의사 깊이맵을 이용한 다중 디코더 기반의 고정밀 분할 딥러닝 모델 개발 및 효율적인 학습 전략1)

김유진[1], 김동영[2], 이정근[1,2,3]
[1]한림대학교 스마트컴퓨팅연구소
[2]한림대학교 컴퓨터공학과
[3]한림대학교 소프트웨어학부
gaimuj32315@gmail.com, kimdongyoung0218@hallym.ac.kr, jeonggun.lee@hallym.ac.kr

Multi-Decoder DNN Model for High Accuracy Segmentation using Pseudo Depth-Map and Efficient Training Strategy

Yu-Jin Kim[1], Dongyoung Kim[2], Jeong-Gun Lee[1,2,3]
[1]Smart Computing Laboratory, Hallym University
[2]Dept. of Computer Engineering, Hallym University
[3]Division of Software, Hallym University

요 약

최근 딥러닝 기술이 급속히 발전하며 현대 사회의 다양한 응용분야에서 빠르게 적용되고 있다. 특히 영상 기반의 딥러닝 기술은 자연어 처리와 함께 인공지능 기술의 핵심 연구 분야로 많은 연구가 진행되고 있다. 논문에서는 최근 많은 연구가 진행되고 있는 영상의 의미적 분할 (Semantic Segmentation) 성능을 향상하기 위한 연구를 진행한다. 특히 모델에서 고정밀의 의미적 분할을 수행할 수 있도록 추가적인 정보로써 의사 깊이맵 (Pseudo Depth-Map)을 활용하는 방법을 제안하였다. 더불어, 의사 깊이맵을 모델 상에서 효과적으로 학습시키기 위하여 다중 디코더 모델과 학습 효율을 높이는 학습 스케줄링 전략을 제안한다. 의사 깊이맵과 다중 디코더 모델 기반의 제안 모델은 기존 의미적 분할 모델과 비교하여 iIoU 기준 2%의 성능 향상을 보였다.

1. 서론

2016년 알파고 쇼크 이후, 딥러닝 기반의 인공지능 기술은 전 세계적으로 다양한 분야에 적용되어 급진적인 기술적 진보를 만들고 있다. 최근 Chat-GPT로 대표되는 거대 언어 모델은 딥러닝 기반의 인공지능 기술로써 현대 사회의 다양한 분야에서 인류를 지원하는 실무적인 기술로 폭넓게 활용되며 인공지능 기술의 정점에 서 있다. 언어모델과 함께 영상분야에서도 딥러닝 기반의 DINOv2와 같은 파운데이션 모델 (Foundation Model)이 발표되며 영상인지 기술을 선도하고 있다. 이미지 분류, 객체 탐지 등에서 더욱 확대되어 최근 고정밀의 섬세한 의미적 분할에 관한 연구 역시 영상분야의 핵심 주제로 인식되며 많은 연구가 진행되고 있다.

이러한 거대 딥러닝 모델 기반의 고성능 모델 개발에도 불구하고, 임베디드 시스템 및 에지 장치에서 시각 지능 서비스를 제공해야 하는 경우와 같이 거대 모델을 활용하기 어려운 상황에서 에지 장치에서 효과적으로 운영될 수 있는 소규모 딥러닝 모델을 활용하여 의미적 분할의 정확도를 높이는 연구가 필요하다. 특히, 자동차 및 공장 등에서 센서의 형태로 제공되는 의미적 분할을 위해서는 경량의 모델 개발이 필수적이다.

본 논문에서는 경량의 딥러닝 모델에서 의미적 분할을 고정밀로 수행하기 위하여 추가적인 깊이맵 (Depth-Map) 정보를 활용하는 방법을 제안한다. 깊이맵 정보를 의미적 분할에 활용하기 위하여 다중 디코더 모델을 제한한다. 제안된 다중 디코더 모델은 전형적인 학습 방식으로 고정밀의 모델로 학습시키기가 쉽지 않기 때문에, 제안된 다중 디코더 모델에서 의미적 분할 정보과 깊이맵 정보를 효과적으로 학습시키기 위한 새로운 형태의 모델 스케줄링 전략을 제안한다.

