인코더-디코더 사이의 특징 융합을 통한 멀티 모달 네트워크의 의미론적 분할 성능 향상

손 찬 영　　호 요 성
광주과학기술원 전기전자컴퓨터공학부
{chanyoungsohn, hoyo}@gist.ac.kr

Improved Semantic Segmentation in Multi-modal Network Using Encoder-Decoder Feature Fusion

Sohn, Chan-Young　Ho, Yo-Sung
Gwangju Institute of Science and Technology (GIST)

요 약

Fully Convolutional Network(FCN)은 기존의 방법보다 뛰어난 성능을 보였지만, FCN은 RGB 정보만을 사용하기 때문에 세밀한 예측이 필요한 장면에서는 다소 부족한 성능을 보였다. 이를 해결하기 위해 인코더-디코더 구조를 이용하여 RGB와 깊이의 멀티 모달을 활용하기 위한 FuseNet이 제안되었다. 하지만, FuseNet에서는 RGB와 깊이 브랜치 사이의 융합은 있지만, 인코더와 디코더 사이의 특징 지도를 융합하지 않는다. 본 논문에서는 FCN의 디코더 부분의 업샘플링 과정에서 이전 계층의 결과와 2배 업샘플링한 결과를 융합하는 스킵 레이어를 적용하여 FuseNet의 모달리티를 잘 활용하여 성능을 개선했다. 본 실험에서는 NYUDv2와 SUNRGBD 데이터 셋을 사용했으며, 전체 정확도는 각각 77%, 65%이고, 평균 IoU는 47.4%, 26.9%, 평균 정확도는 67.7%, 41%의 성능을 보였다.

1. 서론

최근 영상인식 분야에서 기계학습을 통한 여러가지 방법들이 제안되고 있다. 특히, 딥러닝을 기반으로 하는 다양한 네트워크 구조들은 기존의 성능을 뛰어넘으며, 가장 좋은 성능을 내는 방법으로 발표되고 있다. Fully Convolutional Network(FCN)도 신경망 구조의 마지막 계층까지 합성곱을 사용하여 의미론적 분할(Semantic segmentation) 분야에서 인상적인 결과를 보여주었다[1]. 구조의 마지막 계층까지 합성곱 연산을 수행하며 기존의 Convolutional Neural Network(CNN)의 제한점을 극복하고 임의 크기의 입력 영상을 사용할 수 있게 되었다. 그러나 FCN은 특징 지도가 입력 영상 해상도의 1/32까지 줄어든 뒤, 업샘플링하는 과정을 포함하고 있어 세밀한 화소 단위의 객체 구분이 어려웠다.

이 단점을 해결하기 위해서 SegNet에서는 맥스풀링 시에 선택된 화소의 위치를 저장하고, 업샘플링 연산 시에 저장된 첨자를 이용하는 메모라이즈드 풀링이 제안되었다[2]. 이 방법은 FCN이 사용하던 스킵레이어의 단점을 해결하고, 보다 세밀한 업샘플링을 가능하도록 했다. 이를 확장하여 FuseNet은 인코더-디코더 구조를 사용하면서 RGB와 깊이를 융합하는 네트워크 구조를 제안했다[3]. 그러나 FuseNet은 RGB와 깊이 브랜치 간 특징 지도 융합은 수행하지만, 인코더-디코더 사이의 융합은 이루어지지 않는다.

본 논문에서는 기존에 제안된 FuseNet을 기반으로 하여, RGB와 깊이의 특징 지도 융합을 수행한다. 더불어 SegNet에서 제안한 메모라이즈드 풀링을 사용하면서, 제안하는 방법인 인코더-디코더 간의 특징 융합을 사용하여 네트워크의 의미론적 분할 성능 개선을 도모한다.

2. 제안하는 방법

2.1. 스킵 레이어

스킬레이어란 FCN에서 풀링 연산 과정에서 손실된 정보를 보완하고자 제안된 방식이다. FCN에서는 각 계층마다 합성곱과 풀링을 반복한다. 입력 이미지를 합성곱하고 풀링하는 과정을 5회 반복하는데, 4번 풀링한 특징 지도는 5번 풀링한 특징 지도보다 상세한 정보를 포함하고 있다. 즉, 더 적은 풀링연산을 할 수록, 더 세밀한 정보를 담고 있다. 이에 착안하여 4번 풀링한 특징 지도와, 5번 풀링하고 해상도를 맞추기 위해 2배만큼 업샘플링한 특징 지도와 융합한다. 이 때, 요소 합 연산을 사용하여 융합한다.

