KIPS Trans. Softw. and Data Eng.
Vol.12, No.2 pp.91~98
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.2.91

Human-Object Interaction Detection Data Augmentation Using Image Concatenation

Sang-Baek Lee† · Kyu-Chul Lee††

ABSTRACT

Human-object interaction(HOI) detection requires both object detection and interaction recognition, and requires a large amount of data to learn a detection model. Current opened dataset is insufficient in scale for training model enough. In this paper, we propose an easy and effective data augmentation method called Simple Quattro Augmentation(SQA) and Random Quattro Augmentation(RQA) for human-object interaction detection. We show that our proposed method can be easily integrated into State-of-the-Art HOI detection models with HICO-DET dataset.

Keywords : Human-object Interaction Detection, Data augmentation, Image Processing, Object Detection

이미지 이어붙이기를 이용한 인간-객체 상호작용 탐지 데이터 증강

이 상 백† · 이 규 철††

요 약

인간-객체 상호작용 탐지는 객체 탐지와 상호작용 인식을 함께 풀어야하는 분야로 탐지 모델의 학습을 위해서 많은 데이터를 필요로 한다. 현재 공개된 데이터셋은 규모가 부족하여 데이터 증강 기법에 대한 요구가 커지고 있으나, 대부분의 연구에서 기존의 객체 탐지, 이미지 분할 분야에서 활용하는 증강 기법을 활용하고 있는 실정이다. 이에 본 연구에서는 인간-객체 상호작용 탐지 분야에서 활용하는 데이터셋의 특성을 파악하고, 이를 통해 인간-객체 상호작용 탐지 모델 성능 향상에 효과적인 데이터 증강 기법을 제안한다. 본 연구에서 제안한 증강 기법에 대한 검증을 위하여 실험 환경을 구축하고, 기존의 학습 모델에 적용하여 증강 기법을 적용할 경우에 탐지 모델의 성능 향상이 가능함을 확인하였다.

키워드 : 인간-객체 상호작용 탐지, 데이터 증강, 이미지 처리, 객체 탐지

1. 서 론

인간-객체 상호작용이란 장면에 등장하는 사람과 객체간의 관계를 파악하는 작업(task)으로써, 컴퓨터비전 분야에서 매우 도전적인(challenging) 문제로 알려져 있다. 이미지 분류(Image classification), 분할(Segmentation), 객체 탐지(Object detection) 등 다양한 비전 작업들이 존재한다. 그러나 그 중에서도 인간-객체 상호작용 탐지가 어려운 이유는 몇 가지가 있는데, 먼저, 객체 탐지 작업에서 목표로 하는 이미지에서 사람과 물체를 정확히 찾아내야 하고, 이후에 찾아낸 객체간의 관계를 분류해야 한다는 것이다. 비전 분야의 다른 작업들에서 각각 목표로 수행하는 작업들을 인간-객체 상호작용 탐지에서는 동시에 풀어야하기 때문에 보다 복잡한 연산 과정이 필요한 것이다. 또 다른 하나의 이유는 데이터에서 찾을 수 있는데, 인간-객체 상호작용 학습 데이터의 경우에는 이미지 내에 사람과 객체의 위치와 그 둘 사이의 관계를 함께 표현해야하기 때문에 기존의 객체 탐지 분야에서의 학습 데이터 구축보다 복잡하고 어려운 과정을 거치게 된다. 현재 인간-객체 상호작용 탐지 연구 분야에서 주로 사용하는 데이터셋은 HICO-DET[4]과 V-COCO[5]가 있다. 현존하는 인간-객체 상호작용 탐지 데이터셋은 객체 탐지나 이미지 분류 작업에서 사용되는 데이터셋에 비해 규모가 굉장히 작은 편에 속한다. HICO-DET을 살펴보면 전체 인간-객체 상호작용을 표현하는 클래스가 600개인데, 이 중에 130여개의 클래스는 학습 데이터 샘플수가 10개 미만에 불과하다. 복잡한 문제를 해결하기 위한 모델의 학습을 위해서는 충분한 데

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2022-0-00817).
† 정 회 원 : 충남대학교 컴퓨터융합학부 석·박사통합과정
†† 정 회 원 : 충남대학교 컴퓨터융합학부 교수
Manuscript Received : August 30, 2022
First Revision : October 11, 2022
Accepted : October 21, 2022
* Corresponding Author : Kyu-Chul Lee(kclee@cnu.ac.kr)

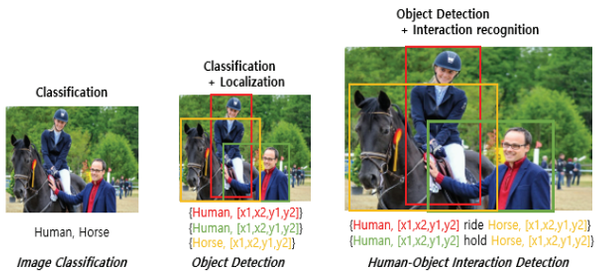

Fig. 1. (Left) Image Classification Example (Middle) Object Detection Example (Right) Human-Object Interaction Detection Example

이터셋이 필요한데 반해 인간-객체 상호작용 학습 데이터셋은 그렇지 못한 상황이다. 또한, 전체 600개 클래스에 대한 데이터 분포를 살펴볼 경우에 긴 꼬리 분포(Long-tailed distribution)의 경향을 보이는데, 이는 학습 모델의 성능 저하를 불러오게 된다. 부족한 데이터로부터 학습된 모델은 학습과정에서 희귀 클래스(Rare classes)에 속하는 경우에 대한 특징 추출(Feature extraction)과 분류를 해내는 능력을 배울 수 있는 기회 자체가 적기 때문에 이에 대한 학습이 제대로 이루어지지 않는다.

