J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

공간 클래스 단순화를 이용한 의미론적 실내 영상 분할☆

Semantic Indoor Image Segmentation using Spatial Class Simplification

김 정 환[1] 최 형 일[1*]
Jung-hwan Kim Hyung-il Choi

요 약

본 논문에서는 실내 공간 이미지의 의미론적 영상 분할을 위해 배경과 물체로 재설계된 클래스를 학습하는 방법을 제안한다. 의미론적 영상 분할은 이미지의 벽이나 침대 등 의미를 갖는 부분들을 픽셀 단위로 나누는 기술이다. 기존 의미론적 영상 분할에 대한 연구들은 신경망을 통해 이미지의 다양한 객체 클래스들을 학습하는 방법들을 제시해왔고, 긴 학습 시간에 비해 정확도가 부족하다는 문제가 지적되었다. 그러나 물체와 배경을 분리하는 문제에서는, 다양한 객체 클래스를 학습할 필요가 없다. 따라서 우리는 이 문제에 집중해, 클래스를 단순화 후에 학습하는 방법을 제안한다. 학습 방법의 실험 결과로 기존 방법들보다 정확도가 약 5~12% 정도 높았다. 그리고 같은 환경에서 클래스를 달리 구성했을 때 학습 시간이 약 14 ~ 60분 정도 단축됐으며, 이에 따라 물체와 배경을 분리하는 문제에 대해 제안하는 방법이 효율적임을 보인다.

☞ 주제어 : 의미론적 영상 분할, 실내 공간 구조, 기계 학습

ABSTRACT

In this paper, we propose a method to learn the redesigned class with background and object for semantic segmentation of indoor scene image. Semantic image segmentation is a technique that divides meaningful parts of an image, such as walls and beds, into pixels. Previous work of semantic image segmentation has proposed methods of learning various object classes of images through neural networks, and it has been pointed out that there is insufficient accuracy compared to long learning time. However, in the problem of separating objects and backgrounds, there is no need to learn various object classes. So we concentrate on separating objects and backgrounds, and propose method to learn after class simplification. The accuracy of the proposed learning method is about 5 ~ 12% higher than the existing methods. In addition, the learning time is reduced by about 14 ~ 60 minutes when the class is configured differently In the same environment, and it shows that it is possible to efficiently learn about the problem of separating the object and the background.

☞ keyword : Semantic image segmentation, Indoor space structure, Machine Learning

1. 서 론

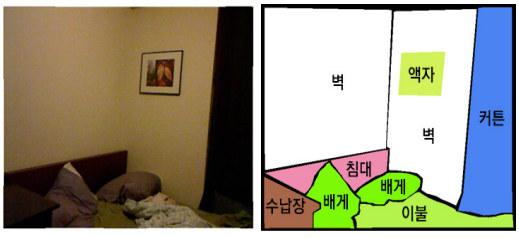
현재 인공지능의 발전 속도 상승으로 인해 기존 IT분야의 소프트웨어와 하드웨어가 본래 보유한 알고리즘에 기계 학습 기능을 더하면서 성능을 더욱 끌어올리고 있다. 따라서 과거에는 구현이 불가능했던 기술이 현재에 이르러 상용화가 가능할 정도로 개발이 되어 서비스를 제공하고 있다. 콘텐츠 분야에서도 글자와 사진, 동영상을 뛰어 넘어 직접 유사 체험이 가능한 AR(Augmented Reality)이나 VR(Virtual Reality)과 같은 기술을 사용한 콘텐츠가 주목받으며 그 중요성이 날이 갈수록 증가되고 있는 추세이다. 다국적 가구 및 인테리어 전문 기업 IKEA에서는 'IKEA Place'라는 모바일 AR 앱을 상용화하여 서비스 하고 있는데, 이 앱은 사용자가 실내 공간을 촬영해 AR 기술을 사용하여 직접 가상의 가구를 배치하는 시뮬레이션을 수행할 수 있다. 그리고 카메라 화면상에서 다음 방향을 제시해주는 실내 내비게이션 기능의 AR 앱도 여럿 존재한다. 이러한 AR과 VR은 컴퓨터 비전 기술을 통해 실내 공간 정보를 활용하여 다양한 방식으로 정보를 제공할 수 있기 때문에 그 바탕이 되는 실내 공간에 대해 인식하고 중요한 정보를 추출하는 기술이 중요하다.

