딥 러닝 기반 이미지 생성 모델을 활용한 객체 인식 사례 연구

*강다빈, *홍지수, 김재홍, 송민지, 김동휘, 박상효
경북대학교 컴퓨터학부
zefeni@knu.ac.kr

A Case Study of Object detection via Generated image Using deep learning model based on image generation

*Dabin Kang, *Jisoo Hong, Jaehong Kim, Minji Song, Dong-hwi Kim, and Sang-hyo Park
School of Computer Science and Engineering, Kyoungpook National University

요약

본 논문에서는 생성된 이미지에 대한 YOLO 모델의 객체 인식의 성능을 확인하고 사례를 연구하는 것을 목적으로 한다. 최근 영상 처리 기술이 발전함에 따라 적대적 공격의 위험성이 증가하고, 이로 인해 객체 인식의 성능이 현저히 떨어질 수 있는 문제가 발생하고 있다. 본 연구에서는 앞서 언급한 문제를 해결하기 위해 text-to-image 모델을 활용하여 기존에 존재하지 않는 새로운 이미지를 생성하고, 생성된 이미지에 대한 객체 인식을 사례 별로 연구한다. 총 8가지의 동물 카테고리로 분류한 후 객체 인식 성능을 확인한 결과 86.46%의 정확도로 바운딩 박스를 생성하였고, 동물에 대한 116개의 60.41%의 정확도를 보여주었다.

1. 서론

최근 영상처리 기술이 발전함에 따라 적대적 공격(adversarial attack)의 위험성이 증가하였다. 이러한 적대적 공격은 딥 러닝 모델의 내부적 취약점을 이용하여 만든 특정 노이즈값을 이용해 의도적으로 오분류를 이끌어내는 입력값을 만들어내는 것을 의미한다. 적대적 공격으로 특별히 머신러닝 시스템을 속이기 위해 제작된 입력 값을 적대적 사례(adversarial example)라고 한다[2]. 그러므로 적대적 사례는 원본 데이터에 최적의 노이즈를 추가하여 생성되며, 사람이 보기에는 문제가 없지만 딥러닝 모델에 의해서 잘못 오인식되는 데이터를 의미한다[1].

적대적 사례에는 몇 가지 주요 예시가 있다. 워싱턴 대학의 타다요시 코노 박사 연구팀은 AEs를 이용해, 도로 교통 표지판에 작은 스티커를 붙여 “정지”신호를 “신호 45마일 속도 제한”신호로 오분류하게 하는 적대적 공격 방법을 발견했다[11]. 또한 통상적으로 사용되는 적대적 사례의 생성방법인 FastGradient Sign Method (FGSM)[2], Deepfool[3], Jacobian-BasedSaliency Map Attack (JSMA)[4], carlini wagner (CW)[5]이 벤치마크로 사용된다. 이외에도 한 개의 픽셀을 조작하여 오인식을 일으키는 one pixel 공격[6]과 decision boundary 측면에서 black box attack일지라도 노이즈를 주면서 찾아가는 방법[7]이 있다. 최근에는 backward pass differentiable approximation (BPDA) 방법[8]과 adaptive 방법[9]이 최신 공격 방법으로써, 적대적 사례의 방어 대책들에 대해서도 상당히 높은 확률의 공격 성공률을 보여준다. 위의 언급된 방법 말고도 많은 적대적 사례 연rn들이 소개되고 있다[1].

이러한 적대적 사례들로 인해 객체 인식의 성능이 현저히 떨어질 수 있는 문제가 발생하고 있다. 객체 인식의 취약점은 causative attack과 exploratory attack으로 분류되는데, 적대적 사례는 그 중 exploratory attack의 대표적인 예시이다. exploratory attack[10]은 이미 학습이 끝난 모델에 대하여 테스트 데이터를 조작해 모델의 오인식을 유발하는 공격 방법이다[1].