일반적으로 공개 데이터셋이 긴 꼬리 분포 특성을 가지는 경우에는 학습 데이터와 테스트 데이터 모두 동일한 분포를 띄기 때문에 희귀 클래스에 대한 성능이 전체 성능 평가에 큰 영향을 끼치지는 않다고 생각할 수 도 있다. 평가 단계에서 희귀 클래스 중 일부를 맞추지 못하더라도, 비희귀 클래스(Non-rare classes)에 속하는 정답을 많이 맞추면 전체 평균 성능은 높게 달성할 수 있기 때문이다. 그러나 현실 세계에서는 희귀 클래스에 속하는 경우에 대한 문제 해결을 요구하는 경우가 많기 때문에 이에 대한 성능도 중요한 요소로 작용한다. 실제 학습 모델을 활용하는 입장에서는 오히려 희귀 클래스에 해당하는 경우를 분류해내는 것이 해당 시스템의 역량이라고 생각할 수 있기 때문이다. 그래서 각각의 클래스별 성능을 고르게 달성하기 위해서는 학습된 모델이 희귀 클래스에 대한 정답도 잘 맞추는 방향으로 학습이 되어야 한다. 이처럼 학습 데이터의 규모가 모델을 충분히 학습시키기에 부족한 경우를 해결하기 위해 데이터 증강을 활용한다. 데이터 증강을 할 때 가장 고려되어야 하는 점은 데이터의 특성을 제대로 파악하는 것이다. 모든 이미지 데이터에 증강 기법을 적용한다고 해서 효과가 있는 것은 아니다. 오히려 일부 태스크에서는 일반적인 이미지 증강 기법이 성능 저하를 야기할 수 있다고 알려져 있다[1]. 예를 들어 자동차나 건물과 같은 이미지에 회전 기법을 사용하여 뒤집은 이미지를 만든다면 이는 현실에서 발생할 수 없는 상황이기 때문에 오히려 객체 탐지를 할 때 역효과가 날 수 있다. 일반적으로 객체 탐지 영역에서 증강을 하는 경우, 전체 이미지의 분할 정보를 이용하여 객체의 위치를 무작위로 조정하여 정답 데이터를 증강하는 방법을 사용한다[7, 10, 11]. 그러나 이와 같은 증강 기법을 인간-객체 상호작용 탐지에서 사용할 경우에는 문제가 될 수 있다. 예를 들어 스노우보드를 타는 장면이나 자전거를 타는 장면 등 인간과 객체 사이의 거리가 중요한 상호작용의 경우에는 인간-객체 분할 정보를 무작위로 옮겼을 경우에 둘 사이에 상호작용의 의미가 사라질 수가 있다. 이에 몇몇 선행 연구[2]에서는 인간-객체 상호작용에 특화된 방법의 증강 기법을 제안하고 있다. 제안하고 있는 방법들에 대해 살펴보면 이미지에서 특정 위치의 객체 클래스를 대체하여 새로운 인간-객체 관계를 생성해내는 방법이 있다[2]. 이를 위해서는 분할을 통해 객체의 정확한 윤곽을 찾고, 해당 부분을 다른 객체 분할 정보를 통해 대체하는 노력이 필요하다. 이처럼 인간-객체 상호작용 분야에서 특화된 데이터 증강 방법들은 기존 샘플을 이용하여 이미지 내에 새로운 인간-객체 상호작용 관계를 생성하거나, 기존의 인간-객체 상호작용 관계를 변형하는 방식을 채택하고 있다. 직관적으로 생각하였을 때, 학습 모델로 하여 인간-객체 상호작용을 찾는 경험을 풍부하게 해주기 위해서는 학습 과정에서 입력으로 사용되는 각각의 이미지가 조금 더 많은 인간-객체 정보와 상호작용 정보를 담고 있다면 복잡한 관계를 찾는 능력을 갖출 수 있을 것이다. 그러나 이미지에 기존의 인간-객체 상호작용 정보가 아닌 새로운 인간-객체 상호작용 관계를 추가하는 것은 쉬운 일이 아니다. 새롭게 추가된 인간-객체 상호작용에 대하여 실제 세계에서 나타날 수 있는 정보인지 검증하는 것 자체가 어렵다.

이에 본 연구에서는 기존 샘플 이미지를 이용해서 보다 많은 인간-객체 상호작용 관계를 담을 수 있는 방법을 제안하고자 한다. 본 연구팀에서는 기존의 샘플 이미지를 4-분할 이미지의 한 장면으로 사용하여 이어붙이기를 수행하였다. 이를 통해 새롭게 증강시킨 샘플 이미지에는 기존 이미지 대비 많은 인간-객체 상호작용 관계가 생성될 수 있다. 또한, 이어붙이기를 통해 생성되는 이미지의 다양성을 높이고, 다양한 인간-객체 관계를 함께 적용할 수 있도록 기존의 인간-객체 상호작용 관계를 이용하여 하나의 장면에 등장할 수 있는 가능성이 존재하는 샘플 이미지끼리 이어붙여서 새로운 이미지 증강을 하는 방법도 소개한다. 이 방법은 단순한 증강 기법이라고 할 수 있지만, 인간-객체 상호작용 탐지 태스크 특성상 이미지 내의 수많은 객체들 간의 맥락 정보(Contextual information)를 학습해야 하는 문제에서는 하나의 샘플 이미지 내에 다양한 인간-객체 정보가 존재한다면, 보다 풍부한 맥락 정보를 학습할 수 있을 것이라는 직관에서 본 연구를 수행하였다.