컴퓨터 비전 분야에서 실내 공간에 대한 연구는 과거부터 현재까지 꾸준히 진행되고 있는 주제이다. [1]은 다중 영상을 활용해 실내 공간상의 소실점을 추출하여 실

1 School of Media, Soongsil Univ, Seoul, 156-881, Korea
* Corresponding author (hic@ssu.ac.kr)
[Received 5 November 2018, Reviewed 20 November 2018(R2 22 February 2019, R3 11 April 2019), Accepted 17 May 2019]
☆ 본 연구는 2017년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업입니다.
(NRF-2017R1D1A1B03034114)

내 구조 Layout을 예측했다. [2]는 객체 검출을 통해 사전에 학습된 부분을 바탕으로 실내 공간을 분석하여 화장실, 사무실, 학교 등 어떤 장면의 사진인지 판단했다. [3]은 실내 공간 이미지에서 객체의 특징을 추출해 객체별로 나누어 알고리즘을 바탕으로 이미지를 의미론적 영상 분할하였다. 이러한 방법들은 과거 다양한 알고리즘을 통해 발전해 오다가, 최근에는 기계학습과 연계되면서 성능을 끌어올리는 연구들이 활발하다.

실내 공간 정보를 분석하는 방법 중의 하나인 의미론적 영상 분할(Semantic Image Segmentation)은 그림 1과 같이 이미지에서 벽이나 침대, 액자와 같이 의미를 지니고 있는 부분을 픽셀 단위로 분할하는 기술로, 원본 이미지로부터 새로운 형태의 객체 맵(Object Map)을 생성할 수 있어서 실내, 실외 구분 없이 이미지를 '이해'하는 문제에 대해 폭 넓고 효과적이어서 AR 분야에도 적용 가능하다. 현재 의미론적 영상 분할은 CNN(Convolutional Neural Network)과 같은 깊은 신경망 기반의 학습을 통한 문제 해결이 돌파구로 떠오르고 있으며, 그 정확도 또한 발전하는 중이다.



(그림 1) 의미론적 영상 분할 예시
(Figure 1) Example of Semantic image segmentation

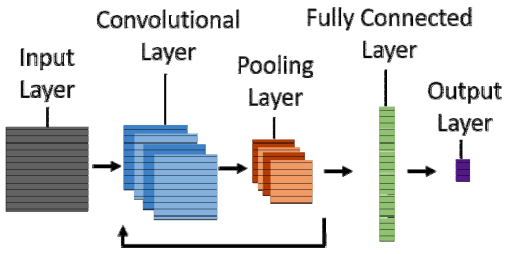
앞서 언급한 실내 기반 AR 앱에서 서비스 중인 기능들은 대부분 독자적인 알고리즘과 GPS 등에 의한 공간 구조 분석에 의해 이루어진 것으로, 고화질 동영상에 대한 처리 속도나 정확도에 대한 문제가 다소 존재한다. 물론 기계 학습을 적용한 경우에도 같은 문제가 발생할 수 있으나, 실제 적용을 가정해본다면 다르다. 실제 제공 중인 서비스에서 최우선적으로 필요한 요소는 촬영한 화면상에서 무엇이 배경이고 물체인지 빠르고 정확하게 인식하는 것이고, 그 이후에 가상 시뮬레이션을 제공해야 한다. 즉, 구체적으로 어떤 물체인지까지는 학습할 필요가 없다. 이 경우에는 학습할 특징이 줄어 학습 시간과 정확도 향상을 기대해볼 수 있다.

따라서 현재 지속적인 성장을 보여주고 있는 기계 학습 기술을 접목시키되, 실내 공간상에서 배경과 물체의 분리에 대해 학습시킨다면 앞의 사례와 같은 경우에 높은 효율을 보일 것이다. 또한 레이아웃 검출에 방해가 되는 물체를 배제할 수 있어서 정확도를 향상시킬 수 있고, 이를 응용해 레이아웃을 3D로 재구축하는 문제에도 효율적으로 작용할 수 있다. 본 논문에서는 '배경'과 '객체'의 분리에 집중하여, 이미지 안의 객체의 클래스를 단순화한 후 신경망 학습을 통해 의미론적 영상 분할을 수행하여 기존의 방법보다 특정 경우에 효율적이고 성능을 향상시킬 수 있는 방법을 제안한다.

2. 본 론

2.1. CNN 관련 기존 연구

최근 의미론적 영상 분할을 위해 CNN을 활용한 학습 네트워크를 설계해 학습하는 방법이 주로 쓰이고 있다. 기존 DNN(Deep Neural Network)을 사용해 이미지를 학습시키면 이미지 고유의 2차원 공간 정보가 사라지는 문제에 대해 해결책으로 떠오른 학습 방법이다.



