**SEMANTIC IMAGE SEGMENTATION WITH DEEP CONVOLUTIONAL NETS AND FULLY CONNECTED CRFS**

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille

**ABSTRACT**

Deep Convolutional Neural Networks (DCNNs) have recently shown state of the art performance in high level vision tasks, such as image classification and object detection. This work brings together methods from DCNNs and probabilistic graphical models for addressing the task of pixel-level classification (also called ”semantic image segmentation”). We show that responses at the final layer of DCNNs are not sufficiently localized for accurate object segmentation. This is due to the very invariance properties that make DCNNs good for high level tasks. We overcome this poor localization property of deep networks by combining the responses at the final DCNN layer with a fully connected Conditional Random Field (CRF). Qualitatively, our “DeepLab” system is able to localize segment boundaries at a level of accuracy which is beyond previous methods. Quantitatively, our method sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 71.6% IOU accuracy in the test set. We show how these results can be obtained efficiently: Careful network re-purposing and a novel application of the ’hole’ algorithm from the wavelet community allow dense computation of neural net responses at 8 frames per second on a modern GPU.

DCNN(Deep Convolutional Neural Networks)은 최근 이미지 분류 및 객체 감지와 같은 높은 수준의 비전 작업에서 최첨단 성능을 보여주었습니다. 이 작업은 픽셀 수준 분류("의미론적 이미지 분할"이라고도 함) 작업을 처리하기 위해 DCNN과 확률적 그래픽 모델의 방법을 결합합니다. 우리는 DCNN의 최종 레이어에서의 응답이 정확한 객체 분할을 위해 충분히 지역화되지 않았음을 보여줍니다. 이는 DCNN을 높은 수준의 작업에 적합하게 만드는 매우 불변 속성 때문입니다. 최종 DCNN 계층의 응답을 완전히 연결된 CRF(Conditional Random Field)와 결합하여 심층 네트워크의 이러한 열악한 현지화 속성을 극복합니다. 질적으로 우리의 "DeepLab" 시스템은 이전 방법을 능가하는 정확도 수준에서 세그먼트 경계를 지역화할 수 있습니다. 양적으로, 우리의 방법은 PASCAL VOC-2012 시맨틱 이미지 분할 작업에서 새로운 최첨단을 설정하여 테스트 세트에서 71.6% IOU 정확도에 도달했습니다. 우리는 이러한 결과를 어떻게 효율적으로 얻을 수 있는지 보여줍니다. 신중한 네트워크 용도 변경 및 웨이블릿 커뮤니티의 '구멍' 알고리즘의 새로운 적용을 통해 최신 GPU에서 초당 8프레임의 신경망 응답을 고밀도로 계산할 수 있습니다.

**1 INTRODUCTION**

Deep Convolutional Neural Networks (DCNNs) had been the method of choice for document recognition since LeCun et al. (1998), but have only recently become the mainstream of high-level vision research. Over the past two years DCNNs have pushed the performance of computer vision systems to soaring heights on a broad array of high-level problems, including image classification (Krizhevsky et al., 2013; Sermanet et al., 2013; Simonyan & Zisserman, 2014; Szegedy et al., 2014;

DCNN(Deep Convolutional Neural Networks)은 LeCun et al. 이후로 문서 인식을 위한 선택 방법이었습니다. (1998), 그러나 최근에야 높은 수준의 비전 연구의 주류가 되었습니다. 지난 2년 동안 DCNN은 이미지 분류를 포함한 광범위한 고수준 문제에서 컴퓨터 비전 시스템의 성능을 크게 향상시켰습니다(Krizhevsky et al., 2013; Sermanet et al., 2013; Simonyan & Zisserman, 2014 Szegedy et al., 2014;

Papandreou et al., 2014), object detection (Girshick et al., 2014), fine-grained categorization (Zhang et al., 2014), among others. A common theme in these works is that DCNNs trained in an end-to-end manner deliver strikingly better results than systems relying on carefully engineered representations, such as SIFT or HOG features. This success can be partially attributed to the built-in invariance of DCNNs to local image transformations, which underpins their ability to learn hierarchical abstractions of data (Zeiler & Fergus, 2014). While this invariance is clearly desirable for high-level vision tasks, it can hamper low-level tasks, such as pose estimation (Chen & Yuille, 2014; Tompson et al., 2014) and semantic segmentation - where we want precise localization, rather than abstraction of spatial details.

Papandreou et al., 2014), 물체 감지(Girshick et al., 2014), 세분화된 분류(Zhang et al., 2014) 등이 있습니다. 이러한 작업에서 공통적인 주제는 엔드 투 엔드 방식으로 교육된 DCNN이 SIFT 또는 HOG 기능과 같이 신중하게 설계된 표현에 의존하는 시스템보다 훨씬 더 나은 결과를 제공한다는 것입니다. 이러한 성공은 부분적으로 DCNN이 로컬 이미지 변환에 내재된 불변성으로 인해 데이터의 계층적 추상화를 학습할 수 있게 되었습니다(Zeiler & Pergus, 2014). 이러한 불변성은 높은 수준의 비전 작업에는 분명 바람직하지만, 공간 세부 사항을 추상화하지 않고 정확한 현지화를 원하는 포즈 추정(Chen & Yuille, 2014; Tompson et al., 2014)과 같은 낮은 수준의 작업을 방해할 수 있습니다.

There are two technical hurdles in the application of DCNNs to image labeling tasks: signal downsampling, and spatial ‘insensitivity’ (invariance). The first problem relates to the reduction of signal resolution incurred by the repeated combination of max-pooling and downsampling (‘striding’) performed at every layer of standard DCNNs (Krizhevsky et al., 2013; Simonyan & Zisserman, 2014; Szegedy et al., 2014). Instead, as in Papandreou et al. (2014), we employ the ‘atrous’ (with holes) algorithm originally developed for efficiently computing the undecimated discrete wavelet transform (Mallat, 1999). This allows efficient dense computation of DCNN responses in a scheme substantially simpler than earlier solutions to this problem (Giusti et al., 2013; Sermanet et al., 2013).

이미지 라벨링 작업에 DCNN을 적용하는 데는 두 가지 기술적인 장애물이 있습니다. 신호 다운샘플링과 공간적 '무감각'(불변성)입니다. 첫 번째 문제는 표준 DCNN(Krizhevsky et al., 2013; Simonyan & Zisserman, 2014; Szegedy et al., ., 2014). 대신 Papandreou et al. (2014), 우리는 원래 undecimated 이산 웨이블릿 변환을 효율적으로 계산하기 위해 개발된 'atrous'(구멍 포함) 알고리즘을 사용합니다(Mallat, 1999). 이를 통해 이 문제에 대한 이전 솔루션보다 훨씬 단순한 방식으로 DCNN 응답을 효율적으로 조밀하게 계산할 수 있습니다(Giusti et al., 2013; Sermanet et al., 2013).

The second problem relates to the fact that obtaining object-centric decisions from a classifier requires invariance to spatial transformations, inherently limiting the spatial accuracy of the DCNN model. We boost our model’s ability to capture fine details by employing a fully-connected Conditional Random Field (CRF). Conditional Random Fields have been broadly used in semantic segmentation to combine class scores computed by multi-way classifiers with the low-level information captured by the local interactions of pixels and edges (Rother et al., 2004; Shotton et al., 2009) or superpixels (Lucchi et al., 2011). Even though works of increased sophistication have been proposed to model the hierarchical dependency (He et al., 2004; Ladicky et al., 2009; Lempitsky et al., 2011) and/or high-order dependencies of segments (Delong et al., 2012; Gonfaus et al., 2010; Kohli et al., 2009; Chen et al., 2013; Wang et al., 2015), we use the fully connected pairwise CRF proposed by Krahenb ¨ uhl & Koltun (2011) for its efficient computation, and ability to capture fine edge details ¨ while also catering for long range dependencies. That model was shown in Krahenb ¨ uhl & Koltun ¨ (2011) to largely improve the performance of a boosting-based pixel-level classifier, and in our work we demonstrate that it leads to state-of-the-art results when coupled with a DCNN-based pixel-level classifier.

