Swin Transformer를 이용한 항공사진에서 다중클래스 차량 검출

이기춘* · 정유석 · 이창우

군산대학교

The Detection of Multi-class Vehicles using Swin Transformer

Ki-chun Lee* · Yu-seok Jeong · Chang-woo Lee

Kunsan National University

E-mail : accforaus@kunsan.ac.kr

요 약

도시 상태를 탐지하기 위해서는 운송 수단 수, 교통 흐름등이 필수적으로 파악되어야 할 요소이다. 본 논문에서는 기존의 Mask R-CNN을 이용하여 다양한 차량의 형태를 학습하고, 드론으로 촬영한 도시 항공 영상에서 특정 유형의 차량 들을 검출하는 시스템을 오늘날 NLP 분야에서 널리 쓰이게 된 Transformer 모델을 컴퓨터 비전 문제에 도입하여 기존의 컨볼루션 신경망보다 높은 성능을 보여준 Swin Transformer 모델을 이용하여 기존의 연구에서 보여주었던 검출 시스템 능력을 향상시켰다.

ABSTARCT

In order to detect urban conditions, the number of means of transportation and traffic flow are essential factors to be identified. This paper improved the detection system capabilities shown in previous studies using the SwinTransformer model, which showed higher performance than existing convolutional neural networks, by learning various vehicle types using existing Mask R-CNN and introducing today's widely used transformer model to detect certain types of vehicles in urban aerial images.

키워드

artificial intelligence, computer vision, object detection, instance segmentation

Ⅰ. 서 론

컴퓨터 비전의 모델링은 오랫동안 컨볼루션 신경망(CNN)에 의해 지배되었다. AlexNet[1]과 ImageNet 이미지 분류 문제에 대한 혁신적인 성능을 시작으로, CNN 아키텍쳐는 더 큰 규모[2, 3], 더 광범위한 연결[4] 및 더 정교한 형태의 CNN[5, 6, 7]을 통해 점점 더 강력해지도록 발전해 왔다. CNN이 다양한 컴퓨터 비전 작업의 백본 네트워크 역할을 함에 따라 이러한 CNN 아키텍쳐의 발전은 전체 분야를 광범위하게 끌어올린 성능 향상으로 이어졌다. 반면에, 자연어 처리(NLP)에서 모델 아키텍쳐의 진화는 다른 방향으로 발전하였으며, 오늘날 널리 사용되는 아키텍처는 Transformer[8] 모델이다. 시퀀스 모델링 및 변환 작업을 위해 설계된 Transformer는 언어의 long-range 종속성 문제를 해결하기 위해 Attention 기법을 적용하여 해결한 것으로 유명하다. NLP 분야에서 엄청난 성공을 거두어 연구원들은 컴퓨터 비전에 사용할 수 있도록 연구하였으며, 최근에는 이미지 분류[9] 및 다양한 컴퓨터 비전 문제에 도입하게 되었다. 본 논문에서는 최근 CNN을 대체할 컴퓨터 비전 문제에서의 큰 가능성을 보여준 Swin Transformer[10]를 이용해 기존의 Mask R-CNN[11]을 이용한 다중 차량 클래스 검출 연구[12]의 성능을 더욱 향상 시키는 시스템을 제안한다.

* corresponding author

II. 데이터 셋 전처리 및 레이블링

2.1 데이터 셋 전처리

획득한 항공 영상은 시스템에서의 원활한 학습을 위해 각 영상을 리사이징 다음과 같이 가로 길이를 1333으로 고정한 후 다양한 세로 길이(480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800)을 적용하여 여러 해상도에 따라 검출이 잘 되도록 리사이징을 진행하고, 데이터 증강을 통해 데이터 셋의 과적합을 방지한다. 이 중 전체 데이터셋 중 75장은 훈련 데이터 셋, 75장은 검증 데이터 셋, 50장은 실험 데이터 셋으로 사용한다.