(그림 2) 초기 CNN 구조 예시
(Figure 2) Example of CNN Structure

그림 2는 [4]에서 제안한 초기 CNN인 AlexNet의 구조를 간략화 한 그림이다. 그림의 구조와 같이 CNN은 이미지를 입력 데이터로 받아 최종적으로 어떤 클래스인지 판별해내는 학습 신경망이다. CNN이 기존 DNN과 다른 부분은 입력 이미지의 공간 정보를 유지하기 위해 영상처리 기법인 convolution을 이용한 계층과 이미지의 크기를 줄이는 pooling 계층을 삽입한 점이다. Convolution에 사용되는 필터는 구성된 가중치(weight)에 따라 에지를 검출하거나 이미지에서 고주파나 저주파 성분을 제거할 수 있으며 특정 패턴을 가진 부분을 검출해낼 수 있다. 그 크기는 3x3, 5x5, 11x11 등 크기가 다양하고 크기가 클

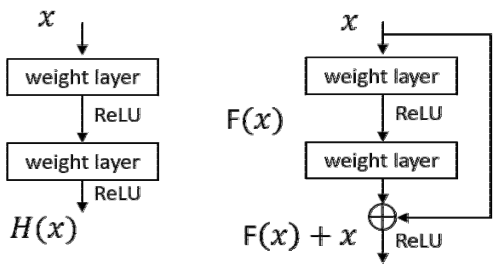
수록 넓은 범위에서 필터에 해당하는 패턴을 추출해낼 수 있다. CNN은 필터의 가중치를 학습시켜 특정 패턴을 검출해낼 수 있도록 설계되었고 학습시킨 필터에 입력 이미지를 convolution 연산했을 때, 학습된 패턴과 유사할 수록 높은 가중치를 가진 특징 맵(Feature map)이 출력된다. 이후 반복해서 여러 필터를 거친 값들이 완전 연결 계층(Fully-Connected Layer)에서 종합되어 최종적으로 결과인 클래스가 출력되는 구조이다. 실제로 AlexNet은 convolution과 pooling 계층을 여러 번 반복하여 총 8계층으로 신경망을 구성하여 풍부한 특징을 학습할 수 있도록 했다.

초기 CNN 구조인 AlexNet이 제시된 이후 해마다 더 신경망을 깊게 하여 성능을 향상시킨 연구가 발표되면서, 신경망을 깊게 구성할수록 학습시킬 가중치가 비례적으로 늘어나게 되었다. 그로 인해 한정된 데이터셋에만 제대로 학습이 되는 현상인 과적합(Overfitting)뿐 아니라 학습 오류 또한 발생할 가능성이 증가된다는 문제가 제기되었다.

[5]는 이에 대한 문제의 해결책으로 잔여 학습(Residual Learning)의 개념을 제시했다. 잔여 학습의 핵심적인 의미는 그림 3과 같이 표현할 수 있으며, 왼쪽 그림은 보통의 CNN, 오른쪽 그림은 잔여 학습이 적용된 CNN으로 각각 신경망의 일부를 나타낸다. 왼쪽 신경망에서는 x를 입력받아 2개의 계층과 활성화 함수 ReLU를 거쳐 H(x)를 출력으로 내며, 학습을 통해 최적의 H(x)를 얻는 것이 목표이고 weight layer의 가중치도 그에 맞게 결정되어야 한다. 하지만 왼쪽 신경망처럼 H(x)를 얻는 것이 목표가 아니라 H(x) - x를 얻는 것으로 목표를 수정한다면, 입력과 출력간의 차이를 얻도록 학습하게 된다. 2개의 weight layer가 H(x) - x를 구하도록 학습이 되어야 할 때 F(x) = H(x) - x라면, 출력 H(x) = F(x) + x가 된다. 그로 인해 그림 3의 왼쪽 신경망은 오른쪽 신경망의 구조와 같이 바뀔 수 있고, 이 형태가 Residual Learning의 기본 구조가 된다. 왼쪽 구조와 비교했을 때 변화한 점은 입력에서 출력으로 바로 연결된 shortcut이 생겼고, 연산량의 측면에서는 덧셈이 추가되는 것 이외에는 차이가 없다.

이로 인해 기존에는 H(x)를 구하기 위해 학습했다면 이제 H(x) - x를 구하도록 학습하게 되며, 최적의 경우에는 F(x)가 0이 되기 때문에 학습할 방향이 미리 결정되어 효율적인 연산이 가능하다. 또한 입력과 같은 x가 그대로 출력에 연결이 되기 때문에 파라미터의 수에 영향이 없으며, 몇 개의 계층을 건너뛰면서 입력과 출력이 연결되기 때문에 학습이 간단해지는 효과를 얻을 수 있다. 그 결과로 간단히 깊은 신경망의 최적화가 가능해지고, 깊어진 신경망으로 인해 더 깊고 세밀한 특징 필터를 학습할 수 있게 되었다. 추가적으로 convolution에 사용되는 5x5 크기의 필터를 3x3 크기의 필터 2개로 대체하는 등의 필터를 분해하는 기술을 추가하였는데, 이 또한 연산량을 줄이는 동시에 정확도를 높이는 역할을 했다.