1) 본 논문은 한림대학교 3단계 산학연협력 선도대학 육성사업 (LINC 3.0)의 2024년도 산학공동 기술개발과제, "차량 라이브뷰 영상을 활용한 스마트 도로안전 모니터링 시스템 개발", 지원을 받았습니다.

그림 1. 의미적 분할 예시

2. 사전 연구

2.1 의미적 분할 (Semantic Segmentation)

초기 이미지상에서 진행된 딥러닝 연구는 이미지 분류 및 객체 탐지에 집중되어 진행되었으나, 이후 보다 복잡도가 높은 "의미적 분할"을 딥러닝을 통해 진행되고 있다 [1]. 의미적 분할은 주어진 이미지에서 각 객체를 "픽셀" 단위로 구분하여 분류하는 작업을 의미한다. 그림 1은 의미적 분할의 예시를 보여주는 것으로 버스와 사람이 같은 색의 픽셀값으로 분할되어 있음을 알 수 있다.

최근 더욱 발전된 "인스턴스 분류" (Instance Segmentation)에서는 객체의 형 (type)뿐만 아니라 객체의 인스턴스까지 구분하여 분류하고 있다.

2.2 깊이맵 (Depth-Map)

깊이맵은 특정 시점과 기준점에서 물체들과의 거리 정보를 이미지화한 것이다. 각 픽셀에는 기준점으로부터 해당 픽셀의 거리를 나타내는 값이 할당되어 주어진 이미지에 대한 3차원 형태의 깊이 정보를 제공한다.

3. 깊이맵과 의미적 분할을 동시에 고려한 다중 디코더 모델

3.1 다중 디코더 모델 구조

본 논문에서 의미적 분할의 정확도를 높이기 위해 깊이맵 정보를 활용하며, 기존의 의미적 분할 모델에 깊이맵 정보의 추가를 통해 인코더의 특징 추출 역량을 강화하기 위해 다중 디코더 모델을 제안한다. 제안된 모델의 구조는 그림 2와 같다. 그림 2에서 보듯이 기본적으로 입력되는 이미지 (I) 는 공유 인코더 (Shared Encoder)를 통해서 특징을 추출하게 되며, 공유 인코더를 통해서 만들어진 특징 벡터는 의미적 분할 이미지 생성을 위한 디코더 (Semantic Decoder)와 깊이맵 생성을 위한 디코더 (Depth Decoder)로 분기되어 각각의 이미지 생성을 위해 사용된다.

제안된 방식과 유사한 구조 모델을 활용하여 의미적 분할과 깊이맵을 동시에 학습시키는 연구가 진행

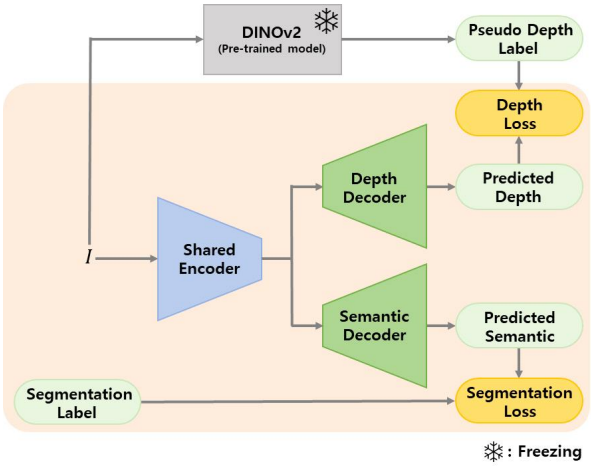

그림 2. 모델 구조

된 바 있으나 [2], 이 경우 손실함수의 가중치를 조절하는 형태로 학습을 진행하였으며, 본 논문은 손실함수의 가중치 조절을 통한 다중 디코더 모델 학습의 어려움을 파악하고, 이를 해결하기 위한 학습 스케줄링 기법을 제안한다.

