전자동 생성형 이미지 증강 시스템에 대한 연구

정성윤, 정남훈, 조규태
LIG 넥스원

seongyun.jeong@lignex1.com, namhoon.jung@lignex1.com, kyutae.cho2@lignex1.com

A Study on the Fully Automated Generative Image Augmentation System

Seongyun Jeong, Namhoon Jung, Kyutae Cho
LIG Nex1

요 약

최근 생성형 모델의 발전으로 이미지 생성 기술이 빠르게 혁신되고 있으며, 특히 생성기반의 이미지 데이터 증강 기법에 대한 연구가 활발히 이루어지고 있다. 이러한 기법은 컴퓨터 비전 과제에서 분류 모델의 일반화 성능을 향상시키고 클래스 불균형 문제를 완화하는 데 기여하고 있다. 그러나 정밀하고 자연스러운 데이터 증강을 위해서는 목적에 맞는 정교한 프롬프트와 시각적 가이드와 같은 요소를 수동으로 준비해야 하며, 이는 대규모 증강에 적용하기에 시간과 작업량 측면에서 비효율적이다. 본 논문에서는 이러한 한계를 해결하기 위해 전자동 생성형 이미지 증강 시스템을 제안한다. 제안하는 시스템은 사용자가 수행했던 프롬프트 작성 및 시각적 특징 추출 과정을 언어 모델과 비전 모델들을 활용해 자동화함으로써, 이미지 입력만으로도 사용자의 개입 없이 다양한 시나리오가 반영된 고품질 이미지를 생성할 수 있도록 한다. 또한, 본 시스템을 활용해 생성한 데이터를 이미지 분류 모델 학습에 적용한 결과, 분류 성능이 향상되는 것을 실험을 통해 확인하였다.

Ⅰ. 서 론

최근 Stable Diffusion [1]과 같은 생성형 인공지능(Generative AI)의 등장으로 이미지 생성 분야에서 큰 기술적 혁신이 이루어지고 있다. 이러한 고성능 생성 모델들은 다양한 컴퓨터 비전 과제에 활발히 활용되고 있으며, 특히 데이터 증강(Data Augmentation) 분야에서 주목받고 있다. 생성형 AI 는 클래스 불균형 문제를 완화하고, 수집이 어려운 희귀 데이터를 보완하는 데 효과적으로 사용된다.

최신 이미지 생성 모델들은 입력 방식에 따라 크게 두 가지로 나뉜다. 텍스트를 기반으로 이미지를 생성하는 Text-to-Image 방식과, 이미지를 기반으로 새로운 이미지를 생성하는 Image-to-Image 방식이다. 텍스트 프롬프트는 주로 배경, 객체, 시나리오 설정 등 전반적인 컨셉을 제시하는 데 사용되며, 이미지 가이드는 객체의 형태, 위치, 구도 등 시각적 특성을 제어하는 데 활용된다. 이 두 방식은 생성 목적에 따라 혼용되며, 사용자가 의도한 이미지에 보다 가깝게 생성될 수 있도록 돕는다.

그러나 사용자가 원하는 고품질 이미지를 얻기 위해서는 정교하게 작성된 프롬프트와 명확한 시각 정보가 요구된다. 예를 들어, 프롬프트에는 배경과 객체 간의 상호작용을 포함한 구체적인 설명이 요구되며, 배경의 완전한 변환을 위해 필요한 객체와 배경의 분리에는 배경 제거 작업이 선행되어야 한다. 이러한 작업이 수작업으로 이루어지다 보니, 실제 데이터 증강 파이프라인 활용은 시간과 비용 측면에서 제약이 따른다.

본 논문에서는 위의 수작업 과정을 자동화하기 위한 전자동 이미지 증강 시스템을 제안한다. 제안하는 시스템은 사용자가 증강하고자 하는 이미지를 입력으로 받아, 클래스 이름을 자동으로 추출하고 주요 객체를 분리한 후, 객체로부터 Edge, Depth 와 같은 시각 단서들을 생성한다. 이어서 언어 모델이 자동으로 이미지 생성을 위한 프롬프트를 작성하고, 최종적으로 다양한 경우의 수에서 새로운 이미지를 생성한다. 사용자는 단순히 이미지를 입력하기만 하면, 시스템이 객체의 시각적 특징을 추출하고 자동으로 시나리오를 설계해 반영된 이미지를 생성하는 구조이다.

제안하는 시스템은 대규모 데이터셋 증강에 적용 가능하며, 복잡한 수작업 없이도 다양한 조건의 데이터를 효과적으로 생성할 수 있다. 본 논문에서는 시스템의 전체 구조를 상세히 설명하고, 생성된 데이터를 이미지 분류 모델 학습에 활용했을 때의 성능 향상 효과를 실험을 통해 입증한다.

Ⅱ. 본론

1. 구조

제안하는 전자동 생성형 이미지 증강 시스템은 먼저 입력 이미지를 Vision-Language Model(VLM)을 통해 분석하여 주요 키워드를 추출하고 (본 연구에서는 WD14 Tagger 사용), 이를 클래스 이름으로 활용한다. 추출된 클래스 이름은 GroundingDINO [2] 모델의 입력으로 사용되어 객체의 위치가 Annotated Image 형태로 출력되며, 이는 다시 SAM [3] 모델에 전달되어 객체의 Mask 이미지를 생성한다. 이 과정을 통해 입력 이미지로부터 객체를 분리한 후, Rotation, Flipping과 동의 이미지 변환 기법을 적용하여 다양한 시점(View Point)의 이미지를 생성한다. 다음으로 생성한 이미지를 Canny Edge 알고리즘과 Depth Anything [4] 모델을 통해 중요 시각적 특징들을 추출한다. 이는 이미지

