4주차. Dual Attention

references	https://openaccess.thecvf.com/content_CVPR_2019/papers/Fu_Dual_Attention_Network_for_Scene_Segmentati

< 발표자 이은경 >

0. Abstract

- Self Attention Mechanism 을 기반으로 다양한 상황 의존성을 캡처하여 Scene Segmentation 수행
- multi-scale feature fusion 으로 context를 포착하는 이전 논문(ICEnet, ...) 과 달리, local feature 를 global dependencies과 **적응적(adaptively)**으로 통합할 수 있는 DANet (Dual Attention Network)을 제안 (position + channel Attention)
- 두 가지 유형의 Attention 모듈을 **dilated FCN 모듈** 위에 추가
    - position 및 channel 에서 각각 semantic interdependencies을 모델링

1) position Attention module : 모든 position에서 feature 의 가중치 합계를 사용하여 각 position의 feature 를 선택적으로 집계.

- 유사한 특징은 거리에 관계 없이 서로 연관

2) channel Attention module : 모든 channel map 사이에 관련 feature 을 통합하여 상호 의존적인 channel map을 선택적으로 강조

> 두 Attention 모듈의 출력을 합산하여 feature representation을 개선하여 보다 정확한 Segmentataion 결과에 기여

- 세 종류의 Scene Segmentation 데이터 세트(Cityscapes, PASCAL Context 및 COCO Stuff)에서 SOTA 성능 달성.
    - 특히 Cityscapes 테스트셋에서 coarse 데이터를 사용하지 않고도 81.5% 의 Mean IoU 를 얻음

1. Introduction

- Scene Segmentation 작업을 효과적으로 수행하기 위해서는 혼란스러운 범주를 구별하고 외관이 다른 객체를 고려해야 함.

  예) '밭'과 '풀'의 영역은 종종 구별할 수 없으며, '자동차'의 물체는 scales, occlusion, illumination에 의해 영향을 받음.
    - 픽셀 수준 인식을 통해 feature 표현의 discriminative ability을 강화할 필요성이 있음.
    - Fully Convolutional Networks (FCNs)[13] 모델이 state-of-the-art 성능 보임.
- 선행연구

  1) multi-scale context fusion 활용 : Deeplab, PSPNet...

  2) long-range dependencies 모형화를 위한 recurrent neural network : 2D-LSTM(local features에 대한 풍부한 spatial dependencies을 포착하기 위해 directed acyclic graph를 사용하여 recurrent neural networks을 구축.

- natural scene image 분할을 위한 새로운 프레임워크 DANet(Dual Attention Network) 제안(그림 2 참조).

- spatial and channel dimensions의 각각 feature dependencies 을 포착하기 위한 self-Attention mechanism을 도입.
- dilated FCN 위에 두 개의 병렬 Attention 모듈을 추가.

1) position Attention module : feature map의 두 position 사이의 공간 의존성을 캡처하기 위한 self-Attention mechanism. ($N \times N$)

- 가중치는 해당 두 위치 사이의 유사도.
- 두 position 간 거리에 유사도는 관계없음.

2) channel Attention module : Self-Attention mechanism을 사용하여 두 채널 맵 사이의 종속성을 캡처. ($C \times C$)

3) 이 두 Attention 모듈의 출력은 feature 표현을 더욱 강화하기 위해 **Sum fusion.**

- 복잡하고 다양한 장면을 다룰 때 이전 방법[4, 29]보다 더 효과적이고 유연.
  예) 그림1 의 길거리 장면.
  1) 첫 번째 사진의 일부 '사람'과 '신호등'은 조명과 시야로 인해 눈에 띄지 않거나 불완전한 물체. 큰 물체(예: 자동차, 건물)의 맥락이 눈에 띄지 않는 물체 라벨링을 해칠 수 있음.
  - 반대로, Attention 모델은 눈에 띄지 않는 객체의 유사한 특징을 선별적으로 취합하여 특징 표현을 강조하고 큰 물체의 영향을 피함.

  2) '자동차'와 '사람'의 scale은 다양하며, 그러한 다양한 대상을 인식하기 위해서는 다른 scale 의 상황 정보가 필요.
  - 즉, 서로 다른 scale의 feature들은 동일한 semantic를 나타내기 위해 동등하게 다루어져야 함. Attention mechanism이 있는 우리 모델은 global view에서 어떤 규모로든 유사한 feature 을 적응적으로 통합하는 것을 목표

  3) position 과 channel관계를 explicitly하게고려
  - long-range dependencies에 유익

Figure 1: The goal of scene segmentation is to recognize each pixel including stuff, diverse objects. The various scales, occlusion and illumination changing of objects/stuff make it challenging to parsing each pixel.