이를 다른 계층에 확장하여, 3번 풀링한 특징 지도는 역시 4번 풀링한 특징 지도보다 세밀한 정보를 포함한다. 따라서, 3번 풀링한 특징 지도를, 4번 풀링한 특징 지도와 2배 업샘플링한 특징 지도의 융합한 결과를 업샘플링한 결과와 요소 합한다.

이런 특징 지도 간 융합을 누적하여 2회 풀링한 특징 지도까지 사용하게 된다.

2.2. FuseNet

FuseNet은 인코더-디코더 구조를 가진다. 그림 1은 FuseNet의 인코더 부분을 도식화한 그림이다. 인코더는 2개 브랜치로 구성되며, RGB 인코더와 깊이 인코더로 구성되어 있다. 두 브랜치에 각각 RGB와 깊이 이미지가 들어가면 합성곱, 배치 정규화, ReLU(CBR)를 거친다. 이후 두 개의 브랜치를 요소 합을 통해 융합하고, RGB 인코더는 융합한 특징 지도에 풀링 연산을 적용하여 인코딩한다. 깊이 인코더는 CBR블록을 지난 데이터를 RGB 인코더에 전달하고, 풀링연산을 수행한 뒤 다시 CBR 블록을 수행한다. 이 과정은 각각 5번 반복된다.

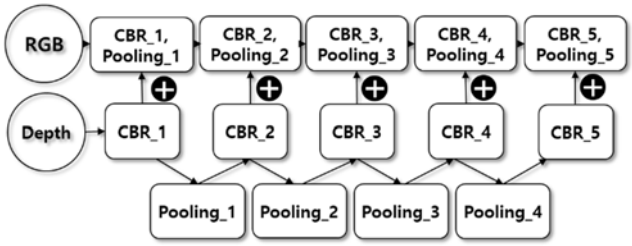

그림 1. FuseNet 의 인코더

디코더는 RGB-D 디코더로 불리며 인코더 부분과 대칭적으로 구성된다. 디코더의 입력은 인코더의 출력인 RGB와 깊이가 융합된 특징 지도이다. 인코더의 출력 해상도는 입력 영상에 비해 작기 때문에, 각 화소마다 예측을 하기 위해서 원 이미지와 동일한 크기의 해상도로 복원해야 한다. 이 때문에, 인코더의 아웃풋에 대해 언풀링을 적용하며, CBR블록을 통과한다. 언풀링 연산 시에는 인코더에서 풀링 연산을 수행할 때 저장한 위치를 사용한다.

2.3. 인코더-디코더 간 특징 융합

기존 FuseNet의 디코더는 단순히 RGB와 깊이가 융합된 특징 지도를 입력으로 한다. 그러나 이러한 방식은 RGB와 깊이 간 융합은 있지만 인코더와 디코더 사이의 특징 융합은 이루어지지 않는다. 이러한 점에 착안하여, FCN의 스킵 레이어를 활용한 인코더-디코더 간의 특징 융합을 적용한다. 그림 2는 제안된 방법의 전체 구조이며, 인코더 부분의 각 단계를 스테이지 블록으로 단순화하여 표현하였다. 각 스테이지 블록은 그림 1에서 나타난 RGB브랜치의 CBR+Pooling 블록과, 깊이 브랜치의 CBR+Pooling 블록을 하나로 나타낸 블록이다. 이 블록 내에서는 RGB와 깊이 특징 지도가 융합된다. 인코더-디코더 간 첫번째 융합은 인코더 부분에서 4회 풀링된 융합된 특징 지도를 디코더의 2회 언풀링한 특징 지도와 요소 합하는 과정이다. 이 결과를 다시 언풀링하여 인코더 부분에서 3회 풀링된 융합된 특징 지도와 융합한다. 이 과정을 디코더의 스테이지 4부터 스테이지 1까지 적용한다. 이를 통해 색 정보와 깊이 정보의 멀티 모달리티를 유지하면서 업샘플링 시 특징 융합을 통해 성능 향상을 꾀한다.

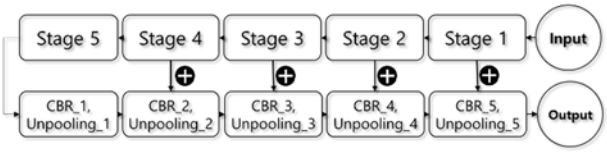

그림 2. 제안된 인코더-디코더 간 특징 융합 방법

기본적인 신경망 구조가 되는 FuseNet은 깊이 입력, HHA 입력을 기반으로 하므로 각 입력에 따른 성능 비교를 고려할 수 있다. 또한, 기존의 FCN에서는 특징 지도 간 융합을 수행할 때, 첫번째 풀링 결과의 특징 지도는 융합하지 않았다. 따라서 본 논문에서도 스테이지 1부터 스테이지 4까지 각각 대칭되는 디코더 부분과 융합하는 모든 융합 방법과, 스테이지 1 부분을 제외한 스테이지 2부터 스테이지 4까지 대칭되는 디코더 부분과 융합하는 일부 융합 방법을 고려한다.