[5]에서 발표된 ResNet은 잔여 학습의 제시와 필터의 분리를 효과적으로 활용하여 기존 방법보다 이례적으로 더 많은 계층인 100계층 이상의 신경망 학습이 가능하도록 설계한 구조를 발표했고, 오류율이 3.58% 수준으로 감소하게 되어 뛰어난 성능을 보였고, 본 논문에서는 이 ResNet구조를 활용하여 학습 신경망을 구성했다.



(그림 3) 잔여 학습의 구조 예시
(Figure 3) Example of residual learning structure

2.2. 의미론적 영상 분할 관련 기존 연구

의미론적 영상 분할을 수행하기 위해서 CNN을 그대로 사용하면 문제가 발생한다. 보통 CNN 후반부에는 완전 연결 계층이 존재하여 모든 데이터가 연결되어 종합되기 때문에 2차원 이미지의 위치 정보가 최종적으로 사라지기 때문이다.

[6]에서는 CNN상에서 완전 연결 계층에 도달하기 전 얻어진 정보에 이미 분류가 가능할 정도의 충분한 특징 패턴이 있고, 그 위치에 대한 정보도 지금까지 convolution 및 pooling만을 거쳤기 때문에 유지하고 있다는 점에 집중했다. 그래서 마지막 계층인 완전 연결 계층을 없애고 전부 convolution 및 pooling 계층으로 구성한 FCN(Fully Convolutional Layer)를 발표했다. 그러나 CNN의 특성 상 convolution과 pooling을 거치게 되면 특징 맵의 크기가 줄어들게 된다. 원본 이미지의 크기와 같은 사이즈로 픽셀 단위의 세밀한 영상 분할을 하려면 줄어든 특징 맵의 결과를 다시 키우는 과정을 거쳐야 하는데, 그 과정을

Up-scale 등으로 부른다.

가장 간단한 Up-scale 방법은 양선형 보간법(Bilinear Interpolation)을 수행해 크기를 늘리는 방법이 있으나, 그 방법만을 사용한다면 원본을 복원하기엔 세밀함이 떨어진다. 따라서 FCN에서는 세밀한 정보를 보강하기 위해 각 convolution 계층별로 남아있는 pooling이 수행되기 이전에 조금 더 큰 크기의 중간 결과(특징 맵)를 참고하여 원본의 세밀한 부분을 살려 정교하게 예측하고자 했다.

그러나 이후 원본 크기로 복원한 영상 분할의 세밀함이 부족하다는 문제가 제기되면서 그 이후 해마다 성능을 향상시킨 방법을 제안하는 논문들이 다수 발표되었다. [7]은 데이터셋에 포함되어 있는 깊이 맵의 정보를 같이 학습하여 두 학습의 결과를 합친 방법을 사용했고 [8]은 Deeplab과 ResNet 101 계층 버전을 조합하여 학습했다. [9]는 pooling시에 핵심적인 부분의 매핑(Mapping) 정보를 저장하고 활용하여 세밀한 복원을 하도록 구성한 신경망을 구성했고, [10]은 잔여 학습을 활용한 Encoder-Decoder 구조를 각 사이즈별로 설계하는 등의 연구들이 진행되었다. 이와 같이 최근 발표된 학습을 통한 의미론적 영상 분할 분야의 연구는 주로 ResNet과 같은 신경망을 기반으로 하거나 새로운 신경망 구조로 대체하는 연구, 추가적으로 영상 분할 맵을 세밀하게 복원하는 방법을 제안하는 추세이다.

본 논문에서는 최근까지 계속 업그레이드 하면서 성능을 향상시킨 구조인 Deeplab[11, 12]을 본 논문에서 활용한다. Deeplab은 핵심적으로 Atrous Convolution, ASPP (Atrous Spatial Pyramid Pooling), Encoder - Decoder 구조 등을 이용해 성능을 향상시켜왔다. Atrous convolution이란 convolution시에 필터 일부분만 사용하고 나머지는 0으로 채워 연산하는 방법으로, 학습시킬 필터의 가중치 개수가 줄어드는 효과를 얻을 수 있다. 그로 인해 기존에 연산량 때문에 적용하지 못했던 큰 크기의 필터를 사용할 수 있는 이점을 얻었다. ASPP는 Atrous convolution을 활용해 여러 크기의 필터를 연산하고 이를 다시 하나의 특징 맵으로 합쳐주는 방법이며, 더 넓은 범위의 특징을 연산의 증가 없이 검출할 수 있게 되었다[10]. Encoder - Decoder 구조는 Encoder 부분에 CNN을 배치하여 중간에서 최종적 특징 맵을 추출한다. Decoder 부분에는 Encoder와 대칭이 되는 신경망을 배치하고 대칭되는 각 convolution 계층에 남아있던 특징 맵을 참고로 하여 Up-Scale하게 하여 보다 세밀한 영상 분할에 초점을 두었다. 또한 추가적으로 pooling 계층을 일부 삭제해서 특징 맵이 갈수록 줄어드는 부분을 줄였다[11]. 이러한 요소들로 인해 Deeplab은 현재 우수한 의미론적 영상 분할 성능을 보여주고 있다.

