Journal of KIIT. Vol. 18, No. 11, pp. 89-96, Nov. 30, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.11.89

변형 WASPP를 이용한 Segmentation 성능 개선 연구

김종식*, 강대성**

A Study on the Improvement of Segmentation Performance using Modified WASPP

Jong-Sik Kim*, Dae-Seong Kang**

이 논문은 2017 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2017R1D1A1B04030870)

요 약

본 논문에서는 DeepLab V3 알고리즘의 핵심인 ASPP(Atrous Spatial Pyramid Pooling)를 대신하기 위해 선행 연구한 WASPP(Wavelet Atrous Spatial Pyramid Pooling)의 mIOU 결과는 epoch=100에서 90.2%로 기존 ASPP 대비 유사하거나 약간 낮아서 성능 향상에는 미흡한 결과가 나왔다. 이를 개선하기 위해 웨이블릿 특성을 활용하면서 기존 ASPP보다 Semantic segmentation 객체 검출 성능 개선에 대해 검토하였다. 그 결과 기존 ASPP 대비 Loss 및 mIOU 부분에서 1%이상의 성능 향상이 이루어졌으며, image_6 (Monitor prediction) 영상에서는 DeepLab v3는 epoch=200 기준으로 여전히 인식 오류가 존재하나 변형 WASPP에서는 epoch=50에서 완벽히 개선되었다. 이것은 특정 영상에서 기존보다 200%이상 성능이 개선된 결과이다.

Abstract

In this paper, the WASPP (Wavelet Atrous Spacial Pyramid Pooling), which was previously studied on behalf of ASPP (Atrous Spacial Pyramid Pooling), which is the core of DeepLab V3 algorithm, was similar or slightly lower than the existing ASPP, with mIOU from epoch=100 to 90.2%, and the results were insufficient to improve performance. To improve this, the performance improvement of semantic segmentation object detection was reviewed over the existing ASPP while utilizing wavelet characteristics. As a result, more than 1% performance improvement has been achieved in loss and mIOU compared to the existing ASPP, and in image_6 (Monitor prediction) photographs, DeepLab v3 still has recognition errors based on epoch=200, but it has been completely improved in epoch=50 in modified WASPP. This is the result of a performance improvement of more than 200% in a particular image.

Keywords
convolution, deep learning, semantic segmentation, ASPP

* 동아대학교 전자공학과 박사과정
- ORCID: https://orcid.org/0000-0002-1459-1943
** 동아대학교 전자공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Aug. 08, 2020, Revised: Sep. 15, 2020, Accepted: Sep. 18, 2020
· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University, 37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. 서 론

시멘틱 세그멘테이션(Semantic segmentation)은 컴퓨터비전 분야에서 가장 중요한 분야 중에 하나다. 단순히 사진을 보고 분류하는 것이 아니라, 장면을 완벽하게 분석해야하는 고차원적인 문제이다. 시멘틱 세그멘테이션은 이미지 내에 있는 물체들을 의미 있는 단위로 객체 분할하는 것이다. 좀 더 구체적으로 설명하면, 이미지의 각 픽셀이 어느 클래스에 속하는지 예측하는 것이다. 이렇게 이미지 내 모든 객체에 대해서 예측을 진행하기 때문에 이 기술을 Dense prediction이라고 부르기도 한다. 어떤 영상 이미지에는 사람, 자동차, 자전거, 고양이, 컴퓨터, 비행기 등 여러 종류의 물체가 포함되어 있을 수 있다. 이렇게 서로 다른 종류의 객체들을 깔끔하게 분할해내는 것이 시멘틱 세그멘테이션이다[1].

시멘틱 세그멘테이션의 성능을 높이기 위한 방법 중 하나로, Spatial pyramid pooling 기법이 많이 활용되고 있는 추세다. 대표적인 알고리즘은 DeepLab 시리즈의 ASPP(Atrous Spatial Pyramid Pooling) 기법이다 [2-4]. 특징맵으로 부터 여러 개의 rate가 다른 Atrous 컨볼류션(Convolution)을 병렬로 적용한 뒤, 이를 다시 연결(Concatenate) 하는 방식이다[3]-[5].