2. Relate work

1) multi-scale feature fusion

▼ Deeplabv2 [3]와 Deeplabv3 [4]

Contextual 정보를 포함하기 위해 atrous spatial pyramid pooling을 채택, 이는 서로 다른 dilated 속도를 가진 parallel dilated 컨볼루션으로 구성.

▼ PSP-Net [29]

서로 다른 scales정보를 포함하는 효과적인 Contextual prior를 위해 pyramid pooling module을 설계.

▼ 인코더-디코더 구조[6, 8, 9]

다른 scale 컨텍스트를 얻기 위해 중간 레벨 및 높은 레벨의 semantic feature 을 결합.

2) 로컬 feature에 대한 Contextual 의존성을 학습하는 것도 feature 표현에 기여.

▼ DAG-RNN [18]

풍부한 Contextual 의존성을 포착하기 위해 반복 신경망을 가진 순환 그래프를 연출.

▼ PSANet [30]

공간 차원의 상대 위치 정보와 컨볼루션 레이어를 기준으로 픽셀 단위 관계를 캡처.

▼ EncNet [27]

글로벌 컨텍스트를 포착하기 위한 채널 Attention mechanism을 도입.

3) self- Attention 모듈은 장거리 의존성을 모델링할 수 있어 광범위하게 적용 [11, 12, 17, 19–21].

▼ [21]

입력의 global dependencies을 도출하기 위한 self-Attention mechanism을 제안하는 첫 번째 작업이며 기계 번역에 이를 적용.

▼ [28]

Attention Module은 이미지에 점점 더 많이 적용. 더 나은 이미지 생성기를 학습하기 위한 self Attention mechanism을 도입.

▼ [23]

주로 영상과 이미지에 대한 시공간 차원에서 non-local 효과를 탐구.

3. Dual Attention Network

- 네트워크의 일반적인 프레임워크를 제시. position 및 channel 에서 각각 장거리 상황 정보를 캡처하는 두 가지 Attention 모듈을 소개.

<3.1 Overview>

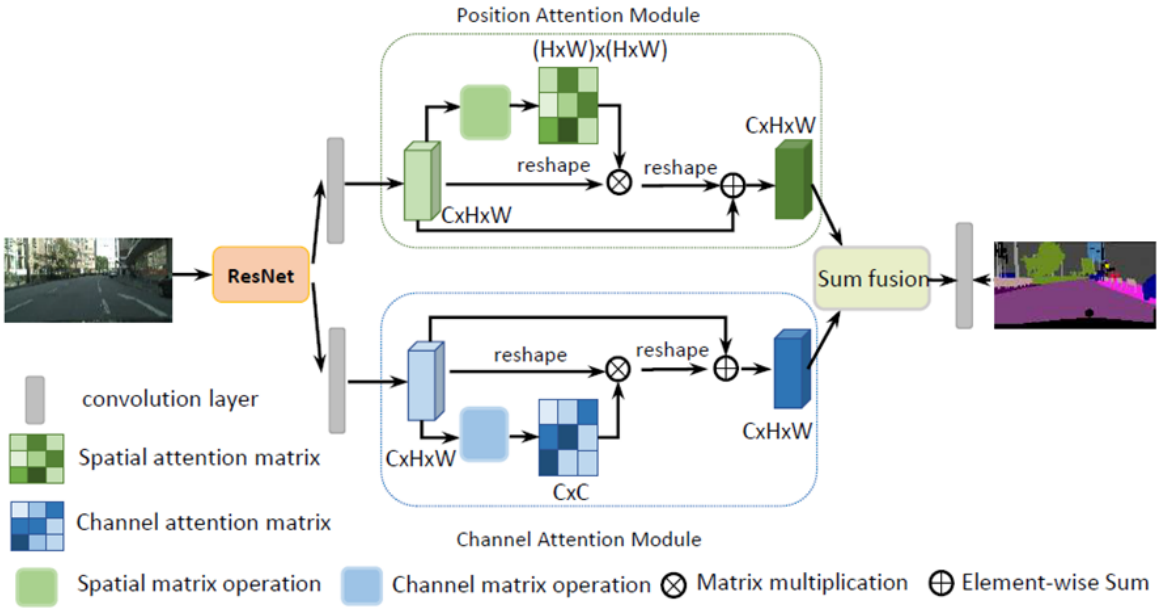
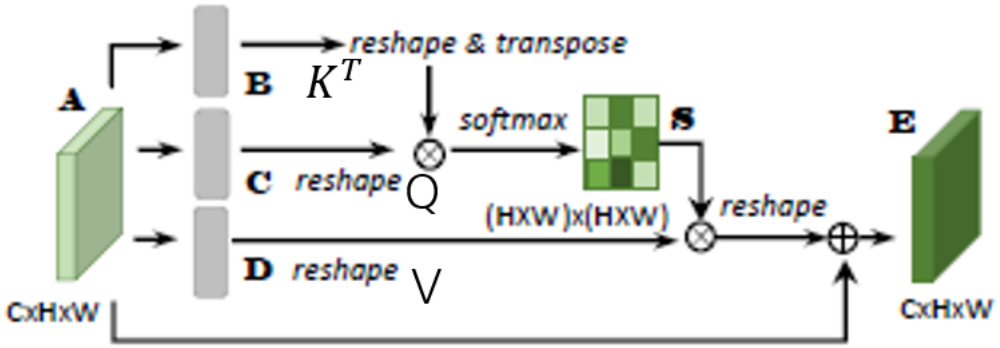


Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

- 그림 2. dilated residual network에 의해 생성된 local feature를 통해 global context를 위해 두 가지 유형의 Attention 모듈으로 더 나은 feature 표현을 얻고자 함.

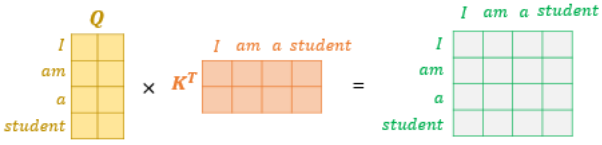
<3.2 Position Attention Module>



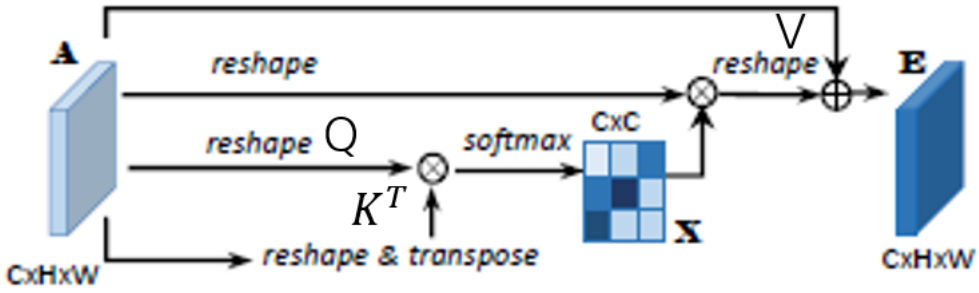
A. Position attention module

- Discriminant feature representations은 scene understanding 를 위해 필수적, 이는 long-range contextual information를 얻을 수 있음.
- 그러나 많은 연구[15, 29]에서 기존 FCN에서 생성된 local features 가 객체를 잘못 분류할 수 있다고 주장. local features 에 대한 풍부한 컨텍스트 관계를 모델링하기 위해 position attention module 도입.
- position Attention module 은 보다 광범위한 상황 정보를 local feature으로 인코딩하여 표현 능력을 향상시킨 후 적응적으로 spatial 컨텍스트를 집계하는 프로세스를 자세히 설명
- 그림.3(A)에 예시된 바와 같이, local feature $A \in \mathbb{R}^{C\times H\times W}$가 주어졌을 때, 두 개의 새로운 feature maps B와 C를 생성하기 위해 convolution layer 적용 $B \in \mathbb{R}^{C\times N}$으로 reshape.
  - $N = H \times W$
- 그런 다음 C와 B의 전치 사이에 행렬 곱셈을 수행하고 소프트맥스 레이어를 적용하여 spatial attention map $S \in \mathbb{R}^{N\times N}$을 계산. $s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)}\ i, j \in \{1, \cdots, N\}$
- 여기서 s_{ji}는 i번째 위치가 j번째 위치에 미치는 영향을 측정. 두 위치의 특성이 더 유사할수록 두 위치간의 상관 관계가 더 커짐.
- 새로운 feature 맵$D \in \mathbb{R}^{C\times H\times W}$를 생성하기 위해 feature A를 convolution layer 적용하여 $\mathbb{R}^{C\times N}$으로 reshape
- 그런 다음 D와 S의 전치 사이의 행렬 곱셈을 수행
  하고 그 결과를 $\mathbb{R}^{C\times H\times W}$ 로 reshape.
- 마지막으로 스케일 파라미터 α를 featuresA의 element-wise sum.
- 최종 output $E \in \mathbb{R}^{C\times H\times W}$ 는 $E_j = \alpha \sum_{i=1}^{N}(s_{ji}D_i) + A_j$ 와 같이 계산
  - α는 0으로 초기화 후 사용. 점차 더 많은 가중치를 할당하는 방법을 학습 [28].
- 각 위치의 결과 feature E는 모든 위치와 원래 feature에 걸친 feature의 가중치 합
  - 따라서 global contextual view를 가지고 있으며 spatial attention map에 따라 선택적으로 컨텍스트를 집계.
- similar semantic features은 상호 이익을 달성하여 intra-class compact 및 semantic consistency 향상.

