**Focal Loss for Dense Object Detection**

Tsung-Yi Lin Priya Goyal Ross Girshick Kaiming He Piotr Dollar

**Abstract**

The highest accuracy object detectors to date are based on a two-stage approach popularized by R-CNN, where a classifier is applied to a sparse set of candidate object locations. In contrast, one-stage detectors that are applied over a regular, dense sampling of possible object locations have the potential to be faster and simpler, but have trailed the accuracy of two-stage detectors thus far. In this paper, we investigate why this is the case. We discover that the extreme foreground-background class imbalance encountered during training of dense detectors is the central cause. We propose to address this class imbalance by reshaping the standard cross entropy loss such that it down-weights the loss assigned to well-classified examples. Our novel Focal Loss focuses training on a sparse set of hard examples and prevents the vast number of easy negatives from overwhelming the detector during training. To evaluate the effectiveness of our loss, we design and train a simple dense detector we call RetinaNet. Our results show that when trained with the focal loss, RetinaNet is able to match the speed of previous one-stage detectors while surpassing the accuracy of all existing state-of-the-art two-stage detectors. Code is at: <https://github.com/facebookresearch/Detectron>.

지금까지의 최고 정확도 객체 검출기는 R-CNN에 의해 대중화된 2단계 접근 방식을 기반으로 하며, 여기서 분류기는 희박한 후보 객체 위치 집합에 적용된다. 반대로 가능한 물체 위치의 정기적이고 밀도 높은 표본 추출에 적용되는 1단계 검출기는 더 빠르고 단순할 가능성이 있지만, 지금까지 2단계 검출기의 정확도를 추적해왔다. 이 논문에서는 왜 이런 일이 일어나는지 조사합니다. 우리는 고밀도 검출기 훈련 중 발생하는 극단적인 전경-배경 등급 불균형이 중심 원인이라는 것을 발견한다. 우리는 표준 교차 엔트로피 손실을 재구성하여 잘 분류된 예제에 할당된 손실을 감소시킴으로써 이러한 클래스 불균형을 해결할 것을 제안합니다. 우리의 새로운 초점 손실은 희박한 일련의 하드 사례에 초점을 맞추고 훈련 중에 수많은 쉬운 음성이 검출기를 압도하는 것을 방지합니다. 손실의 효과를 평가하기 위해, 우리는 RetinaNet이라고 불리는 간단한 밀도 감지기를 설계하고 훈련시킵니다. 우리의 결과에 따르면, 초점 손실을 가지고 훈련했을 때, RetinaNet은 기존의 모든 최첨단 2단계 검출기의 정확도를 능가하면서 이전의 1단계 검출기의 속도를 맞출 수 있다. 코드는 https://github.com/facebookresearch/Detectron에 있습니다.

1. **Introduction**

Current state-of-the-art object detectors are based on a two-stage, proposal-driven mechanism. As popularized in the R-CNN framework [11], the first stage generates a sparse set of candidate object locations and the second stage classifies each candidate location as one of the foreground classes or as background using a convolutional neural network. Through a sequence of advances [10, 28, 20, 14], this two-stage framework consistently achieves top accuracy on the challenging COCO benchmark [21].

현재 최첨단 물체 감지기는 2단계의 제안 기반 메커니즘을 기반으로 한다. R-CNN 프레임워크[11]에서 대중화되었듯이, 첫 번째 단계는 희박한 일련의 후보 객체 위치를 생성하고 두 번째 단계는 각 후보 위치를 전경 클래스 중 하나 또는 컨볼루션 신경 네트워크를 사용하여 배경으로 분류한다. 일련의 발전을 통해 [10, 28, 20, 14], 이 2단계 프레임워크는 도전적인 COCO 벤치마크에서 최고의 정확도를 일관되게 달성합니다 [21].

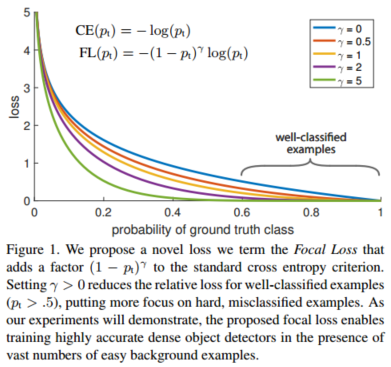


Figure 1. We propose a novel loss we term the Focal Loss that adds a factor (1 − pt) γ to the standard cross entropy criterion. Setting γ > 0 reduces the relative loss for well-classified examples (pt > .5), putting more focus on hard, misclassified examples. As our experiments will demonstrate, the proposed focal loss enables training highly accurate dense object detectors in the presence of vast numbers of easy background examples.

그림 1 표준 교차 엔트로피 기준에 인자(1 - pt) δ를 추가하는 초점 손실이라고 하는 새로운 손실을 제안한다. > 0 을 설정하면 잘 분류된 예(pt > .5)의 상대적인 손실이 줄어들기 때문에 하드하고 잘못 분류된 예에 더욱 초점을 맞출 수 있습니다. 우리의 실험에서 입증될 것처럼, 제안된 초점 손실은 수많은 쉬운 배경 예제가 존재하는 상황에서 고정밀 물체 검출기를 훈련시킬 수 있다.

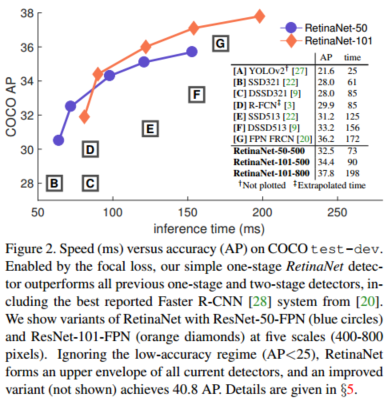


Figure 2. Speed (ms) versus accuracy (AP) on COCO test-dev. Enabled by the focal loss, our simple one-stage RetinaNet detector outperforms all previous one-stage and two-stage detectors, including the best reported Faster R-CNN [28] system from [20]. We show variants of RetinaNet with ResNet-50-FPN (blue circles) and ResNet-101-FPN (orange diamonds) at five scales (400-800 pixels). Ignoring the low-accuracy regime (AP<25), RetinaNet forms an upper envelope of all current detectors, and an improved variant (not shown) achieves 40.8 AP. Details are given in §5.

