경계 검출을 이용한 의미론적 영역 분할 네트워크의 정확도 향상

김재선, 허용석
아주대학교 전자공학과
jskim159@ajou.ac.kr, ysheo@ajou.ac.kr

요 약

본 논문은 경계 검출 모듈을 사용하여 의미론적 영역 분할 정확도를 보다 향상시키는 기법을 제안한다. 기존의 의미론적 영역 분할 네트워크, 특히 파라미터 수가 매우 적은 경량화된 네트워크는 효율적으로 뉴런의 수용영역을 넓히기 위해 팽창 합성곱을 사용하는 것이 일반적이다. 하지만 팽창 합성곱을 사용할 경우 얇은 물체를 제대로 검출하지 못하는 경우가 발생한다. 본 논문은 얇은 물체 검출을 보다 정확하게 수행하기 위해 네트워크 학습시에 경계 검출 보조 모듈을 부착하여 물체의 경계를 검출하고, 손실함수를 추가하여 이를 이용하는 기법을 제안한다. 제안하는 기법은 Cityscapes 데이터셋에 대하여 기존의 의미론적 영역 분할 네트워크보다 좋은 결과를 보인다.

1. 서론

의미론적 영역 분할은 입력 디지털 영상의 모든 픽셀을 각각 미리 정의된 클래스 중 하나로 분류하는 것이다. 분류된 픽셀 단위의 의미론적 정보는 자율 주행 [1, 2], 원격 탐사 [3], 의료 영상 분석 [4] 등 다양한 응용에서 장면을 완전히 이해할 수 있도록 하는데 중요한 역할을 한다.

최근 다른 컴퓨터 비전 문제와 마찬가지로, 의미론적 영역 분할 문제도 심층 합성곱 신경망 (딥 뉴럴 네트워크, deep neural network)을 이용한 기법들이 높은 정확도를 보이고 있다 [5-9]. 기존의 높은 정확도를 보이는 네트워크들은 대체로 파라미터와 연산량이 많아 실행 속도가 느린데, 이를 효율적으로 줄인 기법들도 많이 제안되고 있다 [10-12].

파라미터와 연산량을 줄여 경량화된 네트워크들은 합성곱 층을 깊게 쌓지 않고 효율적으로 뉴런의 수용영역을 넓히기 위해 팽창 합성곱을 사용하는 것이 일반적이다 [10-12]. 하지만, 팽창 합성곱을 사용할 경우 작고 얇은 물체를 제대로 분류하지 못하는 경우가 발생할 수 있다 [13].

본 논문에서는 작고 얇은 물체 분류를 보다 정확하게 수행하기 위해, 네트워크가 물체의 경계를 더욱 정확히 인지하도록 학습하는 기법을 제안한다. 더욱 구체적으로, 네트워크의 파라미터를 공유하면서 의미론적 영역 분할과 객체 경계 검출을 동시에 하는 다중 작업 (Multi-task) 학습을 제안한다. 다중 작업 학습을 위해 본 논문에서는 의미론적 영역 분할 네트워크 학습시에 경계 검출 보조 모듈을 부착하여 물체의 경계를 검출하고, 손실함수를 추가하여 이를 이용하는 기법을 제안한다.

Cityscapes 데이터셋 [14]에 대해 기존 네트워크에 제안하는 기법을 적용하여 학습한 결과, 정확도가 향상되었다.

2. 경계 검출 보조 모듈을 이용한 의미론적 영상 분할

2.1 ESCNet

ESCNet [12]은 네트워크의 파라미터와 연산량을 효율적으로 줄이기 위해 공간-채널 팽창 합성곱 기반의 ESC 모듈을 이용하는 것이 특징이다. 기존의 S. Mehta et al. 이 제안한 ESPNet [11] 구조에 ESC 모듈을 적용하여 네트워크의 파라미터와 연산량을 대폭 줄이면서 정확도도 향상시켰다.