그림 1은 DeepLab v3의 전체 구조이다. ASPP의 단점은 중요한 특징점을 잘 활용하지 못하고 무작위적 특징점 추출에 있다. 제안된 웨이버릿(Wavelet) pooling 방식을 활용하면 다양한 특징점들의 평균을 활용하여 Spatial pyramid pooling 기법의 활용 폭을 확대할 수 있다[6]. 이번에 제안된 변형 WASPP 방법을 활용하여 시멘틱 세그멘테이션의 분리성능과 객체 추론의 성능 향상에 관해 연구하였다.

Ⅱ. 관련 이론

2.1 DeepLab V3의 ASPP

ASPP는 DeepLab V3에서 소개된 Pooling 방법이다. 기존의 Spatial pyramid pooling에서, 각 컨볼류션을 Atrous 컨볼류션으로 바꾼 형태이다.

Spatial pyramid는 그림 2에서 보는 것과 같이 여러 Grid scale에서 Pooling을 진행해서 특징맵을 모두 연결시키는 형태이다[3]. 그림 2에는 DeepLab의 ASPP의 구조를 상세히 표현하였다.

ASPP는 여러 개의 확장 비율(rate=6, 8, 16)을 사용해서 컨볼류션을 한 후 마지막에 연결시킨다. 확장 계수를 6부터 18까지 변화시키므로 다양한 Receptive field를 볼 수 있다. 이렇게 하면 연산 효율적 측면에서 큰 이득을 얻을 수 있다. Atrous 컨볼류션은 기존 컨볼류션 레이어에 대해 다양한 확장 계수에 따라 필터에 0을 삽입함으로써 필터의 수용적 영역을 유연하게 확장할 수 있다[7]. 나아가 ASPP 구조는 Atrous 컨볼류션 레이어를 병렬로 사용하며, 그 산출물을 결합하여 다양한 수용 분야와 추출한 정보를 연결하는 형태이다[4].

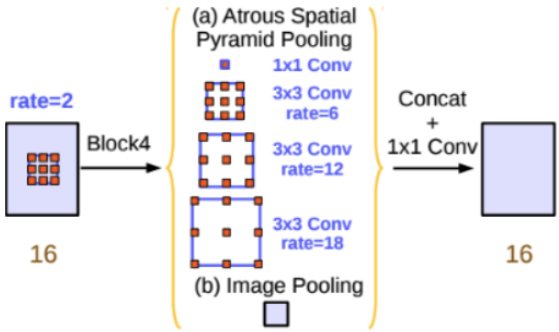

그림 2. ASPP (Atrous + Spatial pyramid pooling)
Fig. 2. ASPP (Atrous + Spatial pyramid pooling)

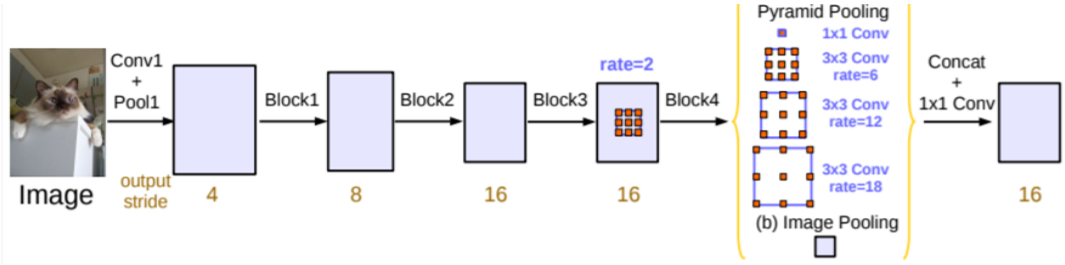

그림 1. DeepLab V3 구조 및 ASPP
Fig. 1. DeepLab V3 architecture and ASPP

2.2 웨이블릿의 분해

그림 3에서 x[n]은 분해될 원래 신호이고 h[n] 및 g[n]이 각각 저역 통과 및 고역 통과 필터를 통한 절차를 보여준다. 웨이블릿의 분해는 모든 레벨에서 필터링 및 서브 샘플링은 샘플 수의 절반(시간 분해능의 절반)과 통과된 주파수 대역의 절반(주파수 분해능의 두 배)을 초래한다[8].

