편광 셀프어텐션의 공간정보 강조 모듈을 결합한 HRNet 모델 설계 및 구현

김진성[1], 박준[1], 정세훈[2], 심춘보[1]
[1]국립순천대학교 IT-Bio 융합시스템전공
[2]국립순천대학교 컴퓨터공학과
k456kille@naver.com, todnehfdl@naver.com, shjung@scnu.ac.kr, cbsim@sunchon.ac.kr

Design and Implementation of HRNet Model Combined with Spatial Information Attention Module of Polarized Self-attention

Jin-Seong Kim[1], Jun Park[1], Se-Hoon Jung[2], Chun-Bo Sim[1]
[1]Interdisciplinary Program in IT-Bio Convergence System, Sunchon National University
[2]Department of Computer Engineering, Sunchon National University

요 약

컴퓨터 비전의 하위 태스크(Task)인 의미론적 분할(Semantic Segmentation)은 자율주행, 해상에서 선박찾기 등 다양한 분야에서 연구되고 있다. 기존 FCN(Fully Conovlutional Networks) 기반 의미론적 분할 모델은 다운샘플링(Dowsnsampling)과정에서 공간정보의 손실이 발생하여 정확도가 하락했다. 본 논문에서는 공간정보 손실을 완화하고자 PSA(Polarized Self-attention)의 공간정보 강조 모듈을 HRNet(High-resolution Networks)의 합성곱 블록 사이에 추가한다. 실험결과 파라미터는 3.1M, GFLOPs는 3.2G 증가했으나 mIoU는 0.26% 증가했다. 공간정보가 의미론적 분할 정확도에 영향이 미치는 것을 확인했다.

1. 서론

컴퓨터 비전(Computer Vision)의 하위 태스크(Task)인 의미론적 분할(Semantic Segmentation)은 자율주행, 해상에서 선박 찾기 등의 다양한 분야에서 연구되고 있다. 기존 FCN(Fully Convolutional Networks)[1] 기반 의미론적 분할 모델은 특징을 추출하는 다운샘플링(Downsampling) 과정에서 피라미드 형태의 특성상 특징맵의 해상도 감소문제로 공간정보 손실이 발생한다. 공간정보는 픽셀의 위치를 파악하기 위해 중요한 역할을 하므로 의미론적 분할에서 매우 중요하다.

본 논문에서는 다운샘플링 과정에서 손실되는 공간정보를 보존하기 위해 편광(Polarization) 기법과 셀프어텐션(Self-attention)[2]기법이 적용된 PSA(Polarized Self-attention)[3]의 공간정보 강조 모듈을 사용한다. 공간정보 강조 모듈은 합성곱 블록(Convolution Block) 중간에 추가한다. 제안하는 방법은 높은 내부 해상도를 유지하며 각 블록 단위로 높은 비선형성을 부여해 신경망 층이 깊어짐에 따라 발생하는 기울기 손실문제를 줄인다.

2. 관련연구

2.1 공간정보

피라미드 형태를 띠는 FCN 기반 의미론적 분할 모델의 특성상 필연적으로 공간정보 손실이 발생한다. 공간정보는 픽셀의 정확한 위치를 분할하기 위해서 중요한 역할을 한다. 고해상도는 풍부한 공간정보를 포함하고 있어 높은 해상도를 유지 시켜주는 것이 중요하다.

HRNet[4]은 이러한 문제를 해결하기 위해 고해상도를 포함하는 평행한 4개의 다중 해상도 라인을 가진 구조를 사용해 고해상도가 분할 성능에 유의미한 영향을 미치는 것을 입증했다.

2.2 어텐션 모듈

어텐션(Attention) 기법은 자연어 처리에서 주로 사용됐다. 하지만, 어텐션의 개념이 셀프어텐션[2]으로 확장되면서 SENet(Squeeze-and-Excitation Networks)[5], CBAM(Convolutional Block Attention Module)[6]과 같이 이미지 처리 분야에서도 널리 사용됐다. 셀프어텐션은 기존의 어텐션