2.3. 제안하는 방법

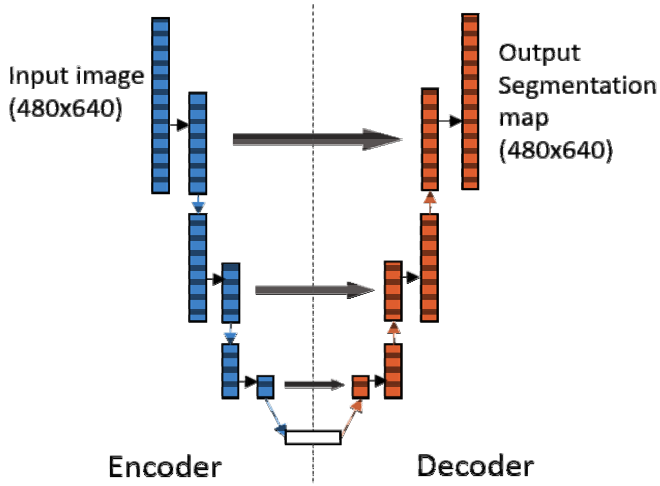
기존 방법들은 보통 의미론적 영상 분할을 수행할 때 비슷한 특성을 지닌 클래스를 묶어 다양한 항목들을 학습한다. 하지만 앞서 설명한 AR 앱의 경우처럼 실내 공간 이미지를 배경과 물체만으로 나누어 응용하는 실내 공간 정보 분석 문제의 경우에, 그 물체가 어떤 것인지 알 필요가 없다. 따라서 본 논문에서는 배경과 객체를 분리하는 문제에 집중하여, 클래스를 배경과 물체로 대폭 줄여 학습시킨다.

(표 1) 제안하는 의미론적 영상 분할 알고리즘
(Table 1) Proposed semantic image segmentation algorithm

1. 실내 이미지(480x640)를 Encoder(ResNet)에 입력해 convolution 및 pooling 수행
2. {배경, 물체, 기타}로 분류된 각 클래스 별로 Encoder의 최종 특징 맵 획득
3. 특징 맵을 원본 크기로 복원하기 위해 Decoder에 입력하여 Up-Scale 수행
4. 정답 영상 분할 맵과 비교하여 오차를 측정한 후 오차를 줄이도록 Encoder, Decoder의 필터 값 갱신
5. 1~4번 항목을 정해진 횟수만큼 반복

'배경' 클래스에는 천장, 바닥, 벽 클래스를 할당했고, '물체' 클래스에는 배경을 제외한 물체 클래스들, 그리고 그 외에 분류되지 않은 부분을 '기타' 클래스를 할당하여 총 3개의 클래스로 설계했다. 그 결과로 Encoder와 Decoder에 입출력되는 Feature 맵의 종류가 대폭 감소하기 때문에 기존의 다양한 클래스로 구성하는 방법보다 연산 시간이 감소하는 효과 또한 기대할 수 있다.

신경망 구조로는 앞서 설명한 Deeplab 팀의 가장 성능이 좋았던 최신 버전인 v3+ 버전을 활용했다. ResNet의 100 계층 이상 깊은 망을 가진 버전들은 크기가 작은 데이터셋에 대해 과적합이 발생할 수 있기 때문에 Encoder에 사용되는 CNN 구조는 ResNet의 50 계층 버전을 사용했다. 학습에 사용된 Hyper-Parameter인 활성화 함수, 손



(그림 4) 제안하는 의미론적 영상 분할 전체 구조
(Figure 4) A brief structure of proposed semantic image segmentation

실 함수, learning rate는 여러 실험을 진행한 결과 본 학습에서는 다른 값들이나 함수와 큰 차이가 없어 이후 표 1의 설명할 부분대로 전반적으로 준수한 성능을 보이는 함수와 값을 사용했다. 표 1은 본 논문에서 제안하고자 하는 의미론적 영상 분할 방법 알고리즘을 순서대로 표현한 것이고, 그림 4는 그 알고리즘 구조를 그림으로 간략하게 표현한 것이다.