하지만 ESCNet 은 복수의 팽창 합성곱 층을 쌓은 구조이기 때문에, 작고 얇은 물체를 제대로 분류하지 못하는 경우가 발생할 수 있다 [13]. 이는 그림 1.(e) 를 보면 확연히 알 수 있다.

2.2 경계 검출 보조 모듈

본 논문에서는 의미론적 영역 분할 문제에서 작고 얇은 물체 분류를 보다 정확하게 수행하기 위해, 물체의 경계 정보를 학습에 이용하는 기법을 제안한다.

구체적으로, 경량화된 의미론적 영역 분할 네트워크인 ESCNet [12]에 그림 **2** 와 같이 경계 검출 모듈을 부착하여 네트워크의 파라미터를 공유하면서 의미론적 영역 분할과 객체 경계 검출을 동시에 하는 다중 작업 학습을 제안한다.

학습에 의미론적 영역 분할 손실함수와 객체 경계 검출 손실함수를 함께 사용하여 공유되는 네트워크가 객체의 경계를 검출하면서 의미론적 영역 분할을 하도록 유도한다. 객체 경계 검출 모듈은 학습시에만 이용하고, 학습이 끝나면 제거할 수 있도록 네트워크를 설계하였다.

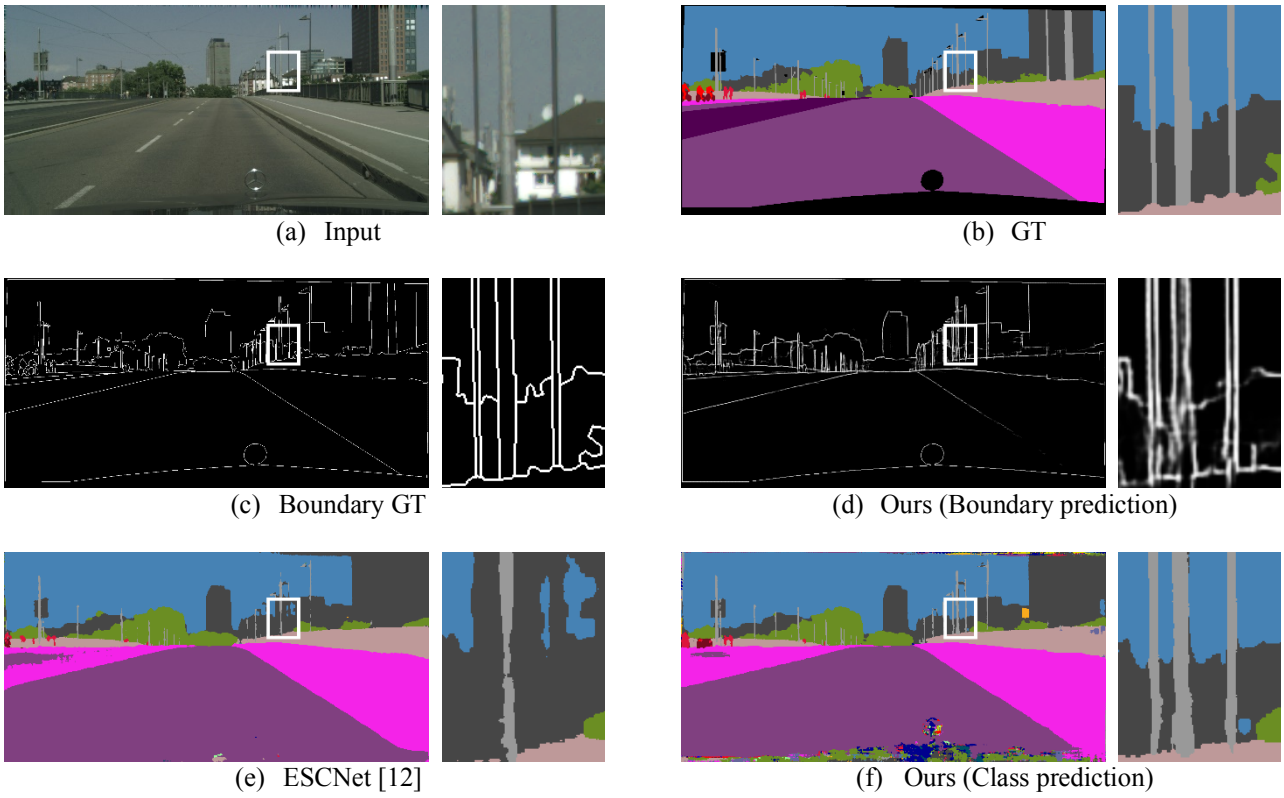

그림 1. 입력, 참값 영상과 ESCNet [12], 제안하는 기법의 결과.

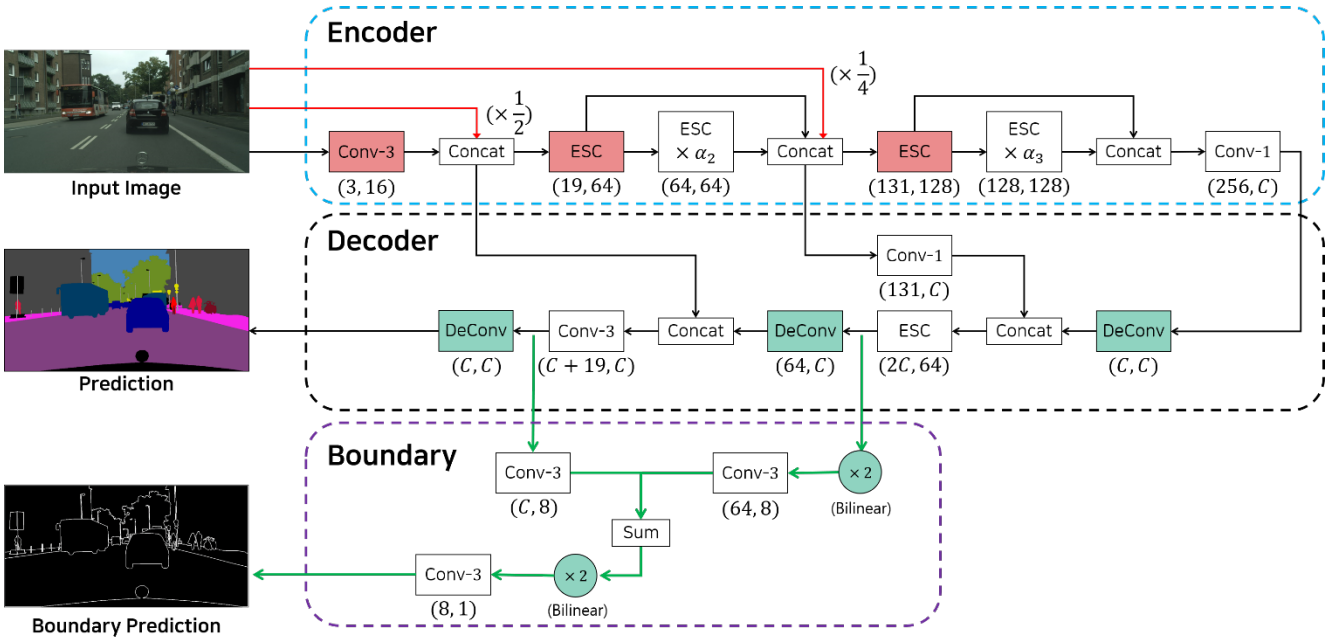

그림 2. 제안하는 경계 검출 보조 모듈을 부착한 네트워크 구조.

기본 네트워크는 ESC 모듈 [12]을 사용하는 ESCNet 으로 구성하였다. Conv-n 은 $n \times n$ 합성곱, DeConv 는 역합성곱을 의미한다. 각 블록의 (#, #) 안에 표시된 숫자는 각각 입력, 출력 특징지도의 채널의 수다. 빨간색과 초록색 블록은 각각 입력 특징지도의 공간 크기를 2 배씩 작게, 크게 한다.