▼ NLP에서의 Attention



<3.3. Channel Attention Module>



B. Channel attention module

- 상위 수준의 각 channel map은 class-specific response으로 재평가될 수 있으며 different semantic responses들이 연관
- channel map 간의 interdependencies을 이용하여, 우리는 상호의존적인 feature 맵을 강조하고 specific semantics의 feature 표현을 개선할 수 있음.
- 따라서, 채널간 상호의존성을 명시적으로 모델링하기 위해 channel Attention 모듈을 구축
- channel Attention 모듈의 구조는 그림 3(B).
- position Attention 모듈과는 달리 channel Attention map$X \in \mathbb{R}^{C\times C}$를 원래의 feature $A \in \mathbb{R}^{C\times H\times W}$로 직접 계산
- 특히 A에서 $\mathbb{R}^{C\times N}$ 으로 모양을 변경한 후 A와A^T사이에 행렬 곱셈을 수행.
- 마지막으로, channel Attention map $X \in \mathbb{R}^{C\times C}$를 얻기 위해 소프트 맥스 레이어를 적용

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)}, i, j \in \{1, \cdots, C\}$$

- x_{ji} : i번째 채널이 j번째 채널에 미치는 영향을 측정
- X와 A^T 사이에 행렬 곱셈을 하여 그 결과를 $\mathbb{R}^{C\times H\times W}$ 로 reshape, 그 결과에 척도 파라미터 β를 곱하여 A와 element-wise-sum을 수행하여 최종 출력 $E \in \mathbb{R}^{C\times H\times W}$ 를 구함
  - 여기서 β는 0부터 점차 가중치를 학습.
- 방정식 4는 각 채널의 최종 특징이 모든 채널의 특징과 원래의 특징에 대한 가중치 합이라는 것을 보여주며, feature 맵 사이의 장거리 의미 의존성을 모델링 > feature 차별성을 높이는 데 도움.
- 두 채널의 관계를 계산하기 전에 feature을 embed하는 데 Convolution를 사용하지 않음.
  - 이는 서로 다른 채널 맵 간의 관계를 유지할 수 있기 때문
- 또한 global pooling 또는 encoding layer에 의해 채널 관계를 탐색하는 최근 연구[27]와는 달리, 채널 상관 관계를 모델링하기 위해 모든 대응 위치의 공간 정보를 활용

<3.4. Attention Module Embedding with Networks>

- 두 가지 Attention 모듈의 feature을 통합.
- 특히, 두 개의 Attention 모듈의 출력을 Convolution 계층에 의해 변환하고 요소별 합계를 수행하여 feature fusion을 수행하면 변환 레이어가 따라 최종 예측 맵을 생성.
- 더 많은 GPU 메모리가 필요한 계단식 작업을 채택하지 않음.
- Attention 모듈은 단순해서 FCN과 같은 기존 모듈에 직접 삽입 가능.
- 매개 변수를 너무 많이 늘리지 않고 feature 표현을 효과적으로 강화.

4. Experiments

- 평가를 위해 , Cityscapes 데이터 세트[5], PASCAL VOC2012[7], PASCAL Context 데이터 세트 [14] 및 COCO Stuff 데이터 세트에 대한 포괄적인 실험 수행
- 실험 결과에 따르면 DANet 은 세 개의 데이터 세트에서 state of the art performance 달성

<4.1. Datasets and Implementation Details>

- Cityscapes : 50 개 도시에서 캡처된 5,000 개의 이미지. 각 이미지는 2048 × 1024 픽셀, 19 개 semantic 클래스의 고품질 픽셀 레벨 레이블 존재. training 에는 2,979 개의 영상이 있고 , validation 에는 500 개의 영상이 있으며 , test 세트에는 1,525 개의 영상이 있음. coarse 데이터를 사용하지 않음. (이 경우 N = 2048 * 1024 ? or ResNet output dim ? $O(N^2)$)
- PASCAL VOC 2012 : training 이미지 10,582 개 , validation 이미지 1,449 개 , test 이미지 1,456 개가 포함 . 여기에는 20 개의 foreground 객체 클래스와 1 개의 background 클래스가 포함
- PASCAL Context : 전체 씬에 대한 자세한 semantic 레이블 제공. training 4,998 개의 이미지와 test 5,105 개의 이미지가 포함 . 가장 빈도가 높은 59 개 클래스에서 하나의 배경 범주 총 60 개 클래스와 함께 평가
- COCO Stuff : training 이미지 9,000 개와 test 이미지 1,000 개가 포함. 각 픽셀에 80 개의 개체와 91 개의 항목에 주석을 달아 171 개의 범주에 대한 결과 보고