그림 2 COCO 테스트 개발에서의 속도(ms) 대 정확도(AP) 초점 손실에 의해 활성화된 우리의 단순한 1단계 RetinaNet 검출기는 [20]의 가장 잘 보고된 빠른 R-CNN [28] 시스템을 포함하여 이전의 모든 1단계 및 2단계 검출기를 능가한다. ResNet-50-FPN(블루 서클)과 ResNet-101-FPN(오렌지 다이아몬드)을 사용한 RetinaNet의 변형을 5개의 스케일(400-800픽셀)로 나타냅니다. RetinaNet은 저정확도 시스템(AP < 25)을 무시하고 모든 현재 검출기의 상부 엔벨로프를 형성하며 개량된 변형(표시되지 않음)은 40.8 AP를 달성한다. 자세한 것은 5파운드에 기재되어 있다.

Despite the success of two-stage detectors, a natural question to ask is: could a simple one-stage detector achieve similar accuracy? One stage detectors are applied over a regular, dense sampling of object locations, scales, and aspect ratios. Recent work on one-stage detectors, such as YOLO [26, 27] and SSD [22, 9], demonstrates promising results, yielding faster detectors with accuracy within 10-40% relative to state-of-the-art two-stage methods.

2단계 검출기의 성공에도 불구하고, 자연스러운 질문은 다음과 같습니다. 단순한 1단계 검출기가 유사한 정확도를 달성할 수 있을까요? 객체 위치, 축척 및 가로 세로 비율의 정기적이고 조밀한 샘플링에 한 단계 검출기가 적용됩니다. YOLO[26, 27] 및 SSD[22, 9]와 같은 1단계 검출기에 대한 최근 연구는 유망한 결과를 보여주며, 첨단 2단계 방법에 비해 10-40% 이내의 정확도로 더 빠른 검출기를 산출한다.

This paper pushes the envelop further: we present a onestage object detector that, for the first time, matches the state-of-the-art COCO AP of more complex two-stage detectors, such as the Feature Pyramid Network (FPN) [20] or Mask R-CNN [14] variants of Faster R-CNN [28]. To achieve this result, we identify class imbalance during training as the main obstacle impeding one-stage detector from achieving state-of-the-art accuracy and propose a new loss function that eliminates this barrier.

본 논문은 한계를 더 확장한다. 우리는 기능 피라미드 네트워크(FPN)[20] 또는 마스크 R-CNN[14] 변형 고속 R-CN[28]과 같은 보다 복잡한 2단계 검출기의 최신 COCO AP와 처음으로 일치하는 1단계 객체 검출기를 제시한다. 이러한 결과를 달성하기 위해, 우리는 훈련 중의 계급 불균형을 1단계 검출기가 최첨단 정확성을 달성하는 것을 방해하는 주요 장애물로 식별하고 이 장벽을 제거하는 새로운 손실 함수를 제안한다.

Class imbalance is addressed in R-CNN-like detectors by a two-stage cascade and sampling heuristics. The proposal stage (e.g., Selective Search [35], EdgeBoxes [39], DeepMask [24, 25], RPN [28]) rapidly narrows down the number of candidate object locations to a small number (e.g., 1-2k), filtering out most background samples. In the second classification stage, sampling heuristics, such as a fixed foreground-to-background ratio (1:3), or online hard example mining (OHEM) [31], are performed to maintain a manageable balance between foreground and background.

클래스 불균형은 2단계 계단식 및 샘플링 휴리스틱에 의해 R-CNN 유사 검출기에서 해결된다. 제안 단계(예: 선택 검색 [35], EdgeBoxes [39], DeepMask [24, 25], RPN [28])는 후보 객체 위치의 수를 적은 수(예: 1-2k)로 빠르게 좁혀 대부분의 배경 샘플을 필터링합니다. 두 번째 분류 단계에서는 전경 대 배경 고정 비율(1:3)이나 온라인 하드 예제 마이닝(OHEM) [31)과 같은 표본 발견학을 수행하여 전경과 배경 사이의 관리 가능한 균형을 유지한다.

In contrast, a one-stage detector must process a much larger set of candidate object locations regularly sampled across an image. In practice this often amounts to enumerating ∼100k locations that densely cover spatial positions, scales, and aspect ratios. While similar sampling heuristics may also be applied, they are inefficient as the training procedure is still dominated by easily classified background examples. This inefficiency is a classic problem in object detection that is typically addressed via techniques such as bootstrapping [33, 29] or hard example mining [37, 8, 31].

반대로 1단계 검출기는 이미지 전체에 걸쳐 정기적으로 샘플링되는 훨씬 큰 세트의 후보 객체 위치를 처리해야 합니다. 실제로 이는 공간 위치, 척도 및 가로 세로 비율을 촘촘히 포함하는 ~100k개의 위치를 열거하는 데 해당합니다. 유사한 표본 추출 휴리스틱도 적용될 수 있지만, 교육 절차는 여전히 쉽게 분류되는 배경 사례에 의해 지배되기 때문에 비효율적입니다. 이러한 비효율성은 일반적으로 부트스트래핑[33, 29] 또는 하드 예제 마이닝[37, 8, 31]과 같은 기법을 통해 해결되는 객체 감지에서 전형적인 문제입니다.

In this paper, we propose a new loss function that acts as a more effective alternative to previous approaches for dealing with class imbalance. The loss function is a dynamically scaled cross entropy loss, where the scaling factor decays to zero as confidence in the correct class increases, see Figure 1. Intuitively, this scaling factor can automatically down-weight the contribution of easy examples during training and rapidly focus the model on hard examples. Experiments show that our proposed Focal Loss enables us to train a high-accuracy, one-stage detector that significantly outperforms the alternatives of training with the sampling heuristics or hard example mining, the previous state-ofthe-art techniques for training one-stage detectors. Finally, we note that the exact form of the focal loss is not crucial, and we show other instantiations can achieve similar results.