2.3 전체 손실함수

의미론적 영역 분할과 객체 경계 검출을 동시에 하는 다중 작업 학습을 하기 위해, 의미론적 영역 분할 손실함수와 객체 경계 검출 손실함수를 함께 학습에 사용하였다. 의미론적 영역 분할 손실함수 $\mathcal{L}_{seg}$는 Cross-Entropy (CE) 손실함수를 사용하였고, 객체 경계 검출 손실함수 $\mathcal{L}_{boundary}$는 L. Zhou et al. [15] 과 같이 Dice 계수와 이진 Cross-Entropy 손실함수 합한 손실함수를 사용하였다.

전체 손실함수 $\mathcal{L}_{total}$은 다음 수식과 같이 정의된다:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \beta\mathcal{L}_{boundary}, \quad (1)$$

이 때, β는 전체 손실함수 $\mathcal{L}_{total}$에서 객체 경계 검출 손실함수 $\mathcal{L}_{boundary}$ 의 영향력을 조절하는 가중치이다.

3. 실험 결과 및 분석

3.1 Cityscapes 데이터셋

제안하는 기법의 성능을 검증하기 위해 Cityscapes 데이터셋 [14]을 사용하였다. Cityscapes 데이터셋은 도심 교통 장면 이해를 위한 다량의 의미론적 영역 분할 데이터셋으로, 5000 개의 입력 영상과 그에 대응하는 정교한 참값 영상으로 구성되어 있다. 참값 영상의 각각의 픽셀은 미리 정의된 19 개의 클래스 중 하나로 할당되어 있으며, 학습용 영상 2975 개, 검증용 영상 500 개, 테스트용 영상 1525 개로 나뉘어져 있다.

3.2 실험 환경

본 논문에서는 PyTorch 를 이용하여 실험하였다. 하나의 Titan Xp GPU 를 사용하였고, [11] 에서 공개한 코드를 사용하여 학습했기 때문에 학습의 세부 사항은 [11] 과 동일하다.

제안하는 기법을 실험하기 위해서는 기존 Cityscapes 데이터셋의 참값 외에 객체 경계 참값 영상이 필요하기 때문에 T.-C. Wang et al. [16] 이 사용한 방법으로 생성하였다.

3.3 평가 지표

본 논문에서는 의미론적 영역 분할 기법의 성능을 평가하기 위해 클래스별 IoU (Intersection over Union)과 전체 클래스의 평균 IoU (mIoU)를 평가 지표로 사용한다.

표 1. 다양한 β 값에 따른 제안하는 기법의 성능 분석

Method	β	mIoU (Validation set)
Baseline (ESCNet [12] + CE Loss)	-	62.16
Ours	0.3	62.05
	0.5	62.95
	0.7	62.95
	1.0	**63.44**
	1.3	61.62

3.4 실험 결과

먼저 전체 손실함수 $\mathcal{L}_{total}$에서 객체 경계 검출 손실함수 $\mathcal{L}_{boundary}$ 의 영향력을 조절하는 가중치인 β의 적절한 값을 찾기 위해 표 1.과 같이 다양한 β에 따른 제안하는 기법의 성능 분석하였다.

Cityscapes 데이터셋 [14]의 검증용 데이터셋에 대하여 ESCNet [12]은 62.16 mIoU 의 정확도를 보인다. 반면, 제안하는 기법은 β가 1.0 일 때, 63.44 mIoU 의 정확도를 보이는 것을 확인할 수 있다.

표 2. 는 Cityscapes 검증용 데이터셋의 미리 정의된 19 개의 클래스에 대한 각각의 IoU 를 나타낸다. 제안하는 기법은 작고 얇은 물체인 Pole 등의 클래스에 대해 기존의 ESCNet 보다 향상된 정확도를 보인다.