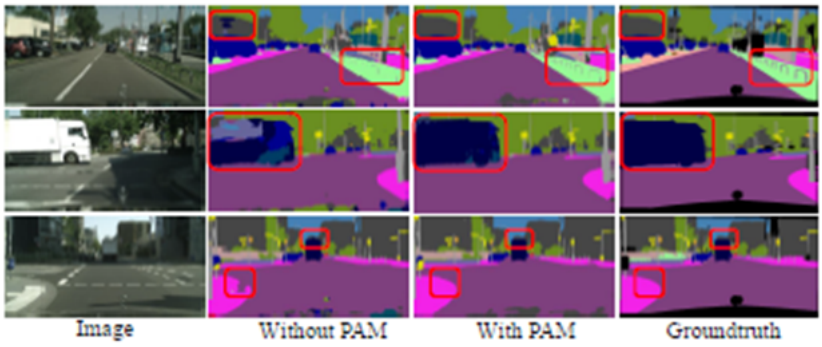


Figure 4: Visualization results of position attention module on Cityscapes val set.

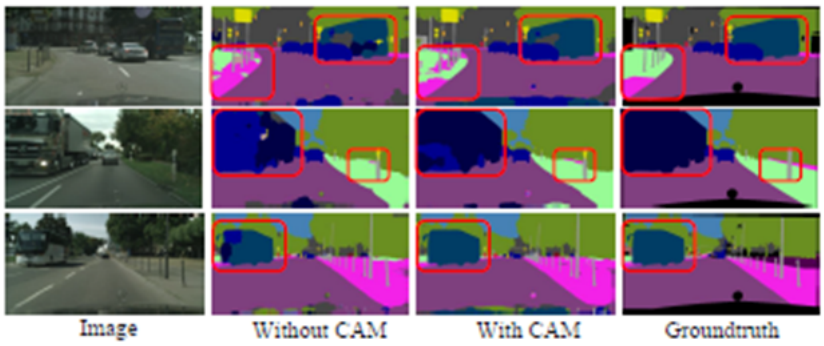


Figure 5: Visualization results of channel attention module on Cityscapes val set.

- Pytorch 기반 구현
- [4,27] 에 따라 , 우리는 초기 학습률에학습률에$(1-\frac{iter}{tot_liter})^{0.9}$를곱하는 poly learning rate policy 를 채택
- Cityscapes 데이터셋의 경우 기본 학습률은 0.01 로 설정. Momentum 와 weight decay coefficients 는 각각 0.9 와 0.0001 로 설정 .
- Synchronized BN 을 사용하여 모델을 training. 배치 크기는 Cityscapes 의 경우 8 로 설정되고 다른 데이터셋의 경우 16 으로 설정
- multi scale augmentation 을 채택할 때 training 시간을 COCO Stuff 의 경우 180Epoch, 기타 데이터셋의 경우 240Epoch 로 설정 .
- Deeplab에 이어 , 우리는 두 개의 Attention 모듈이 모두 사용될 때 네트워크 끝에서 다중 손실을 채택

- 데이터 augmentation 을 위해 Cityscapes 데이터 세트에 대한 ablation study training 중에 무작위 cropping (cropsize 768) 와 random left right flipping 을 적용

< 4.2.1 Ablation Study for Attention Modules >

- better scene understanding 의 장거리 의존성을 캡처하기 위해 dilated 네트워크 위에 이중 Attention 모듈을 사용 .
- Attention 모듈의 성능을 검증하기 위해 표 1 의 다양한 설정으로 실험을 수행
- 표1 과 같이 Attention 모듈은 성능을 향상시킴 .

Method	BaseNet	PAM	CAM	Mean IoU%
Dilated FCN	Res50			70.03
DANet	Res50	✓		75.74
DANet	Res50		✓	74.28
DANet	Res50	✓	✓	76.34
Dilated FCN	Res101			72.54
DANet	Res101	✓		77.03
DANet	Res101		✓	76.55
DANet	Res101	✓	✓	77.57

Table 1: Ablation study on Cityscapes val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

- 기본 FCN(ResNet50) 과 비교했을 때 position Attention 모듈을 채택하면 Mean IoU 에서 75.74% 의 결과를 얻을 수 있어 5.71% 의 개선 효과를 얻을 수 있음 .
- channel contextual module 을 개별적으로 채택하는 것은 기준치를 4. 25% 이상 증가 .
- 두 Attention 모듈을 함께 통합하면 성능이 76.34% 로 더욱 향상 .
- 또한 , 사전 교육된 ResNet 101 를 채택할 경우 , 두 개의 Attention 모듈을 갖춘 네트워크는 기준 모델에 비해 분할 성능을 5.03% 향상시킵니다 .
- position Attention 모듈의 효과는 그림 4에 시각화 .