앞서 언급한 문제를 해결하기 위해 본 논문에서는 생성된 이미지에 대한 YOLOv5[17]의 객체 인식 성능을 사례 별로 연구하고자 한다. 또한 이미지 생성에는 text-to-image의 성능이 좋은 DALL-E-2[18, 19]를 활용한다. 최근 들어 OpenAI의 DALL-E-2, Google의 Imagen, 그리고 Noble AI 등의 텍스트에 해당하는 이미지를 생성하는 인공지능의 성능이 우수해지고 있다. 특히 기존 방법론들과 DALL-E-2의 결과를 정성적으로 비교한 결과, MS-COCO 데이터에 대해 DALL-E-2 모델은 FID 점수에 있어 기존의 최고 모델의 2point 이내로 차이나는 좋은 성적을 거두었다. 또한 모델에 대한 humanevaluation 결과 MS-COCO 데이터에 대해 다섯 개 중 가장 현실적인 이미지를 고르도록 했을 때 DALL-E-2는 90%의 확률로 선택되었고, 캡션과 가장 매칭 되는 이미지를 고르도록 했을 때 93.3% 선택되었다[12]. 그러므로 본 연구에서는 이미지 재생성을 위해 DALL-E-2를 선정하였다.

DALL-E-2를 통해 새로운 이미지를 생성하여 활용한 기존 연구들은 다음과 같다. 문제 도메인에 대한 데이터 집합을 생성하기 위해 텍스트 이미지 생성 프로세스를 적용하고, 객체 범주 및 자연 장면으로 구성된 짧은 텍스트 입력을 제공하는 합성 이미지를 생성하여 문제 도메인 데이터 집합의 확장에 활용되었다[14]. 또한 text-to-image 합성을 위한 최첨단 방법에 대한 연구를 수행하고, 이러한 방법을 평가하기 위한 프레임워크를 제안하였다[15]. 이렇듯 선행된 연구들에서 minGPT나 DALL-E-2 등의 모델을 이용하여 text-to-image 방식으로 새로운 이미지를 생성하지만, 이렇게 생성된 이미지를 객체 인식의 사례 별 연구로 확장한 경우는 없었음을 확인할 수 있다.

*: Authors are equally contributed for this paper

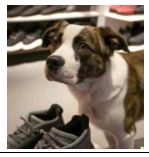
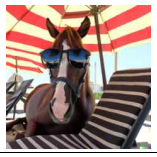
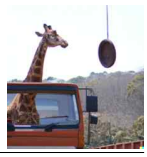
카테고리	cat	dog	horse	sheep	cow	elephant	bear	giraffe
문구	A photo of a cat putting on lipstick in front of the mirror	A photo of a dog tried on shoes at a shoe store	A photo of a horse wearing sunglasses and lying on a sunbed under a paraso	A photo of a sheep truckload of apples and oranges with a salad and wine	A photo of a cow that opens the refrigerator and takes out a sandwich	A photo of a Elephant looking at the clock while putting flowers in the pot	A photo of a bear sitting on a chair and eating a sandwich	A photo of a giraffe that throws a discus in a truck
이미지								

표 1: 카테고리별 DALL-E-2에 넣은 문구와 해당 문구에 의해 생성된 샘플.

따라서 본 연구에서는 DALL-E-2 모델을 이용하여 새로운 데이터 집합을 생성하고 이렇게 생성된 데이터 집합을 객체 인식에 활용하여 각 사례별 객체 인식률을 점검하고자 한다.

2. 사례 연구

본 논문에서는 DALL-E-2를 이용하여 텍스트를 새로운 이미지로 변환하고 카테고리별로 이미지를 분류하여 YOLOv5 모델로 객체 인식을 수행한다. 그러므로 이 과정을 이미지를 생성하고 객체 인식을 테스트하는 2가지 파트로 나누어 설명한다. 그림 1은 본 연구의 구조도를 나타낸다.