두 번째 문제는 분류기에서 객체 중심 결정을 얻으려면 공간 변환에 대한 불변성이 필요하고 본질적으로 DCNN 모델의 공간 정확도를 제한한다는 사실과 관련이 있습니다. 우리는 완전히 연결된 CRF(Conditional Random Field)를 사용하여 세부 사항을 캡처하는 모델의 능력을 향상시킵니다. Conditional Random Fields는 의미론적 분할에서 광범위하게 사용되어 multi-way classifier에 의해 계산된 클래스 점수와 픽셀과 에지의 로컬 상호 작용에 의해 캡처된 낮은 수준의 정보를 결합합니다(Rother et al., 2004; Shotton et al., 2009). 또는 슈퍼픽셀(Lucchi et al., 2011). 계층적 종속성(He et al., 2004; Ladicky et al., 2009; Lempitsky et al., 2011) 및/또는 세그먼트의 고차 종속성을 모델링하기 위해 정교함을 높이는 작업이 제안되었지만(Delong et al. , 2012; Gonfaus et al., 2010; Kohli et al., 2009; Chen et al., 2013; Wang et al., 2015), 우리는 Krahenb ¨ uhl & Koltun(2011)이 제안한 완전 연결 쌍별 CRF를 사용합니다. 효율적인 계산, 그리고 장거리 종속성을 충족하면서 미세한 에지 세부 사항을 캡처하는 기능. 그 모델은 Krahenb ¨ uhl & Koltun ¨(2011)에서 부스팅 기반 픽셀 수준 분류기의 성능을 크게 향상시키는 것으로 나타났으며, 우리의 작업에서 우리는 이 모델이 DCNN 기반 픽셀 수준 분류기.

The three main advantages of our “DeepLab” system are (i) speed: by virtue of the ‘atrous’ algorithm, our dense DCNN operates at 8 fps, while Mean Field Inference for the fully-connected CRF requires 0.5 second, (ii) accuracy: we obtain state-of-the-art results on the PASCAL semantic segmentation challenge, outperforming the second-best approach of Mostajabi et al. (2014) by a margin of 7.2% and (iii) simplicity: our system is composed of a cascade of two fairly well-established modules, DCNNs and CRFs.

"DeepLab" 시스템의 세 가지 주요 이점은 (i) 속도입니다. 'atrous' 알고리즘 덕분에 고밀도 DCNN은 8fps에서 작동하는 반면 완전히 연결된 CRF에 대한 평균 필드 추론에는 0.5초가 필요합니다. (ii) 정확도: 우리는 Mostajabi et al.의 두 번째로 우수한 접근 방식을 능가하는 PASCAL 의미론적 세분화 문제에 대한 최신 결과를 얻습니다. (2014) 7.2%의 마진과 (iii) 단순성: 우리 시스템은 DCNN과 CRF라는 상당히 잘 정립된 두 모듈의 캐스케이드로 구성되어 있습니다.

**2 RELATED WORK**

Our system works directly on the pixel representation, similarly to Long et al. (2014). This is in contrast to the two-stage approaches that are now most common in semantic segmentation with DCNNs: such techniques typically use a cascade of bottom-up image segmentation and DCNN-based region classification, which makes the system commit to potential errors of the front-end segmentation system. For instance, the bounding box proposals and masked regions delivered by (Arbelaez et al., ´ 2014; Uijlings et al., 2013) are used in Girshick et al. (2014) and (Hariharan et al., 2014b) as inputs to a DCNN to introduce shape information into the classification process. Similarly, the authors of Mostajabi et al. (2014) rely on a superpixel representation. A celebrated non-DCNN precursor to these works is the second order pooling method of (Carreira et al., 2012) which also assigns labels to the regions proposals delivered by (Carreira & Sminchisescu, 2012). Understanding the perils of committing to a single segmentation, the authors of Cogswell et al. (2014) build on (Yadollahpour et al., 2013) to explore a diverse set of CRF-based segmentation proposals, computed also by (Carreira & Sminchisescu, 2012). These segmentation proposals are then re-ranked according to a DCNN trained in particular for this reranking task. Even though this approach explicitly tries to handle the temperamental nature of a front-end segmentation algorithm, there is still no explicit exploitation of the DCNN scores in the CRF-based segmentation algorithm: the DCNN is only applied post-hoc, while it would make sense to directly try to use its results during segmentation.

우리 시스템은 Long et al.과 유사하게 픽셀 표현에서 직접 작동합니다. (2014). 이것은 DCNN을 사용한 의미론적 분할에서 현재 가장 일반적으로 사용되는 2단계 접근 방식과 대조됩니다. 이러한 기술은 일반적으로 상향식 이미지 분할 및 DCNN 기반 영역 분류의 캐스케이드를 사용하므로 시스템이 프론트 엔드 세분화 시스템. 예를 들어, (Arbelaez et al., ' 2014; Uijlings et al., 2013)에 의해 전달된 경계 상자 제안 및 마스크 영역은 Girshick et al.에서 사용됩니다. (2014) 및 (Hariharan et al., 2014b) DCNN에 대한 입력으로 사용하여 분류 프로세스에 모양 정보를 도입합니다. 유사하게, Mostajabi et al. (2014) 슈퍼픽셀 표현에 의존합니다. 이러한 작업의 유명한 non-DCNN 전구체는 (Carreira et al., 2012) (Carreira & Sminchisescu, 2012)에 의해 전달된 지역 제안에 레이블을 할당하는 2차 풀링 방법입니다. Cogswell et al.의 저자는 단일 세분화에 대한 위험을 이해합니다. (2014) (Yadollahpour et al., 2013)를 기반으로 다양한 CRF 기반 세분화 제안을 탐색하며 (Carreira & Sminchisescu, 2012)도 계산했습니다. 그런 다음 이러한 세분화 제안은 특히 이 재지정 작업에 대해 훈련된 DCNN에 따라 재지정됩니다. 이 접근 방식이 명시적으로 프론트 엔드 분할 알고리즘의 변덕스러운 특성을 처리하려고 시도하지만 CRF 기반 분할 알고리즘에서 DCNN 점수의 명시적인 활용은 여전히 ​​없습니다. DCNN은 사후에만 적용되는 반면 세분화하는 동안 결과를 직접 사용하려는 감각.

Moving towards works that lie closer to our approach, several other researchers have considered the use of convolutionally computed DCNN features for dense image labeling. Among the first have been Farabet et al. (2013) who apply DCNNs at multiple image resolutions and then employ a segmentation tree to smooth the prediction results; more recently, Hariharan et al. (2014a) propose to concatenate the computed inter-mediate feature maps within the DCNNs for pixel classification, and Dai et al. (2014) propose to pool the inter-mediate feature maps by region proposals. Even though these works still employ segmentation algorithms that are decoupled from the DCNN classifier’s results, we believe it is advantageous that segmentation is only used at a later stage, avoiding the commitment to premature decisions.

우리의 접근 방식에 더 가까운 작업으로 이동하면서 여러 다른 연구자들이 밀집된 이미지 라벨링을 위해 컨볼루션 계산된 DCNN 기능의 사용을 고려했습니다. 첫 번째는 Farabet et al. (2013) 여러 이미지 해상도에서 DCNN을 적용한 다음 예측 결과를 평활화하기 위해 분할 트리를 사용합니다. 보다 최근에는 Hariharan et al. (2014a) 픽셀 분류를 위해 DCNN 내에서 계산된 중간 특징 맵을 연결하는 것을 제안하고 Dai et al. (2014)는 지역 제안별로 중간 특징 맵을 풀링할 것을 제안합니다. 이러한 작업은 여전히 ​​DCNN 분류기의 결과에서 분리된 분할 알고리즘을 사용하지만, 우리는 분할이 나중 단계에서만 사용되어 조기 결정에 대한 약속을 피하는 것이 유리하다고 생각합니다.

More recently, the segmentation-free techniques of (Long et al., 2014; Eigen & Fergus, 2014) directly apply DCNNs to the whole image in a sliding window fashion, replacing the last fully connected layers of a DCNN by convolutional layers. In order to deal with the spatial localization issues outlined in the beginning of the introduction, Long et al. (2014) upsample and concatenate the scores from inter-mediate feature maps, while Eigen & Fergus (2014) refine the prediction result from coarse to fine by propagating the coarse results to another DCNN.

보다 최근에(Long et al., 2014; Eigen & Fergus, 2014)의 무분할 기술은 DCNN을 슬라이딩 윈도우 방식으로 전체 이미지에 직접 적용하여 DCNN의 마지막 완전 연결 레이어를 컨볼루션 레이어로 대체합니다. 도입의 시작 부분에서 설명된 공간적 지역화 문제를 다루기 위해 Long et al. (2014) 중간 특성 맵의 점수를 업샘플링하고 연결하는 반면 Eigen & Fergus(2014)는 거친 결과를 다른 DCNN으로 전파하여 거친 결과를 미세하게 예측 결과를 수정합니다.