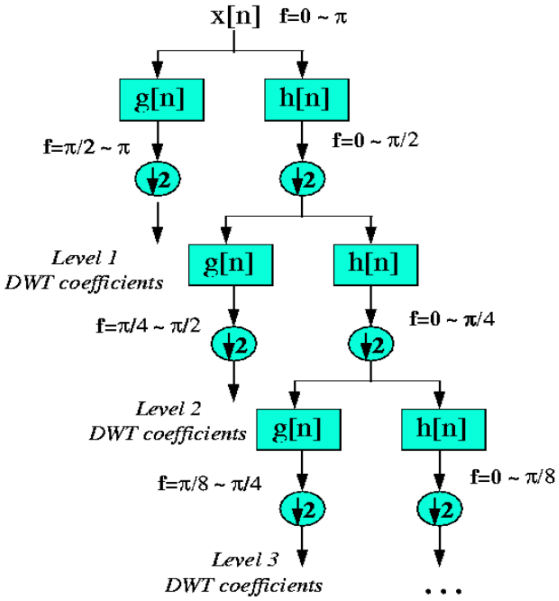

그림 3. 웨이블릿 분해과정
Fig. 3. Wavelet decomposition process

좀 더 자세히 설명을 하면 첫 번째 단계를 진행하면 저주파성분 L과 고주파 성분 H로 나뉘고, 두 번째 단계에서는 이 L, H 성분을 재차 필터링하여 LL, LH, HL, HH 4개의 부 영상을 얻는다. LL 대역의 영상은 해상도가 반으로 줄어든 저주파 성분 이면서 에너지 집중도가 높고 중요한 정보를 유지하고 있다. 나머지 LH, HL, HH 대역의 영상은 수직, 수평, 대각 성분에 대한 Edge 특징을 가지고 있는 고주파 성분에 해당한다.

2.3 WASPP(Wavelet ASPP)의 구현

그림 4는 ASPP를 WASPP로 변경한 블록다이어그램이다. WASPP는 DWT의 변형된 형태로 DWT의 웨이블릿이 진행되면서 특징점 중에서 중요한 정보에 해당하는 저역 부분을 추출하는 것이다[9]. 현재의 ASPP는 컨볼류션 레이어에 대해 다양한 확장성을 확보한 반면 무작위적인 Rate 변화로 인해 주요 특징점이 약해지는 부분이 있다. 이를 개선하기 위해 특징점 중에서 중요한 요소를 모두 포함하고 있는 저역 부분만을 추출하여 다양한 확장성이 강화되도록 WASPP를 적용하였다. 구동 원리를 간단히 설명하면 PWPC의 구조 중 Block 4에서 입력받은 특징점을 1×1 컨볼류션을 처리한 후 LFP(Low Frequency Propagation)의 저주파 성분들(Line LL)에 대해 DWT를 연속 수행한다. 각각의 피라미드 레벨에서, 추출된 Line LL 성분은 1×1 컨볼류션 레이어로 변환되고, 그 후 피라미드 입력과 동일한 공간 분해능, 즉 Block 5로 이중 선형 업 샘플링 된다[9]. 그런 다음 이러한 업 샘플링 된 특징 맵을 연결하여 다른 규모로 캡처된 글로벌 컨텍스트를 연결하는 블록다이어그램이다.

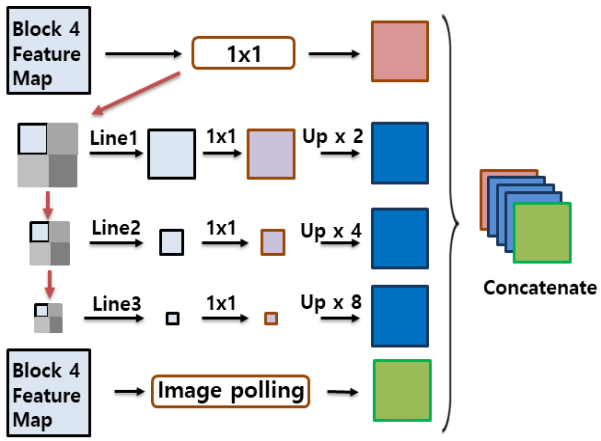

그림 4. WASPP 구조
Fig. 4. Structure of WASPP

그림 5는 WASPP의 기본 구조이다. Backbone은 FCN 기반 모델의 입력 이미지에서 특징을 추출하기 위한 기본 구조이다. DeepLab V3는 기본적으로 ResNet을 Backbone으로 사용한다. ResNet은 Residual Network를 사용하는 깊은 네트워크이다. ResNet과 같은 고전적인 분류 네트워크를 채택하는 이유는 두 가지가 있다. 첫째, 이러한 네트워크는 ImageNet 대규모 시각적 인식 경쟁에서 우수한 성능을 발휘한 것이다. 둘째, 사전 훈련된 모델로 네트워크의 미세 조정이 수월하기 때문이다[10].