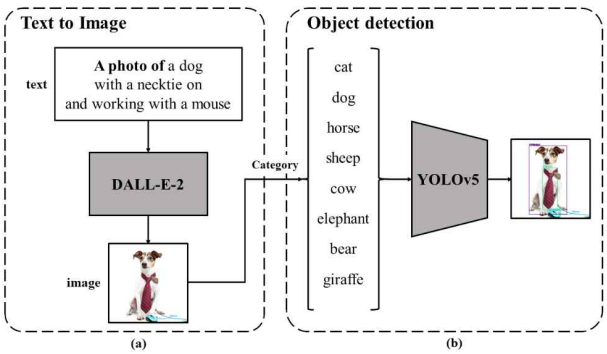


그림 1: 본 연구의 구조도. (a) text to image 파트를 설명한다. (b) 객체 인식 파트를 설명한다.

2.1 이미지 생성

DALL-E-2를 활용하여 생성된 이미지의 데이터 집합을 생성하기 위해 입력할 문구를 선행 제작하였다. 해당 문구는 COCO val 2017 데이터 집합에 속한 동물 객체 하나를 위주로 객체의 동작과 주변사물을 함께 포함하여 구성하였다.

적대적 공격 사례에 활용되는 이미지들은 현실에 존재하는 이미지를 생성한 것이다. 그러므로 DALL-E-2를 활용하여 이미지를 생성할 때 작화 형식이 아닌 실제 현실과 유사한 사진 형식으로 이미지를 생성했다. 'in a photorealistic style'을 문구의 뒤에 삽입하는 방식으로 이미지를 생성했을 때 작화 형식의 이미지가 다수 포함된다. 따라서 우리는 이미지 생성 시 'A photo of'를 문구의 앞에 삽입하여 작화 형식이 아닌 사진 형식의 데이터 집합을 생성하였다.

그림 2: 같은 내용의 문구에 'in a photorealisitc style'과 'A photo of' 를 다르게 하여 생성한 이미지 비교

해당 과정으로 문구 61개를 생성한 후 DALL-E-2에 입력해 이미지를 생성했다. 그 중 사진 형식으로, 문구의 내용을 크게 벗어나지 않게 생성된 데이터로 필터링하여 먼저 234개의 이미지를 도출하였다.

카테고리는 문구 생성시 주로 고려했던 COCO val 2017 데이터 집합의 동물 객체들 중 8가지로 선정하였다. 본래 COCO val 2017 데이터 집합에 속한 동물 객체는 bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe로 총 10가지이다. 연구의 정확도를 높이고자 카테고리 분류 시 각 카테고리별 데이터 집합의 개수가 10개 이하인 bird, zebra 카테고리를 제외하였다. 총 8개의 카테고리를 주요 객체로 해서 구성한 문구로 생성한 이미지를 데이터 집합으로 선정했다. 그 결과 총 185개의 최종 결과 이미지 데이터 집합을 도출하였다. 이후 엑셀로 각 이미지 데이터에 해당 생성 문구를 캡션으로 달고, 카테고리별로 이미지를 분류했다. 8개의 카테고리에 대해 각각 cat 37개, dog 52개, horse 12개, sheep 10개, cow 19개, elephant 12개, bear 38개, giraffe 12개로 총 192개의 객체를 도출하였다.

2.2 객체 인식률 점검

선행 학습된 YOLOv5x 모델을 활용하여 총 8가지의 카테고리에 대해 객체 인식을 수행하였다. mAP(mean average precision)는 각 객체 카테고리당 구한 AP(average precision)를 합하여 객체 카테고리의 총 개수로 나누는 성능 지표이다. YOLOv5x 모델은 다른 학습된 모델들인 YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l보다 mAP 값이 높으므로 이를 선택하였다[18].

카테고리	이미지 개수(개)	맞춘 개수(개)	바운딩 박스 확률(%)	정확도 (%)
cat	37	27	83.78	72.97
dog	52	40	90.38	76.92
horse	12	9	91.67	75
sheep	10	6	70	60
cow	19	11	68.42	57.89
elephant	12	8	100	66.67
bear	38	6	**58.93**	**15.79**
giraffe	12	9	100	75

표 2: 카테고리에 대한 실험 결과. 가장 낮은 정확도와 가장 낮은 바운딩 박스 확률을 굵게 표시함.