The main difference between our model and other state-of-the-art models is the combination of pixel-level CRFs and DCNN-based ‘unary terms’. Focusing on the closest works in this direction, Cogswell et al. (2014) use CRFs as a proposal mechanism for a DCNN-based reranking system, while Farabet et al. (2013) treat superpixels as nodes for a local pairwise CRF and use graph-cuts for discrete inference; as such their results can be limited by errors in superpixel computations, while ignoring long-range superpixel dependencies. Our approach instead treats every pixel as a CRF node, exploits long-range dependencies, and uses CRF inference to directly optimize a DCNN-driven cost function. We note that mean field had been extensively studied for traditional image segmentation/edge detection tasks, e.g., (Geiger & Girosi, 1991; Geiger & Yuille, 1991; Kokkinos et al., 2008), but recently Krahenb ¨ uhl & Koltun (2011) showed that the inference can be very efficient for ¨ fully connected CRF and particularly effective in the context of semantic segmentation.

우리 모델과 다른 최첨단 모델의 주요 차이점은 픽셀 수준 CRF와 DCNN 기반 '단항 항'의 조합입니다. 이 방향에서 가장 가까운 연구에 초점을 맞추면서 Cogswell et al. (2014) CRF를 DCNN 기반 순위 재지정 시스템에 대한 제안 메커니즘으로 사용하는 반면 Farabet et al. (2013) 슈퍼픽셀을 로컬 pairwise CRF의 노드로 처리하고 이산 추론을 위해 그래프 컷을 사용합니다. 따라서 장거리 슈퍼픽셀 종속성을 무시하면서 슈퍼픽셀 계산의 오류로 인해 결과가 제한될 수 있습니다. 대신 우리의 접근 방식은 모든 픽셀을 CRF 노드로 취급하고 장거리 종속성을 활용하며 CRF 추론을 사용하여 DCNN 기반 비용 함수를 직접 최적화합니다. 평균 필드는 예를 들어(Geiger & Girosi, 1991; Geiger & Yuille, 1991; Kokkinos et al., 2008) 전통적인 이미지 분할/에지 감지 작업에 대해 광범위하게 연구되었지만 최근에는 Krahenb ¨ uhl & Koltun(2011) )는 추론이 ¨ 완전히 연결된 CRF에 대해 매우 효율적일 수 있으며 특히 의미론적 분할의 맥락에서 효과적일 수 있음을 보여주었습니다.

After the first version of our manuscript was made publicly available, it came to our attention that two other groups have independently and concurrently pursued a very similar direction, combining DCNNs and densely connected CRFs (Bell et al., 2014; Zheng et al., 2015). There are several differences in technical aspects of the respective models. Bell et al. (2014) focus on the problem of material classification, while Zheng et al. (2015) unroll the CRF mean-field inference steps to convert the whole system into an end-to-end trainable feed-forward network.

우리 원고의 첫 번째 버전이 공개된 후 DCNN과 조밀하게 연결된 CRF를 결합하여 두 개의 다른 그룹이 독립적으로 동시에 매우 유사한 방향을 추구했음을 알게 되었습니다(Bell et al., 2014; Zheng et al., 2015). 각 모델의 기술적 측면에는 몇 가지 차이점이 있습니다. Bell et al. (2014) Zheng et al. (2015) CRF 평균 필드 추론 단계를 펼쳐 전체 시스템을 종단 간 학습 가능한 피드포워드 네트워크로 변환합니다.

We have updated our proposed “DeepLab” system with much improved methods and results in our latest work (Chen et al., 2016). We refer the interested reader to the paper for details.

우리는 제안된 "DeepLab" 시스템을 훨씬 개선된 방법과 최신 작업 결과로 업데이트했습니다(Chen et al., 2016). 자세한 내용은 관심 있는 독자에게 해당 논문을 참조하십시오.

**3 CONVOLUTIONAL NEURAL NETWORKS FOR DENSE IMAGE LABELING**

Herein we describe how we have re-purposed and finetuned the publicly available Imagenetpretrained state-of-art 16-layer classification network of (Simonyan & Zisserman, 2014) (VGG-16) into an efficient and effective dense feature extractor for our dense semantic image segmentation system.

여기에서 우리는 (Simonyan & Zisserman, 2014) (VGG-16)의 공개적으로 사용 가능한 Imagenet 사전 훈련된 최첨단 16계층 분류 네트워크를 우리의 조밀한 의미 체계를 위한 효율적이고 효과적인 조밀한 특징 추출기로 용도를 변경하고 미세 조정한 방법을 설명합니다. 이미지 분할 시스템.

**3.1 EFFICIENT DENSE SLIDING WINDOW FEATURE EXTRACTION WITH THE HOLE ALGORITHM**

Dense spatial score evaluation is instrumental in the success of our dense CNN feature extractor. As a first step to implement this, we convert the fully-connected layers of VGG-16 into convolutional ones and run the network in a convolutional fashion on the image at its original resolution. However this is not enough as it yields very sparsely computed detection scores (with a stride of 32 pixels). To compute scores more densely at our target stride of 8 pixels, we develop a variation of the method previously employed by Giusti et al. (2013); Sermanet et al. (2013). We skip subsampling after the last two max-pooling layers in the network of Simonyan & Zisserman (2014) and modify the convolutional filters in the layers that follow them by introducing zeros to increase their length (2×in the last three convolutional layers and 4×in the first fully connected layer). We can implement this more efficiently by keeping the filters intact and instead sparsely sample the feature maps on which they are applied on using an input stride of 2 or 4 pixels, respectively. This approach, illustrated in Fig. 1 is known as the ‘hole algorithm’ (‘atrous algorithm’) and has been developed before for efficient computation of the undecimated wavelet transform (Mallat, 1999). We have implemented this within the Caffe framework (Jia et al., 2014) by adding to the im2col function (it converts multichannel feature maps to vectorized patches) the option to sparsely sample the underlying feature map. This approach is generally applicable and allows us to efficiently compute dense CNN feature maps at any target subsampling rate without introducing any approximations.

조밀한 공간 점수 평가는 조밀한 CNN 특징 추출기의 성공에 중요한 역할을 합니다. 이를 구현하기 위한 첫 번째 단계로 VGG-16의 완전 연결 레이어를 컨볼루션 레이어로 변환하고 원래 해상도의 이미지에서 컨볼루션 방식으로 네트워크를 실행합니다. 그러나 이것은 매우 드물게 계산된 감지 점수(보폭 32픽셀)를 산출하기 때문에 충분하지 않습니다. 8픽셀의 목표 보폭에서 점수를 더 조밀하게 계산하기 위해 이전에 Giusti 등이 사용했던 방법의 변형을 개발합니다. (2013); Sermanet et al. (2013). 우리는 Simonyan & Zisserman(2014) 네트워크에서 마지막 두 개의 최대 풀링 레이어 이후의 서브샘플링을 건너뛰고 길이를 늘리기 위해 0을 도입하여 뒤따르는 레이어의 컨볼루션 필터를 수정합니다(마지막 세 개의 컨볼루션 레이어에서 2x, × 첫 번째 완전 연결 계층). 필터를 그대로 유지하고 대신 각각 2픽셀 또는 4픽셀의 입력 스트라이드를 사용하여 필터가 적용되는 기능 맵을 드물게 샘플링하여 이를 보다 효율적으로 구현할 수 있습니다. 그림 1에 설명된 이 접근 방식은 '홀 알고리즘'('atrous 알고리즘')으로 알려져 있으며 데시메이션되지 않은 웨이블릿 변환의 효율적인 계산을 위해 이전에 개발되었습니다(Mallat, 1999). 기본 기능 맵을 드물게 샘플링하는 옵션을 im2col 기능(다중 채널 기능 맵을 벡터화된 패치로 변환)에 추가하여 Caffe 프레임워크(Jia et al., 2014) 내에서 이것을 구현했습니다. 이 접근 방식은 일반적으로 적용 가능하며 근사를 도입하지 않고도 모든 대상 서브샘플링 속도로 고밀도 CNN 기능 맵을 효율적으로 계산할 수 있습니다.

텍스트, 손목시계이(가) 표시된 사진

자동 생성된 설명

Figure 1: Illustration of the hole algorithm in 1-D, when kernel size = 3, input stride = 2, and output stride = 1.