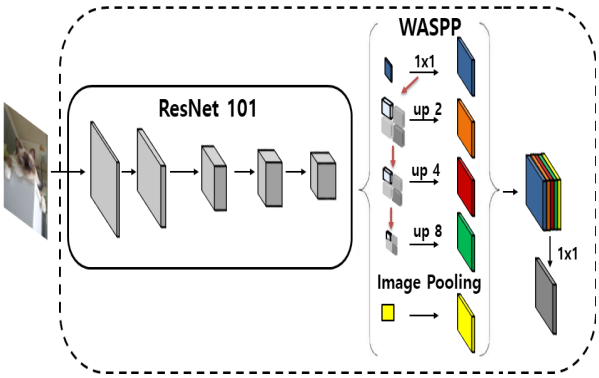

그림 5. WASPP가 포함된 FCN 아키텍처
Fig. 5. Architecture of fully convolutional network with the WASPP

ResNet 101에서는 입력 이미지(512×512)에 대해 그림 6과 같이 계층구조가 설계가 되어있다. 모두 5번의 컨볼류션을 수행하며 Conv 5에서 특징점 출력 값은 입력대비 1/16로 줄어든 32×32×2048가 된다. 그리고 WASPP의 각 계층구조는 그림 6에 자세히 기록되어있다. WASPP의 각 레이어, 1×1 컨볼류션 이후, 또는 Global average pooling을 한 후 최종적으로 32×32×1280으로 연결된다.

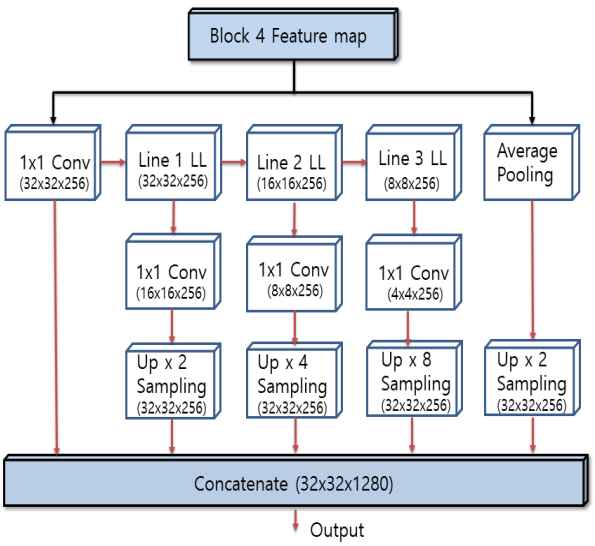

그림 6. WASPP의 동작원리
Fig. 6. Operation of WASPP

그림 6에는 WASPP의 동작에 대해 블록다이어그램으로 제시하였다. 입력은 컨볼류션 5(Block 4)에서 시작되며 Haar 웨이블릿을 활용하여 Line 3까지 Low frequency 필터를 적용하여 1/2 ~ 1/8만큼 다운 샘플링된 저역 특징점을 먼저 추출하는 방식이다. Pyramid 진행과 웨이블릿 특징점 추출은 Line 3까지 구분하여 진행하였다. 그리고 웨이블릿을 진행한 후 각각의 Line에서 1×1 컨볼류션을 진행한 후 연결하여 전체 출력에 대해 진행한다.

III. 제안하는 방법

그림 7은 변형된 WASPP의 전체 구조다. 기존 WASPP와 동일한 FCN 기반 모델의 입력 이미지에서 특징을 추출하여 세그멘테이션하는 구조이다. DeepLab V3와 동일한 ResNet을 기본 구조로 사용했다. 2.3절에서 설명한 것처럼 ResNet 101 + mWASPP + 연결 구조로 되어있다. 변형 WASPP의 경우는 기존 DeepLab v3의 ASPP의 구조에서 주요 특징점을 추출하는 Atrous 컨볼류션 부분에 대해서만 웨이블릿을 추가하는 특징을 가지는 구조이다.