2.2 COCO 데이터 셋

Swin Transformer 모델을 이용하여 학습을 진행하기 위해서는 기존의 레이블링된 데이터 파일 포맷(VGG[13])에서 MS COCO[14] 데이터 셋 으로 변환과정이 필요하다. COCO 포맷은 이미지의 레이블링 데이터와 어노테이션 정보가 저장된다. COCO 포맷의 Info 영역은 데이터 셋의 정보를 담고 있고, Licenses 부분은 이미지들의 라이센스 정보를 담고 있다. 또한 Categories는 이미지의 유형, 본 논문에서는 V01, V02, V03, V04 정보가 들어가게 된다. Annotations 부분은 각 이미지별 객체 검출과 사물 영역의 정보를 담고 있다.

하나의 COCO 데이터셋의 JSON파일엔 다음과 같은 형식으로 제공된다.

```
{
        “info”: info,
        “licenses”: [licenses],
        “categories”: [categories],
        “images”: [images],
        “annotations”: [annotations]
}
```

III. 실험 및 검출

3.1 학습 방법

차량 검출에 대한 Swin Transformer 모델은 학습은 라벨링 장업을 통해 얻어진 COCO 데이터 셋을 이용하여 차량 유형에 대한 가중치를 가진 모델 파일(.pth)을 생성한다. 학습 환경은 Ubuntu 20.04.2 운영체제, CPU AMD Ryzen 9 3900X, RAM 64GB GPU NVIDIA GeForce RTX 2080Tix2를 사용하였다. 가중치 학습을 위해 EPOCH값은 12으로 설정 하고 SPE(Step Per Epoch)값은 132로 설정 후 학습을 수행하였다.

3.2 학습모델 결과 분석

그림 2는 기존 논문의 모델과 본 논문의 학습된 모델을 비교한 결과이다. 그림 2와 같이 기존의 Mask R-CNN(위) 검출 시스템을 이용한 검출 결과와 Swin Transformer(아래) 검출 시스템을 이용한 검출 결과로 구분된다. 기존의 연구와의 차이점은 기존의 연구에서 검출하지 못하는 차량을 검출했다는 것을 보았다. 나무에 가려지거나 주변 차량과 거리가 좁은 차량들에 대해서는 검출 능력이 보다 우수한 것을 확인할 수 있었다.

그림 1. 가중치 모델을 이용항 검출 결과
(위) 기존 Mask R-CNN (아래) Swin Transformer
Fig. 1. Final output of model
(Top) Mask R-CNN (Bottom) Swin Transformer

IV. 결 론

본 논문에서는 기존에 진행했던 다중클래스 차량 검출 연구의 항공사진에서 다중클래스 차량 검출 시스템의 성능을 향상하고자 기존의 Mask R-CNN모델 에서 Swin Transformer모델을 이용하여 항공 상에서의 차량을 검출하고 마스크를 씌워 표시하는 모델의 성능을 개선한 연구를 수행하였다. 기존의 연구와 다르게 본 연구에 사용된 모델은 한 객체에 대하여 다중 레이블이 검출 된다는 문제가 발견되었다. 추후 문제점을 보완하고, 데이터 셋의 크기를 증가하여 문제를 해결하고, 모델의 성능을 더욱 향상 시키고자 한다.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton.Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097−1105, 2012. 1, 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770−778, 2016. 1, 2, 4

[3] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In BMVC, 2016. 1

[4] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4700−4708, 2017. 1, 2

[5] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1492−1500, 2017. 1, 2, 3

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 764−773, 2017. 1, 3

[7] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9308−9316, 2019. 1, 3

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998−6008, 2017. 1, 2, 4

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, irk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, ostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021. 1, 2, 3, 4, 5, 6, 9

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021. 1, 2, 4

[11] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, Z. Zhang, (2019). MMDetection: Open MMLab Detection Toolbox and Benchmark. arXiv preprint arXiv:1906.07155.

[12] 윤형진, 이민혜, 조정원, 정유석, 이창우 “다중 클래스 차량 검출에 관한 연구”

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, May 2015. 2, 4

[14] 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 European conference on computer vision, pages 740 755. Springer, 2014. 5