그림 1: 커널 크기 = 3, 입력 스트라이드 = 2, 출력 스트라이드 = 1일 때 1차원의 구멍 알고리즘 그림.

We finetune the model weights of the Imagenet-pretrained VGG-16 network to adapt it to the image classification task in a straightforward fashion, following the procedure of Long et al. (2014). We replace the 1000-way Imagenet classifier in the last layer of VGG-16 with a 21-way one. Our loss function is the sum of cross-entropy terms for each spatial position in the CNN output map (subsampled by 8 compared to the original image). All positions and labels are equally weighted in the overall loss function. Our targets are the ground truth labels (subsampled by 8). We optimize the objective function with respect to the weights at all network layers by the standard SGD procedure of Krizhevsky et al. (2013).

Imagenet 사전 훈련된 VGG-16 네트워크의 모델 가중치를 미세 조정하여 Long et al의 절차에 따라 간단한 방식으로 이미지 분류 작업에 적용합니다. (2014). VGG-16의 마지막 레이어에 있는 1000-way Imagenet 분류기를 21-way로 교체합니다. 손실 함수는 CNN 출력 맵의 각 공간 위치에 대한 교차 엔트로피 항의 합입니다(원본 이미지와 비교하여 8로 서브샘플링됨). 모든 위치와 레이블은 전체 손실 함수에서 동일하게 가중됩니다. 우리의 목표는 정답 레이블(8로 서브샘플링됨)입니다. Krizhevsky et al.의 표준 SGD 절차에 따라 모든 네트워크 계층의 가중치에 대한 목적 함수를 최적화합니다. (2013).

During testing, we need class score maps at the original image resolution. As illustrated in Figure 2 and further elaborated in Section 4.1, the class score maps (corresponding to log-probabilities) are quite smooth, which allows us to use simple bilinear interpolation to increase their resolution by a factor of 8 at a negligible computational cost. Note that the method of Long et al. (2014) does not use the hole algorithm and produces very coarse scores (subsampled by a factor of 32) at the CNN output. This forced them to use learned upsampling layers, significantly increasing the complexity and training time of their system: Fine-tuning our network on PASCAL VOC 2012 takes about 10 hours, while they report a training time of several days (both timings on a modern GPU).

테스트하는 동안 원본 이미지 해상도의 클래스 점수 맵이 필요합니다. 그림 2에 설명되어 있고 섹션 4.1에서 자세히 설명된 것처럼 클래스 점수 맵(로그 확률에 해당)은 매우 부드럽습니다. 이를 통해 간단한 쌍선형 보간법을 사용하여 무시할 수 있는 계산 비용으로 해상도를 8배 증가시킬 수 있습니다. Long et al. (2014)은 구멍 알고리즘을 사용하지 않으며 CNN 출력에서 ​​매우 거친 점수(32의 인수로 부분 샘플링됨)를 생성합니다. 이로 인해 학습된 업샘플링 레이어를 사용해야 했고 시스템의 복잡성과 교육 시간이 크게 증가했습니다. PASCAL VOC 2012에서 네트워크를 미세 조정하는 데는 약 10시간이 소요되지만 교육 시간은 며칠로 보고됩니다(두 시간 모두 최신 GPU에서 ).

**3.2 CONTROLLING THE RECEPTIVE FIELD SIZE AND ACCELERATING DENSE COMPUTATION WITH CONVOLUTIONAL NETS**

Another key ingredient in re-purposing our network for dense score computation is explicitly controlling the network’s receptive field size. Most recent DCNN-based image recognition methods rely on networks pre-trained on the Imagenet large-scale classification task. These networks typically have large receptive field size: in the case of the VGG-16 net we consider, its receptive field is 224×224 (with zero-padding) and 404×404 pixels if the net is applied convolutionally. After converting the network to a fully convolutional one, the first fully connected layer has 4,096 filters of large 7×7 spatial size and becomes the computational bottleneck in our dense score map computation.

고밀도 점수 계산을 위해 네트워크를 용도 변경하는 또 다른 핵심 요소는 네트워크의 수용 필드 크기를 명시적으로 제어하는 ​​것입니다. 가장 최근의 DCNN 기반 이미지 인식 방법은 Imagenet 대규모 분류 작업에 대해 사전 훈련된 네트워크에 의존합니다. 이러한 네트워크는 일반적으로 큰 수용 필드 크기를 갖습니다. 우리가 고려하는 VGG-16 네트의 경우 수신 필드는 224×224(제로 패딩 포함) 및 넷이 컨볼루션으로 적용되는 경우 404×404 픽셀입니다. 네트워크를 완전 컨볼루션으로 변환한 후 첫 번째 완전 연결 계층은 7×7 공간 크기의 4,096개 필터를 가지며 고밀도 점수 맵 계산에서 계산 병목 현상이 됩니다.

We have addressed this practical problem by spatially subsampling (by simple decimation) the first FC layer to 4×4 (or 3×3) spatial size. This has reduced the receptive field of the network down to 128×128 (with zero-padding) or 308×308 (in convolutional mode) and has reduced computation time for the first FC layer by 2 − 3 times. Using our Caffe-based implementation and a Titan GPU, the resulting VGG-derived network is very efficient: Given a 306×306 input image, it produces 39×39 dense raw feature scores at the top of the network at a rate of about 8 frames/sec during testing. The speed during training is 3 frames/sec. We have also successfully experimented with reducing the number of channels at the fully connected layers from 4,096 down to 1,024, considerably further decreasing computation time and memory footprint without sacrificing performance, as detailed in Section 5. Using smaller networks such as Krizhevsky et al. (2013) could allow video-rate test-time dense feature computation even on light-weight GPUs.

우리는 첫 번째 FC 레이어를 4x4(또는 3x3) 공간 크기로 공간적으로 서브샘플링(단순 데시메이션)하여 이 실용적인 문제를 해결했습니다. 이것은 네트워크의 수신 필드를 128x128(제로 패딩 포함) 또는 308x308(컨볼루션 모드)으로 줄였으며 첫 번째 FC 계층에 대한 계산 시간을 2 - 3배 단축했습니다. Caffe 기반 구현과 Titan GPU를 사용하여 결과적으로 VGG 파생 네트워크는 매우 효율적입니다. 306×306 입력 이미지가 주어지면 약 8의 비율로 네트워크 상단에서 39×39 조밀한 원시 기능 점수를 생성합니다. 테스트 중 프레임/초. 훈련 중 속도는 3 프레임/초입니다. 또한 섹션 5에 자세히 설명된 대로 완전 연결 계층의 채널 수를 4,096개에서 1,024개로 줄이는 실험을 성공적으로 수행했습니다. (2013)은 경량 GPU에서도 비디오 속도 테스트 시간 밀도 기능 계산을 허용할 수 있습니다.

**4 DETAILED BOUNDARY RECOVERY: FULLY-CONNECTED CONDITIONAL RANDOM FIELDS AND MULTI-SCALE PREDICTION**

**4.1 DEEP CONVOLUTIONAL NETWORKS AND THE LOCALIZATION CHALLENGE**

As illustrated in Figure 2, DCNN score maps can reliably predict the presence and rough position of objects in an image but are less well suited for pin-pointing their exact outline. There is a natural trade-off between classification accuracy and localization accuracy with convolutional networks: Deeper models with multiple max-pooling layers have proven most successful in classification tasks, however their increased invariance and large receptive fields make the problem of inferring position from the scores at their top output levels more challenging.

그림 2에서 볼 수 있는 것처럼 DCNN 점수 맵은 이미지에서 개체의 존재와 대략적인 위치를 안정적으로 예측할 수 있지만 정확한 윤곽을 가리키는 데는 적합하지 않습니다. 컨볼루션 네트워크에서는 분류 정확도와 현지화 정확도 사이에 자연스러운 균형이 있습니다. 다중 최대 풀링 레이어가 있는 더 깊은 모델은 분류 작업에서 가장 성공적인 것으로 입증되었지만, 증가된 불변성과 큰 수용 필드로 인해 점수에서 위치를 추론하는 문제가 발생합니다. 그들의 최고 출력 수준에서 더 도전적입니다.

Recent work has pursued two directions to address this localization challenge. The first approach is to harness information from multiple layers in the convolutional network in order to better estimate the object boundaries (Long et al., 2014; Eigen & Fergus, 2014). The second approach is to employ a super-pixel representation, essentially delegating the localization task to a low-level segmentation method. This route is followed by the very successful recent method of Mostajabi et al. (2014).

