Deformable Convolution 기반 어텐션 모듈을 사용한 의미론적 분할 모델 설계

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

Design of a Semantic Segmentation Model Using an Attention Module Based on Deformable Convolution

Jin-Seong Kim[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

의미론적 분할(Semantic Segmentation)은 이미지 내의 객체 및 배경을 픽셀 단위로 분류하는 작업으로 정밀한 탐지가 요구되는 분야에서 활발히 연구되고 있다. 기존 어텐션 기법은 의미론적 분할의 다운샘플링(Downsampling) 과정에서 발생하는 정보손실을 완화하기 위해 널리 사용됐지만 고정된 Convolution 필터의 형태 때문에 객체의 형태에 따라 유동적으로 대응하지 못했다. 본 논문에서는 이를 보완하고자 Deformable Convolution과 셀프어텐션(Self-attention) 구조기반 어텐션 모듈을 사용한 의미론적 분할 모델을 제안한다.

1. 서론

의미론적 분할은 이미지 내의 객체 및 배경을 픽셀 단위로 분류하는 작업으로 자율주행과 같은 정밀한 탐지가 요구되는 분야에서 활발히 연구되고 있다[1]. 의미론적 분할에서 성능을 향상할 수 있는 주요 방법은 다운샘플링(Downsampling) 과정에서 발생하는 공간 및 전역적 문맥 정보의 손실을 줄이거나 업샘플링(Upsampling)의 품질을 높이는 것이다.

정보의 손실은 다운샘플링 과정에서 특징 맵의 크기가 줄어 해상도가 감소하면서 발생한다. 이를 방지하기 위해선 고해상도부터 저해상도까지의 정보를 적절히 융합하거나 어텐션 기법을 이용해 손실되는 정보를 강조하는 방법이 있다[2][3]. 하지만, 기존의 연구는 Convolution 필터의 형태가 고정적이기 때문에 객체의 형태에 따라 유동적으로 대응하지 못했다.

본 논문에서는 업샘플링보다 다운샘플링 과정에서 손실되는 정보 보존에 집중한다. 객체의 다양한 형태에 유동적으로 대응하여 특징을 추출하기 위해, Deformable Convolution 기반의 어텐션 모듈을 활용한 의미론적 분할 모델을 제안한다.

2. 관련 연구

2.1 Deformable Convolution

기존 CNN(Convolutional Neural Networks)에서 이미지의 기하학적 변형은 대부분 데이터 어그먼테이션(Data Augmentation)에 의존하였다. 또한, 고정된 형태의 필터 때문에 너무 크거나 학습되지 못한 변형에서는 성능이 저하됐다. 이를 해결하기 위해 Deformable Convolution(DC)[4] 은 필터에 Offset을 추가하고 가중치를 업데이트하면서 주변 픽셀의 특징 맵 정보를 활용하여 Offset을 업데이트한다. 업데이트된 Offset은 입력 이미지의 픽셀 위치를 조정하여 필터의 형태를 객체의 형태에 따라 유동적으로 변하게 한다.

2.2 어텐션

어텐션(Attention)은 자연어 처리에서 다른 문장 간의 유사도에 따라 강조해야 할 부분을 다르게 처리하기 위해 사용됐다. 더 나아가 셀프어텐션(Self-attention)은 같은 문장 내에서 단어 간의 유사도를 계산함으로써 유사도 계산의 품지를 높여 문장 내의 동음의 어도 구분할 수 있게 하였다.

이러한 개념들이 컴퓨터 비전에도 적용되어 픽셀

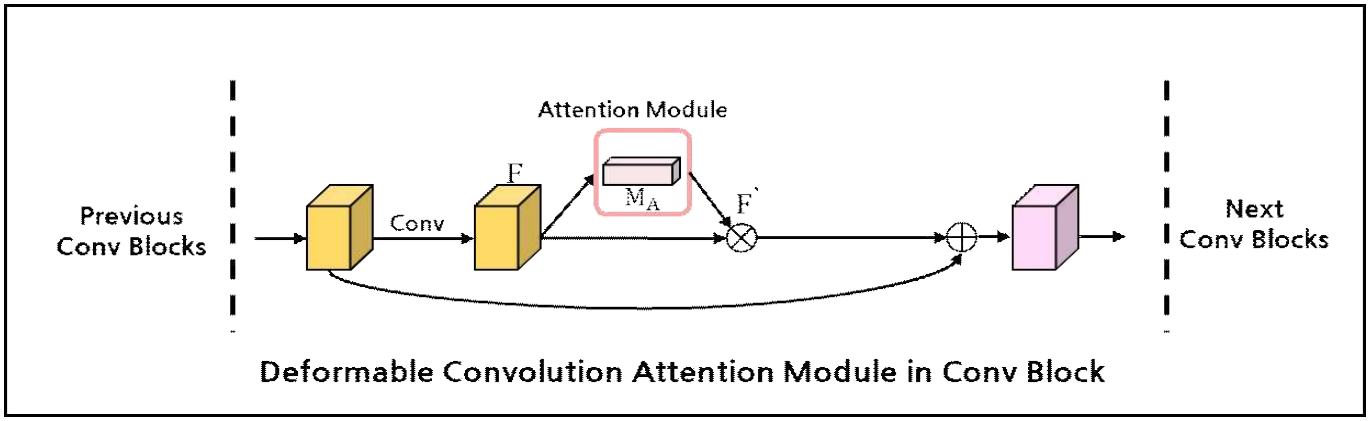

그림 1 Deformable Convolution 기반 어텐션 모듈이 삽입된 합성곱 블록 구조

간의 유사도를 계산해 더 중요한 특징을 강조했다. 대표적으로 SENet은 채널 정보를 스칼라값으로 압축하여 더 중요한 채널을 강조하였다[5]. 또한, 의미론적 분할을 상정한 비선형성을 반영하기 위해 편광 기법과 셀프어텐션을 사용한 PSA도 발표됐다[6]. 본 논문의 어텐션 모듈 구조는 의미론적 분할에 더 최적화된 PSA의 셀프어텐션 구조를 채택한다.