표 2에서는 DALL-E-2를 사용하여 생성한 객체 별 192개의 이미지들을 8개의 동물 카테고리로 나누어 분석한 결과를 보여준다. 표 2에서 맞춘 개수는 각각의 동물 객체에 바운딩 박스를 올바르게 표시하고 정확히 맞춘 개수를 나타내며, 바운딩 박스 확률은 동물 객체에 바운딩 박스를 올바르게 처리한 것을 모두 구한 확률이다. 또한 정확도는 바운딩 박스를 동물 객체에 올바르게 표시하고 동물을 정확히 맞춘 것만을 구한 확률이다.

전체 192개의 이미지 중 정확하게 바운딩 박스를 표시한 이미지는 116개로, 60.41%의 확률로 정확하게 바운딩 박스를 표시했다. 모든 카테고리 중 특히 Bear 카테고리는, 바운딩 박스 확률 58.93%, 정확도 15.79%로 제일 낮은 정확도를 보였다.

본 연구를 통해, 생성된 이미지에 대한 정확도는 60.41%로, 이는 일반 이미지에 비해 낮은 정확도를 보인다. 이를 통해 YOLOv5x가 생성된 이미지에 대한 객체 인식에 취약함을 보여준다.

정확도를 낮추며 객체를 오인식하는 케이스는 1. 주요 객체를 COCO val 2017 데이터 집합 내의 다른 객체로 인식하는 경우 2. 객체를 인식하지 못하는 경우로 나눌 수 있는데, 특히 본 연구에서 주목할 점은 1. 주요 객체를 COCO val 2017 데이터 집합 내의 다른 객체로 인식하는 경우 중, '동물 카테고리의 객체를 사람으로 인식하는 점'이다. DALL-E-2를 이용한 이미지 생성 결과, 육안으로 봤을 때에는 명확히 카테고리에 해당하는 이미지임에도, 어떤 행동을 하는지나, 어떤 사물과 함께 인식되는지의 여부에 따라 YOLOv5x 모델은 사람으로 인식하는 경우가 매우 빈번했다.

Bear 카테고리에 대한 인식률이 낮은 이유는 1. Bear를 Teddy bear로 인식하는 점 2. Bear를 Person이나 다른 객체로 인식하는 점으로 파악할 수 있다. 특히 1.에서 COCO val 2017 데이터 집합 내에 있는 카테고리인 Teddy bear로 인식하는 오류를 통해서 객체가 이미지화 되어도 해당 객체로 인식한 것과 차이점을 고려할 수 있다. 하지만 이러한 고려 사항에도 Bear를 Person이나 Dog를 위주로 다른 객체로 빈번하게 인식한 결과, 타 객체에 비해 확연히 낮은 정확도를 보였다.

그림 3의 왼쪽 그림은 DALL-E-2에서 'a photo of cow playing skateboard with dog'을 입력하여 생성한 이미지이다. YOLOv5x 모델은 COCO val 2017 데이터 집합에 있는 소와 개에 대해 학습하였고, 테스트 결과 위 이미지의 주요 객체를 모두 잘 분석하였음을 보여준다.

문구를 제작할 때에는 주요 객체가 하나 있는 문장을 위주로 생성하였는데, 위와 같이 주요 객체가 두 개 있는 문장도 생성해 어떻게 동작하는지를 살펴보고자 하였다. 동물 객체가 두 개 있는 이미지 12개를 생성하였고, 위와 같이 YOLOv5x 모델에서 타 이미지들과 같이 큰 특이 사항 없이 인식되는 모습을 확인했다.