최근 작업에서는 이러한 현지화 문제를 해결하기 위해 두 가지 방향을 추구했습니다. 첫 번째 접근 방식은 객체 경계를 더 잘 추정하기 위해 컨볼루션 네트워크의 여러 계층에서 정보를 활용하는 것입니다(Long et al., 2014; Eigen & Fergus, 2014). 두 번째 접근 방식은 슈퍼 픽셀 표현을 사용하는 것인데, 기본적으로 로컬라이제이션 작업을 저수준 분할 방법에 위임하는 것입니다. 이 경로는 Mostajabi et al.의 매우 성공적인 최근 방법으로 이어집니다. (2014).

In Section 4.2, we pursue a novel alternative direction based on coupling the recognition capacity of DCNNs and the fine-grained localization accuracy of fully connected CRFs and show that it is remarkably successful in addressing the localization challenge, producing accurate semantic segmentation results and recovering object boundaries at a level of detail that is well beyond the reach of existing methods.

섹션 4.2에서 우리는 DCNN의 인식 능력과 완전히 연결된 CRF의 세분화된 위치 정확도를 결합하여 새로운 대안 방향을 추구하고 이것이 위치 문제를 해결하고 정확한 의미론적 분할 결과를 생성하며 객체를 복구하는 데 매우 성공적임을 보여줍니다. 기존 방법의 범위를 훨씬 넘어선 세부 수준의 경계.

**4.2 FULLY-CONNECTED CONDITIONAL RANDOM FIELDS FOR ACCURATE LOCALIZATION**

텍스트, 모니터, 화면, 자연이(가) 표시된 사진

자동 생성된 설명

Figure 2: Score map (input before softmax function) and belief map (output of softmax function) for Aeroplane. We show the score (1st row) and belief (2nd row) maps after each mean field iteration. The output of last DCNN layer is used as input to the mean field inference. Best viewed in color.

그림 2: Aeroplane에 대한 점수 맵(softmax 함수 이전 입력) 및 신념 맵(softmax 함수 출력). 각 평균 필드 반복 후에 점수(첫 번째 행)와 신념(2행) 맵을 표시합니다. 마지막 DCNN 레이어의 출력은 평균 필드 추론에 대한 입력으로 사용됩니다. 색상으로 가장 잘 보입니다.

Traditionally, conditional random fields (CRFs) have been employed to smooth noisy segmentation maps (Rother et al., 2004; Kohli et al., 2009). Typically these models contain energy terms that couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. Qualitatively, the primary function of these short-range CRFs has been to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features.

전통적으로 조건부 랜덤 필드(CRF)는 잡음이 있는 분할 맵을 부드럽게 하기 위해 사용되었습니다(Rother et al., 2004; Kohli et al., 2009). 일반적으로 이러한 모델에는 인접 노드를 연결하는 에너지 항이 포함되어 있어 공간적으로 근접한 픽셀에 동일한 레이블 할당을 선호합니다. 질적으로, 이러한 단거리 CRF의 주요 기능은 로컬 손으로 엔지니어링된 기능 위에 구축된 약한 분류기의 스퓨리어스 예측을 정리하는 것입니다.

Compared to these weaker classifiers, modern DCNN architectures such as the one we use in this work produce score maps and semantic label predictions which are qualitatively different. As illustrated in Figure 2, the score maps are typically quite smooth and produce homogeneous classification results. In this regime, using short-range CRFs can be detrimental, as our goal should be to recover detailed local structure rather than further smooth it. Using contrast-sensitive potentials (Rother et al., 2004) in conjunction to local-range CRFs can potentially improve localization but still miss thin-structures and typically requires solving an expensive discrete optimization problem.

이러한 약한 분류기와 비교하여 우리가 이 작업에서 사용하는 것과 같은 최신 DCNN 아키텍처는 질적으로 다른 점수 맵과 의미 레이블 예측을 생성합니다. 그림 2에서 볼 수 있듯이 점수 맵은 일반적으로 매우 매끄럽고 균질한 분류 결과를 생성합니다. 이 체제에서 단거리 CRF를 사용하는 것은 해로울 수 있습니다. 우리의 목표는 더 매끄럽게 하기보다는 상세한 로컬 구조를 복구하는 것이어야 하기 때문입니다. 로컬 범위 CRF와 함께 대비에 민감한 전위(Rother et al., 2004)를 사용하면 잠재적으로 현지화를 개선할 수 있지만 여전히 얇은 구조를 놓치고 일반적으로 값비싼 이산 최적화 문제를 해결해야 합니다.

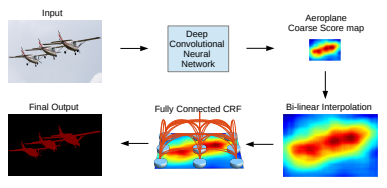


Figure 3: Model Illustration. The coarse score map from Deep Convolutional Neural Network (with fully convolutional layers) is upsampled by bi-linear interpolation. A fully connected CRF is applied to refine the segmentation result. Best viewed in color.

그림 3: 모델 일러스트레이션. Deep Convolutional Neural Network(완전한 컨볼루션 레이어 포함)의 거친 점수 맵은 이중 선형 보간법으로 업샘플링됩니다. 완전 연결 CRF는 세분화 결과를 구체화하기 위해 적용됩니다. 색상으로 가장 잘 보입니다.

To overcome these limitations of short-range CRFs, we integrate into our system the fully connected CRF model of Krahenb ¨ uhl & Koltun (2011). The model employs the energy function

단거리 CRF의 이러한 한계를 극복하기 위해 우리는 Krahenb ¨ uhl & Koltun(2011)의 완전히 연결된 CRF 모델을 시스템에 통합합니다. 모델은 에너지 함수를 사용합니다.



where x is the label assignment for pixels. We use as unary potential , where is the label assignment probability at pixel i as computed by DCNN. The pairwise potential is , and zero otherwise (i.e., Potts Model). There is one pairwise term for each pair of pixels i and j in the image no matter how far from each other they lie, i.e. the model’s factor graph is fully connected. Each is the Gaussian kernel depends on features (denoted as f) extracted for pixel i and j and is weighted by parameter . We adopt bilateral position and color terms, specifically, the kernels are

여기서 x는 픽셀에 대한 레이블 할당입니다. 단항 전위로 사용합니다. 여기서 는 DCNN에서 계산한 픽셀 i에서의 레이블 할당 확률입니다. 쌍방향 전위는 , 여기서 이고 그렇지 않으면 0입니다(즉, Potts 모델). 서로 얼마나 멀리 떨어져 있든 이미지의 각 픽셀 i 및 j 쌍에 대해 하나의 쌍별 항이 있습니다. 즉, 모델의 요인 그래프가 완전히 연결되어 있습니다. 각 은 픽셀 i와 j에 대해 추출된 특징(f로 표시)에 의존하는 가우시안 커널이며 매개변수 에 의해 가중치가 부여됩니다. 우리는 양측 위치 및 색상 용어를 채택합니다. 특히 커널은 다음과 같습니다.



where the first kernel depends on both pixel positions (denoted as p) and pixel color intensities (denoted as I), and the second kernel only depends on pixel positions. The hyper parameters control the “scale” of the Gaussian kernels.

여기서 첫 번째 커널은 픽셀 위치(p로 표시)와 픽셀 색상 강도(I로 표시)에 모두 의존하고 두 번째 커널은 픽셀 위치에만 의존합니다. 하이퍼 매개변수 parameters 및 는 가우스 커널의 "규모"를 제어합니다.

Crucially, this model is amenable to efficient approximate probabilistic inference (Krahenb ¨ uhl & ¨ Koltun, 2011). The message passing updates under a fully decomposable mean field approximation can be expressed as convolutions with a Gaussian kernel in feature space. High-dimensional filtering algorithms (Adams et al., 2010) significantly speed-up this computation resulting in an algorithm that is very fast in practice, less that 0.5 sec on average for Pascal VOC images using the publicly available implementation of (Krahenb ¨ uhl & Koltun, 2011).