그림 2의 딥러닝 모델은 기본적으로 "Unet" 구조를 활용하였다. 공유 인코더는 ImageNet 데이터셋을 통해 사전 학습된 ResNet50으로 구성하였으며, 깊이맵 이미지를 생성하는 디코더는 Unet의 디코더 형태로 구성하였다.

본 논문에서 학습 및 추론에 사용한 데이터셋은 PASCAL VOC 2012 [3]이다. 의미적 분할 이미지를 생성하는 디코더는 총 20개의 클래스를 분류하게 된다. 의미적 분할에 대한 라벨 정보를 제공하는 데이터셋은 많지 않아 본 논문에서 제안하는 방법론을 제한하게 되기 때문에, 본 논문에서는 최근 발표된 DINOv2 [4]를 활용하여 의사 깊이맵을 생성하여 학습에 활용하였다. 그림 2의 상단에 DINOv2는 사전 학습된 모델로써 주어진 이미지, I,에 대한 의사 깊이맵 이미지를 생성하며, DINOv2에 의해 생성된 깊이맵 이미지가 학습에 활용된다.

3.2 의미적 분할과 깊이맵을 활용한 손실함수

두 개의 디코더로부터 생성된 의미적 분할 이미지와 깊이맵 이미지는 데이터셋에서 제공되는 정답 분할 이미지 및 DINOv2에 의해 생성된 깊이맵 이미지와 함께 손실함수를 통해 손실 계산하여 학습에 사용된다.

그림 2에 기술된 "Segmentation Loss"와 "Depth Loss"는 각각 Focal 손실함수 (L_{Seg})와 MSE 손실함

수 (L_{depth})로 표현되며 식-1과 식-2와 같다. 모델 디코더에 의해 생성된 의미적 분할 이미지와 깊이맵 이미지에 대한 세부적인 손실함수는 다음과 같다.

$$L_{CE} = CE(I_{pred}, I_{target}) \quad \text{..... (식-1)}$$
$$L_{Seg} = \alpha(1-\exp(-L_{CE}))^{\gamma} L_{CE}$$
$$L_{depth} = mean(\| I_{pred} - I_{target} \|_2^2) \quad \text{..... (식-2)}$$

Focal 손실함수는 Cross Entropy(CE) 손실함수에서 예측이 잘 되는 픽셀에 대해서 γ(focusing parameter)를 통해 손실 값을 줄이는 함수이다. MSE(Mean Squared Error) 손실함수는 예측값과 정답값의 차이를 제곱하여 평균을 구한다.

3.3 학습 스케줄링을 통한 성능 향상

본 논문에서는 의미적 분할에 대한 성능을 평가하기 위하여 성능 지표로 “Mean Intersection Over Union” (평균 IoU)와 “Instance-level Intersection Over Union” (iIoU) 그리고 정확도 (accuracy)를 사용하였다. 평균 IoU란 각 예측한 클래스별로 정답 레이블과 예측 레이블 간의 겹친 정도를 파악한 수치이다. iIoU는 배경을 제외한 객체에 한하여 IoU를 계산한 값이다. 이를 기준으로 영상 분할 작업의 성능이 얼마만큼 올랐는지를 평가하였다. 학습에 사용된 하이퍼-파라미터는 다음 표. 1과 같다.

표 1 학습을 위한 하이퍼-파라미터

Hyperparameter	설정
Optimizer	Adam
Learning rate	0.0001
γ in focal loss	2

표 2 의미적 분할 성능 비교 평가

사용 모델	mIoU	iIoU	acc	다중 작업
단일-디코더 모델	0.7703	0.48	0.986	×
다중-디코더 모델 + 기존 학습방식	0.6412	0.15	0.977	O
다중-디코더 모델 + 제안 학습방식	**0.7789**	**0.50**	**0.997**	O

제안된 모델의 다중 디코더를 손실함수의 가중치를 두어 동시에 학습시켰을 때 모델이 제대로 학습을 진행하지 못하는 것을 확인하였다. 두 개의 디코더가 동시에 학습을 하면서 다른 디코더의 학습을 방해하며 학습 시에 수렴이 잘 진행되지 않았다. 표 2에서 보는 바와 같이 “다중-디코더 모델 + 기존 학습방식”의 경우가 이에 해당하며 mIoU가 0.7459로 단일 디코더를 사용한 의미적 분할 모델보다 더 낮은 성능을 보여주었다.