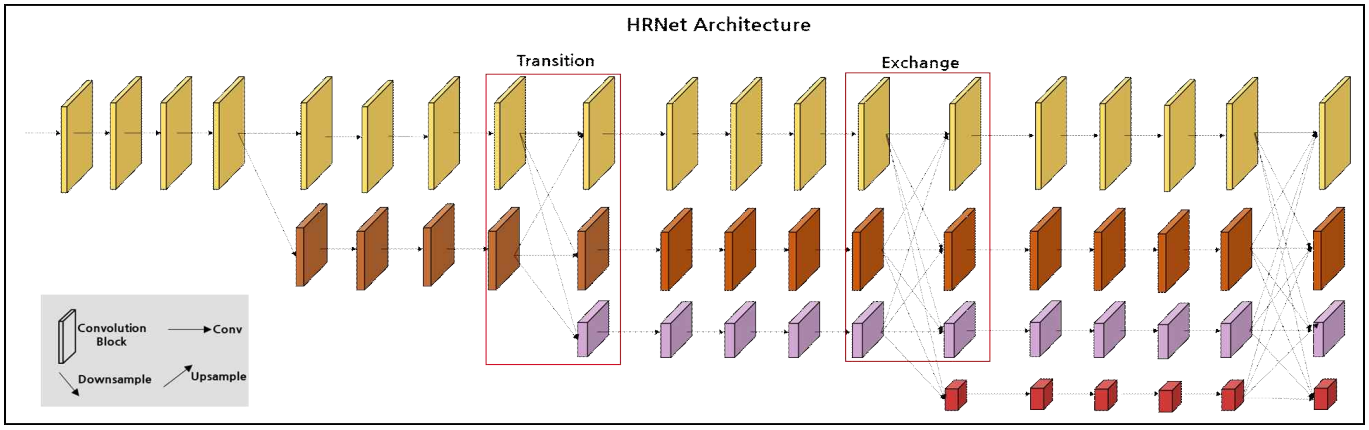

그림 1 HRNet 구조

기법과 유사하나 차이점은 입력되는 값이 모두 동일하다는 점이다. 셀프어텐션은 입력된 특징맵내에서 픽셀 간 혹은 채널 간의 관계성을 재조정하는 데 유용하다.

하지만 기존의 어텐션 모듈은 의미론적 분할을 상정하지 않아 비선형성을 제대로 반영하지 못하는 문제점이 있다. PSA[3]는 사진학에서 모티브를 얻은 편광 기법과 셀프어텐션을 사용하여 높은 내부 해상도를 유지하고 비선형성을 반영하여 기존의 문제점을 완화했다. 편광 기법은 어텐션 연산에서 채널과 공간 특징 중 하나를 붕괴시켜 채널을 공간으로 공간을 채널로 편입시키는 방법이다.

3. 제안하는 방법

HRNet은 기존의 구조와 동일하게 사용한다. 1단계에서 병목(Bottle Neck) 구조를 사용하고 2, 3, 4 단계는 일반적인 합성곱 블록 구조를 사용한다. 공간정보 강조 모듈은 병목 구조에는 사용하지 않는다. HRNet의 구조는 그림 1과 같다. 기존 HRNet은 다중 해상도를 이용해 해상도에 따른 서로 다른 정보를 효율적으로 교환하고 융합해 공간정보 손실을 줄인다.

제안하는 방법은 공간정보의 손실을 보다 줄이고자 합성곱 블록 사이에 공간정보 강조 모듈을 추가하여 높은 내부 해상도와 비선형성을 반영하여 공간정보의 손실을 줄인다. PSA 모듈을 그대로 사용하지 않는 이유는 파라미터 수의 증가를 줄이고 픽셀의 클래스 분류 성능보다 위치를 분할하는 성능에 집중하기 위해서다.

구체적으로 셀프어텐션 기법을 활용해 한 번의 합성곱을 거친 특징맵을 Key, Query, Value로 입력한다. Query를 GAP(Global Average Pooling)와 소프트맥스(Softmax)를 통해 어텐션맵을 만들고 Value와 곱한 다음 시그모이드(Sigmoid)를 거쳐 중요한 정보를 강조한다. 마지막으로 Key와 곱해 공간정보를 강조한다. 그림 2는 공간정보 강조 모듈의 구조다.

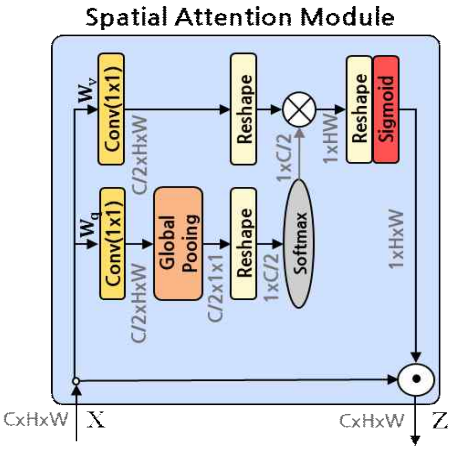

그림 2 공간정보 강조 모듈 구조

4. 실험환경 및 성능평가

4.1 하드웨어 및 소프트웨어 사양

하드웨어 사양의 경우 CPU는 인텔 i7 7700k, 그래픽 카드는 NVIDIA Geforce RTX 3090, 램은 DDR4 32GB를 이용한다. 실험에 이용할 소프트웨어 버전(Version)은 표 1과 같다.