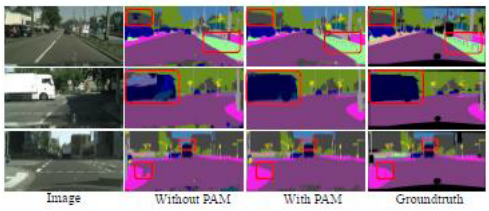


Figure 4: Visualization results of position attention module on Cityscapes val set.

- 첫 번째 행의 'pole' 과 두 번째 행의 'sidewalk'와 같은 곳에 position Attention 모듈을 사용하면 일부 세부사항과 객체 경계가 더 명확
- 로컬 feature 에 대한 선택적 fusion 은 세부사항의 구별을 강화 .

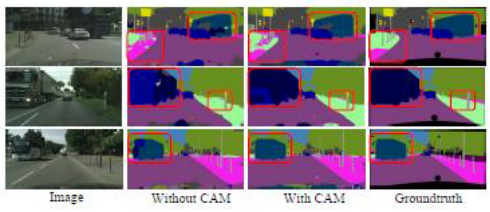


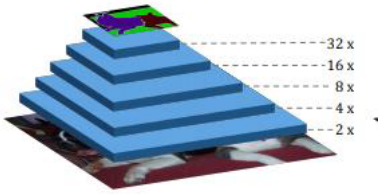
Figure 5: Visualization results of channel attention module on Cityscapes val set.

- 그림 5 는 채널 Attention 모듈을 통해 첫 번째 및 세 번째 행의 bus 와 같이 일부 잘못 분류된 범주가 현재 올바르게 분류되었음을 보여줌 .

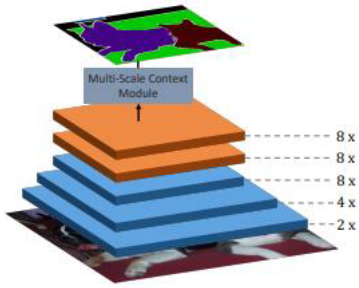
- Channel map 간의 선택적 통합은 컨텍스트 정보를 캡처하는 데 도움 .
- 일관성이 확실히 개선

▼ Dilated FCN

Dilated FCN(https://arxiv.org/pdf/1903.11816.pdf)



(a) FCN [22]



(c) DilatedFCN

파란 layer : downsampling
주황 layer : dilated convolutions

<4.2.2 Ablation Study for Improvement Strategies>

- Deeplabv3[4]에 이어 성능을 더욱 개선하기 위해 동일한 전략을 채택 .

  1) DA(Data augmentation) augmentation): 랜덤 스케일링을 사용한 데이터 augmentation

  2) 다중 그리드 Multi Grid)Grid): 마지막 ResNet 블록에 다양한 크기의 그리드 계층을 적용

  3)MS(Map scaling?): 8 개 이미지 스케일 {0.5, 0.75, 1, 1.25, 1.5, 1.75} 의 분할 확률 맵을 평균화

Method	BaseNet	PAM	CAM	Mean IoU%
Dilated FCN	Res50			70.03
DANet	Res50	✓		75.74
DANet	Res50		✓	74.28
DANet	Res50	✓	✓	76.34
Dilated FCN	Res101			72.54
DANet	Res101	✓		77.03
DANet	Res101		✓	76.55
DANet	Res101	✓	✓	77.57

Table 1: Ablation study on Cityscapes val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

- 랜덤 dilated 을 통한 데이터 dilated 은 성능을 거의 1.26% 향상 .
- 이는 training 데이터의 scale 다양성을 강화함으로써 네트워크 이점을 얻을 수 있음
- 사전 훈련된 네트워크의 더 나은 feature 표현을 얻기 위해 멀티그리드를 채택하고 있으며 , 이는 1.11% 의 추가 개선 .
- 마지막으로 , Segmentation map fusion 은 성능이 81.50% 로 더욱 향상되어 잘 알려진 방법 Deeplabv3[4] (cityscape val set 에서 79.30%) 을 2.20% 능가