표3 제안된 스케줄링에 따른 학습된 모델 성능 평가

우선학습	가중치 복제	학습 순서	mIoU	iIoU	acc
의미적 분할 우선	O	d→b	0.7297	0.3759	0.9840
		b	0.7481	0.4231	0.9852
	X	d→b	0.7380	0.3945	0.9845
		b	0.7475	0.4221	0.9852
깊이맵 우선	O	s→b	**0.7789**	**0.5052**	0.9974
		b	0.0136	0.0061	0.9103
	X	s→b	**0.7737**	**0.5012**	0.9870
		b	0.0085	0.0103	0.9094

• s: 의미적 분할 디코더, d: 깊이맵 디코더
• b:디코더 모두 학습(손실함수 가중치: $0.85L_{Seg}+0.15L_{depth}$)

이러한 문제를 해결하며 모델 내부의 두 디코더 간의 효과적인 학습을 위한 “학습 스케줄링 전략”을 제안한다. 제안된 학습 스케줄링 전략에서는 다중 디코더를 개별적으로 순차적으로 학습하는 방식을 구성하였으며, 하나의 디코더가 학습을 마친 후에 다른 디코더를 학습하도록 하였고 최종적으로 두 개의 디코더를 모두 학습하는 파인튜닝 (fine tuning) 단계를 진행하였다.

1. 다중 디코더 중 어떤 디코더를 먼저 학습시킬 것인가?
2. 먼저 학습시킨 디코더를 다른 디코더의 초기값으로 사용할 것인가?
3. 더불어, 최종적으로 두 개의 모든 디코더를 학습하는 파인튜닝 단계를 진행할 것인가?

이러한 다중 디코더 모델을 스케줄링에 의해 학습시키고자 할 때, 다음과 같은 고려 사항이 발생한다.

이와 같은 학습 스케줄링의 고려 사항을 통하여 다양한 학습 스케줄링 설정을 구성하였으며, 각 설정 별로 실험하여 성능을 평가하였다. 표 3은 다양한 설정에 대한 의미적 분할 성능을 보여준다.

표 3에서 “우선학습”은 어떤 디코더를 먼저 학습시켰는지 나타낸다. 실험 결과표에서 보는 바와 같이 의미적 분할을 위한 디코더를 먼저 학습시키는 경우 깊이맵 디코더를 먼저 학습시킨 모델보다 높은 성능을 보이는 설정이 많았다. “가중치 복제”는 먼저 학습한 디코더의 가중치 값을 그대로 다른 디코더로 전이시킬 것인지를 나타내는 것이다. 결과적으

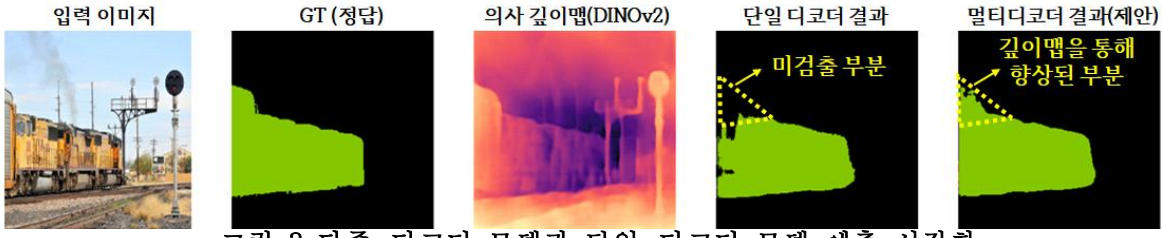

그림 3 다중-디코더 모델과 단일-디코더 모델 예측 시각화

로 복제하는 것과 하지 않는 것은 큰 차이를 보이지 못하였다. 상이한 기능을 하는 디코더이기 때문에 가중치 복제가 큰 도움이 되지 않는 것으로 판단한다.