결정적으로, 이 모델은 효율적인 근사 확률 추론에 적합합니다(Krahenb ¨ uhl & ¨ Koltun, 2011). 완전 분해 가능한 평균 필드 근사 하에서 업데이트를 전달하는 메시지는 기능 공간에서 가우시안 커널을 사용한 컨볼루션으로 표현될 수 있습니다. 고차원 필터링 알고리즘(Adams et al., 2010)은 공개적으로 사용 가능한 (Krahenb ¨ uhl) 구현을 사용하여 Pascal VOC 이미지에 대해 평균 0.5초 미만인 실제로 매우 빠른 알고리즘을 생성하는 이 계산 속도를 크게 높입니다. & 콜툰, 2011).

**4.3 MULTI-SCALE PREDICTION**

Following the promising recent results of (Hariharan et al., 2014a; Long et al., 2014) we have also explored a multi-scale prediction method to increase the boundary localization accuracy. Specifically, we attach to the input image and the output of each of the first four max pooling layers a two-layer MLP (first layer: 128 3x3 convolutional filters, second layer: 128 1x1 convolutional filters) whose feature map is concatenated to the main network’s last layer feature map. The aggregate feature map fed into the softmax layer is thus enhanced by 5 \* 128 = 640 channels. We only adjust the newly added weights, keeping the other network parameters to the values learned by the method of Section 3. As discussed in the experimental section, introducing these extra direct connections from fine-resolution layers improves localization performance, yet the effect is not as dramatic as the one obtained with the fully-connected CRF.

유망한 최근 결과(Hariharan et al., 2014a; Long et al., 2014)에 따라 경계 위치 정확도를 높이기 위한 다중 규모 예측 방법도 탐색했습니다. 구체적으로, 우리는 입력 이미지와 처음 4개의 최대 풀링 레이어 각각의 출력에 2층 MLP(첫 번째 레이어: 128 3x3 컨볼루션 필터, 두 번째 레이어: 128 1x1 컨볼루션 필터)를 첨부합니다. 네트워크의 마지막 레이어 기능 맵. 따라서 softmax 레이어에 제공되는 집계 기능 맵은 5 \* 128 = 640개 채널만큼 향상됩니다. 새로 추가된 가중치만 조정하고 다른 네트워크 매개변수는 섹션 3의 방법으로 학습한 값으로 유지합니다. 실험 섹션에서 논의한 바와 같이 미세 해상도 레이어에서 이러한 추가 직접 연결을 도입하면 로컬라이제이션 성능이 향상되지만 효과는 그렇지 않습니다. 완전히 연결된 CRF로 얻은 것만큼 극적입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 1: (a) Performance of our proposed models on the PASCAL VOC 2012 ‘val’ set (with training in the augmented ‘train’ set). The best performance is achieved by exploiting both multi-scale features and large field-of-view. (b) Performance of our proposed models (with training in the augmented ‘trainval’ set) compared to other state-of-art methods on the PASCAL VOC 2012 ‘test’ set.

표 1: (a) PASCAL VOC 2012 'val' 세트에서 제안된 모델의 성능(증강된 'train' 세트에서 훈련 포함). 다중 스케일 기능과 넓은 시야를 모두 활용하여 최상의 성능을 얻을 수 있습니다. (b) PASCAL VOC 2012 '테스트' 세트에서 다른 최첨단 방법과 비교하여 제안된 모델(증강된 'trainval' 세트에서 훈련 포함)의 성능.

**5 EXPERIMENTAL EVALUATION**

Dataset We test our DeepLab model on the PASCAL VOC 2012 segmentation benchmark (Everingham et al., 2014), consisting of 20 foreground object classes and one background class. The original dataset contains 1, 464, 1, 449, and 1, 456 images for training, validation, and testing, respectively. The dataset is augmented by the extra annotations provided by Hariharan et al. (2011), resulting in 10, 582 training images. The performance is measured in terms of pixel intersectionover-union (IOU) averaged across the 21 classes.

데이터 세트 20개의 전경 객체 클래스와 1개의 배경 클래스로 구성된 PASCAL VOC 2012 분할 벤치마크(Everingham et al., 2014)에서 DeepLab 모델을 테스트합니다. 원본 데이터 세트에는 각각 훈련, 검증 및 테스트를 위한 1, 464, 1, 449 및 1,456개의 이미지가 포함되어 있습니다. 데이터 세트는 Hariharan et al.에서 제공한 추가 주석으로 보강됩니다. (2011), 10,582개의 훈련 이미지가 생성되었습니다. 성능은 21개 클래스에서 평균화된 IOU(픽셀 교차 결합)로 측정됩니다.

Training We adopt the simplest form of piecewise training, decoupling the DCNN and CRF training stages, assuming the unary terms provided by the DCNN are fixed during CRF training.

훈련 우리는 DCNN에서 제공하는 단항 항이 CRF 훈련 중에 고정되어 있다고 가정하고 DCNN과 CRF 훈련 단계를 분리하는 가장 단순한 형태의 조각별 훈련을 채택합니다.

For DCNN training we employ the VGG-16 network which has been pre-trained on ImageNet. We fine-tuned the VGG-16 network on the VOC 21-way pixel-classification task by stochastic gradient descent on the cross-entropy loss function, as described in Section 3.1. We use a mini-batch of 20 images and initial learning rate of 0.001 (0.01 for the final classifier layer), multiplying the learning rate by 0.1 at every 2000 iterations. We use momentum of 0.9 and a weight decay of 0.0005.

DCNN 교육을 위해 ImageNet에서 사전 교육된 VGG-16 네트워크를 사용합니다. 섹션 3.1에 설명된 대로 교차 엔트로피 손실 함수에 대한 확률적 경사 하강법으로 VOC 21방향 픽셀 분류 작업에서 VGG-16 네트워크를 미세 조정했습니다. 20개의 이미지로 구성된 미니 배치와 0.001(최종 분류기 계층의 경우 0.01)의 초기 학습률을 사용하여 2000회 반복할 때마다 학습률을 0.1로 곱합니다. 0.9의 운동량과 0.0005의 가중치 감소를 사용합니다.

After the DCNN has been fine-tuned, we cross-validate the parameters of the fully connected CRF model in Eq. (2) along the lines of Krahenb ¨ uhl & Koltun (2011). We use the default values of ¨ and we search for the best values of by cross-validation on a small subset of the validation set (we use 100 images). We employ coarse-to-fine search scheme. Specifically, the initial search range of the parameters are (MATLAB notation), and then we refine the search step sizes around the first round’s best values. We fix the number of mean field iterations to 10 for all reported experiments.

DCNN이 미세 조정된 후 Eq.에서 완전히 연결된 CRF 모델의 매개변수를 교차 검증합니다. (2) Krahenb ¨ uhl & Koltun(2011)의 선을 따라. ¨ ¨ 및 의 기본값을 사용하고 검증 세트의 작은 하위 집합(100개 이미지 사용)에 대한 교차 검증을 통해 및의 최상의 값을 검색합니다. 우리는 대략적인 검색 방식을 사용합니다. 구체적으로, 매개변수의 초기 검색 범위는 및 (MATLAB 표기법)이며 검색 단계 크기를 세분화합니다. 첫 번째 라운드의 최고의 가치를 중심으로. 보고된 모든 실험에 대해 평균 필드 반복 횟수를 10으로 고정합니다.

Evaluation on Validation set We conduct the majority of our evaluations on the PASCAL ‘val’ set, training our model on the augmented PASCAL ‘train’ set. As shown in Tab. 1 (a), incorporating the fully connected CRF to our model (denoted by DeepLab-CRF) yields a substantial performance boost, about 4% improvement over DeepLab. We note that the work of Krahenb ¨ uhl & Koltun ¨ (2011) improved the 27.6% result of TextonBoost (Shotton et al., 2009) to 29.1%, which makes the improvement we report here (from 59.8% to 63.7%) all the more impressive.

검증 세트에 대한 평가 우리는 PASCAL 'val' 세트에 대해 대부분의 평가를 수행하고, 증강된 PASCAL 'train' 세트에서 모델을 훈련합니다. 탭에 표시된 대로. 1(a), 완전히 연결된 CRF를 우리 모델(DeepLab-CRF로 표시)에 통합하면 DeepLab에 비해 약 4% 개선된 상당한 성능 향상을 얻을 수 있습니다. 우리는 Krahenb ¨ uhl & Koltun ¨(2011)의 작업이 TextonBoost(Shotton et al., 2009)의 27.6% 결과를 29.1%로 개선했으며, 이는 여기에서 보고하는 개선(59.8%에서 63.7%)을 모두 만듭니다. 더 인상적이다.