본 논문에서, 우리는 계급 불균형에 대처하기 위한 이전의 접근법에 대한 보다 효과적인 대안으로 작용하는 새로운 손실 함수를 제안한다. 손실 함수는 동적으로 스케일링된 교차 엔트로피 손실이며, 올바른 클래스의 신뢰가 증가함에 따라 스케일링 계수가 0으로 감소합니다(그림 1 참조). 직관적으로, 이 확장 요소는 교육 중에 쉬운 예제의 기여도를 자동으로 낮추고 모델을 어려운 예제에 신속하게 집중할 수 있습니다. 실험 결과, 제안된 초점 손실은 1단계 검출기를 훈련하기 위한 이전의 첨단 기술인 표본 휴리스틱이나 하드 예제 마이닝으로 훈련의 대안을 크게 능가하는 높은 정확도의 1단계 검출기를 훈련시킬 수 있다는 것을 보여준다. 마지막으로, 우리는 초점 손실에 대한 정확한 형태가 중요하지 않으며 다른 인스턴스화도 유사한 결과를 얻을 수 있음을 보여줍니다.

To demonstrate the effectiveness of the proposed focal loss, we design a simple one-stage object detector called RetinaNet, named for its dense sampling of object locations in an input image. Its design features an efficient in-network feature pyramid and use of anchor boxes. It draws on a variety of recent ideas from [22, 6, 28, 20]. RetinaNet is efficient and accurate; our best model, based on a ResNet-101-FPN backbone, achieves a COCO test-dev AP of 39.1 while running at 5 fps, surpassing the previously best published single-model results from both one and two-stage detectors, see Figure 2.

제안된 초점 손실에 대한 효과를 입증하기 위해 입력 이미지에서 객체 위치의 조밀한 샘플링으로 명명된 RetinaNet이라는 간단한 1단계 객체 검출기를 설계한다. 효율적인 네트워크 내 기능 피라미드와 앵커 박스 사용이 특징입니다. [22, 6, 28, 20]의 다양한 최신 아이디어를 끌어냅니다. RetinaNet은 효율적이고 정확합니다. ResNet-101-FPN 백본을 기반으로 하는 최고의 모델은 5fps로 실행되는 동안 39.1의 COCO 테스트 개발 AP를 달성하여 1단계 및 2단계 검출기 모두에서 이전에 가장 잘 발표된 단일 모델 결과를 능가합니다(그림 2 참조).

1. **Related Work**

**Classic Object Detectors:** The sliding-window paradigm, in which a classifier is applied on a dense image grid, has a long and rich history. One of the earliest successes is the classic work of LeCun et al. who applied convolutional neural networks to handwritten digit recognition [19, 36]. Viola and Jones [37] used boosted object detectors for face detection, leading to widespread adoption of such models. The introduction of HOG [4] and integral channel features [5] gave rise to effective methods for pedestrian detection. DPMs [8] helped extend dense detectors to more general object categories and had top results on PASCAL [7] for many years. While the sliding-window approach was the leading detection paradigm in classic computer vision, with the resurgence of deep learning [18], two-stage detectors, described next, quickly came to dominate object detection.

기본 개체 디텍터: 조밀한 이미지 그리드에 분류기가 적용되는 슬라이딩 윈도우 패러다임은 길고 풍부한 역사를 가지고 있다. 초기 성공 사례 중 하나는 손으로 쓴 숫자 인식에 컨볼루션 신경망을 적용한 LeCun 등의 고전적인 연구입니다 [19, 36]. Viola와 Jones[37]는 얼굴 감지를 위해 부스트 객체 감지기를 사용하였고, 이러한 모델이 널리 채택되었다. HOG [4]와 일체형 채널 기능[5]의 도입으로 보행자 감지를 위한 효과적인 방법이 제시되었습니다. DPM[8]은 고밀도 디텍터를 보다 일반적인 개체 범주로 확장할 수 있도록 지원했으며 수년간 PASCAL[7]에서 상위 결과를 얻었습니다. 슬라이딩 윈도우 접근법이 고전적 컴퓨터 비전에서는 선도적인 검출 패러다임이었지만 딥러닝[18]이 부활하면서 그 다음으로 설명되는 2단계 검출기는 빠르게 객체 감지를 지배하게 되었습니다.

**Two-stage Detectors:** The dominant paradigm in modern object detection is based on a two-stage approach. As pioneered in the Selective Search work [35], the first stage generates a sparse set of candidate proposals that should contain all objects while filtering out the majority of negative locations, and the second stage classifies the proposals into foreground classes / background. R-CNN [11] upgraded the second-stage classifier to a convolutional network yielding large gains in accuracy and ushering in the modern era of object detection. R-CNN was improved over the years, both in terms of speed [15, 10] and by using learned object proposals [6, 24, 28]. Region Proposal Networks (RPN) integrated proposal generation with the second-stage classifier into a single convolution network, forming the Faster RCNN framework [28]. Numerous extensions to this framework have been proposed, e.g. [20, 31, 32, 16, 14].

2단계 디텍터: 현대 물체 감지에서 지배적인 패러다임은 2단계 접근법에 기초한다. 선택적 검색 작업[35]에서 개척된 바와 같이, 첫 번째 단계는 대부분의 부정적인 위치를 필터링하면서 모든 객체를 포함해야 하는 희박한 후보 제안 집합을 생성하며, 두 번째 단계는 제안들을 포그라운드 클래스/배경으로 분류합니다. R-CNN[11]은 2단계 분류기를 컨볼루션 네트워크로 업그레이드하여 정확도가 크게 향상되고 객체 검출의 현대 시대가 도래했습니다. R-CNN은 [15, 10]의 속도와 학습된 객체 제안을 사용하여 수년간 향상되었습니다 [6, 24, 28]. 지역 제안 네트워크 (RPN)는 두 번째 단계 분류자와 제안 생성을 단일 컨볼루션 네트워크에 통합하여 더 빠른 RCN 프레임워크를 형성했습니다 [28]. 이 프레임워크에 대한 수많은 확장이 제안되었습니다(20, 31, 32, 16, 14).