<4.2.3 Visualization of Attention Module>

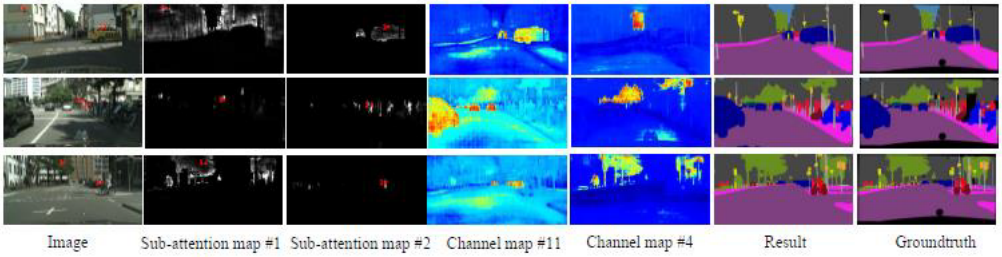


Figure 6: Visualization results of attention modules on Cityscapes val set. For each row, we show an input image, two sub-attention maps ($H \times W$) corresponding to the ponits marked in the input image. Meanwhile, we give two channel maps from the outputs of channel attention module, where the maps are from 4^{th} and 11^{th} channels, respectively. Finally, corresponding result and groundtruth are provided.

- position Attention 를 위해 , 전체적인 self Attention map 는 (H × W) × (H × W) 크기로 , 이미지의 각 특정 지점에 대해 (H × W) 크기의 해당 하위 Attention map 이 있음을 의미 .
- 그림 6 에서는 각 입력 이미지에 대해 두 점 (#1 및 #2 로 표시) 을 선택하고 해당 하위 Attention 맵을 각각 2 열과 3 열에 표시 .
- position Attention 모듈이 명확한 semantic 유사성과 장거리 관계를 포착할 수 있다는 것을 관찰

  예) 첫 번째 행에서 빨간색 포인트 #1 은 건물에 표시되며 Attention map(2 열 는 건물이 있는 대부분의 영역을 강조 . 더욱이, 하위 Attention map 에서, 경계 중 일부는 #1 지점으로부터 멀리 떨어져 있더라도 경계가 매우 명확 .

  예) 포인트 #2 는 Attention map 이 자동차 로 라벨이 지정된 대부분의 위치에 집중 . 두 번째 행에서는 해당 픽셀 수가 적더라도 글로벌 영역의 'traffic 과 ' 을 동일하게 유지 .

  예) 세 번째 행은 'vegetation' 와 ‘person’ class 를 위한 것 . 특히 포인트 #2 는 가까운 ‘rider’ class 에는 대응하지 않지만 , ‘person’ class 에는 대응

<4.2.4 Comparing with State of the art>

Methods	Mean IoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
DeepLab-v2 [3]	70.4	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8
RefineNet [10]	73.6	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70
GCN [15]	76.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DUC [22]	77.6	98.5	85.5	92.8	58.6	55.5	65	73.5	77.9	93.3	72	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8
ResNet-38 [24]	78.4	98.5	85.7	93.1	55.5	59.1	67.1	74.8	78.7	93.7	72.6	95.5	86.6	69.2	95.7	64.5	78.8	74.1	69	76.7
PSPNet [29]	78.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BiSeNet [26]	78.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PSANet [30]	80.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DenseASPP [25]	80.6	**98.7**	**87.1**	93.4	**60.7**	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	**78.0**	**90.3**	80.7	69.7	76.8
DANet	**81.5**	98.6	86.1	**93.5**	56.1	**63.3**	**69.7**	**77.3**	**81.3**	**93.9**	**72.9**	**95.7**	**87.3**	**72.9**	**96.2**	76.8	89.4	**86.5**	**72.2**	**78.2**

Table 3: Per-class results on Cityscapes testing set. DANet outperforms existing approaches and achieves 81.5% in Mean IoU.

- Cityscapes 테스트 세트의 기존 방법과 추가로 비교 . 특히 , 주석이 달린 데이터만으로 DANet 101 을 training 후 테스트 결과를 공식 평가 서버에 제출 .
- DANet 은 dominantly advantage. 를 가진 기존 접근 방식을 능가. 특히 , PSANet은 동일한
  백본 ResNet 101을 사용했음에도 성능이 좋음. 더 강력한 사전 훈련된 모델을 사용하는 DenseASPP 까지 능가

<4.3. Results on PASCAL VOC 2012 Dataset>

Method	BaseNet	PAM	CAM	Mean IoU%
Dilated FCN	Res50			75.7
DANet	Res50	✓	✓	79.0
DANet	Res101	✓	✓	80.4

Table 4: Ablation study on PASCAL VOC 2012 val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

- 추가 효과 평가를 위한 Pascal VOC2012 dataset 에 대한 실험.