Turning to qualitative results, we provide visual comparisons between DeepLab and DeepLab-CRF in Fig. 7. Employing a fully connected CRF significantly improves the results, allowing the model to accurately capture intricate object boundaries.

정성적 결과로 전환하여 그림 7에서 DeepLab과 DeepLab-CRF 간의 시각적 비교를 제공합니다. 완전히 연결된 CRF를 사용하면 결과가 크게 향상되어 모델이 복잡한 객체 경계를 정확하게 캡처할 수 있습니다.

Multi-Scale features We also exploit the features from the intermediate layers, similar to Hariharan et al. (2014a); Long et al. (2014). As shown in Tab. 1 (a), adding the multi-scale features to our 7 DeepLab model (denoted as DeepLab-MSc) improves about 1.5% performance, and further incorporating the fully connected CRF (denoted as DeepLab-MSc-CRF) yields about 4% improvement. The qualitative comparisons between DeepLab and DeepLab-MSc are shown in Fig. 4. Leveraging the multi-scale features can slightly refine the object boundaries.

다중 스케일 기능 우리는 또한 Hariharan et al.과 유사한 중간 계층의 기능을 이용합니다. (2014a); Long et al. (2014). 탭에 표시된 대로. 1(a), 7개의 DeepLab 모델(DeepLab-MSc로 표시)에 다중 스케일 기능을 추가하면 성능이 약 1.5% 향상되고 완전히 연결된 CRF(DeepLab-MSc-CRF로 표시)를 추가로 통합하면 약 4% 향상됩니다. DeepLab과 DeepLab-MSc 간의 정성적 비교는 그림 4에 나와 있습니다. 다중 스케일 기능을 활용하면 개체 경계를 약간 세분화할 수 있습니다.

텍스트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Table 2: Effect of Field-Of-View. We show the performance (after CRF) and training speed on the PASCAL VOC 2012 ‘val’ set as the function of (1) the kernel size of first fully connected layer, (2) the input stride value employed in the atrous algorithm

표 2: 시야의 효과. (1) 첫 번째 완전 연결 계층의 커널 크기, (2) atrous 알고리즘에 사용된 입력 stride 값의 함수로 설정된 PASCAL VOC 2012 'val'에 대한 성능(CRF 이후) 및 훈련 속도를 보여줍니다.

Field of View The ‘atrous algorithm’ we employed allows us to arbitrarily control the Field-ofView (FOV) of the models by adjusting the input stride, as illustrated in Fig. 1. In Tab. 2, we experiment with several kernel sizes and input strides at the first fully connected layer. The method, DeepLab-CRF-7x7, is the direct modification from VGG-16 net, where the kernel size = 7×7 and input stride = 4. This model yields performance of 67.64% on the ‘val’ set, but it is relatively slow (1.44 images per second during training). We have improved model speed to 2.9 images per second by reducing the kernel size to 4×4. We have experimented with two such network variants with different FOV sizes, DeepLab-CRF and DeepLab-CRF-4x4; the latter has large FOV (i.e., large input stride) and attains better performance. Finally, we employ kernel size 3×3 and input stride = 12, and further change the filter sizes from 4096 to 1024 for the last two layers. Interestingly, the resulting model, DeepLab-CRF-LargeFOV, matches the performance of the expensive DeepLabCRF-7x7. At the same time, it is 3.36 times faster to run and has significantly fewer parameters (20.5M instead of 134.3M).

Field of View 우리가 사용한 'atrous algorithm'은 그림 1과 같이 입력 보폭을 조정하여 모델의 Field-ofView(FOV)를 임의로 제어할 수 있습니다. 2, 우리는 첫 번째 완전 연결 계층에서 여러 커널 크기와 입력 스트라이드를 실험합니다. DeepLab-CRF-7x7 방법은 커널 크기 = 7×7 및 입력 보폭 = 4인 VGG-16 net에서 직접 수정한 것입니다. 이 모델은 'val' 세트에서 67.64%의 성능을 산출하지만 다음과 같습니다. 상대적으로 느립니다(훈련 중 초당 1.44개 이미지). 커널 크기를 4×4로 줄여 모델 속도를 초당 2.9개 이미지로 개선했습니다. 우리는 FOV 크기가 다른 DeepLab-CRF 및 DeepLab-CRF-4x4의 두 가지 네트워크 변형을 실험했습니다. 후자는 큰 FOV(즉, 큰 입력 보폭)를 가지며 더 나은 성능을 얻습니다. 마지막으로 커널 크기 3×3 및 입력 stride = 12를 사용하고 마지막 두 계층에 대해 필터 크기를 4096에서 1024로 추가로 변경합니다. 흥미롭게도 결과 모델인 DeepLab-CRF-LargeFOV는 값비싼 DeepLabCRF-7x7의 성능과 일치합니다. 동시에 실행 속도가 3.36배 빠르고 매개변수가 훨씬 적습니다(134.3M 대신 20.5M).

The performance of several model variants is summarized in Tab. 1, showing the benefit of exploiting multi-scale features and large FOV.

여러 모델 변형의 성능이 탭에 요약되어 있습니다. 1, 다중 스케일 기능 및 대형 FOV 활용의 이점을 보여줍니다.

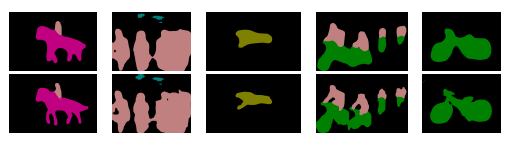


Figure 4: Incorporating multi-scale features improves the boundary segmentation. We show the results obtained by DeepLab and DeepLab-MSc in the first and second row, respectively. Best viewed in color.

그림 4: 다중 스케일 기능을 통합하면 경계 분할이 향상됩니다. 첫 번째 행과 두 번째 행에 각각 DeepLab과 DeepLab-MSc에서 얻은 결과를 보여줍니다. 색상으로 가장 잘 보입니다.

Mean Pixel IOU along Object Boundaries To quantify the accuracy of the proposed model near object boundaries, we evaluate the segmentation accuracy with an experiment similar to Kohli et al. (2009); Krahenb ¨ uhl & Koltun (2011). Specifically, we use the ‘void’ label annotated in val set, ¨ which usually occurs around object boundaries. We compute the mean IOU for those pixels that are located within a narrow band (called trimap) of ‘void’ labels. As shown in Fig. 5, exploiting the multi-scale features from the intermediate layers and refining the segmentation results by a fully connected CRF significantly improve the results around object boundaries.

객체 경계에 따른 평균 픽셀 IOU 객체 경계 근처에서 제안된 모델의 정확도를 정량화하기 위해 Kohli et al.과 유사한 실험으로 분할 정확도를 평가합니다. (2009); Krahenb ¨ uhl & Koltun(2011). 특히, 우리는 일반적으로 객체 경계 주변에서 발생하는 ¨ val set에 주석이 달린 'void' 레이블을 사용합니다. 우리는 'void' 레이블의 좁은 대역(trimap이라고 함) 내에 있는 픽셀에 대한 평균 IOU를 계산합니다. 그림 5에서 볼 수 있듯이 중간 레이어의 다중 스케일 기능을 활용하고 완전히 연결된 CRF로 세분화 결과를 개선하면 객체 경계 주변의 결과가 크게 향상됩니다.

Comparison with State-of-art In Fig. 6, we qualitatively compare our proposed model, DeepLabCRF, with two state-of-art models: FCN-8s (Long et al., 2014) and TTI-Zoomout-16 (Mostajabi et al., 2014) on the ‘val’ set (the results are extracted from their papers). Our model is able to capture the intricate object boundaries.

최신 모델과의 비교 그림 6에서 제안한 모델인 DeepLabCRF를 FCN-8s(Long et al., 2014) 및 TTI-Zoomout-16(Mostajabi et al., 2014)의 두 가지 최신 모델과 정성적으로 비교합니다. al., 2014) 'val' 세트(결과는 해당 논문에서 추출됨). 우리 모델은 복잡한 객체 경계를 포착할 수 있습니다.

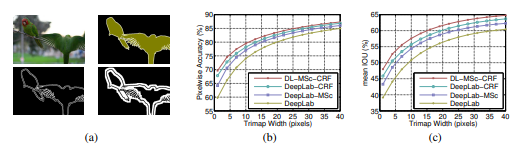


Figure 5: (a) Some trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). Quality of segmentation result within a band around the object boundaries for the proposed methods. (b) Pixelwise accuracy. (c) Pixel mean IOU.