**One-stage Detectors:** OverFeat [30] was one of the first modern one-stage object detector based on deep networks. More recently SSD [22, 9] and YOLO [26, 27] have renewed interest in one-stage methods. These detectors have been tuned for speed but their accuracy trails that of twostage methods. SSD has a 10-20% lower AP, while YOLO focuses on an even more extreme speed/accuracy trade-off. See Figure 2. Recent work showed that two-stage detectors can be made fast simply by reducing input image resolution and the number of proposals, but one-stage methods trailed in accuracy even with a larger compute budget [17]. In contrast, the aim of this work is to understand if one-stage detectors can match or surpass the accuracy of two-stage detectors while running at similar or faster speeds.

1단계 디텍터: OverFeat [30]은 딥 네트워크를 기반으로 한 최초의 현대적인 1단계 물체 감지기 중 하나였다. 보다 최근에 SSD[22, 9]와 YOLO[26, 27]가 1단계 방법에 대한 관심을 새롭게 일으켰습니다. 이 검출기는 속도에 맞게 튜닝되었지만 정확도는 두 단계 방법에 따라 추적됩니다. SSD의 AP는 10-20% 낮은 반면, YOLO는 훨씬 더 극단적인 속도/정확도 균형에 초점을 맞춥니다. 그림 2를 참조하십시오. 최근 연구에서는 입력 이미지 해상도와 제안 횟수를 줄이는 것만으로 2단계 검출기를 빠르게 만들 수 있지만, 1단계 방법은 더 큰 계산 예산에도 불구하고 정확도에서 뒤처졌다[17]. 이와 대조적으로 본 연구의 목적은 1단계 검출기가 유사하거나 더 빠른 속도로 작동하면서 2단계 검출기의 정확도를 일치시키거나 능가할 수 있는지를 이해하는 것이다.

The design of our RetinaNet detector shares many similarities with previous dense detectors, in particular the concept of ‘anchors’ introduced by RPN [28] and use of features pyramids as in SSD [22] and FPN [20]. We emphasize that our simple detector achieves top results not based on innovations in network design but due to our novel loss.

당사의 RetinaNet 검출기 설계는 이전의 고밀도 검출기와 많은 유사점을 공유하며, 특히 RPN[28]에 의해 도입된 '앵커'의 개념과 SSD[22] 및 FPN[20]에서와 같은 기능 피라미드의 사용을 공유한다. 우리는 네트워크 설계의 혁신이 아니라 새로운 손실로 인해 당사의 단순한 검출기가 최고의 결과를 달성한다는 점을 강조합니다.

**Class Imbalance:** Both classic one-stage object detection methods, like boosted detectors [37, 5] and DPMs [8], and more recent methods, like SSD [22], face a large class imbalance during training. These detectors evaluate 104-105 candidate locations per image but only a few locations contain objects. This imbalance causes two problems: (1) training is inefficient as most locations are easy negatives that contribute no useful learning signal; (2) en masse, the easy negatives can overwhelm training and lead to degenerate models. A common solution is to perform some form of hard negative mining [33, 37, 8, 31, 22] that samples hard examples during training or more complex sampling/reweighing schemes [2]. In contrast, we show that our proposed focal loss naturally handles the class imbalance faced by a one-stage detector and allows us to efficiently train on all examples without sampling and without easy negatives overwhelming the loss and computed gradients.

클래스 불균형: 부스트 디텍터[37, 5] 및 DPM[8]과 같은 기존의 1단계 개체 감지 방법과 SSD[22]와 같은 최신 방법 모두 교육 중에 큰 클래스 불균형에 직면합니다. 이러한 디텍터는 이미지당 104-105개의 후보 위치를 평가하지만 일부 위치에만 개체가 포함되어 있습니다. 이러한 불균형은 두 가지 문제를 일으킨다. (1) 대부분의 위치가 유용한 학습 신호도 기여하지 않는 쉬운 부정적인 위치이기 때문에 훈련이 비효율적이며, (2) 쉬운 부정적인 부분이 훈련을 압도하여 퇴보 모델로 이어질 수 있다. 일반적인 해결책은 훈련 중 또는 더 복잡한 표본 추출/재가중 계획[2] 중 딱딱한 예를 표본 추출하는 일종의 하드 네거티브 마이닝[33, 37, 8, 31, 22]을 수행하는 것입니다. 반대로, 우리는 제안된 초점 손실이 1단계 검출기가 직면한 클래스 불균형을 자연스럽게 처리하며, 샘플링을 하지 않고 손실과 계산 구배를 압도하는 쉬운 음수 없이 모든 예제를 효율적으로 훈련할 수 있음을 보여준다.

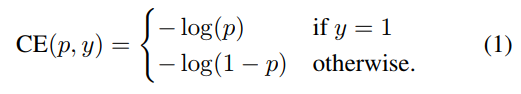
**Robust Estimation:** There has been much interest in designing robust loss functions (e.g., Huber loss [13]) that reduce the contribution of outliers by down-weighting the loss of examples with large errors (hard examples). In contrast, rather than addressing outliers, our focal loss is designed to address class imbalance by down-weighting inliers (easy examples) such that their contribution to the total loss is small even if their number is large. In other words, the focal loss performs the opposite role of a robust loss: it focuses training on a sparse set of hard examples.