3. 제안하는 방법

그림 1은 어텐션 모듈이 삽입된 Convolution 블록의 구조다. 제안하는 방법은 어떠한 모델에도 삽입될 수 있도록 ResNet 기반의 일반적인 Convolution 블록 사이에 어텐션 모듈을 삽입한다. 특징을 추출하기 위한 백본은 일반적인 Convolution 블록을 사용하는 모델이라면 모두 사용할 수 있다. 의미론적 분할 맵을 만들기 위한 업샘플링 방법은 이중선형(Bilinear) 업샘플링을 사용한다.

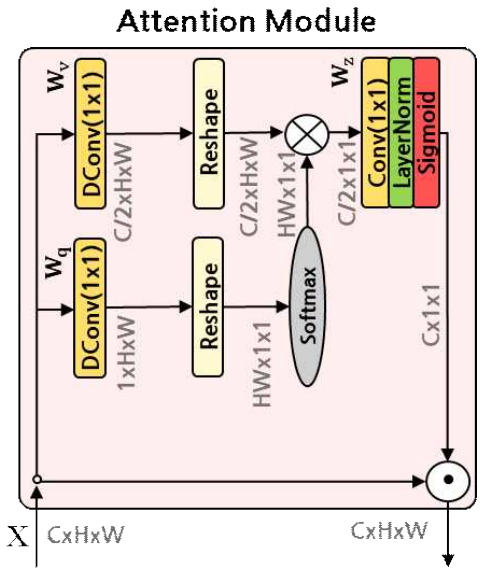

그림 2 어텐션 모듈 구조

그림 2는 어텐션 모듈의 구조다. 구체적으로 입력 이미지를 한 번의 Convolution을 거쳐 Key, Query, Value를 생성한다. 첫 번째로 Query와 Value를 DC에 통과시켜 형태에 따라 유동적으로 학습한다. 이후 Query의 각 채널은 스칼라값으로 압축되고 Softmax를 통해 어텐션 맵을 생성한다. 생성된 어텐션 맵은 Value와 곱한 다음 Sigmoid를 거쳐 중요한 정보는 강조하고 중요하지 않은 정보는 탈락시킨다. 최종적으로 Key와 곱해 정보를 강조한다.

4. 데이터 세트 구성 및 평가지표

성능을 평가하기 위한 데이터 세트는 장면이해를 위한 도시경관 데이터 세트인 Cityscapes[7]를 사용한다. Cityscapes는 5,000개의 고화질 장면 이미지와 정교한 주석이 달린 장면 이미지로 이루어져 있다. 데이터 세트 구성은 훈련, 검증 및 테스트를 위해 2,975, 500, 1,525개의 이미지로 나뉜다. 총 30개의 클래스가 있지만 희박한 클래스를 제외한 19개의 클래스를 훈련과 평가에 사용한다.

평가지표는 모든 클래스의 IoU 값을 평균 낸 mIoU(mean of class-wise Intersection over Union)을 사용한다. IoU(Intersection over Union)는 정답과 예측의 교집합 영역을 합집합의 영역으로 나눈 값이다.

$$IoU = \frac{Overlapping\ Area}{Union\ Area} \quad (1)$$

$$mIoU = \frac{1}{i}\sum_{k=0}^{i} IoU_k \quad (2)$$

5. 결론

기존의 어텐션 모듈은 고정된 형태의 필터 때문에 객체의 형태에 따라 유동적으로 학습하지 못했다. 본 논문에서는 이를 보완하기 위해 객체의 형태

에 따라 유동적으로 학습하는 Deformable Convolution 기반 어텐션 모듈을 사용한 의미론적 분할 모델을 설계한다. 어텐션 모듈은 어떠한 모델에도 삽입될 수 있도록 RsNet 기반의 일반적인 Convolution 블록 사이에 삽입한다. 어텐션 모듈의 구조는 의미론적 분할에 최적화된 PSA의 셀프어텐션 구조를 채택한다. 데이터 세트는 장면이해를 위한 도시경관 데이터 세트인 Cityscapes를, 평가지표는 mIoU를 사용한다.

객체의 형태나 구조가 시시각각 변하는 현실 세계에서는 다양성과 복잡성에 대응하는 모델이 필요하다. 제안하는 방법은 다양한 형태와 구조의 객체들에 더욱 강건하게 대응할 수 있을 것으로 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) and 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] 이준엽, 이영완, “임베디드 보드에서 실시간 의미론적 분할을 위한 심층 심경만 구조,” *정보과학회논문지*, Vol. 45, No. 1, pp. 94-98, 2018.

[2] 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.

[3] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” in *Proceedings of the International conference on machine learning*, PMLR, pp. 2048-2057, 2015.

[4] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu and Y. Wei, “Deformable Convolutional Networks,” in *Proceedings of the IEEE international conference on computer vision 2017*, pp. 764-773, 2017.

[5] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition 2018*, pp. 7132-7141, 2018.

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

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, et al, "The cityscapes dataset for semantic urban scene understanding", in *Proceedings of the IEEE conference on computer vision and pattern recognition 2016*, pp. 3213-3223, 2016.