표 1의 1번에서는 480x640 크기의 실내 공간 이미지를 Encoder인 50 계층의 ResNet에 입력하여 convolution과 pooling을 수행한다. 이 과정에서 Atrous Convolution이 여러 크기별로 수행되어 이미지는 다양한 크기의 필터에서 얻어진 특징 맵으로 변하게 된다. 또한 pooling 전 중간 과정에 활성화 함수 ReLU가 삽입되어 입력된 값을 훼손하지 않도록 도움을 주었다. 그림 4에서 왼쪽 부분의 입력 데이터가 점점 작아지는 부분이 이에 해당한다.

표 1의 2번에서는 여러 크기의 필터를 거쳐 얻어진 특징 맵들에는 각 클래스에 해당하는 부분에 값이 크게 남아있다. 이때 각 필터 별로 데이터를 합치는 ASPP 방법이 사용되어, 종합적으로 각 클래스 필터 별로 다르게 색이 칠해진 최종 특징 맵들을 얻는다. 이 최종 특징 맵들을 하나로 합치게 되면 작은 크기의 초기 영상 분할 맵이 생성된다. 이 부분은 그림 4의 중앙 하단에 위치한 흰색 네모 부분에 해당되며, Encoder의 출력 값이자 Decoder의 입력 값이다.

표 1의 3번에서는 초기 영상 분할 맵을 원본 크기로 늘리기 위해 Encoder와 대칭이 되는 Decoder에 입력하여 Up-Scale을 수행한다. 그 과정에서 up-convolution을 수행하는 것 뿐 아니라 동시에 Encoder에 각각 남아있는 중간 결과인 특징 맵을 참고하여 원본만큼 크기를 증가시켜서 최종적으로 원본 크기와 같은 영상 분할 맵을 출력한다. 그림 4의 오른쪽 부분에서 올라갈수록 점점 크기가 커지는 것이 up-convolution 과정이며, Encoder와 연결된 회색 화살표는 Encoder의 남아있는 특징 맵을 참고하는 것을 표현한다.

표 1의 4번에서는 최종적으로 예측된 영상 분할 맵과 정답 영상 분할 맵을 비교해 오차를 측정하는데, Cross-entropy 기법을 사용한다. 첫 번째 수행일 경우 Encoder와 Decoder의 필터들이 학습되어있지 않기 때문에 큰 오차가 발생할 것이다. 그리고 오차를 줄이는 방향으로 모든 필터들의 가중치를 갱신하는데, 이때는 설정해둔 learning rate에 영향을 받는다. 본 논문에서는 0.0001을 적용해서 learning rate를 설정했다. 표 1의 5번에서는 일련의 과정을 정해둔 횟수만큼 반복하며 오차를 줄이는 학습을 한다.

3. 실험 결과 및 결론

3.1. 실험 환경

실제 학습과 테스트에 사용된 컴퓨터의 CPU는 Intel i5-4690 CPU 3.5GHz, 메모리는 8Gb, VGA는 NVIDIA

(원본 영상) (정답 값) (예측 값)

(그림 5) NYU Depth v2 데이터셋 원본 영상, 정답 값 및 실제 예측 값 예시

(Figure 5) NYU Depth v2 Dataset : Original image, Ground truth and Prediction

GeForce GTX 1050 Ti이며, OS는 Windows 10 Pro 64-bit 환경에서 학습을 수행하고 테스트했다.

실험에는 NYU Depth V2 데이터셋[13]을 사용했고, 이 데이터셋은 480x640 크기의 실내 공간 이미지를 벽, 사람, 침대, 컴퓨터 등의 894 클래스로 구분하여 픽셀단위로 의미기반 영상 분할된 이미지를 정답 값으로 가지고 있다. Depth 정보가 있으나, 실험에서는 사용되지 않았다. 총 1,449장의 이미지로 구성이 되어 있으며, 실험에서 Training image를 889장, Validation image를 224장, Test image를 336장으로 구성하여 학습했다.

학습과 테스트에 사용된 신경망은 ResNet의 50 Layer 구조를 사용한 Deeplab의 v3+ 버전을 사용했고 Tensorflow로 구현한 소스코드를 참조했다. Hyper-Parameter 설정은 이미지 crop size를 256 x 256로 설정하였고, 손실 함수로는 cross_entropy, 활성화 함수를 ReLU, Learning rate는 0.0001에 Adam 기법을 적용해 학습했다. 참조한 소스 코드의 링크는 다음과 같다.