마지막으로 "학습 순서"는 본 논문에서 활용한 두 개의 디코더 학습을 어떤 순서로 학습시키는지를 의미한다. "d→b"의 경우 우선학습이 먼저 진행된 뒤 깊이맵 디코더 (d: depthmap)를 학습시킨 후 마지막으로 두 개의 디코더 (b: both)를 모두 학습시키는 학습 순서를 의미한다.

깊이맵을 위한 디코더를 학습시킨 후에, 의미적 분할 디코더로 전이시켜 학습하고, 최종적으로 두 디코더를 모두 학습한 경우, 본 논문에서 제안한 깊이맵 정보를 활용한 의미적 분할 모델의 mIoU는 0.7789, iIoU는 0.5052로 깊이맵 정보 없이 단일 의미적 분할 모델로 학습한 경우의 mIoU인 0.7703보다 **0.0086 높았으며, 이는 iIoU에서 2%의 성능 향상**을 의미한다.

의사 깊이맵 이미지를 이용하여 의미적 분할의 모델 학습을 강화하여 약 2%의 성능 향상을 얻을 수 있었다. 본 논문에서 제안된 형태의 의미적 분할 모델을 통해 추출된 결과는 그림 3에서 보는 바와 같이 사람이 보기에 효과적인 분할을 하는 것을 알 수 있다. 실제 정답(ground truth)이 아닌 의사 깊이맵을 통해서 향상된 정확도를 얻은 것은 추가적인 깊이맵에 대한 라벨링 없이 얻은 비지도식 학습의 의미 있는 결과라 볼 수 있다. 한편 DINOv2와 같은 고성능 기초(Foundation) 모델의 높은 정확도가 그림 3의 "Depth"처럼 의미 있는 의사 정답을 생성함에 중요한 역할을 한 것으로 판단한다. 더불어, 의미적 분할 정보와 깊이맵 정보를 모두 정답으로 구축하고 있는 데이터셋을 활용하게 된다면 보다 향상된 의미적 분할 정확도를 얻을 수 있을 것으로도 기대한다.

마지막으로, 본 논문에서 제안한 다중 디코더 모델은 최종적으로 현장에 의미적 분할 모델만 배포될 경우, 깊이맵을 위한 디코더는 제거하고 기존의 의미적 분할 모델과 같은 형태의 단일 디코더 모델로 배포되어 연산량 및 메모리 사용 측면에서 오버헤드가 발생하지 않는다.

4. 결론

영상 기반의 딥러닝 기술은 자연어 처리와 함께 인공지능 기술의 핵심 연구 분야로 많은 연구가 진행되고 있다. 본 논문에서는 딥러닝이 적용되어 활발히 연구가 진행되고 있는 영상의 의미적 분할 (semantic segmentation) 성능을 높이는 연구를 진행하였다. 제안된 딥러닝 모델에서는 고정밀의 의미적 분할 인식을 위해 추가로 의사 깊이맵 (Pseudo Depth-Map) 정보를 제공하는 방법을 제안하였다. 더불어, 의사 깊이맵을 모델 상에서 효과적으로 학습시키기 위하여 다중 디코더 모델과 학습 효율을 높이는 학습 스케줄링 전략을 제안한다. 의사 깊이맵과 다중 디코더 모델 기반의 제안 모델은 기존 의미적 분할 모델과 비교하여 2%의 성능 향상을 보였다.

참고문헌

[1] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3431-3440.

[2] Kendall, Alex et al. "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7482-7491.

[3] Everingham, M. and Van-Gool, L. and Williams, C. K. I. and Winn, J. and Zisserman, A. "The PASCAL Visual Object Classes Challenge 2012 VOC2012 Results" http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html

[4] Oquab, Maxime et al. "DINOv2: Learning Robust Visual Features without Supervision." ArXiv abs/2304.07193, 2023.