2. 관련 연구

2.1 인간-객체 상호작용 탐지 데이터셋

인간-객체 상호작용 탐지는 스마트시티, 보안, 헬스 케어 등 컴퓨터비전 기술이 활용되는 다양한 분야에서 인간의 행동을 인식하고 활용할 때 중요한 기술이다. 인간-객체 상호

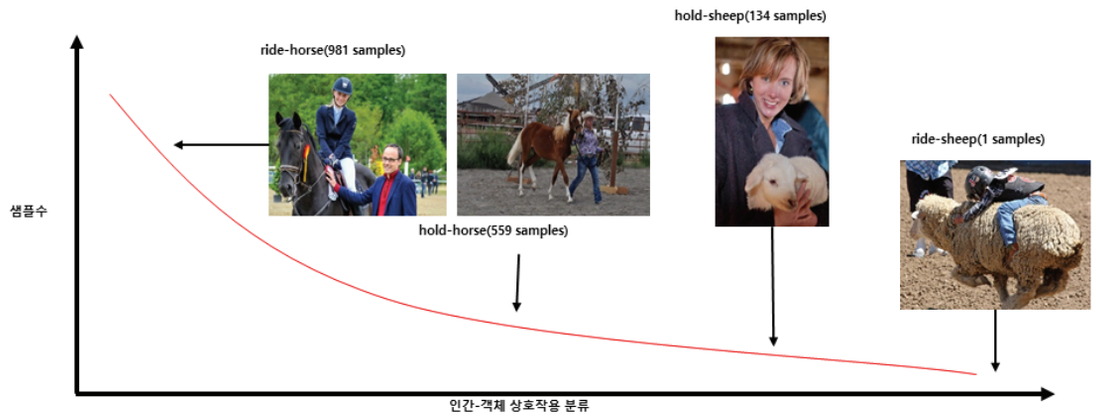

Fig. 2. Long-tailed Distribution Problem for HOI Detection

작용 탐지에 대한 중요성이 강조됨에 따라 HICO-DET[4], V-COCO[5], HCVRD[6] 등 다양한 오픈 데이터셋이 공개되었다. 인간-객체 상호작용 데이터의 특징은 Triplet 구조를 갖는 다는 것이다. Triplet 구조란 {Subject, Object, Predicate}과 같이 세 가지 요소(Tuple)의 조합으로 이루어진 형태를 의미한다. 모든 인간-객체 상호작용 탐지 데이터셋은 {인간 바운딩 박스, 객체 바운딩 박스, 상호작용}으로 구성된 정답 데이터를 가진다. 대부분의 선행 연구에서는 HICO-DET과 V-COCO 두 가지 오픈 데이터를 이용하여 평가를 진행하고 있으며, 그 외에 인간-객체 상호작용 탐지 데이터는 활용도가 크지 않다. HICO-DET과 V-COCO 데이터셋은 모두 마이크로소프트에서 제공하고 있는 MS-COCO 데이터셋 기반으로 생성되었다. MS-COCO[8]는 객체 탐지, 분할, 이미지 캡션 등을 수행하기 위해 만들어진 대규모 데이터셋으로 총 33만 장 이상의 이미지와 20만 개 이상의 정답 데이터, 150만 개의 객체 인스턴스로 구성되어 있다. 정답 데이터의 경우에는 80개의 객체 클래스와 25만 여개의 사람 관절 정보로 이루어진 것이 특징이다.

HICO-DET은 MS-COCO의 80개 객체 클래스와 117개의 동작 클래스를 이용하여 총 600개의 인간-객체 상호작용 클래스로 구성되어 있다. V-COCO도 마찬가지로 MS-COCO 데이터로부터 파생된 데이터셋으로 48개의 객체 클래스와 26개의 동작 클래스로 구성되어 있다. 그러나 현존하는 인간-객체 상호작용 탐지 데이터셋은 공통적인 문제점을 가지고 있다. 이는 서로 다른 상호작용 클래스간의 데이터 불균형 문제로 긴 꼬리 분포 문제라고 불린다. Fig. 2는 HICO-DET의 상호작용 클래스에 따른 데이터 분포를 나타내는데, {Human, Sheep, Ride}처럼 사람이 양을 타는 장면의 경우에 전체 데이터 중에서 단 1개의 데이터 샘플만이 존재한다. 그와 반대로 사람이 말을 타는 장면의 경우 전체 데이터에서 981개나 차지할 정도로 비중이 높은 것을 알 수 있다. 이와 같이 학습 데이터의 분포가 일부분에 치중될 경우에 희귀 클래스에 해당하는 특징은 학습하기가 어렵다. 이는 학습 모델의 전체 성능을 저하시키는 결과를 초래하게 된다.

2.2 이미지 데이터 증강

이미지 데이터 증강은 대부분의 컴퓨터비전 분야에서 사용된다. 이미지 분류, 객체 탐지, 사용자 자세 추정과 같은 컴퓨터비전 분야의 주요 작업부터 손 글씨 인식과 같은 지금은 간단한 작업으로 알려진 것들에 이르기까지 모든 분야에 걸쳐 데이터 증강은 기본적으로 사용하는 것으로 알려져 있다[13]. 대부분의 태스크에서 이미지 증강을 사용하는 가장 큰 이유는 증강 기법을 통해 새로운 학습 데이터를 생성함으로써 학습 모델로 하여금 보다 많은 학습을 통해 성능을 향상하는 것에 있다. 특히 학습 모델의 복잡도나 규모에 비해 학습 데이터의 양이 적은 경우에는 학습 모델의 목적 함수(Object function)의 목표를 달성하지 못하게 된다. 이를 방지하기 위해서 컴퓨터비전 분야에서는 이미지 처리 기법들을 활용한 증강 기법을 많이 사용한다. 이미지 회전(Rotation), 반전(Flip), 변환(Transform), 크기 조절(Scaling), 자르기(Crop)과 같은 기법들은 매우 유명한 이미지 처리 기법들로 이를 활용하면 하나의 이미지로부터 수많은 증강된 이미지를 생성해 낼 수 있다. 이와 같은 전통적인 증강 기법들은 학습 모델의 성능 향상을 위해 도움이 된다고 알려져 있으나, 모든 도메인에서 적용되는 것은 아니다[1]. 예를 들면, 이미지 분류에 사용되는 데이터에 자르기 기법을 무작위로 적용한다면 정답 데이터에 해당하는 부분이 잘려나갈 수 도 있을 것이다. 또한, 보행자의 자세 추정을 하기 위한 태스크에서 회전을 증강 기법으로 활용한다면, 보행자가 90°나 180° 회전하여 정상적인 보행자의 모습이 될 수 없을 것이다. 이처럼 모든 증강 기법이 항상 도움이 된다고 볼 수 없기 때문에 해결하고자 하는 도메인의 특성을 파악하고, 그에 맞는 증강 기법을 적용하는 것이 필요하다.

