거대 텍스트-to-이미지 생성 모델을 활용한 주차 판별 작업 데이터 증강

김남우, 강병우, 한광석, 유환조*

포항공과대학교

{treekim, bykang, gshan, hwanjoyu}@postech.ac.kr

Data Augmentation for Parking Detection Tasks Using Large-Scale Text-to-Image Generation Models

Kim Nam U, Kang Byoung Woo, Han Gwang Seok, Yu Hwan Jo*

POSTECH(Pohang University of Science and Technology)

요 약

본 논문에서는 거대 텍스트-to-이미지 생성 모델의 데이터 증강 능력을 탐구한다. 특히, 스테이블 디퓨전과 ControlNet을 활용하여 데이터의 공간적 구조를 유지하면서 고품질의 실내 CCTV 데이터를 생성하는 방법을 제안하고, 이 방법을 통해 생성된 데이터를 주차 유무 판별하는 작업에 적용했다. 실제 실내 주차장 CCTV 데이터를 사용하여 모델의 성능을 검증하였고, 생성된 이미지만으로 학습한 모델이 실제 데이터로 학습한 모델과 비슷하거나 더 우수한 성능을 보임을 확인했다. 이 결과는 제안된 방법이 데이터 라벨링과 수집 비용을 절감하면서 실제 현장 문제를 해결할 수 있는 효과적인 솔루션을 제공함을 보여준다. 이 연구는 또한 데이터 수집이 어려운 특정 상황에 대한 이미지를 생성하는 미래 연구 방향을 제시하며, 공간적 특성이 중요한 작업에 데이터 증강을 적용할 수 있는 기반을 마련한다.

Ⅰ. 서 론

딥러닝 모델은 충분한 양의 데이터 없이는 학습 데이터에 쉽게 과적합되며, 결과적으로 모델의 일반화 능력이 저하될 수 있다. 충분한 데이터를 확보하는 데에는 높은 라벨링 및 수집 비용이 들고 추가 데이터 수집이 불가능한 문제가 발생한다. 이러한 문제를 해결하기 위해, 데이터 증강 방법이 활용되고 주로 데이터 증강은 기하학적 변환(예: 회전, 뒤집기)과 광도 변환(예: 색상, 밝기 조정)을 결합하여 기존 이미지에서 새로운 이미지를 생성한다[1].

최근 거대 텍스트-to-이미지 생성 모델이 고도로 사실적인 이미지를 생성할 수 있게 되면서, 이러한 모델을 통한 데이터 증강 방법이 컴퓨터 비전 분야에서 활용되고 있다[2,3]. 기존 데이터 증강 방법에 비해 더 다양하고 고차원적인 데이터를 생성할 수 있게 되었지만, 데이터의 공간적 구조가 중요한 작업에서 충분한 성능을 낼 수 있는지에 대한 검증은 아직 이루어지지 않았다.

이러한 한계를 극복하기 위해 본 논문에서는 거대 텍스트-to-이미지 생성 모델 스테이블 디퓨전[4]에 ControlNet[5]을 적용함으로써 원하는 데이터 구조와 형태를 유지하면서 데이터를 증강할 수 있는 방법을 제안한다. 특히, 주차 유무를 판별하는 작업에 적용된 이 방법은 주차장 CCTV 데이터를 생성하여, 생성된 데이터만으로 학습한 모델이 실제 데이터로 학습한 모델만큼의 성능을 내는 것을 보임으로써, 제안 방법이 데이터 라벨링과 수집 비용을 줄이고 실제 현장 문제를 해결할 수 있는 효과적인 방법임을 입증한다.

Ⅱ. 본론

주차 유무를 판별하는 작업을 위해 데이터를 생성할 때, 각 주차 공간을 정확하게 식별하고 위치를 일관되게 유지하는 것은 모델이 주차 공간의 경계를 정확히 학습하고 변화하는 환경에서도 안정적으로 주차 유무를 판단할 수 있게 만드는 데 중요하다.

