 적용하여 학습한 모델의 gradCAM결과(우측)

본 논문이 제안하는 ContextAug 방법은 multi-label classification 문제에서 이미지 데이터의 맥락적 편향을 제거하기 위해 설계된 데이터 증강방식이다. 결과적으로 ContextAug이 맥락적 편향을 제거하는 데에 얼마나 효과가 있었는지 정성적으로 분석하기 위해 gradCAM을 이용해 시각화하였다. [그림4]은 학습된 모델이 레이블을 예측하기 위해 살펴본 receptive field를 시각화한 결과이다. 첫번째 행에서는 ContextAug를 적용하여 학습한 모델이 레이블 “kite”를 예측할 때 ContextAug를 적용하지 않았을 때에 비해 맥락정보보다 객체 자체에 집중한 것을 확인할 수 있고, 두번째 행에서는 레이블 “person”을 예측하기 위해 객체 자체에 초점을 둔 것을 확인할 수 있다. 세번째 행의 경우, ContextAug를 적용했을 때 이미지의 우측에 가려진 “boat”까지 살펴보며 예측한 것을 확인할 수 있다. 이를 통해 우리가 제안한 방법이 설계된 의도에 맞게 맥락적 편향을 잘 제거했다고 결론을 내린다.

5.2 Discussion

Multi-label classification의 다양한 state-of-the-art 모델[18, 19, 20]에 비해 성능(mAP)이 낮게 나타난 이유는 본 실험에 다양한 방식의 데이터 증강 방식을 적용하지 않았기 때문이라고 판단한다. 우리는 baseline과 비교하여 우리가 제안한 ContextAug 메소드가 유효한 성능 향상을 보이는가에 대해 평가하고, 이는 향후 Cutmix, Mixup, RandAug 등과

함께 사용할 경우 유효한 성능 향상을 보일 것을 기대한다. 또한 제안된 ContextAug는 Unrel[21]과 같이 자주 관찰되지 않는 데이터에 대해 더 좋은 성능을 보일 것을 기대한다. 이는 기존의 데이터셋(PascalVOC, MS-COCO)이 포함하고 있는 문맥적 편향을 반증할 수 있는 예시가 될 수 있을 것이다.

ContextAug는 데이터 증강에 객체와 배경 간의 시각적 편향을 포함하지 않기 위해 세그멘테이션 레벨 데이터를 이용한다. 따라서 바운딩 박스와 같이 낮은 수준의 레이블을 활용할 수 없다는 단점이 있어 다양한 데이터셋을 이용하여 적용할 수 없다. 또한 세그멘테이션 레벨 데이터는 occlusion이 있을 경우 stitching된 객체의 형태가 전체 객체를 포함하지 않을 수 있으므로 잘못된 합성을 가져올 수 있다. 이는 [14]와 같이 객체와 배경 간의 문맥적 편향과 객체와 객체 간의 문맥적 편향을 개선하는 각 방식을 동시에 제안함으로써 해결할 수 있으나, 현재까지는 개별적인 데이터 증강 방식을 적용하고 있지 않다. 따라서 낮은 수준의 레이블을 이용한 맥락적 편향을 제거하는 방법을 향후 연구로 제안하는 바이다.

Reference

[1] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive psychology, 1982

[2] A. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. Communications of the ACM, 53(3), 2010

[3] D. Parikh, C. L. Zitnick, and T. Chen. Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 34(10), 2012.

[4] Singh, Krishna Kumar, et al. "Don't judge an object by its context: learning to overcome contextual bias." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[5] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In CVPR, 2015.

[6] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In CVPR, 2009.

[7] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. CVPR, 2019.

[8] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

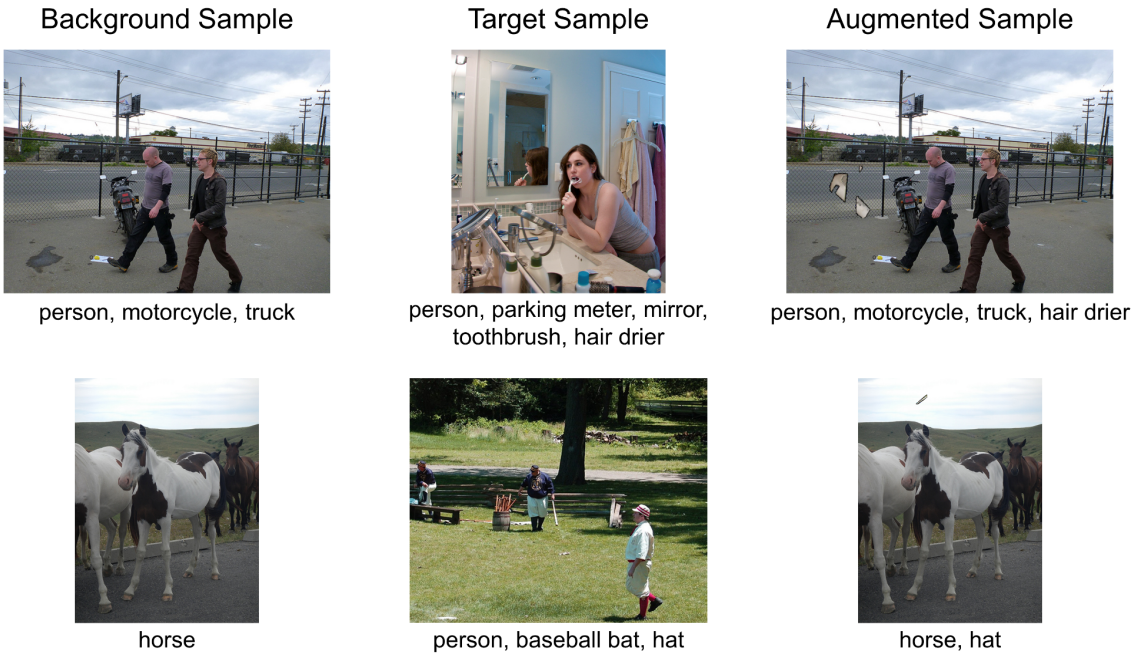
[9] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[10] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[11] Shetty, Rakshith, Bernt Schiele, and Mario Fritz. "Not Using the Car to See the Sidewalk--Quantifying and Controlling the Effects of Context in Classification and

Segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.
[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR, 2009.
[14] Mo, Sangwoo, et al. "Object-aware contrastive learning for debiased scene representation." Advances in Neural Information Processing Systems 34 (2021): 12251-12264.
[15] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).
[16] Yun, Sangdoo, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features." Proceedings of the IEEE/CVF international conference on computer vision. 2019.
[17] K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. “Noise or signal: The role of image backgrounds in object recognition.”, In *International Conference on Learning Representations(ICLR)*, 2021.
[18] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben- Baruch, and Asaf Noy. “Ml-decoder: Scalable and versatile classification head”. *arXiv preprint arXiv:2111.12933*, 2021.
[19] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. “Query2label: A simple transformer way to multi-label classification”, *arXiv preprint arXiv:2107.10834*, 2021.
[20] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. “Imagenet-21k pretraining for the masses”, *In Advances in Neural Information Processing Systems(NeurIPS)* , 2021.
[21] Julia Peyre, Josef Sivic, Ivan Laptev, Cordelia Schmid, “Weakly-Supervised Learning of Visual Relations”, *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5179-5188

APPENDIX



[그림 5] 세그멘테이션 수준의 어노테이션의 한계로 인해 판별하기 어렵게 증강된 샘플 예시