3D Convolution을 이용한, RGB-D Data에 대한 Semantic Segmentation 모델 연구

정만수, 이정우*

서울대학교

tlwh1179@snu.ac.kr, *junglee@snu.ac.kr

Semantic Segmentation Model for RGB-D Data Using 3D Convolution

Jung Man Soo, Lee Jung Woo*

Seoul National Univ.

요 약

본 논문은 RGB값과 Depth Map을 동시에 활용하는 RGB-D Semantic Segmentation Task에서 3D Convolution을 활용하는 방안을 제안 및 실험하였다. 기존에 존재하는 모델들과 달리 본 논문에서는 주어진 RGB와 Depth Map을 이용해 3D Tensor를 재구성하여 Input으로 사용하였으며, 3D Tensor를 다루기 위해 모델에 3D Convolution을 도입하였다. 또한, 모델이 Depth 정보를 좀 더 적극적으로 활용하게 만들기 위해, Depth-wise Attnetion Mechanism을 도입한 모듈을 추가하여 성능 향상을 이끌어냈다. 실험 결과, Input을 3D로 확장하면서 늘어난 Parameter 및 부족한 Dataset의 크기로 인해 Overfitting 현상이 일어났으나, Training mIOU에서 높은 성능을 보였다는 점을 고려했을 때, 충분한 학습 데이터 투입 및 모델 개선을 통해 Overfitting 현상을 줄인다면 Evaluation Phase에서도 상당한 성능을 낼 것으로 기대된다.

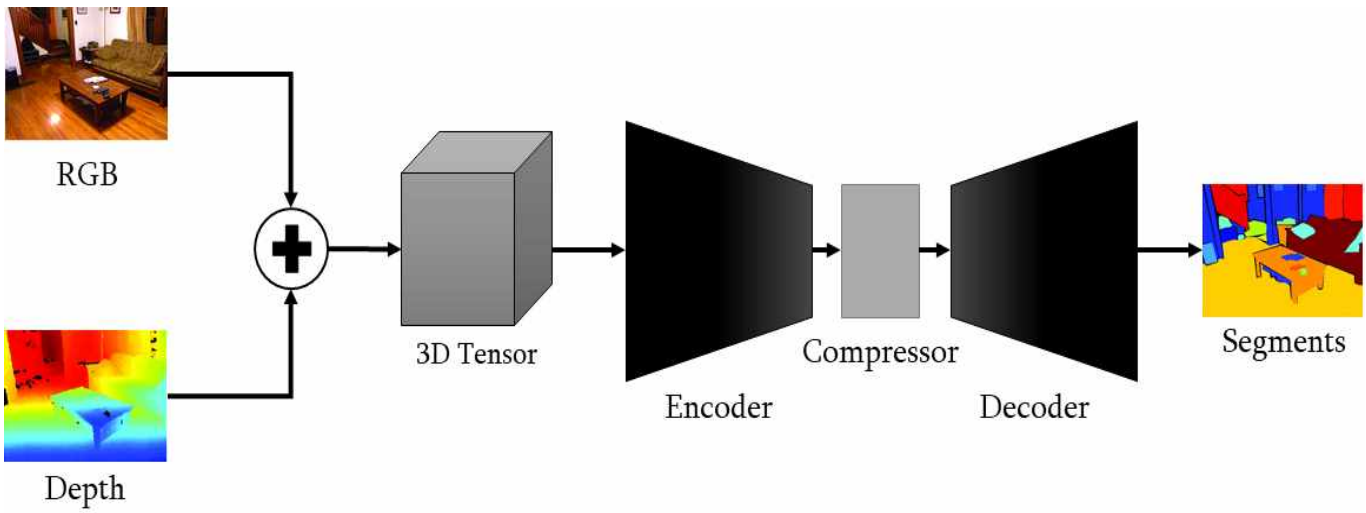

그림 1 - RGB-D Segmentation을 위해 제안된 모델

Dataset	Compressor	mIOU(Train)	mIOU(Eval)
SUN RGBD-37	Average	0.1926	0.1722
SUN RGBD-37	Depth-wise Attention	0.5829	0.2114
NYU v2-40	Average	0.6418	0.1532
NYU v2-40	Depth-wise Attention	0.7228	0.1795