인간-객체 상호작용 탐지 분야의 경우에는 {인간 바운딩 박스, 객체 바운딩 박스, 상호작용}의 Triplet 구조가 유지되는 것이 특징이기 때문에, 이에 대한 간섭이 없는 선에서 데이터 증강이 이루어져야 한다[15]. 또한, 학습 모델이 보다 많은 인간-객체 상호작용 관계를 학습할 수 있어야 하기 때문에 학습 데이터에 존재하는 인간-객체 상호작용 수를 늘려

주는 것이 중요하다. 이에 본 연구팀에서는 기존의 학습 데이터에 대하여 이어붙이기를 통해 인간-객체 상호작용의 수가 늘어난 형태의 새로운 데이터를 생성하는 방법을 시도하였다. 추가로 하나의 장면에서 발생 가능한 인간-객체 상호작용은 한정되어 있기 때문에 동일한 인간-객체 상호작용 관계를 갖는 데이터끼리 이어붙이기를 통해 학습 데이터에 대한 맥락 정보를 유지하도록 하였다.

3. 제안하는 증강 기법

일반적으로 모든 인간-객체 상호작용 데이터셋을 살펴보면 하나의 학습 이미지에 적게는 1~2개, 많게는 5~10개 내외의 인간과 객체가 등장하고, 각각의 인간, 객체는 1개 이상의 인간-객체 상호작용 관계를 가지고 있게 된다. 인간-객체 상호작용 탐지의 대표적인 데이터셋인 HICO-DET을 기준으로 살펴보면 하나의 학습 이미지 내에 존재하는 인간-객체 상호작용수의 분포는 Fig. 3과 같다. 대부분의 학습 이미지는 2~3개 내외의 인간-객체 상호작용 관계로 표현되어 있다는 것을 알 수 있다. 학습 이미지 내에 하나의 객체에는 여러 개의 상호작용이 중복될 수 있고, 이는 각각의 학습 이미지는 그 안의 상호작용수보다 적은 수의 객체 정보를 담고 있다는 것을 의미한다.

이와 같이 대부분의 학습 이미지 샘플들은 한정된 숫자의 인간-객체 상호작용을 가지고 있기 때문에 모델을 한번 학습시킬 때, 한정된 관계만을 학습할 수 밖에 없게 된다. 이를 해결하기 위해 선행연구[3, 7]에서는 샘플 이미지 내의 다양한 인간-객체 상호작용 관계를 전달하기 위해서 다양한 {인간 바운딩 박스, 객체 바운딩 박스, 상호작용}의 Triplet을 생성하기 위한 연구가 진행되었다. 그러나 생성 모델을 통하여 새로운 이미지를 생성해내는 경우에 몇 가지 문제가 발생할 수 있다. 먼저, 생성되는 이미지의 품질(Quality)이 실제 데이터(Ground truth)에 비해 떨어지는 경우이다. 이런 경우는 대부분의 생성 모델이 가지고 있는 문제점으로 볼 수 있다. 또 다른 문제로는 인간-객체 상호작용 데이터셋 특성에 따라 발생하는 문제이다. 인간-객체 상호작용은 인간과 객체 간의 위치, 거리 등이 중요한 요소인데, 생성 모델을 이용하여 생성하는 이미지의 경우에 객체의 위치가 목표하는 인간-객체 상호작용을 나타낼 수 있는 곳이 아닌 엉뚱한 곳에 자리하는 경우가 발생한다. Fig. 4와 같이 객체 탐지 분야에서 주로 사용하는 객체 분할 정보를 활용하여 복사-붙이기 형태의 증강 기법을 사용할 경우에 설명한 것과 같이 인간과 객체가 올바른 위치에 생성되지 않게 되고, 인간-객체 상호작용이 전혀 유지되지 않는 형태로 데이터가 증강될 수 있다[7]. 이를 방지하기 위하여 인간-객체 간의 공간적 상관 관계(Spatial correlation)을 통하여 생성되는 인간과 객체간의 위치를 보정해주는 연구도 존재한다[3]. 인간-객체 간의 공간적 상관 관계를 통해 서로 간의 위치를 보장해주더라도 생성 모델을 이용하여 생성한 이미지의 경우에 아직까지 실제 이미지와

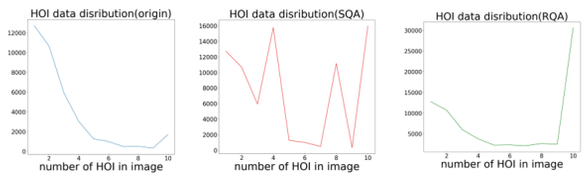

Fig. 3. HOI Data Distribution Per Image (Left) Original Data, (Middle) Simple Quattro Augmentation Data, (Right) Random Quattro Augmentation Data

Original Image　　Instance-level Copy-paste

Fig. 4. An Example of Instance-Level Copy-Paste Augmentation

비교할 경우에 이미지 해상도나 품질에 한계가 있는 것이 사실이다. 이러한 이유에서 본 연구에서는 기존의 샘플 이미지를 이용하여 보다 많은 {인간 바운딩 박스, 객체 바운딩 박스, 상호작용} 관계를 담고 있는 새로운 샘플 이미지를 생성할 수 있는 방법을 제안한다.