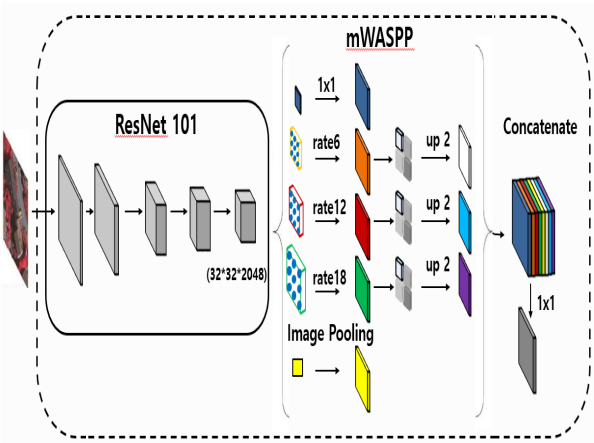

그림 7. 제안된 전체 아키텍처
Fig. 7. Architecture of our proposed

ResNet 101에서는 입력 이미지(512×512)에 대해 5단의 계층구조로 설계가 되어있다. 즉 모두 5번의 컨볼류션을 수행하며 Conv 5에서 특징점 출력 값은 입력대비 1/16로 줄어든 32×32×2048이다. 그리고 변형된 mWASPP의 구조는 그림 7처럼 구성되어있다. 각각 Line 1 레이어, 1×1 컨볼류션, Atrous pooling 결과 그리고 Global average pooling을 한 후 최종적으로 32×32×2048로 연결다.

3.1 변형 WASPP의 구현 방법

그림 8에는 기존 ASPP의 Pyramid 구조를 이용하면서 중요한 Pyramid pooling에 대해서만 중요한 정보를 담고 있는 웨이블릿의 저역 부분 특징점을 활용한 변형 WASPP의 블록다이어그램이다. 웨이블릿 기능을 이용하여 ASPP 특징점 중에서 Rate 특징점 만을 이용하여 진행하였다. 기존 ASPP는 특정 레이어 대해서 Atrous 컨볼류션을 여러 Rate로 진행한 후 결과들을 합치는 구조이다. 여기서 핵심은 Rate 6, 12, 18에서 추출된 특징점이 결국 세그멘테이션 결정하는데 중추적 역할을 한다는 것이다. 그래서 본 연구에서는 기존 ASPP의 Rate 6, 12, 18의 특징점에 대해서만 웨이블릿을 추가하여 각 Rate 특징점 결과 중에서 저역 부분의 특징점을 추가로 한번 더 추출하는 변형 WASPP를 구현하였다.

그림 9는 변형 WASPP 구동방식을 표현한 것이다. Haar 웨이블릿을 활용하여 ASPP의 각 Atrous pooling만 Line 1까지 Low frequency 필터를 적용하여 1/2 만큼 다운된 저역 특징점을 추출한다. Line 1의 경우 Rate 6에 대해서 웨이블릿 저역(중요한 정보) 특징점 추출이 진행되고 1×1 컨볼류션이 진행된다. 이후 2배로 Up-sampling한 특징점들은 연결로 전달된다. 나머지 Rate 12, 18에 대해서도 동일한 방식으로 알고리즘을 구현하였다.

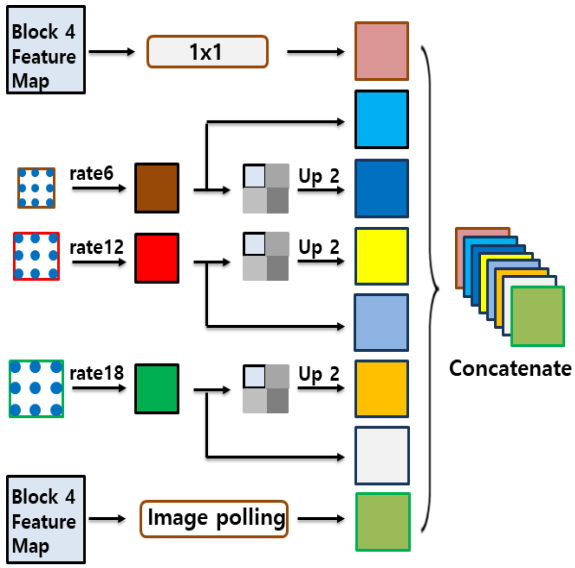

그림 8. 변형 WASPP의 구조
Fig. 8. Structure of modified WASPP

IV. 실험 방법 및 결과

변형 WASPP를 활용한 세그멘테이션 객체 검출 실험은 CPU: AMD Ryzen 7 3700X 8-Core Processor 3.6 GHz, GPU: NVIDIA GeForce RTX 2080TI RAM 32GB 컴퓨터 환경에서 실험을 진행하였다. Dataset은 PASCAL VOC 2012[11]와 SBD(Sementic Boundaries DataSet) 2011[12]을 사용하였다.