표 1. 소프트웨어 버전

구분	세부 사항
Operating System	Ubuntu 20.04.1 LTS
CUDA	11.2.67
cuDNN	8.1.0
Programming Language	Python 3.8.10
Deep Learning Framework	Pytorch 1.8.1

4.2 데이터 세트 구성

PASCAL Context[7] 데이터 세트는 20개의 객체 혹은 물체(Stuff) 클래스와 배경 클래스 1개를 포함하여 400개 이상의 클래스를 포함하지만, 클래스별 이미지 수가 적어 일반적으로 59개 클래스 혹은 배경 클래스를 포함한 60개 클래스만 사용한다. 클래스의 범주는 객체, 물체, 객체와 물체가 섞인 하이브리드 범주로 나뉜다. 총 10.103개 이미지로 구성됐으며, 4998개의 훈련 이미지, 5,105개의 테스트 이미지로 나뉜다. 본 논문에서는 객체와 물체 분할 성능을 평가하기 위해 59개 클래스만 사용한다.

4.3 성능평가

배치 사이즈(Batch Size)는 11로 설정하고 단일 GPU를 사용한다. 옵티마이저(Optimizer)는 SGD(Stochastic Gradient Descent)에 모멘텀(Momentum)을 0.9 추가하여 사용한다. 학습률(Learning Rate)은 0.004, 웨이트 디케이(Weight Decay)는 0.0001이다. 학습 스케쥴러(Scheduler)는 폴리 학습률(Poly Learning Rate) 정책을 사용한다. 성능평가 결과는 표 2와 같다.

표 2. PASCAL Context 59 클래스 의미론적 분할 결과

Method	#Params[M]	GFLOPs[G]	mIoU[%]
HRNetV2-W48	65.9	76.870	50.21
Proposed Method	69.0	80.084	**50.47**

표 2에서 파라미터 수는 기존 모델보다 3.1M, GFLOPs(Giga Floating Operations Per Second)는 3.2G 증가하여 학습 시간이 기존보다 4시간 더 소요됐다. mIoU(mean Intersection over Union)는 0.26% 증가한 것으로 보아 공간정보 강조가 객체와 물체를 분할하는 데 영향이 미치는 것을 확인했다.

5. 결론

본 논문에서는 기존 HRNet의 합성곱 블록에 공간정보 강조 모듈을 추가한 의미론적 분할 모델의 성능을 실험하고 비교한다. 실험결과 파라미터는 3.1M, GFLOPs는 3.2G 증가했으나 mIoU는 0.26% 증가했다. 모델의 복잡도가 과도하게 커지지 않는 선에서 성능이 향상됐다는 점과 공간정보가 의미론적 분할 성능에 영향을 준다는 것을 확인했다는 점에서 의미가 있다고 판단된다.

사사문구

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2023-2020-0-01489) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation). This work was supported by the BK21 plus program through the National Research Foundation (NRF) funded by the Ministry of Education of Korea(5199990214660).

참고문헌

[1] J. Long, E. Shelhamer, and T. Darrell “Fully Convolutional Networks for Semantic Segmentation,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 3431-3440, 2015.

[2] A. Vaswani, N. Shazeer, M. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances Neural Information Processing Systems(NIPS)*, pp. 5998-6008, 2017.

[3] H. Liu, F. Liu, X. Fan, and D. Huang, “Polarized Self-attention: Towards high-quality pixel-wise regression,” *arXiv preprint*, arXiv:2107.00782, 2021.

[4] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, et al., “High-resolution representations for labeling pixels and regions,” *arXiv preprint*, arXiv:1904.04514. 2019.

[5] Jie Hu, Li Shen and Gang sun, "Squeeze-and-Excitation Networks", in *Proceedings of the IEEE conference on computer vision and pattern recognition 2018*, pp. 7132-7141, 2018.

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee and In So Kweon, "CBAM: Convolutional Block Attention Module", in *Proceedings of the Eurpean conference on computer vision(ECCV) 2018*, pp. 3-19, 2018.

[7] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, et al., “The role of context for object detection and semantic segmentation in the wild,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 891-898, 2014.