3. 실험

3.1 실험 환경 및 지표

본 실험에서 데이터 셋은 NYUDv2와 SUNRGBD를 사용한다. 클래스의 개수는 NYUDv2는 13개, SUNRGBD는 38개이다. 그림 3은 SUNRGBD의 색상표이며, 실내 환경에 대한 객체를 분류한다. 성능 측정 지표는 평균 IoU, 평균 정확도를 사용하였다. 전체 정확도는 올바르게 분류된 화소의 비율을 의미하며 다음과 같이 정의된다.

$$\textit{전체 정확도} = \frac{1}{N}\sum_{c} TP_c, c \in \{1, \dots, K\} \quad (1)$$

다음으로, 평균 정확도는 클래스 단위의 정확도의 평균이며 다음과 같이 정의된다.

$$\textit{평균 정확도} = \frac{1}{K}\sum_{c} \frac{TP_c}{TP_c + FP_c} \quad (2)$$

마지막으로 평균 IoU는 예측한 화소와 Ground Truth 간 교집합을 합집합으로 나눈 수치이며, 다음과 같이 정의된다.

$$\textit{평균 IoU} = \frac{1}{K}\sum_{c} \frac{TP_c}{TP_c + FP_c + FN_c} \quad (3)$$

FuseNet의 기본 모델 및 인코더-디코더 간 특징 융합 방법 실험 시, 깊이를 HHA로 인코딩하여 사용하였다[4]. 이 때, RGB, 깊이, HHA 영상을 입력으로 사용할 때, 성능향상을 위해 정규화를 사용하였다. 정규화는 모든 화소 값을 학습 데이터 셋 이미지 전체의 화소값의 평균으로 나누었다.

학습 장비는 NVIDIA Geforce GTX 1080 Ti with 11GB RAM을 사용했다. 또한 학습 시에 확률적 경사도 하강 방법을 사용했으며, 학습속도, 모멘텀, 가중치 감소값은 각각 10^{-5}, 0.99, 5^{-4}를 사용했다. 에폭 수는 100을 사용하였고, 총 학습시간은 약 1일 정도 소요되었다.

그림 3. SUNRGBD 데이터셋의 클래스에 따른 색상표

3.2 실험 결과

FCN의 스킵 레이어에서는 가장 마지막 단계의 풀링 레이어에서 특징 지도의 요소 합을 수행하지 않는다. 따라서 본 실험에서도 인코더 스테이지 1과 디코더 마지막 스테이지 간 특징 융합 여부에 따른 성능 차이를 알아보기 위한 실험을 진행했다.

표 1은 NYUDv2 데이터 셋의 일부 융합 방법과 모두 융합한 방법을 실험한 결과이다. 이때, 입력 영상 중 깊이 영상은 깊이 값과, 정규화 되지 않은 HHA 인코딩 값을 사용하였다. 마지막 스테이지까지 융합을 사용한 모두 융합 방법이 평균 IoU, 평균 정확도에서 일부 융합보다 상대적으로 나은 성능을 보였다. 따라서 이후 실험은 모두 융합한 구조를 사용하여 진행하였다. 또한, 입력 영상 중, 깊이 영상에 대해서는 정규화하지 않은 HHA 인코딩과 깊이 영상 중에, 정규화하지 않은 HHA 인코딩을 사용하는 것이 평균 IoU를 기준으로 약 1.5% 나은 성능을 보이는 것을 확인했다. 결과적으로 NYUDv2 데이터셋에서, 입력 영상을 정규화되지 않은 HHA가 아닌 깊이를 사용하고, 마지막 스테이지까지 융합한 방법이 가장 나은 성능을 보였다.