표 1 - Dataset 및 Compressor에 따른 실험 결과

Ⅰ. 서 론

최근 Convolutional Layer를 활용한 Neural Network가 각종 Computer Vision 분야에서 놀라운 성과를 보여주고 있으며, 특히 Segmentation 분야에서는 독보적인 성능을 보여주고 있다. 그러나 현존하는 대부분의 Segmentation 모델들은 이미지의 RGB 값만을 토대로 계산하기 때문에 물체의 형태나 Depth를 제대로 활용하지 못한다는 단점이 있다. 이를 극복하기 위해, 이미지의 RGB 값과 Depth Map을 동시에 활용하는 RGB-D Segmentation이 새로운 연구 분야로 떠오르고 있다. 이러한 맥락에서, 본 논문에서는 RGB-D Segmentation에 대한 새로운 모델을 제안하고 해당

모델의 성능 및 특성을 분석하였다.

Ⅱ. 본론

Segmentation에서 Depth Map을 활용하는 방법은 여러 가지가 있다. [1]은 RGB에 대한 Encoder와 Depth Map에 대한 Encoder를 따로 두어 Feature Level에서 융합하는 방식을 활용하였다. [2]는 Convolutional Layer를 수정하여 채널 방향으로 Concatenate된 Depth Map을 활용하도록 하였다. 그러나 이러한 방식들은 모두 2D Convolutional Operation을 기반으로 하기 때문에 Depth Map으로 주어지는 3D 정보를 온전히 활용할 수 없다는 한계점이 존재한다.

이러한 점을 해결하기 위해, 본 논문에서는 입력으로 들어온 RGB 값을 Depth 방향으로 Reconsruct하는 방식을 취했다. [그림 1]은 논문에서 제안하는 방법을 보여준다. 2D 정보로 들어온 RGB를 각 Pixel에 대응하는 Depth Value에 따라 배치함으로써 3D 텐서 형태로 변환한다. 변환된 Input Tensor는 Encoder를 통과하여 3D Feature가 된다. Compressor는 3D Tensor의 Depth 정보를 압축하여 2D Tensor로 변환하는 역할을 하며, 2D Tensor를 Decoder를 통과시킴으로써 최종적으로 Segmentation Map을 얻는다.

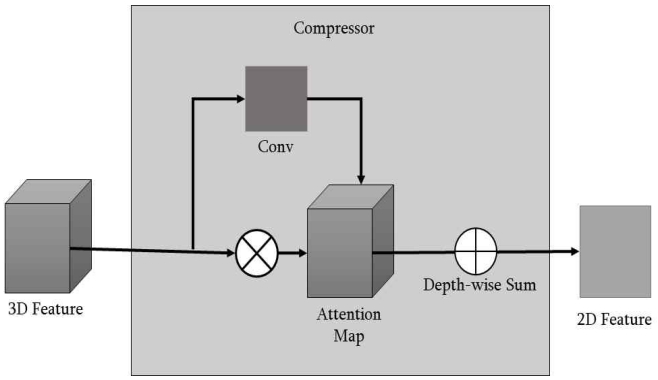

그림 2 - Depth-wise attention이 적용된 Compressor

. 여기서 Compressor는 단순히 Depth 방향으로 Average를 구하는 방법과 Depth-wise Attention을 도입하는 방법이 있다. [그림 2]는 Depth-wise Attention을 도입한 Compressor의 모습이다. Compreesor는 Input으로 들어온 3D Feature에 대해 Convolution 및 Sigmoid연산을 취하여 Attention Map을 구한다. 이때 Attention Map은 Input Feature에서 몇 번째 Depth의 값이 얼마나 중요한지 판별하는 역할을 한다. 이렇게 구한 Attnetion Map을 3D Feature에 곱한 후, Depth 방향으로 더하여 2D Feature를 구한다.