그림 5: (a) 일부 트리맵의 예(왼쪽 위: 이미지. 오른쪽 위: ground-truth. 왼쪽 아래: 2픽셀의 트리맵. 오른쪽 아래: 10픽셀의 트리맵). 제안된 방법에 대한 객체 경계 주변의 밴드 내에서 분할 품질 결과. (b) 픽셀 단위 정확도. (c) 픽셀은 IOU를 의미합니다.

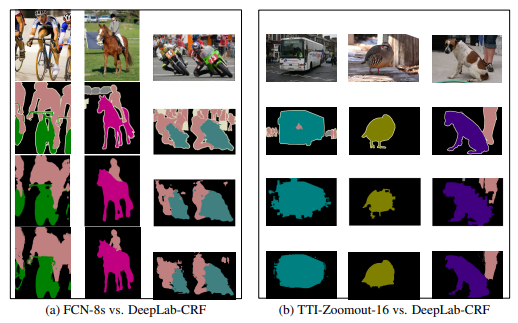


Figure 6: Comparisons with state-of-the-art models on the val set. First row: images. Second row: ground truths. Third row: other recent models (Left: FCN-8s, Right: TTI-Zoomout-16). Fourth row: our DeepLab-CRF. Best viewed in color.

그림 6: val 세트에 대한 최신 모델과의 비교. 첫 번째 행: 이미지. 두 번째 행: 진실. 세 번째 줄: 다른 최신 모델(왼쪽: FCN-8, 오른쪽: TTI-Zoomout-16). 네 번째 행: DeepLab-CRF. 색상으로 가장 잘 보입니다.

Reproducibility We have implemented the proposed methods by extending the excellent Caffe framework (Jia et al., 2014). We share our source code, configuration files, and trained models that allow reproducing the results in this paper at a companion web site https://bitbucket.org/ deeplab/deeplab-public.

재현성 우수한 Caffe 프레임워크를 확장하여 제안한 방법을 구현했습니다(Jia et al., 2014). 이 문서의 결과를 재현할 수 있는 소스 코드, 구성 파일 및 훈련된 모델을 동반 웹 사이트 https://bitbucket.org/deeplab/deeplab-public에서 공유합니다.

Test set results Having set our model choices on the validation set, we evaluate our model variants on the PASCAL VOC 2012 official ‘test’ set. As shown in Tab. 3, our DeepLab-CRF and DeepLabMSc-CRF models achieve performance of 66.4% and 67.1% mean IOU1 , respectively. Our models outperform all the other state-of-the-art models (specifically, TTI-Zoomout-16 (Mostajabi et al., 2014), FCN-8s (Long et al., 2014), and MSRA-CFM (Dai et al., 2014)). When we increase the FOV of the models, DeepLab-CRF-LargeFOV yields performance of 70.3%, the same as DeepLab-CRF7x7, while its training speed is faster. Furthermore, our best model, DeepLab-MSc-CRF-LargeFOV, attains the best performance of 71.6% by employing both multi-scale features and large FOV

테스트 세트 결과 검증 세트에서 모델 선택을 설정한 후 PASCAL VOC 2012 공식 '테스트' 세트에서 모델 변형을 평가합니다. 탭에 표시된 대로. 3, DeepLab-CRF 및 DeepLabMSc-CRF 모델은 각각 66.4% 및 67.1% 평균 IOU1의 성능을 달성합니다. 우리 모델은 다른 모든 최신 모델(특히, TTI-Zoomout-16(Mostajabi et al., 2014), FCN-8s(Long et al., 2014) 및 MSRA-CFM(Dai et al., 2014)을 능가합니다. al., 2014)). 모델의 FOV를 높이면 DeepLab-CRF-LargeFOV는 DeepLab-CRF7x7과 동일한 70.3%의 성능을 제공하지만 훈련 속도는 더 빠릅니다. 또한 최고의 모델인 DeepLab-MSc-CRF-LargeFOV는 다중 스케일 기능과 대형 FOV를 모두 사용하여 71.6%의 최고의 성능을 달성합니다.

선반, 다채로운, 다른이(가) 표시된 사진

자동 생성된 설명

Figure 7: Visualization results on VOC 2012-val. For each row, we show the input image, the segmentation result delivered by the DCNN (DeepLab), and the refined segmentation result of the Fully Connected CRF (DeepLab-CRF). We show our failure modes in the last three rows. Best viewed in color.

그림 7: VOC 2012-val에 대한 시각화 결과. 각 행에 대해 입력 이미지, DCNN(DeepLab)이 전달한 분할 결과, 완전 연결 CRF(DeepLab-CRF)의 세분화된 분할 결과를 보여줍니다. 우리는 마지막 세 행에 실패 모드를 보여줍니다. 색상으로 가장 잘 보입니다.

텍스트, 전자기기, 키보드이(가) 표시된 사진

자동 생성된 설명

Table 3: Labeling IOU (%) on the PASCAL VOC 2012 test set, using the trainval set for training.

표 3: 훈련용 trainval 세트를 사용하여 PASCAL VOC 2012 테스트 세트에서 IOU(%) 레이블 지정.

**6 DISCUSSION**

Our work combines ideas from deep convolutional neural networks and fully-connected conditional random fields, yielding a novel method able to produce semantically accurate predictions and detailed segmentation maps, while being computationally efficient. Our experimental results show that the proposed method significantly advances the state-of-art in the challenging PASCAL VOC 2012 semantic image segmentation task.

우리의 작업은 심층 컨볼루션 신경망과 완전히 연결된 조건부 랜덤 필드의 아이디어를 결합하여 의미론적으로 정확한 예측과 상세한 세분화 맵을 생성할 수 있는 새로운 방법을 산출하면서 계산적으로 효율적입니다. 우리의 실험 결과는 제안된 방법이 도전적인 PASCAL VOC 2012 의미론적 이미지 분할 작업에서 최첨단 기술을 크게 발전시킨다는 것을 보여줍니다.

There are multiple aspects in our model that we intend to refine, such as fully integrating its two main components (CNN and CRF) and train the whole system in an end-to-end fashion, similar to Krahenb ¨ uhl & Koltun (2013); Chen et al. (2014); Zheng et al. (2015). We also plan to experiment ¨ with more datasets and apply our method to other sources of data such as depth maps or videos. Recently, we have pursued model training with weakly supervised annotations, in the form of bounding boxes or image-level labels (Papandreou et al., 2015).

두 가지 주요 구성 요소(CNN 및 CRF)를 완전히 통합하고 Krahenb ¨ uhl & Koltun(2013)과 유사한 종단 간 방식으로 전체 시스템을 훈련하는 등 우리 모델에는 개선하려는 여러 측면이 있습니다. ; Chen et al. (2014); Zheng et al. (2015). 또한 더 많은 데이터 세트로 ¨를 실험하고 깊이 맵이나 비디오와 같은 다른 데이터 소스에 우리의 방법을 적용할 계획입니다. 최근에 우리는 경계 상자 또는 이미지 수준 레이블의 형태로 약한 감독 주석을 사용하여 모델 교육을 추구했습니다(Papandreou et al., 2015).

At a higher level, our work lies in the intersection of convolutional neural networks and probabilistic graphical models. We plan to further investigate the interplay of these two powerful classes of methods and explore their synergistic potential for solving challenging computer vision tasks.

더 높은 수준에서 우리의 작업은 합성곱 신경망과 확률적 그래픽 모델의 교차점에 있습니다. 우리는 이 두 가지 강력한 방법 클래스의 상호 작용을 추가로 조사하고 어려운 컴퓨터 비전 작업을 해결하기 위한 시너지 잠재력을 탐구할 계획입니다.

**ACKNOWLEDGMENTS**

This work was partly supported by ARO 62250-CS, NIH Grant 5R01EY022247-03, EU Project RECONFIG FP7-ICT-600825 and EU Project MOBOT FP7-ICT-2011-600796. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research. We would like to thank the anonymous reviewers for their detailed comments and constructive feedback.

이 작업은 ARO 62250-CS, NIH Grant 5R01EY022247-03, EU 프로젝트 RECONFIG FP7-ICT-600825 및 EU 프로젝트 MOBOT FP7-ICT-2011-600796에 의해 부분적으로 지원되었습니다. 또한 이 연구에 사용된 GPU 기부로 NVIDIA Corporation의 지원에 감사드립니다. 상세한 논평과 건설적인 피드백을 주신 익명의 심사자께 감사드립니다.