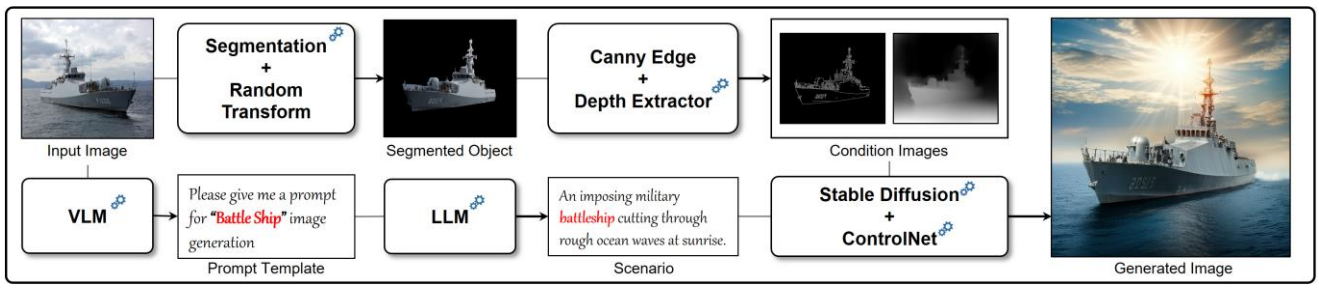

그림 1 전자동 생성형 이미지 증강 시스템의 구조 및 동작 예시

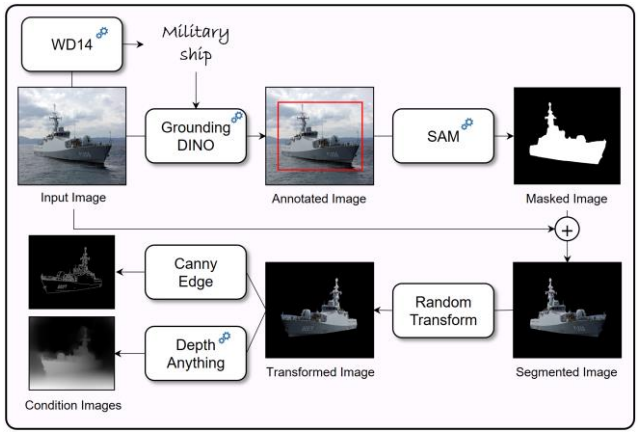

그림 2 시각적 특징 추출 과정

생성시 시각적 가이드로 활용된다.

추출한 키워드는 사전에 정의된 프롬프트 템플릿의 빈칸을 채우는 데 사용되며, 완성된 템플릿은 Large-Language Model(LLM)의 입력으로 활용되어 이미지 생성을 위한 최종 프롬프트를 생성한다. 시스템의 전체 동작 과정은 그림 1에 제시된 파이프라인을 통해 확인할 수 있다. 그림 2를 통해 시각적 특징을 추출하는 자세한 과정을 확인할 수 있다. 그림 3은 해당 작업을 여러 번 수행하였을 때 생성된 데이터의 예시를 보여준다.

그림 3 제안하는 시스템을 활용해 생성한 이미지 예시

2. 실험

제안한 시스템의 이미지 분류 과제에서의 유효성을 검증하기 위해 ResNet-18 모델을 Vehicle-512 데이터셋을 증강해가며 Random initialized된 모델을 50epoch, 0.001 Learning Rate, 512 Batch size로 설정하고, Adam Optimizer를 활용해 학습 시키고 이미지 분류 성능을 확인했다.

학습에 사용한 Vehicle-512 데이터셋은 16개의 클래스 별 200장의 이미지로 구성되어 있고 모든 실험은 NVIDA A100 GPU를 통해 수행되었다.

표 1에서 분류 모델 학습시 데이터의 증강 비율이 증가할수록 정확도가 증가하는 모습을 확인할 수 있다. 이 결과를 통해 제안하는 시스템을 통해 증강한 데이터가 분류 모델의 일반화 성능을 개선할 수 있음을 보였다.

표 1 ResNet-18 이미지 분류 성능 비교표

Synthesis Ratio	Trainset Size	Accuracy	Δ
0%	1x	69.06%	-
25%	1.25x	71.88%	2.82
50%	1.5x	75.00%	5.94
100%	2.0x	78.44%	9.38

Ⅲ. 결론

본 논문에서는 생성형 인공지능을 활용하여 데이터 증강을 자동화하는 이미지 증강 시스템을 제안하였다. 이 시스템은 입력 이미지로부터 클래스 이름과 주요 객체를 자동으로 추출하고, 객체의 시각적 특징을 분석한 뒤, 언어 모델을 통해 프롬프트를 자동으로 생성함으로써 최종적으로 사용자가 원하는 이미지와 유사한 증강 데이터를 생성한다. 이를 통해 기존 방식에서 수작업으로 이루어졌던 프롬프트 작성과 객체 분리 작업의 부담을 획기적으로 줄였다. 제안한 시스템을 이미지 분류 모델 학습에 적용한 결과, 증강된 데이터셋이 모델의 분류 성능을 유의미하게 향상시킴을 확인하였다. 이는 본 시스템이 단순한 자동화 도구를 넘어, 실제 학습 성능 향상에 기여할 수 있는 실용적인 데이터 증강 시스템임을 보여준다.

참 고 문 헌

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1068410695, 2022.

[2] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision, pages 38– 55. Springer, 2024.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4015– 4026, 2023.

[4] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao. Depth anything v2. Advances in Neural Information Processing Systems, 2024.