견실한 추정: 오차가 큰 예(하드 예제)의 손실을 다운가중화하여 특이치의 기여도를 줄이는 강력한 손실 함수(예: Huber 손실 [13])를 설계하는 데 많은 관심이 있어 왔다. 이와 대조적으로, 우리의 초점손실은 특이치를 다루기 보다는 전체손실에 대한 기여도가 크더라도 작도록(쉬운 사례) 인라이어를 다운가중화함으로써 계급 불균형을 해결하도록 설계되었다. 다시 말해, 초점 손실은 강력한 손실의 반대 역할을 수행하며, 훈련은 희박한 일련의 어려운 사례에 초점을 맞춘다.

1. **Focal Loss**

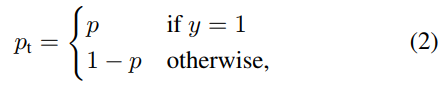
The Focal Loss is designed to address the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes during training (e.g., 1:1000). We introduce the focal loss starting from the cross entropy (CE) loss for binary classification1:

초점 손실은 훈련 중 포그라운드 클래스와 백그라운드 클래스 간에 극심한 불균형이 발생하는 1단계 객체 감지 시나리오(예: 1:1000)를 다루기 위해 설계되었습니다. 바이너리 분류1을 위해 교차 엔트로피(CE) 손실부터 시작하는 초점 손실을 소개합니다.



In the above y ∈ {±1} specifies the ground-truth class and p ∈ [0, 1] is the model’s estimated probability for the class with label y = 1. For notational convenience, we define pt: and rewrite CE(p, y) = CE(pt) = − log(pt).

위의 y { {±1}에서 지면-진실 클래스를 지정하며 p [ [0,1]은 라벨 y = 1인 클래스에 대한 모델의 추정 확률입니다. 표기 편의를 위해 pt:를 정의하고 CE(p, y) = CE(pt) = - log(pt)를 다시 씁니다.



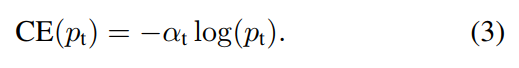
The CE loss can be seen as the blue (top) curve in Figure 1. One notable property of this loss, which can be easily seen in its plot, is that even examples that are easily classified (pt .5) incur a loss with non-trivial magnitude. When summed over a large number of easy examples, these small loss values can overwhelm the rare class.

CE 손실은 그림 1의 파란색(상단) 곡선으로 볼 수 있습니다. 그림에서 쉽게 볼 수 있는 이 손실에 대한 주목할 만한 특성 중 하나는 쉽게 분류되는 예(pt.5)에서도 사소한 크기가 아닌 손실이 발생한다는 것이다. 많은 수의 쉬운 예제에 대해 합산하면 이러한 작은 손실 값이 희귀 클래스를 압도할 수 있습니다.

* 1. **. Balanced Cross Entropy**

A common method for addressing class imbalance is to introduce a weighting factor α ∈ [0, 1] for class 1 and 1−α for class −1. In practice α may be set by inverse class frequency or treated as a hyperparameter to set by cross validation. For notational convenience, we define αt analogously to how we defined pt. We write the α-balanced CE loss as:

클래스 불균형을 해결하는 일반적인 방법은 클래스 1에 가중 인자 α α[0,1]를 도입하고 클래스 -1에 1-α를 도입하는 것이다. 실제로 α는 역 등급 주파수에 의해 설정되거나 교차 검증에 의해 설정되는 초매개변수로 처리될 수 있다. 표기상의 편의를 위해 pt 정의 방법과 유사하게 αt를 정의합니다. α-균형 CE 손실을 다음과 같이 쓴다.



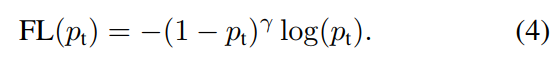
This loss is a simple extension to CE that we consider as an experimental baseline for our proposed focal loss.

이러한 손실은 제안된 초점 손실에 대한 실험 기준으로서 우리가 고려하는 CE에 대한 단순한 확장입니다.

* 1. **. Focal Loss Definition**

As our experiments will show, the large class imbalance encountered during training of dense detectors overwhelms the cross entropy loss. Easily classified negatives comprise the majority of the loss and dominate the gradient. While α balances the importance of positive/negative examples, it does not differentiate between easy/hard examples. Instead, we propose to reshape the loss function to down-weight easy examples and thus focus training on hard negatives. More formally, we propose to add a modulating factor (1 − pt)γ to the cross entropy loss, with tunable focusing parameter γ ≥ 0. We define the focal loss as:

우리의 실험에서 알 수 있듯이, 고밀도 검출기를 훈련하는 동안 발생하는 큰 클래스 불균형은 교차 엔트로피 손실을 압도합니다. 쉽게 분류되는 부정이 손실의 대부분을 차지하며, 그 기울기를 지배한다. α는 긍정적/부정적 사례의 중요성의 균형을 유지하지만, 쉬운/어려운 사례를 구별하지 않는다. 대신, 우리는 쉬운 사례들을 다운 웨이트하여 단단한 네거티브에 훈련을 집중하도록 손실 함수를 재구성할 것을 제안합니다. 좀 더 공식적으로, 우리는 조정 가능한 초점 매개변수 ≥ ≥ 0으로 교차 엔트로피 손실에 변조 인자(1 - pt)를 추가할 것을 제안한다. 우리는 초점 손실을 다음과 같이 정의한다.

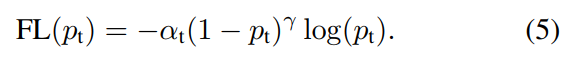


The focal loss is visualized for several values of γ ∈ [0, 5] in Figure 1. We note two properties of the focal loss. (1) When an example is misclassified and pt is small, the modulating factor is near 1 and the loss is unaffected. As pt → 1, the factor goes to 0 and the loss for well-classified examples is down-weighted. (2) The focusing parameter γ smoothly adjusts the rate at which easy examples are downweighted. When γ = 0, FL is equivalent to CE, and as γ is increased the effect of the modulating factor is likewise increased (we found γ = 2 to work best in our experiments).