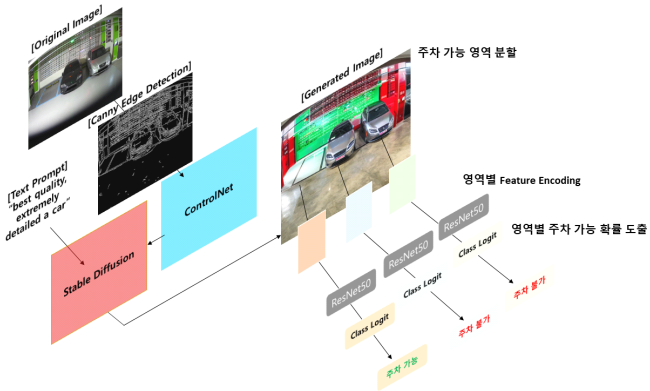

그림 1 데이터 증강 및 주차 판별 전체 구조도

본 논문에서는 ㈜한맥아이피에스가 제공한 실제 실내 주차장 CCTV 데이터를 사용하여 제안된 방법론의 적합성을 검증했다. 총 10대의 CCTV에서 데이터가 수집되었으며, 각 CCTV는 3개의 주차 공간을 촬영했다. 평균적으로 CCTV마다 277개의 이미지가 수집되었고, 표준편차는 39개이다. 이 중 2대의 CCTV에서 수집된 595개의 이미지(1,785개의 주차면)를 사용하여 모델을 학습시켰고, 나머지 8대의 CCTV에서 수집된 2,174개의 이미지(6,522개의 주차면)로 모델의 성능을 평가했다. 주차 유무 판별 방법론으로는 [6]에서 사용된 R-CNN과 square pooling 기법을 적용했고, 각 이미지 데이터에서 주차 영역을 구분한 후, ResNet50을 통해 각 영역별 특징을 추출하고 주차 유무를 분류했다.

생성 이미지와 실제 이미지의 성능 차이를 동등하게 비교하기 위해, 학습에 사용된 이미지의 수량과 구조를 동일하게 맞춘 생성 이미지를 제작했다. 이미지 생성 과정은 다음 단계로 진행되었다.

1. Canny edge detection : 실제 이미지에 Canny edge detection을 적용하여 (하한 임계값 100, 상한 임계값 200으로 설정), 이미지의 구조적 요

평가 CCTV	전체 주차면 수	오분류 주차면 수(원본)	오분류 주차면 수(생성)	주차 판별 정확도(원본)	주차 판별 정확도(생성)
CCTV1	906	10	7	0.989	0.992
CCTV2	924	14	7	0.985	0.992
CCTV3	921	29	7	0.969	0.992
CCTV4	654	12	6	0.982	0.991
CCTV5	741	8	10	0.989	0.987
CCTV6	843	4	3	0.995	0.996
CCTV7	876	3	2	0.997	0.998
CCTV8	657	13	4	0.980	0.994

Table 1 학습 데이터에 따른 CCTV별 주차 판별 정확도 결과

소를 명확하게 한다. 이 과정에서 생성된 Canny edge 이미지는 ControlNet의 입력값으로 사용되어, 텍스트-이미지 생성 모델에서 원하는 구조를 유지하는 데 필수적인 역할을 한다.

2. ControlNet Canny 모델 : Canny edge detection 이미지를 입력 값으로 받아, 해당 edge의 형태를 유지하면서 텍스트-이미지 생성 모델이 입력받은 프롬프트에 따른 이미지를 생성하도록 제어하는 역할을 한다.

3. 텍스트-이미지 생성 모델 : 유저가 입력한 프롬프트와 ControlNet의 임베딩 값을 받아 최종 이미지를 생성하는 역할을 한다. 스테이블 디퓨전 1.5 버전, UniPC[7] 스케쥴러를 적용했고 80번의 추론 단계를 거치도록 설정했다. 또한, 각 이미지 생성 시 1부터 100,000 사이의 무작위 seed를 할당하여 생성 이미지의 다양성을 높이도록 했다. 긍정적 프롬프트로는 "best quality, extremely detailed a car"를, 부정적 프롬프트로는 "monochrome, lowres, worst quality, low quality"를 사용하여 높은 품질과 상세한 자동차 이미지 생성을 목표로 했다.