3.1 Simple Quattro Augmentation(SQA)

Simple Quattro Augmentation(SQA)는 타겟 샘플 이미지를 이어붙여서 새로운 이미지를 생성해내는 증강 방법이다. SQA의 생성 방법은 Fig. 5와 같다. 타겟 샘플이 들어오면 동일한 이미지를 수평 연결(HCONCAT), 수직 연결(VCONCAT)을 하여 Fig. 6과 같은 형태의 증강 이미지를 생성한다. 정답 데이터의 경우에도 증강된 이미지에 맞춰서 증강을 해주어야 학습 과정에서 인간과 객체의 위치와 상호작용에 대한 학습이 가능하다. 이를 위해 정답 데이터를 복사하고, 이미지 크기만큼 조절하여 추가하는 작업(ANNOTATION_MODIFY)이 필요하다. 이미지를 이어붙이는 방법을 적용하는 것은 2분할, 3분할 등 다양한 형태로 응용을 할 수 있지만, 본 연구에서는 입력데이터의 높이, 너비에 대한 비율이 깨지지 않도록 하기 위하여 4분할 데이터를 사용하였다. 또한, 비율을 고려하여 4분할이 아닌 9분할, 16

Algorithm 1: Simple Quattro Augmentation(SQA)

Input: $I = \{I_1, I_2, ..., I_N\}$ #N is number of dataset
Output: $O = \{O_1, O_2, ..., O_N\}$
for $I_i \in I$ **do**
  $\mathbf{O}_i = VCONCAT(I_i, I_i)$
  $\mathbf{O}_i = HCONCAT(O_i, O_i)$
  ANNOTATION_MODIFY($\mathbf{I}_i$)
end

Fig. 5. Simple Quattro Augmentation Pseudo-code

Fig. 6. Simple Quattro Augmentation(SQA) Example

분할 등을 적용하게 되면 하나의 증강된 이미지 내의 객체 정보가 기존 대비 매우 작아지는 문제점이 있기 때문에 4분할 데이터를 사용하였다. 이와 같이 4분할 이미지로 증강을 할 경우에 각 이미지간의 연결 부분이 아티팩트(Artifact)로 작용하여 사람의 눈으로 보기에는 매끄럽지 못 할 수도 있다. 그러나 본 연구에서는 아티팩트를 감안하더라도 증강된 이미지에 인간-객체 상호작용 수를 늘리는 부분에 초점을 맞추어 실험을 진행하였다. 이와 같이 SQA 기법을 통해 증강을 하면 기존 이미지 대비 4배 많은 인간-객체 상호작용이 생긴다는 장점이 있다. 추가적으로 각각의 인간-객체 상호작용 간에는 상호작용 관계가 없기 때문에 학습 과정에서 정답이 아닌 데이터(Negatively labeled data)로써 학습 모델의 가중치 업데이트에 사용될 수 있다. Fig. 6의 결과에서 볼 수 있듯이 동일한 이미지를 이어붙이기 때문에 각각의 인간, 객체 바운딩 박스는 위치만 다를 뿐 동일한 모습을 띄게 된다. 학습 모델은 해당 이미지를 학습하는 과정에서 동일한 모습의 인간, 객체 특징들을 통해 상호작용 관계를 학습하기 때문에 인간, 객체 각각의 개별 특징뿐만 아니라 객체 간의 위치나 거리 등의 맥락 정보를 학습할 필요가 있다는 것을 알 수 있게 된다.

3.2 Random Quattro Augmentation(RQA)

Random Quattro Augmentation(RQA)은 3.1절에서 제안한 SQA 기법보다 생성된 이미지 내에 다양한 인간-객체 상호작용을 가질 수 있도록 제안한다. SQA 기법은 동일한 이미지를 이어붙여서 새로운 이미지를 생성해내는 것이기 때문에 최대 생성 가능한 데이터가 기존의 학습 데이터 수에 한정된다. 또한, 증강을 함으로써 이미지 내에 인간-객체 상호작용의 숫자가 늘어나기는 하지만 각각의 생성 이미지 내에 새로운 형태의 인간-객체 상호작용이 생성되는 것은 아니다. 이에 본 연구팀에서는 서로 다른 이미지간의 이어붙이기를 통해 새로운 생성 이미지를 증강해내는 방법을 고안하였다. RQA의 생성 방법은 Fig. 8과 같다. 각각의 인간-객체 상호작용 학습 데이터는 하나 이상의 인간-객체 상호작용 정보를 담고 있다. 먼저 학습 데이터가 표현할 수 있는 모든 인간-객체 상호작용을 담을 수 있는 리스트(List)를 생성하고, 해당 인간-객체 상호작용이 존재하는 이미지 인덱스를 리스트에 담도록 한다(RANDOM_INDEX_SEARCH). 이후에 모든 이

Fig. 7. Random Quattro Augmentation(RQA) Example

Algorithm 2: Random Quattro Augmentation(RQA)

Input: $I = \{I_1, I_2, ..., I_N\}$
Manipulate_data: $M = \{M_1, M_2, ..., M_M\}$ # M is number of HOI
$\mathrm{M}_M \leftarrow RANDOM\_INDEX\_SEARCH(I_n)$
for $I_i \in I$ **do**
  $\mathrm{Target}_4 \leftarrow RANDOM\_SELECT(M_i, k = 4)$
  $\mathrm{tmp}_1 = VCONCAT(T_1, T_2)$
  $\mathrm{tmp}_2 = VCONCAT(T_3, T_4)$
  $\mathrm{O}_i = HCONCAT(tmp_1, tmp_2)$
  ANNOTATION_MODIFY(Target_4)
end