초점 손실은 그림 1의 figure [ [0, 5]의 여러 값에 대해 시각화됩니다. 우리는 초점손실의 두 가지 특성을 주목한다. (1) 예를 잘못 분류하고 pt가 작을 때 변조계수는 1에 가깝고 손실은 영향을 받지 않는다. 부분 → 1인 경우 인자는 0이 되고 잘 분류된 예제의 손실은 하향 가중됩니다. (2) 집중 모수 θ는 쉬운 예제의 하향 가중치 비율을 부드럽게 조정합니다. α = 0일 때 FL은 CE와 동일하며, α가 증가하면 변조 인자의 효과도 증가합니다(실험에서 가장 잘 작동하는 것은 α = 2임을 확인함).

Intuitively, the modulating factor reduces the loss contribution from easy examples and extends the range in which an example receives low loss. For instance, with γ = 2, an example classified with pt = 0.9 would have 100× lower loss compared with CE and with pt ≈ 0.968 it would have 1000× lower loss. This in turn increases the importance of correcting misclassified examples (whose loss is scaled down by at most 4× for pt ≤ .5 and γ = 2).

직관적으로 변조 계수는 손쉬운 사례에서 손실 기여도를 줄이고 사례가 낮은 손실을 받는 범위를 확장합니다. 예를 들어, θ = 2인 경우, pt = 0.9로 분류된 예는 CE에 비해 100배, pt ≥ 0.968인 경우에는 1000배 더 낮은 손실을 가집니다. 이는 다시 잘못 분류된 예들을 교정하는 것의 중요성을 증가시킵니다(손실은 pt ≤.5 및 θ = 2에 대해 최대 4배까지 축소됨).



In practice we use an α-balanced variant of the focal loss: We adopt this form in our experiments as it yields slightly improved accuracy over the non-α-balanced form. Finally, we note that the implementation of the loss layer combines the sigmoid operation for computing p with the loss computation, resulting in greater numerical stability.

실제로 우리는 초점손실의 α-균형 변형을 사용한다. 우리는 이 형태를 실험에서 채택했습니다. 이것이 비α-균형 형태에 비해 정확도가 약간 향상되었기 때문입니다. 마지막으로, 손실 계층의 구현은 p를 계산하기 위한 시그모이드 연산을 손실 계산과 결합하여 더 큰 수치 안정성을 가져온다는 점에 주목한다.

While in our main experimental results we use the focal loss definition above, its precise form is not crucial. In the appendix we consider other instantiations of the focal loss and demonstrate that these can be equally effective.

주요 실험 결과에서는 위의 초점 손실 정의를 사용하지만 정확한 형태는 중요하지 않습니다. 부록에서 우리는 초점손실에 대한 다른 인스턴스화를 고려하고 이러한 인스턴스화가 똑같이 효과적일 수 있음을 입증한다.

* 1. **Class Imbalance and Model Initialization**

Binary classification models are by default initialized to have equal probability of outputting either y = −1 or 1. Under such an initialization, in the presence of class imbalance, the loss due to the frequent class can dominate total loss and cause instability in early training. To counter this, we introduce the concept of a ‘prior’ for the value of p estimated by the model for the rare class (foreground) at the start of training. We denote the prior by π and set it so that the model’s estimated p for examples of the rare class is low, e.g. 0.01. We note that this is a change in model initialization (see §4.1) and not of the loss function. We found this to improve training stability for both the cross entropy and focal loss in the case of heavy class imbalance.

이항 분류 모형은 기본적으로 y = -1 또는 1의 출력 확률이 동일하도록 초기화됩니다. 이러한 초기화 하에서, 계층 불균형이 존재하는 상황에서, 잦은 계층으로 인한 손실이 전체 손실을 지배하고 조기 훈련에서 불안정성을 야기할 수 있다. 이에 대응하기 위해, 우리는 교육을 시작할 때 희귀 클래스(전경)에 대해 모형에 의해 추정된 p 값에 대한 '선순위' 개념을 도입한다. 우리는 α로 이전을 나타내며, 희귀 클래스의 예에 대한 모형의 추정 p가 낮도록 설정합니다(예: 0.01). 우리는 이것이 모델 초기화(§ 4.1 참조)의 변경이며 손실 함수의 변경은 아니라는 점에 주목한다. 우리는 이를 통해 계층 불균형이 심한 경우 교차 엔트로피 및 초점 손실 모두에 대한 훈련 안정성을 향상시켰습니다.

* 1. **. Class Imbalance and Two-stage Detectors**

Two-stage detectors are often trained with the cross entropy loss without use of α-balancing or our proposed loss. Instead, they address class imbalance through two mechanisms: (1) a two-stage cascade and (2) biased minibatch sampling. The first cascade stage is an object proposal mechanism [35, 24, 28] that reduces the nearly infinite set of possible object locations down to one or two thousand. Importantly, the selected proposals are not random, but are likely to correspond to true object locations, which removes the vast majority of easy negatives. When training the second stage, biased sampling is typically used to construct minibatches that contain, for instance, a 1:3 ratio of positive to negative examples. This ratio is like an implicit α-balancing factor that is implemented via sampling. Our proposed focal loss is designed to address these mechanisms in a one-stage detection system directly via the loss function.

2단계 검출기는 종종 α-밸런싱이나 우리가 제안한 손실을 사용하지 않고 교차 엔트로피 손실에 대해 훈련된다. 대신, (1) 2단계 계단식 및 (2) 편향된 미니배치 표본 추출의 두 가지 메커니즘을 통해 클래스 불균형을 해결한다. 첫 번째 계단식 단계는 객체 제안 메커니즘[35, 24, 28]으로, 가능한 객체 위치의 거의 무한한 집합을 1,000개 또는 2,000개로 줄입니다. 중요한 것은 선택된 제안들이 무작위로 이루어지는 것이 아니라, 실제 대상 위치에 해당할 가능성이 높기 때문에 대부분의 쉬운 부정적인 요소들이 제거된다는 것이다. 두 번째 단계를 교육할 때 편향된 표본 추출은 일반적으로 양의 예와 음의 예제의 1:3 비율을 포함하는 미니 배치를 구성하는 데 사용됩니다. 이 비율은 샘플링을 통해 구현되는 암묵적 α-밸런싱 인자와 같다. 우리가 제안한 초점 손실은 손실 기능을 통해 1단계 감지 시스템에서 이러한 메커니즘을 직접 다루기 위해 고안되었습니다.