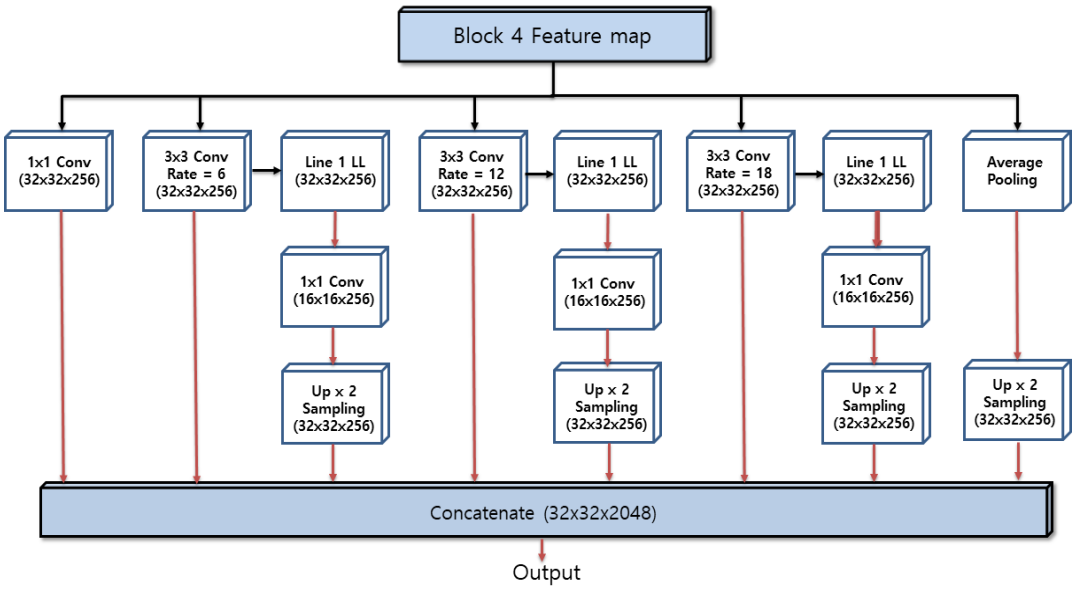

그림 9. 변형 WASPP의 구현
Fig. 9. Implementation of modified WASPP

데이터 세트 정보는 표 1에 제시하였다.

표 1. 데이터 세트 정보
Table 1. Information of dataset

Dataset	Train data	Validation data	Test data
PASCAL VOC 2012	1464	1449	17,125
SBD 2011	8498	2858	165,482

PASCAL VOC 2012는 Dataset으로 17,125개를 사용하고, Train 이미지는 1464개와 공개적으로 사용할 수 있는 주석을 가진 1449개의 이미지를 사용하였다. SBD 2011는 비교 수행 목적으로 165,482개를 사용하였다. 정량적 평가의 경우, 세그멘테이션과 Object detection에서 가장 빈번하게 사용되는 성능척도인 mIoU(Mean Intersection over Union)를 사용하였다. IoU는 Ground Truth(GT) 대비 예측 마스크가 서로 얼마나 '겹쳐지는지'를 수치화 한 값이다.

4.1 실험 결과

기존 Deeplab v3, WASPP 와 변형 WASPP를 비교 실험한 결과는 표 2에는 epoch = 100의 결과를 표시하였다.

표 2. num_epoch =100에서 변형 WASPP test 결과
Table 2. Modified WASPP test result when num_epoch =100

Method	Performance on PASCAL VOC 2012			
	t-loss(%)	mIoU(%)	v-loss(%)	mIoU(%)
DeepLab V3	1.28	90.93	29.49	70.27
WASPP	1.38	90.15	30.41	68.41
Modified WASPP	**1.23**	**91.29**	**26.81**	**71.89**

표 2의 epoch=100에서 변형 WASPP는 DeepLab v3 대비 mIOU 1% 정도의 개선된 효과를 보인다. 특히 표 3의 image_6(모니터 prediction) 사진에서는 컴퓨터를 모니터로 오 인식하는 부분이 완전히 개선된 결과를 얻었다. 이것은 epoch=50의 결과로 기존 DeepLab v3는 epoch=200에서도 여전히 오인식이 나타나고 있다.