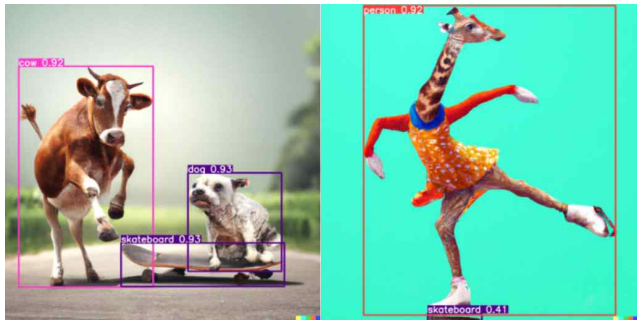


그림 3: YOLOv5 모델을 통해 객체 인식이 잘 된 이미지와 잘못 인식된 이미지 예시

그림 3의 오른쪽 그림은 DALL-E-2에서 'a photo of a bear sitting on a chair and eating a sandwich'를 입력하여 생성한 이미지이다. YOLO5x 모델은 곰에 대해 학습하였지만 테스트 결과 위 이미지를 개로 인식하였음을 보여준다.

3. 결론

기존의 현실 데이터를 기반으로 하여 객체를 인식하고 인식률을 높이기 위한 연구는 활발히 진행되고 있지만 text-to-image 모델을 사용하여 생성된 이미지에 대한 객체 인식 연구는 활발히 진행되고 있지 않다. 본 연구에서는 DALL-E-2를 사용하여 이미지를 생성하였고, COCO val 2017 데이터 집합으로 선행 학습된 YOLOv5x 모델을 활용하여 생성한 이미지들에 대한 객체 인식 사례 연구를 진행하였다.

실험 결과 총 8개 동물에 대한 이미지 185개를 생성하였다. 그 안에서 객체 192개 중 166개의 동물을 바운딩 박스로 감지하여 86.46% 확률로 바운딩 박스를 만들었으나 그 중 116개의 동물만을 정확히 맞추어서 60.41% 정답률을 보여준다. 그 결과 생성된 이미지에 대한 객체 인식률이 일반 이미지에 대한 객체 인식률보다 낮은 것을 알 수 있다.

본 논문은 실험 결과를 통해 생성된 이미지에 대한 객체 인식률이 일반 이미지에 대한 객체 인식률보다 저조함을 보여준다. 선행 학습된 모델을 사용하여 바운딩 박스 카운트를 활용한 분석을 하였기에 mAP 지수를 분석하지 못하였다는 한계가 있다. 하지만 생성된 객체에 대한 인식률을 높이기 위한 추가 연구가 필요함을 제시한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학 사업의 연구 결과로 수행되었음(2021-0-01082). 또한, 본 연구는 2020년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2020R1I1A3072227).

참고 자료 (Reference)

[1]: 권현, 김용철. (2021). 딥러닝 모델에 대한 적대적 사례 기술 동향.

정보보호학회지, 31(2), 5-12.

[2]: Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[3]: Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574-2582).

[4]: Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372-387). IEEE.

[5]: Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy (sp)* (pp. 39-57). Ieee.

[6]: Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, *23*(5), 828-841.

[7]: He, W., Li, B., & Song, D. (2018, February). Decision boundary analysis of adversarial examples. In *International Conference on Learning Representations*.

[8]: Athalye, A., Carlini, N., & Wagner, D. (2018, July). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning* (pp. 274-283). PMLR.

[9]: Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, *33*, 1633-1645.

[10]: Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, *6*, 14410-14430.

[11]: Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625-1634).

[12]: Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821-8831). PMLR.

[14]: Liang, S., Liu, A., Liang, J., Li, L., Bai, Y., & Cao, X. (2022, October). Imitated Detectors: Stealing Knowledge of Black-box Object Detectors. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4839-4847).

[15]: Dinh, T. M., Nguyen, R., & Hua, B. S. (2022). TISE: Bag of Metrics for Text-to-Image Synthesis Evaluation. In *European Conference on Computer Vision* (pp. 594-609). Springer, Cham.

[16]: Nepal, U., & Eslamiat, H. (2022). Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. *Sensors*, *22*(2), 464.

[17]: https://github.com/ultralytics/yolov5

[18]: Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125.

[19]: https://openai.com/dall-e-2/