Fig. 8. Random Quattro Augmentation Pseudo-code

미지에 대하여 순차적으로 타겟이 되는 이미지에 가장 많이 존재하는 인간-객체 상호작용 인덱스를 뽑고, 이를 통해 리스트에 해당 인간-객체 상호작용을 포함하고 있는 타겟 이미지 4장을 무작위로 뽑도록 하였다. 뽑힌 4장의 이미지를 수평 연결, 수직 연결을 하게 되면 증강 이미지 생성이 된다. 이와 같이 무작위 선택을 하게 되면, 전체 데이터셋에서 인간-객체 상호작용의 숫자가 많은 경우에는 보다 다양한 경우의 수로 증강 이미지가 생성될 것이고, 인간-객체 상호작용의 숫자가 적은 경우라도 SQA에서 생성한 증강 이미지보다 다양성을 가질 수 있게 된다. Fig. 7에서 확인할 수 있듯이 3.1절에서 제안한 SQA 방법과 비교하였을 때, RQA 방법을 적용하면 보다 다양성을 갖는 증강 데이터를 획득할 수 있다. 증강을 하고자 하는 목표 인간-객체 상호작용을 포함하고 있는 모든 이미지를 후보군으로 두고 무작위로 4장의 이미지를 선택하여 이어붙이기를 수행하기 때문에 모든 증강 이미지에는 동일한 이미지가 포함될 수도 있다.

4. 실험 및 결과

본 장에서는 제안하는 증강 기법에 대한 실험을 위해 사용된 학습 데이터와 성능 측정 지표, 학습 모델에 대하여 설명하고, 실험 결과를 보인다. 본 연구에서 제안하는 SQA와 RQA 기법에 대한 검증을 위하여 인간-객체 상호작용 탐지 모델 중 가장 높은 SOTA(State-of-the-Art) 성능을 달성하고 기준 모델(baseline model)을 선정하고 기본 학습 데이터로 학습하였을 때의 성능과 증강된 데이터를 이용하여 학습하였을 때의 성능을 비교, 평가하였다. Fig. 9에서 확인할

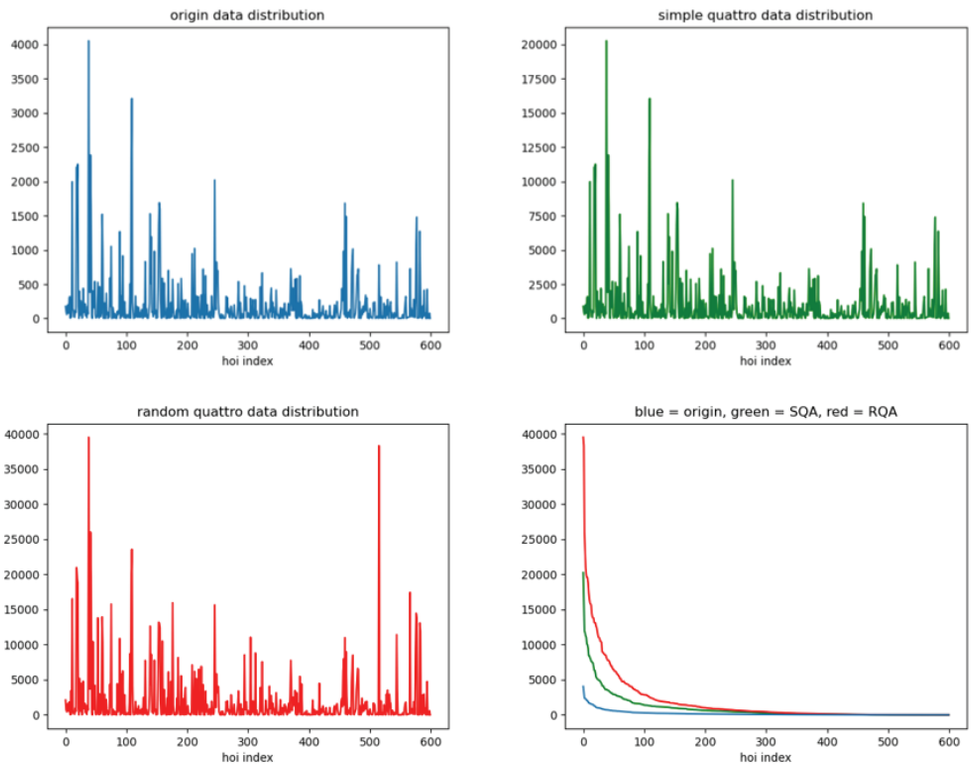

Fig. 9. HICO-DET Data Distribution (Top Left) Original (Top Right) SQA (Bottom Left) RQA (Bottom Right) Comparison Between Distribution by Descending Order

수 있듯이 SQA 방법을 통해 학습 데이터를 증강하게 되면 기존의 학습 데이터의 인간-객체 상호작용보다 4배 많은 인간-객체 상호작용 관계가 생성된다. 그에 반해 RQA 방법을 통해 학습 데이터를 증강하는 경우에는 전체 학습 데이터에 대하여 동일한 인간-객체 상호작용 관계를 갖는 이미지를 무작위로 선정하여 데이터 증강을 하기 때문에 증강된 데이터 분포가 기존 데이터 분포와 다른 형태를 띄는 것을 알 수 있다. 실제 전체 데이터셋에 대한 인간-객체 상호작용 분포를 살펴보더라도 기존 데이터셋 대비 많은 인간-객체 상호작용 데이터가 생성된 것을 확인하였다.