(https://github.com/GeorgeSeif/Semantic-Segmentation-Suite)

3.2. 실험 결과

정확도를 검사할 때 IoU(Intersection over Union)라는 기준을 사용했는데, 이는 실제 정답 값과 예측 값을 비교하여 측정하는 방법으로 수식 1과 같이 연산되어 정확한 위치에 영상 분할했는가에 대한 정확도를 검사한다. A와 B는 정답 값과 예측 값을 의미하며 둘의 교집합/합집합을 수식화한 것이다. 그 외에 픽셀 정확도는 예측한 영상 분할 맵을 정답 영상 분할 맵과 비교해서 올바른 클래스로 분류된 픽셀의 비율을 측정한 정확도이다. 정확도 비교에 참조된 [7, 8, 9, 10]은 NYU Depth v2 데이터셋을 이용해 클래스는 40개로 학습 후 의미론적 영상 분할을 수행한 실험 결과를 발표했고, 각 논문의 가장 성능이 좋았던 결과를 토대로 비교했다.

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (1)$$

표 2는 제안하는 방법으로 학습 후, 의미론적 영상 분할을 테스트 했을 때의 정확도를 기존 방법과 비교한 결과이다. 각각 픽셀 정확도, 평균 IoU, 그리고 학습에 사용된 기반 구조를 포함하고 있으며, 기반 구조의 앞 영문은 사용된 신경망 명칭의 약자이고 뒤의 숫자는 신경망 계층의 수이다. 제안하는 방법으로 클래스를 단순화하여 학습을 수행했을 때 평균 IoU 수치와 픽셀 정확도가 각각 51.9%와 80.5%로 기존 방법들보다 정확도가 높았다.

(표 2) 의미론적 영상 분할 정확도 테스트 결과 비교
(Table 2) Comparison of semantic segmentation accuracy test results

방법	픽셀 정확도	평균 IoU	기반 구조
[7]	69.0	39.8	ResNet-152
[8]	70.9	41.8	ResNet-101
[9]	-	45.9	VGGNet-16
[10]	73.6	46.5	ResNet-152
Ours	**80.5**	**51.9**	**ResNet-50**

표 3은 같은 실험 환경에서 클래스를 달리 구성하여 각 1회씩 학습을 했을 경우의 소요 시간을 나타낸다. 클래스의 개수를 원본대로 894개와 기존 연구된 논문들이 구성했던 40개, 본 논문의 클래스 3개로 설정하여 수행한 학습 시간을 비교했을 때, 제안하는 방법이 각각 약 60분, 14분 정도 적게 소요되었다.

그림 5는 NYU Depth v2 데이터셋의 샘플 이미지들로, 원본 이미지와 정답 값인 영상 분할된 이미지와 실제 논문에서 제안한 방법으로 학습하고 테스트한 이미지의 예시이다.

(표 3) 클래스 개수별 학습 시간 비교
(Table 3) Comparison of learning time by class number

클래스 개수	학습 횟수	학습시간
894	1	63분 11초
40	1	16분 52초
3	**1**	**2분 24초**

3.3. 결론

본 논문에서는 세밀하게 각 객체 별로 분류하는 것이 아닌 객체와 배경으로 구별하는 문제에 집중하여, 실내 공간 이미지에 대해 ResNet 기반 Deeplab 구조를 활용한 의미론적 영상 분할 학습 네트워크를 3 클래스로 구성해서 학습을 수행했다. 실험 결과 기존 방법보다 평균 IoU와 픽셀 정확도 수치를 비교했을 때 높은 수치를 보였고, 따라서 제안하는 방법이 수치상으로 기존 방법보다 더 정확하게 의미론적 영상 분할을 수행했음을 보였다.

또한 클래스의 수를 줄였기 때문에 학습에 걸리는 시간을 크게 단축해서 높은 하드웨어 성능을 보유하지 않더라도 현실적으로 학습이 여러 번 가능했다. 이로 인해 앞에 제시된 AR 기술의 경우처럼 '객체와 배경의 분리'가 효과적인 실내 공간 정보 분석 문제에서 제안하는 방법이 실험에 비교된 기존 방법보다 더 효율적으로 작용할 것이다. 추후에 연산량 감소 부분이 더 연구된다면 모바일 환경에서의 학습 가능성이 생긴다. 더 나아가 실제 사용자가 서비스를 이용하면서, 실시간으로 학습하며 중심 신경망을 보조하는 모바일 환경의 보조 신경망을 구성할 수 있을 것이다.