[표 1]은 Dataset 및 Compressor의 타입을 바꿔가며 실험한 결과이다. 실험 데이터셋은 SUN RGBD-37 [3]과 NYU v2-40 [4]를 사용하였다. SUN RGBD-37은 10,335장의 RGB 이미지와 이에 대응되는 Depth Map으로 이루어졌고, NYU v2-40은 1,449장의 RGB 이미지와 이에 대응되는 Depth Map으로 이루어져 있다. 실험 결과, 전반적으로 Training mIOU는 높게 나타났지만, Evaluation mIOU가 낮게 나오는 Overfitting 현상이 일어났음을 볼 수 있다. 해당 현상을 완화하기 위해 Data Augmentation, Dropout, Separable Convolution [5] 등 여러 방법을 시도해보았으나 성능 향상을 얻어낼 순 없었다. 이는 본래 2D 형태였던 Image를 3D로 변환하고 Convolution Kernel도 2D에서 3D로 변환하면서 연산과정이 전반적으로 Over-paramterized된 반면, 실험에서 사용한 데이터셋의 크기가 상대적으로 작기 때문에 일어난 것으로 추정된다. 실제로, 좀 더 크기가 큰 SUN RGBD-37 데이터셋에서 Training mIOU와 Evaluation mIOU의 차이가 NYU v2-40 데이터셋에서의 차이보다 더 작음을 확인할 수 있다.

Compreesor의 타입별로 보았을 때, 단순히 Depth 방향으로 Averaging out하기 보다는 Depwth-wise Attention을 도입하여 Weighted Sum하는 방식이 더 높은 mIOU를 보임을 볼 수 있다. 이를 통해 우리는 Depth-wise Attention Mechanism을 적용할 경우 모델이 좀 더 Depth 정보를 효율적으로 이용할 수 있음을 확인할 수 있다.

Ⅲ. 결론

본 논문에서는 Semantic Segmentation Task에서 Depth Value를 효율적으로 활용하기 위해, Input 이미지를 RGB값과 Depth Map을 활용하여 3D 텐서로 변환하는 방법을 실험해보았다. 또한, 신경망이 Depth Value를 좀 더 적극적으로 활용할 수 있도록 Depth-wise Attention을 도입한 Comressor 구조를 제안하였다. 실험 결과, Depth-wise Attention을 도입한 Compressor가 다른 구조에 비해 더 높은 성능을 보여준다는 것을 확인하였으나, 3D 형태로 확장하면서 늘어난 Parameter 수로 인해 Overfitting 현상이 크게 일어남을 확인할 수 있었다. 이로 인해 Evalution mIOU에서는 좋은 성능을 내지 못하였지만, Training mIOU는 상당히 높게 기록된다는 점을 보았을 때, 추후 연구에서 더 많은 데이터를 투입하고 모델을 개선하는 등 Overfitting 현상 해결에 주력한다면 Evaluation phase에서도 좋은 성능을 보여줄 것이다.

ACKNOWLEDGMENT

This work is in part supported by Institute of Information & communications Technology Planning & Evaluation (IITP, 2021-0-00106) grant funded by the Ministry of Science and ICT (MSIT), INMAC, and BK21-plus.

참 고 문 헌

[1] Seichter, Daniel, et al. "Efficient rgb-d semantic segmentation for indoor scene analysis." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

[2] Cao, Jinming, et al. "Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[3] Lichtenberg, Samuel. "SUN RGB-D: An RGB-D Scene Understanding Benchmark Suite." (2015).

[4] Silberman, Nathan, et al. "Indoor segmentation and support inference from rgbd images." European conference on computer vision. Springer, Berlin, Heidelberg, 2012.

[5] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.