표 1. 모두 융합과 일부 융합 방법에 따른 성능 비교

NYUDv2	HHA (정규화 안함)		깊이	
	모두 융합	일부 융합	모두 융합	일부 융합
전체 정확도	0.763	0.764	**0.773**	0.761
평균 IoU	0.468	0.432	**0.483**	0.466
평균 정확도	0.668	0.657	**0.677**	0.656

표 2에서 볼 수 있듯이, 정규화하지 않은 HHA를 사 용하는 것 보다 정규화된 HHA를 입력으로 사용하는 것이 더 좋은 성능을 발휘했다. 결과적으로 NYUDv2, SUNRGBD 데이터 셋에 대하여, 정규화된 HHA, 깊이, 정규화하지 않은 HHA 순으 로 성능이 좋은 것을 확인했다. 또한, 정규화된 HHA를 입력으로 할 때의 제안된 방법과 기존 FuseNet을 비교하면, 제안된 방법을 사용할 때 더 나은 성능을 보였다. 결과적으로 정규화된 HHA를 사용하며 제안된 인코더-디코더 간 특징 융합 방법을 사용하면 최선의 성능을 발휘할 수 있다. 평균 IoU를 기준으로, 제안된 방식과, FuseNet을 비교하였을 때, 정규화한 입력을 사용하게 되면 SUNRGBD에서 약 2.3%정도 차이를 보인다.

표 2. 제안된 방식과 FuseNet 성능 비교 표

데이터 셋	측정	제안된 방식		FuseNet	
		정규화	정규화 안함	정규화	정규화 안함
NYUDv2	전체 정확도	**0.77**	0.763	0.759	0.758
	평균 IoU	**0.474**	0.468	0.468	0.462
	평균 정확도	**0.677**	0.668	0.668	0.648
SUNRGBD	전체 정확도	0.65	**0.651**	0.64	0.641
	평균 IoU	**0.269**	0.256	0.246	0.264
	평균 정확도	**0.41**	0.378	0.361	0.378

그림 4는 SUNRGBD 데이터 셋으로 실험한 결과의 일부를 시각화한 결과다. FuseNet보다 제안된 방법을 사용할 때, 객체 내부의 색상을 보다 잘 추론한 것을 확인했다. 첫번째 행에서 볼 수 있듯이, FuseNet의 경우 추론한 오브젝트의 내부가 하나의 색으로 분명하게 나타나지 않는다. 그러나 제안된 방법에서는 FuseNet 결과와 비교하였을 때, 하나의 오브젝트 내부를 하나의 색으로 추론한 결과를 볼 수 있다.

또한, 두번째 행의 각 영상을 비교하면, 보라색으로 표현되는 의자 클래스를 제안된 방식이 잘 예측한 것을 확인할 수 있다. 입력 이미지에서 피아노 앞에 위치한 의자를 Ground Truth에서 검정색으로 표시하였지만, 제안된 방법에서는 보라색으로 잘 추론하였으며, 피아노 우측의 의자 또한 기존 FuseNet에 비해 정확하게 예측했다.

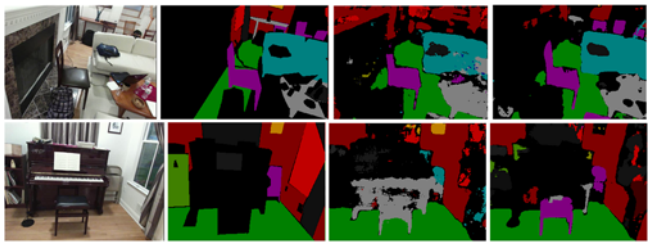

그림 4. SUNRGBD 데이터 셋 실험 시각화 결과 일부

4. 결론

본 논문은 FCN에서 업샘플링할 때 예측의 성능을 높이기 위해 적용된 스킵 레이어를 사용하였다. 이 방법은 적은 풀링 연산을 수행한 특징 지도일수록 더 세밀한 정보를 가진다는 점에 착안하였으며, FuseNet의 인코더와 디코더 사이에 특징 지도 융합에 적용했다. 제안된 방식으로 3개 스테이지에 대한 융합과, 4개 스테이지에 대한 융합을 실험한 결과, 기존의 FuseNet보다 평균 IoU, 평균 정확도가 상승한 것을 확인할 수 있었다. 또한 인코더-디코더 간 특징 융합을 하지 않은 FuseNet보다, 제안된 방법을 적용할 때 성능이 향상되는 것을 확인할 수 있었다. 결과적으로 4개 스테이지에 대한 융합을 수행하며, 정규화환 HHA를 입력으로 사용하여 학습할 때 가장 좋은 성능을 보였다.

감사의 글

본 논문은 민 · 군기술협력사업(Civil-Military Technology Cooperation Program)으로부터 지원을 받아 수행된 연구임

참 고 문 헌

[1] J. Long, E. Shelhamer, and T. Darrell, " Fully Convolutional Networks for Semantic Segmentation," CVPR, 2015.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, " SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," arXiv preprint arXiv:1511.00561, 2015.

[3] C. Hazirbas, L. Ma, C. Domokos and D. Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture," ACCV, 2016.

[4] S. Gupta, R. Girshick, P. Arbelaez, J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," ECCV, 2014.