- Pascal VOC 의 Quantitative result 는 2012 년 val 세트에서 보여줌 . DANet 50 은 3.3% 를 초과하는 성능향상
- 더 깊은 ResNet 101 모델 채택시 mean IoU 80.4% 을 달성
- [4, 27, 29] 에 이어 PASCAL VOC 2012 training 세트에서도 모델을 더 잘 fine tuning. 테스트 세트에 대한 PASCAL VOC 2012 의 결과는 표 5 에 나와 있습니다

Method	Mean IoU%
FCN [13]	62.2
DeepLab-v2(Res101-COCO) [3]	71.6
Piecewise [11]	75.3
ResNet38 [10]	82.5
PSPNet(Res101) [29]	82.6
EncNet (Res101) [27]	**82.9**
DANet(Res101)	82.6

Table 5: Segmentation results on PASCAL VOC 2012 testing set.

<4.4. Results on PASCAL Context Dataset>

Method	Mean IoU%
FCN-8s [13]	37.8
Piecewise [11]	43.3
DeepLab-v2 (Res101-COCO) [3]	45.7
RefineNet (Res152) [10]	47.3
PSPNet (Res101) [29]	47.8
Ding et al.(Res101) [6]	51.6
EncNet (Res101) [27]	51.7
Dilated FCN(Res50)	44.3
DANet (Res50)	50.1
DANet (Res101)	**52.6**

Table 6: Segmentation results on PASCAL Context testing set.

- PASCAL Context 에 대한 실험을 수행하여 방법의 효과를 추가로 평가 .
- PASCAL VOC 2012 에 대해 동일한 training 및 test설정을 채택 .
- 기준 (dilated FCN 50) 은 평균 IOU 44.3% 를 달성. DANet50 은 성능을 50.1% 로 향상. 깊이 있는 pre-training 네트워크 ResNet101 을 통해 , 우리 모델 결과는 이전 방법들을 큰 차이로 능가하는 Mean IoU 52.6% 를 달성 .
- Deeplab v2 와 RefineNet 은 서로 다른 방식의 변환 또는 다른 단계의 인코더에 의한 멀티스케일 feature fusion 을 채택. 또한 추가 COCO 데이터로 모델을 training하거나 Segmentation 결과를 개선하기 위해 심층 모델을 채택 .
- 기존 방식과 달리 , global dependencies을 명시적으로 포착하기 위해 Attention 모듈을 도입하고 , 더 나은 성능 달성

<4.5. Results on COCO Stuff Dataset>

Method	Mean IoU%
FCN-8s [13]	22.7
DeepLab-v2(Res101) [3]	26.9
DAG-RNN [18]	31.2
RefineNet (Res101) [10]	33.6
Ding et al.(Res101) [6]	35.7
Dilated FCN (Res50)	31.9
DANet (Res50)	37.2
DANet (Res101)	**39.7**

Table 7: Segmentation results on COCO Stuff testing set.

- 제안된 네트워크의 일반화를 검증하기 위해 COCO Stuff 에 대한 실험도 수행 .
- 그 결과 , 우리의 모델은 이러한 방법을 큰 차이로 능가하는 Mean IoU 에서 39.7% 를 달성 .

- 비교 방법 중 DAG RNN[18]은 2D 이미지용 chain RNN 을 활용하여 풍부한 position dependencies을 모델링
- Ding et al.[6] 눈에 띄지 않는 객체 및 배경 물질 segmentation을개선하기 위해 디코더 단계에서 gating mechanism 채택 .
- 우리의 방법은 보다 효과적으로 long-range context information를 포착하고 scene segmentation에서 더 나은 feature representation을 배울 수 있음

5. conclusion

- 우리는 self Attention mechanism 을 이용하여 local semantic features 을 적응적으로 통합하는 Scene Segmentation 를 위한 DANet(Dual Attention Network) 을 제시 .
- 특히 , spatial and channel dimensions 의 global dependencies 을 각각 포착하기 위한 position attention module 과 channel attention module 을 소개 .
- ablation experiments 에서는 dual Attention 모듈이 long range contextual information 를 효과적으로 캡처하고 보다 정밀한 분할 결과를 제공한다는 것을 보여줌 .
- Attention 네트워크는 4 개의 Scene Segmentation 데이터셋(Cityscapes, InPascal VOC 2012, Pascal Context, COCO Stuff) 에서 일관되게 뛰어난 성능을 달성
- 추가로 , 컴퓨팅 복잡성을 줄이고 모델의 견고성을 향상시키는 것이 중요하며 , 향후 작업에서 연구 할 것