1. **RetinaNet Detector**

RetinaNet is a single, unified network composed of a backbone network and two task-specific subnetworks. The backbone is responsible for computing a convolutional feature map over an entire input image and is an off-the-self convolutional network. The first subnet performs convolutional object classification on the backbone’s output; the second subnet performs convolutional bounding box regression. The two subnetworks feature a simple design that we propose specifically for one-stage, dense detection, see Figure 3. While there are many possible choices for the details of these components, most design parameters are not particularly sensitive to exact values as shown in the experiments. We describe each component of RetinaNet next.

RetinaNet은 백본 네트워크와 두 개의 작업별 하위 네트워크로 구성된 단일 통합 네트워크입니다. 백본은 전체 입력 이미지에 걸쳐 컨볼루션 피쳐 맵을 계산하는 역할을 하며, 자체에서 벗어난 컨볼루션 네트워크입니다. 첫 번째 서브넷은 백본의 출력에 대해 컨볼루션 객체 분류를 수행하고, 두 번째 서브넷은 컨볼루션 경계 상자 회귀 분석을 수행합니다. 두 개의 하위 네트워크는 특별히 1단계 밀도 탐지를 위해 제안하는 단순한 설계를 특징으로 합니다(그림 3 참조). 이러한 성분의 세부 정보에 대해 여러 가지 옵션을 사용할 수 있지만 대부분의 설계 모수는 실험에서와 같이 정확한 값에 특별히 민감하지 않습니다. 다음은 RetinaNet의 각 구성 요소를 설명합니다.

**Feature Pyramid Network Backbone:** We adopt the Feature Pyramid Network (FPN) from [20] as the backbone network for RetinaNet. In brief, FPN augments a standard convolutional network with a top-down pathway and lateral connections so the network efficiently constructs a rich, multi-scale feature pyramid from a single resolution input image, see Figure 3(a)-(b). Each level of the pyramid can be used for detecting objects at a different scale. FPN improves multi-scale predictions from fully convolutional networks (FCN) [23], as shown by its gains for RPN [28] and DeepMask-style proposals [24], as well at two-stage detectors such as Fast R-CNN [10] or Mask R-CNN [14].

피쳐 피라미드 네트워크 백본: 우리는 [20]의 Feature Pyramid Network(FPN)를 RetinaNet의 백본 네트워크로 채택합니다. 간단히 말해, FPN은 하향식 경로와 횡방향 연결로 표준 콘볼루션 네트워크를 강화하여 네트워크가 단일 해상도 입력 이미지에서 풍부한 멀티스케일 기능 피라미드를 효율적으로 구성합니다(그림 3(a)-(b). 피라미드의 각 레벨은 서로 다른 척도의 물체를 감지하는 데 사용할 수 있습니다. FPN은 Fast R-CNN[10] 또는 Mask R-CNN[14]과 같은 2단계 검출기에서 RPN[28]과 DeepMask-style 제안[24]에 대한 이득에서 알 수 있듯이 완전 콘볼루션 네트워크(FCN)[23]의 멀티스케일 예측을 개선한다.

Following [20], we build FPN on top of the ResNet architecture [16]. We construct a pyramid with levels P3 through P7, where l indicates pyramid level (Pl has resolution2l lower than the input). As in [20] all pyramid levels have C = 256 channels. Details of the pyramid generally follow [20] with a few modest differences.2 While many design choices are not crucial, we emphasize the use of the FPN backbone is; preliminary experiments using features from only the final ResNet layer yielded low AP.

[20]에 이어 ResNet 아키텍처[16]를 기반으로 FPN을 구축합니다. P3에서 P7까지의 레벨로 피라미드를 구성하며, 여기서 l은 피라미드 레벨을 나타냅니다(Pl은 입력보다 분해능 2l 낮음). [20]에서와 같이 모든 피라미드 레벨은 C = 256 채널을 가집니다. 피라미드의 세부사항은 일반적으로 약간의 차이와 함께 [20]를 따른다.2 많은 설계 선택이 중요한 것은 아니지만 FPN 백본의 사용은 다음과 같다고 강조합니다. 최종 ResNet 계층의 기능만 사용한 예비 실험에서 낮은 AP를 보였습니다.

**Anchors:** We use translation-invariant anchor boxes similar to those in the RPN variant in [20]. The anchors have areas of 322 to 5122 on pyramid levels P3 to P7, respectively. As in [20], at each pyramid level we use anchors at three aspect ratios {1:2, 1:1, 2:1}. For denser scale coverage than in [20], at each level we add anchors of sizes {20, 21/3, 22/3} of the original set of 3 aspect ratio anchors. This improve AP in our setting. In total there are A = 9 anchors per level and across levels they cover the scale range 32 - 813 pixels with respect to the network’s input image.

앵커: [20]의 RPN 변형과 유사한 변환 불변 앵커 박스를 사용한다. 앵커는 피라미드 레벨 P3 ~ P7에서 각각 322 ~ 5122의 면적을 가집니다. [20]와 같이, 각 피라미드 레벨에서는 세 가지 가로 세로 비율 {1:2, 1:1, 2:1}의 앵커를 사용합니다. [3]보다 더 밀도 있는 축척 적용범위를 위해, 각 레벨에서 3개의 가로 세로 비율 앵커의 원래 세트의 크기 {20, 21/3, 22/3}의 앵커를 추가합니다. 이렇게 하면 AP 설정이 향상됩니다. 모두 A = 레벨당 9개의 앵커가 있으며, 레벨에 걸쳐 네트워크 입력 이미지와 관련하여 32 - 813픽셀의 스케일 범위를 포함합니다.