표 3은 epoch = 50, 100, 200에 대한 실험 결과이며, 여기에는 DeepLab v3와 변형 WASPP를 비교한 결과이다.

표 3. image_6(Monitor 예측)의 비교 결과
Table 3. Comparison of image_6 (Monitor prediction)

	RGB	Ground truth
Image _6		
epochs	DeepLab V3	Modified WASPP
50		
100		
200		

V. 결 론

실험 결과 변형 WASPP 알고리즘은 기존 DeepLab V3 알고리즘의 핵심인 ASPP를 대신하기 위해 DWT의 Haar 웨이블릿의 Low pass 특성을 활용하여 이미지의 특징점을 1/8까지 다운하는 평균 pooling 방식을 활용하였다. 그 결과 변형 WASPP는 기존 Deeplab v3 대비 t-loss(1.28% → 1.23%), mIOU (90.9% → 91.3%), v-loss(29.49% → 26.81%), mIOU (70.3% → 71.89%)로 epoch=100 기준으로 개선된 실험결과를 얻었다. 그리고 시멘틱 세그멘테이션 image 정확도 예측은 기존 DeepLab v3가 image_6 (컴퓨터 prediction) 사진에서 epoch=200 기준으로 여전히 오류가 존재하나, 변형 WASPP는 epoch=50에서 오류가 완전히 개선된 결과를 얻었다. 이것은 기

존보다 200% 성능 개선된 결과이다. 다만 변형 WASPP는 Atrous 컨볼류션 이후에 웨이블릿 컨볼류션을 추가로 진행하므로 실행시간이 긴 단점이 있다. 이를 개선하기 위해서는 WASPP에서는 웨이블릿 입력을 현재 Conv 5 보다 앞선 Conv 4에서 특징점 추출을 진행하는 방법 등 다양한 방법으로 컨볼류션 시간을 단축하는 방법을 시도하거나 기존 DeepLab V3+ 보다 더 개선된 성능을 위한 다양한 시도가 요구된다.

References

[1] Hee Il Hahn, "Proposal of Image Segmentation Technique using Persistent Homology", Journal of IIBC, Vol. 18, No. 1, pp. 223-229, Jan. 2018.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs", Computer Vision and Pattern Recognition, arXiv preprint arXiv:1412.7062v4, Jun. 2016.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, pp. 834-848, Apr, 2018.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation", Computer Vision and Pattern Recognition, arXiv preprint arXiv:1706.05587v3, Dec. 2017.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", The European Conference on Computer Vision (ECCV), pp. 801-818, Aug. 2018.

[6] Jong-Sik Kim and Dae-Seong Kang, "Improved Segmentation Object Detection Using Discrete Wavelet Transform(DWT)", Journal of Korean Institute of Information Technology Vol. 17, No. 11, pp. 249-251, Nov. 2019.

[7] Fisher Yu and Vladlen Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions" Published as a conference paper at ICLR 2016, arXiv:1511.07122v3, Apr. 2016.

[8] Robi Polikar, "Multiresolution Analysis: The Discrete Wavelet Transform", http://users.rowan.edu/~polikar/WTpart4.html, [accessed: May 28, 2020]

[9] Lingni Ma, Jörg Stückler, Tao Wu, and Daniel Cremers, "Detailed Dense Inference with Convolutional Neural Networks via Discrete Wavelet Transform", arXiv preprint arXiv:1808.01834v1, Aug, 2017.

[10] Wang Yuhao, Liang Binxiu, Ding Meng, and Li Jiangyun, "Dense Semantic Labeling with Atrous Spatial Pyramid Pooling and Decoder for High-Resolution Remote Sensing Imagery", Remote Sensing, Vol. 11, No. 1, pp. 20, Jan. 2019.

[11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge", International Journal of Computer Vision, Vol. 88, pp. 303-338, Sep. 2009.

[12] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Semantic contours from inverse detectors", 2011 International Conference on Computer Vision, 10.1109/ICCV.2011.6126343, Nov. 2011.

저자소개

김 종 식 (Jong-Sik Kim)

2020년 8월 : 동아대학교
전자공학과 (공학석사)
2020년 9월 ~ 현재 : 동아대학교
전자공학과 박사과정
관심분야 : 영상처리, AI

강 대 성 (Dae-Seong Kang)

1994년 5월 : Texas A&M 대학교
전자공학과(공학박사)
1995년 ~ 현재 : 동아대학교
전자공학과 교수
관심분야 : 영상처리, 패턴인식