스테이블 디퓨전 구조는 크게 인코더 블록, 미들 블록, 디코더 블록으로 나뉜다. 이 중 인코더와 미들 블록은 ControlNet에서 컨트롤 이미지를 인코딩하는 데 사용될 수 있도록 동일한 구조로 복제된다. Θ_c 는 복제된 파라미터를 나타내며, $Z(\cdot;\cdot)$는 1x1 컨볼루션 레이어를 의미한다. $y = F(x;\Theta)$는 원본 스테이블 디퓨전의 디코더 블록의 출력을 의미하고 $y_c = F(x;\Theta) + Z(F(x + Z(c;\Theta_{z1});\Theta_c);\Theta_{z2})$ 는 원래의 스테이블 디퓨전 출력에 복제된 인코더와 제로 컨볼루션을 통해 추가적인 처리가 더해진 결과를 의미한다. 본 논문에서는 사전 훈련된 1.5버전의 스테이블 디퓨전과 ControlNet canny 모델을 사용하여 도출된 y_c 값을 기반으로 이미지를 생성했다.

위 단계를 거쳐 생성된 이미지만으로 학습한 모델과 실제 이미지만으로 학습한 모델로 나누어 주차 유무 판별 성능을 측정 및 비교했다. 학습에 사용되지 않은 CCTV에서 수집한 데이터로 평가를 진행했으며, 결과는 Table 1에서 확인할 수 있다. 이 결과에 따르면 생성된 이미지로 학습한 모델은 실제 데이터로 학습한 모델만큼이나 좋거나 더 우수한 성능을 보여주었다. 생성 과정을 통해 데이터의 다양성을 높임으로써 더욱 강건한 모델 학습이 가능해진 것으로 생각할 수 있다.

Ⅲ. 결론

본 논문은 텍스트-to-이미지 생성 모델, 특히 ControlNet과 스테이블 디퓨전을 활용하여 고품질의 주차장 이미지를 생성하는 새로운 방법을 제안하고, 이를 주차 유무를 판별하는 작업에 적용함으로써 실제 문제 해결에 기여할 수 있음을 보여준다. 이 방법론은 높은 라벨링 비용과 수집 비용 없이 원하는 구조의 데이터를 생성할 수 있게 하며, 공간적 특성이 중요한 작업에 데이터 증강을 효과적으로 적용할 수 있는 근거를 제시한다. 또한, 데이터 수집이 어려운 특정 상황(비오는 날, 야간 등)에 대한 이미지를 생성하는 미래 연구 방향에 대해서도 고려해 볼 수 있을 것이다.

ACKNOWLEDGMENT

This work was supported by the Digital Innovation Hub project supervised by the Daegu Digital Innovation Promotion Agency(DIP) grant funded by the Korea government(MSIT and Daegu Metropolitan City) in 2024 (No. DBSD1-07).

※ MSIT: Ministry of Science and ICT.

참 고 문 헌

[1] Cheung, T. H., & Yeung, D. Y. (2023). A survey of automated data augmentation for image classification: Learning to compose, mix, and generate. IEEE Transactions on Neural Networks and Learning Systems.

[2] He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., ... & Qi, X. (2022). Is synthetic data from generative models ready for image recognition?. arXiv preprint arXiv:2210.07574.

[3] Trabucco, B., Doherty, K., Gurinas, M., & Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. arXiv preprint arXiv:2302.07944.

[4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

[5] Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3836-3847).

[6] Marek, M. (2021). Image-based parking space occupancy classification: Dataset and baseline. arXiv preprint arXiv:2107.12207.

[7] Zhao, W., Bai, L., Rao, Y., Zhou, J., & Lu, J. (2024). Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. Advances in Neural Information Processing Systems, 36.