4.1 데이터셋과 평가 지표

본 연구에서는 제안한 증강 기법에 대한 평가를 위하여 HICO-DET 데이터셋을 활용하였다. HICO-DET은 마이크로소프트에서 공개하고 있는 객체 탐지, 자세 추정, 객체 분할 등의 분야에서 활용되고 있는 MS-COCO에서 일부분(Subset)을 추출하고 생성한 학습 데이터이다. 최초 공개된 데이터셋은 HICO[17]라는 이름으로 발표가 되었으나, 일부 객체에 대한 레이블링이나 학습 이미지의 부재 등에 의하여 발전된 형태로 공개된 데이터셋이 HICO-DET이다. HICO-DET은 총 47,776장의 이미지(38,118장의 학습 데이터, 9,658 장의 평가 데이터)로 이루어져 있으며, 600개의 인간-객체 상호작용 카테고리, 80개의 객체 정보, 117개의 상호작용으로 구성되어 있다.

평가 지표로는 주로 객체 탐지 분야에서 활용이 되는 mAP(Mean Average Precision)을 사용하였다. 현재 모든 인간-객체 상호작용 탐지 선행 연구에서는 탐지 모델의 성능 평가를 위해서 mAP를 사용한다. mAP는 다음의 두 가지 조건이 모두 충족하였을 때 올바른 정답(True Positive)로 간주한다. 1) 인간-객체 상호작용 예측이 정확하게 되었을 경우와 2) 인간과 객체의 예측된 바운딩 박스와 실제 정답간의 IoU(Interaction of Union)가 일정 임계치(보통 0.5 이상)를 넘었을 때이다.

4.2 상세 평가 환경 설명

1) 학습 모델

본 연구에서는 인간-객체 상호작용 탐지 모델에서 SOTA 성능을 달성하고 있는 UPT[9] 모델과 SCG[10] 모델을 이용하여 성능 평가를 수행하였다. UPT 모델은 트랜스포머(Transformer) 기반의 모델로 객체 탐지 분야에서 높은 성능을 달성한 DETR[14]을 변형한 인간-객체 상호작용 탐지 모델이다. SCG 모델은 객체 탐지를 먼저 수행하고, 탐지된 객체 바운딩 박스로부터 특징 추출을 통해 인간, 객체, 상호작용 각각에 대한 분류를 수행하는 2-Stage 방법에서 높은 성능을 달성한 모델이다. SCG 모델의 경우에 일반적으로 인간-객체 상호작용에서 중요한 요소라고 여겨지는 시각적 특징(Visual feature), 공간적 특징(Spatial feature)을 함께 활용하기 위하여 콘볼루션 그래프 네트워크(Convolutional Graph Net-

work)를 이용하여 탐지 모델을 설계한 것이 특징이다.

데이터 증강 기법의 정확한 성능 평가를 위하여 기준이 되는 두 가지 모델에 대하여 동일한 환경에서 기본 학습 데이터셋(HICO-DET)과 증강된 학습 데이터셋을 각각 학습을 진행하고 성능 측정을 하여 비교, 평가하였다.

2) 하이퍼파라미터 설정

본 연구에서는 평가 기준 모델인 UPT 모델과 SCG 모델의 저자가 공개하고 있는 최고 성능 달성이 가능한 하이퍼파라미터 설정값을 사용하여 학습을 진행하였다. 이 설정 값에는 객체 탐지 사전 학습 모델(Pre-trained model), 학습 횟수(Iteration), 배치 크기(Batch size), 학습율(Learning rate), 가중치 감쇠(Weight decay) 등이 사용되었다.

4.3 실험 결과

제안한 증강 기법을 통해 HICO-DET 데이터셋에 적용하여 증강한 학습 데이터셋을 통해 평가 모델에 적용한 결과 RQA 기법을 이용할 경우 SCG 모델 기준 약 0.6 mAP, UPT 모델 기준 약 0.12 mAP 정도의 성능 향상을 확인하였다. 인간-객체 상호작용 데이터셋의 경우에 긴 꼬리 분포를 갖는 특성에 의해 전체 데이터셋에 대한 mAP와 샘플수가 많은 경우인 비희귀 클래스와 샘플수가 적은 경우인 희귀 클래스를 구분하여 계산하고 비교한다. 긴 꼬리 분포에서 샘플이 많은 부분에 속하는 비희귀 클래스에 대한 성능 평가에서는 SCG 모델 기준 약 1.4 mAP, UPT 모델 기준 약 0.4 mAP 의 성능 향상을 달성한 것에 반해 희귀 클래스에서의 성능 지표는 크게 향상되지 않았음을 확인하였다. 이는 본 연구에서 제안한 증강 기법이 인간-객체 상호작용 기준으로 증강을 수행하여 긴 꼬리 분포 문제를 해결하기에는 희귀 클래스에 해당하는 데이터에 대한 증강이 충분하게 되지 않았다는 것을 보여준다. 결과적으로 Table 1에서 볼 수 있듯이 제안한 증강 기법을 적용할 경우에 기준 모델 기준으로 전체(Full)/희귀(Rare)/비희귀(Non-rare) 각각의 지표에서 높은 mAP를 달성한 것을 확인하였다. 또한, Fig. 10에서와 같이 SQA, RQA 각각의 증강된 데이터셋으로 학습하였을 때, 기존의 원본 데이터셋으로 학습한 모델이 탐지하지 못하는 샘플들을 탐지할 수 있으며, 동일하게 탐지하더라도 기존보다 높은 스코어를 기록하는 것을 확인하였다. 동일하게 탐지하는 경우라 할지라도 높은 스코어를 달성하는 것이 중요한 이유는 일반적으로 탐지 모델을 활용하는데 있어서 특정 역치(Threshold) 이상을 달성하는 경우 탐지가 성공적으로 되었다고 판단하기 때문이다.