제안하는 기법은 그림 **1**.(d) 에서 보이는 바와 같이 작고 얇은 물체의 경계를 검출할 수 있도록 학습 되었기때문에, 그림 **1**.(f) 와 같이 기존의 ESCNet (그림 **1**.(e)) 보다 작고 얇은 물체를 정확하게 분류할 수 있다.

4. 결론

본 논문에서는 팽창 합성곱을 사용하는 경량화된 의미론적 네트워크가 작고 얇은 물체를 보다 정확하게 분류할 수 있도록 네트워크의 파라미터를 공유하면서 의미론적 영역 분할과 객체 경계 검출을 동시에 하는 다중 작업 (Multi-task) 학습을 제안한다. 의미론적 영역 분할 네트워크 학습시에 경계 검출 보조 모듈을 부착하여 물체의 경계를 검출하고, 손실함수를 추가하여 이를 이용하는 기법을 제안한다. 제안한 기법은 Cityscapes 데이터셋에 대해 기존보다 향상된 정확도를 보였으며, 얇고 작은 물체를 보다 정확하게 분류할 수 있음을 보였다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2019R1C1C1007446).

표 2. Cityscapes 검증 데이터셋에 대한 클래스별 IoU 와 mIoU.

Method	mIoU	Bicycle	Buildings	Bus	Car	Fence	Motorcycle	Person	Pole	Rider	Road	Sidewalk	Sky	Terrain	Traffic light	Traffic sign	Train	Truck	Vegetation	Wall
Baseline (ESCNet [12])	62.16	**59.37**	88.58	**64.22**	87.62	43.67	**30.87**	65.77	51.47	**43.02**	**95.50**	**73.14**	**92.85**	**57.91**	**50.59**	63.32	33.28	55.63	**90.16**	33.96
Ours	**63.44**	59.31	**89.21**	58.72	**88.33**	**48.60**	24.80	**66.29**	**54.51**	39.07	95.45	72.39	92.41	55.05	50.10	**63.92**	**58.41**	**58.84**	89.96	**39.90**

참고문헌

[1] B. Li, S. Liu, W. Xu and W. Qiu, “Real-time object detection and semantic segmentation for autonomous driving,” in: Proceedings of the MIPPR 2017 (Automatic Target Recognition and Navigation), 10608, International Society for Optics and Photonics, pp. 106080P, 2018.

[2] Y.-H. Tseng and S.-S. Jan, “Combination of computer vision detection and seg- mentation for autonomous driving,” in Proc. IEEE/ION Position, Location and Navigation Symposium (PLANS), IEEE, pp. 1047–1052, 2018.

[3] R. Kemker, C. Salvaggio and C. Kanan, “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,” ISPRS J. Photogramm. Remote Sens, vol. 145, pp. 60–77, 2018.

[4] F. Jiang, A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, K. Adil and S. Liu, “Medical image semantic segmentation based on deep learning,” Neural Comput. Appl., vol. 29, pp. 1257–1265, 2018.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3431-3440, 2015.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6230-6239, 2017.

[7] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5168-5177, 2017.

[8] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 833-851, 2018.

[9] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu, "Dual attention network for scene segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3146-3154, 2019.

[10] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: efficient spatial pyramid of dilated convolutions for semantic segmentation," in Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 552-568, 2018.

[11] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: efficient residual factorized ConvNet for real-time semantic segmentation," IEEE Trans. Intell. Transp. Syst., vol. 19, no. 1, pp. 263-272, Jan. 2018.

[12] J. Kim and Y. S. Heo, “Efficient semantic segmentation using spatio-channel dilated convolutions,” in IEEE Access, vol. 7, pp. 154239-154252, 2019.

[13] R. Hamaguchi, A. Fujita, K. Nemoto, T. Imaizumi, and S. Hikosaka. ‘‘Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery.’’ [Online]. Available: https://arxiv.org/abs/1709.00179, 2017.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3213-3223, 2016.

[15] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 192–1924, 2018.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8798-8807, 2018.