한계점으로, 원본 데이터셋 자체에 미 분류된 값이 다소 존재하기 때문에 추후에 궁극적으로 세밀한 영상 분할을 수행하기 위해서는 미 분류된 부분을 알고리즘을 통해 전처리를 하거나, 더욱 세밀한 영상 분할 맵을 정답 값으로 가지는 데이터셋을 사용해야 한다. 또한 실험에 사용된 데이터셋의 클래스 수는 894개인데 학습에 사용된 이미지 수는 총 1,449장 중 889장으로, 비교적 적은 데이터 때문에 충분한 학습을 하지 못했을 것이다. 따라서 더 많은 데이터와 충분히 세밀한 영상 분할 맵을 정답 값으로 가진 데이터셋을 추가로 학습한다면 더 좋은 효과를 볼 수 있을 것이다.

참고문헌(Reference)

[1] Chang-Hyung L, Hyung-Il C, "Vanishing point detection method using multiple initial vanishing points", The Journal of the Korea Contents Association, Vol.18, No.22, pp.231-239, 2018. https://doi.org/10.5392/JKCA.2018.18.02.231

[2] Espinace P, Kollar T, Soto A, Roy N, "Indoor scene recognition through object detection", In Robotics and Automation (ICRA), IEEE International Conference, pp.1406-1413, 2010. https://doi.org/10.1109/ROBOT.2010.5509682

[3] Gupta S, Arbelaez P, Girshick R., Malik J, "Indoor scene understanding with rgb-d images: Bottom-up

segmentation, object detection and semantic segmentation", International Journal of Computer Vision, Vol.112, No.2, pp.133-149, 2015. https://doi.org/10.1007/s11263-014-0777-6

[4] Krizhevsky A, Sutskever I, Hinton G E, "Imagenet classification with deep convolutional neural networks", In Advances in neural information processing systems, pp.1097-1105, 2012. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-network

[5] He K., Zhang X., Ren S, Sun J, "Deep residual learning for image recognition", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770-778, 2016. http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[6] Long J, Shelhamer E, Darrell T, "Fully convolutional networks for semantic segmentation", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3431-3440, 2015. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[7] Liu J, Wang Y, Li Y, Fu J, Li J, Lu H, "Collaborative Deconvolutional Neural Networks for Joint Depth Estimation and Semantic Segmentation", IEEE Transactions on Neural Networks and Learning Systems, Vol.29, No.11, pp.5655-5666, 2018. https://doi.org/ 10.1109/TNNLS.2017.2787781

[8] Herranz-Perdiguero C, Redondo-Cabrera C, López-Sastre R J, "In pixels we trust: From Pixel Labeling to Object Localization and Scene Categorization", IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.355-361, 2018. https://doi.org/10.1109/IROS.2018.8593736

[9] Cheng Y, Cai R., Li Z, Zhao X., Huang K, "Localitysensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1475-1483, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Cheng_Locality-Sensitive_Deconvolution_Networks_CVPR_2017_paper.html

[10] Lin G., Milan A., Shen C, Reid I, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation", In Cvpr, Vol.1, No.2, pp.1925-1934, 2017. http://openaccess.thecvf.com/content_cvpr_2017/html/Lin_RefineNet_Multi-Path_Refinement_CVPR_2017_paper.html

[11] Chen L C, Papandreou G, Kokkinos I, Murphy K., Yuille A L, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", IEEE transactions on pattern analysis and machine intelligence, Vol.40 No.4, pp.834-848, 2018. https://doi.org/10.1109/TPAMI.2017.2699184

[12] Chen L C, Zhu Y, Papandreou G, Schroff F, Adam H, "Encoder-decoder with atrous separable convolution for semantic image segmentation", Proceedings of the European Conference on Computer Vision (ECCV), pp. 801-818, 2018. https://doi.org/10.1007/978-3-030-01234-2_49

[13] Silberman N, Hoiem D, Kohli P, Fergus R, "Indoor segmentation and support inference from rgbd images", In European Conference on Computer Vision, pp. 746-760 2012. https://doi.org/10.1007/978-3-642-33715-4_54

◐ 저 자 소 개 ◑

김 정 환(Jung-hwan Kim)
2017년 숭실대학교 컴퓨터공학과(공학사)
2017년~현재 숭실대학교 대학원 미디어학과 석사과정
관심분야 : 컴퓨터 비전, 기계 학습, 실내 공간 인식
E-mail : 96junghwan@naver.com

최 형 일(Hyung-il Choi)
1979년 연세대학교 전자공학과(공학사)
1983년 미시간대학 전기전산학과(공학석사)
1987년 미시간대학 전기전산학과(공학박사)
1989년~1999년 숭실대학교 컴퓨터학부 교수
2000년~현재 숭실대학교 미디어학과 교수
관심분야 : 컴퓨터 비전, 패턴인식, 증강현실
E-mail : hic@ssu.ac.kr