Table 1. Results on HICO-DET : SCG and UPT Models

Method	Backbone	Full	Rare	Non-rare
SCG (baseline)	ResNet-50	30.85	21.19	32.18
SCG_ours (SQA)	ResNet-50	31.32	24.27	32.68
SCG_ours (RQA)	ResNet-50	**31.42**	**24.73**	**33.55**
UPT (baseline)	ResNet-50	30.92	25.13	32.64
UPT_ours (SQA)	ResNet-50	30.96	**25.20**	32.68
UPT_ours (RQA)	ResNet-50	**31.04**	24.65	**32.94**

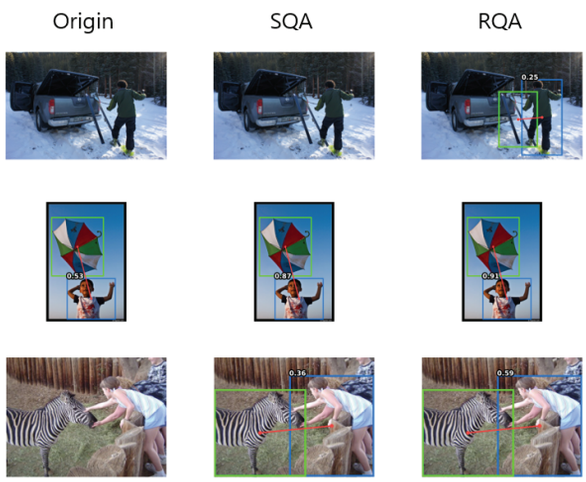

Fig. 10. HOI Prediction Result on HICO-DET by UPT
Model (First Column) HOI : {human, repair, skis}
(second column) HOI : {human, lose, umbrella}
(third column) HOI : {human, hold, zebra}

5. 결 론

본 연구에서는 인간-객체 상호작용 탐지 연구 분야에서 활용 가능한 데이터 증강 기법을 제안하였다. 기존의 이미지 증강 기법은 인간-객체 상호작용을 유지하면서 증강을 하기에 어려움이 있어 기존의 학습 데이터에 존재하는 이미지를 서로 이어붙이는 방법을 통하여 인간-객체 상호작용을 유지하면서 새롭게 생성되는 이미지에서 보다 많은 인간-객체 상호작용을 제공할 수 있도록 하였다. 제안한 방법을 통해 실제 SOTA 학습 모델에 적용하여 성능 향상이 되는 것을 확인하였다. 그러나 증강 기법을 적용하였음에도 불구하고 인간-객체 상호작용 탐지 분야에서 사용하는 데이터셋에서 공통적으로 발생하는 긴 꼬리 분포 문제를 해결하기에는 희귀 클래스에 해당하는 샘플의 증강이 충분히 이루어지지 않았다. 희귀 클래스에 해당하는 데이터셋에 대한 증강을 위해서는 해당 클래스에 속하는 인간-객체 상호작용 정보를 추가적으로 획득할 수 있는 방법에 대한 연구가 이루어져야 한다. 이에 본 연구팀에서는 향후 연구로 긴 꼬리 분포 문제를 해결하면서 인간-객체 상호작용 탐지 모델의 성능을 향상시킬 수 있는 증강 기법에 대한 연구를 수행할 예정이다. 또한, 본 연구에서 확인한 것과 같이 학습 데이터에 인간-객체 상호작용이 많아질수록 학습 성능이 향상되는 점에서 착안하여 기존에

선행 연구된 인간-객체 상호작용 탐지 모델의 구조에 대한 분석을 수행하고 새로운 모델 설계를 통해 인간-객체 상호작용 탐지 모델에 대한 연구, 개발도 수행할 예정이다.

References

[1] I. Kostrikov, D. Yarats, and R. Fergus. "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *arXiv preprint arXiv:2004.13649*, 2020.

[2] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," *arXiv preprint arXiv:2112.07383*, 2021.

[3] H. S. Fang, Y. Xie, D. Shao, Y. L. Li, and C. Lu, "DecAug: Augmenting HOI detection via decomposition," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.35, No.2, pp.1300-1308, 2021.

[4] Y. W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018.

[5] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

[6] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. Hengel, "HCVRD: A benchmark for large-scale human-centered visual relationship detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.32, No.1, 2018.

[7] H. S. Fang, J. Sun, R. Wang, M. Gou, Y. L. Li, and C. Lu, "Instaboost: Boosting instance segmentation via probability map guided copy- pasting," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[8] T. Y. Lin et al., "Microsoft coco: Common objects in context," *European Conference on Computer Vision*, Springer, Cham, 2014.

[9] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[10] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[11] S. Li, K. Gong, C. H. Liu, Y. Wang, F. Qiao, and X. Cheng, "Metasaug: Meta semantic augmentation for long-tailed visual recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[12] F. Z. Zhang, D. Campbell, and S. Gould. "Spatially conditioned graphs for detecting human-object interactions," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[13] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, Vol.6, No.1, pp.1-48, 2019.

[14] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[15] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol.36, No.2, 2022.

[16] A. Zhang et al., "Mining the benefits of two-stage and one-stage hoi detection," *Advances in Neural Information Processing Systems*, Vol.34, pp.17209-17220, 2021.

[17] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

이 상 백

https://orcid.org/0000-0002-4440-1792
e-mail : roy881020@cnu.ac.kr
2014년 충남대학교 컴퓨터공학과(학사)
2014년 ~ 현 재 충남대학교 컴퓨터융합학부 석·박사통합과정
관심분야 : 딥러닝, 지식베이스, 인간-객체 상호작용 탐지

이 규 철

https://orcid.org/0000-0003-0857-807X
e-mail : kclee@cnu.ac.kr
1984년 서울대학교 컴퓨터공학과(학사)
1986년 서울대학교 컴퓨터공학과(석사)
1990년 서울대학교 컴퓨터공학과(박사)
1990년 ~ 현 재 충남대학교 컴퓨터융합학부 교수
관심분야 : 데이터베이스, 딥러닝, 빅데이터, 융합기술