Each anchor is assigned a length K one-hot vector of classification targets, where K is the number of object classes, and a 4-vector of box regression targets. We use the assignment rule from RPN [28] but modified for multiclass detection and with adjusted thresholds. Specifically, anchors are assigned to ground-truth object boxes using an intersection-over-union (IoU) threshold of 0.5; and to background if their IoU is in [0, 0.4). As each anchor is assigned to at most one object box, we set the corresponding entry in its length K label vector to 1 and all other entries to 0. If an anchor is unassigned, which may happen with overlap in [0.4, 0.5), it is ignored during training. Box regression targets are computed as the offset between each anchor and its assigned object box, or omitted if there is no assignment.

각 앵커에는 분류 대상의 길이 K 1-hot 벡터가 할당되며, 여기서 K는 객체 클래스의 수이고 상자 회귀 대상의 4-벡터가 할당됩니다. RPN [28]의 할당 규칙을 사용하지만 다중 클래스 탐지 및 조정된 임계값으로 수정되었습니다. 특히, 앵커는 0.5의 교차로-오버-유니온(IoU) 임계값을 사용하여 지면-진실 객체 박스에 할당되며, 앵커의 IoU가 [0, 0.4]에 있는 경우 배경에 할당됩니다. 각 앵커는 최대 하나의 객체 상자에 할당되므로 길이 K 레이블 벡터의 해당 항목을 1로 설정하고 다른 모든 항목을 0으로 설정합니다. [0.4, 0.5]에서 겹칠 수 있는 앵커가 할당되지 않은 경우, 훈련 중에 무시됩니다. 상자 회귀 분석 대상은 각 고정 장치와 할당된 개체 상자 사이의 간격띄우기로 계산되거나 할당이 없는 경우 생략됩니다.

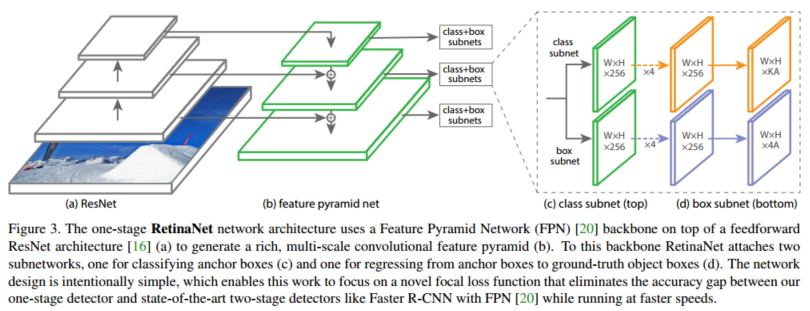


Figure 3. The one-stage RetinaNet network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying anchor boxes (c) and one for regressing from anchor boxes to ground-truth object boxes (d). The network design is intentionally simple, which enables this work to focus on a novel focal loss function that eliminates the accuracy gap between our one-stage detector and state-of-the-art two-stage detectors like Faster R-CNN with FPN [20] while running at faster speeds.

그림 3 1 스테이지의 RetinaNet 네트워크 아키텍처에서는 피드포워드 ResNet 아키텍처[16](a) 위에 Feature Pyramid Network (FPN; 기능 피라미드 네트워크)[20]백본을 사용하여 풍부한 멀티스케일 컨볼루션피처 피라미드를 생성합니다(b). 이 백본에 RetinaNet은 2개의 서브네트워크를 접속합니다.하나는 앵커박스(c)를 분류하기 위한 것이고, 다른 하나는 앵커박스에서 지상 진실 객체 박스(d)로 회귀하기 위한 것입니다. 네트워크 설계는 의도적으로 단순하며, 이를 통해 1단계 검출기와 FPN [20]을 사용하는 Faster R-CNN과 같은 최첨단 2단계 검출기 사이의 정확도 차이를 제거하는 새로운 초점 손실 기능에 초점을 맞출 수 있다.

Classification Subnet: The classification subnet predicts the probability of object presence at each spatial position for each of the A anchors and K object classes. This subnet is a small FCN attached to each FPN level; parameters of this subnet are shared across all pyramid levels. Its design is simple. Taking an input feature map with C channels from a given pyramid level, the subnet applies four 3×3 conv layers, each with C filters and each followed by ReLU activations, followed by a 3×3 conv layer with KA filters. Finally sigmoid activations are attached to output the KA binary predictions per spatial location, see Figure 3 (c). We use C = 256 and A = 9 in most experiments.

분류 서브넷: 분류 서브넷은 각 A 앵커 및 K 객체 클래스에 대한 각 공간 위치에서 객체가 존재할 확률을 예측합니다. 이 서브넷은 각 FPN 레벨에 연결된 작은 FCN입니다. 이 서브넷의 매개 변수는 모든 피라미드 레벨에서 공유됩니다. 디자인은 단순하다. 주어진 피라미드 레벨에서 C 채널이 있는 입력 피쳐 맵을 취하면, 서브넷은 각각 C 필터를 가진 4개의 3×3 콘볼 레이어를 적용하며 각각 ReLU 활성화에 이어 KA 필터를 가진 3×3 콘볼 레이어를 적용합니다. 마지막으로 시그모이드 활성화는 공간 위치별 KA 이진 예측을 출력하기 위해 부착됩니다(그림 3(c) 참조). 대부분의 실험에서 C = 256 및 A = 9를 사용합니다.

In contrast to RPN [28], our object classification subnet is deeper, uses only 3×3 convs, and does not share parameters with the box regression subnet (described next). We found these higher-level design decisions to be more important than specific values of hyperparameters.

RPN[28]과 달리, 객체 분류 서브넷은 더 깊고 3×3 초만 사용하며 상자 회귀 서브넷과 매개변수를 공유하지 않습니다(다음 설명). 우리는 이러한 상위 수준의 설계 결정이 하이퍼 파라미터의 특정 값보다 더 중요하다는 것을 발견했습니다.

**Box Regression Subnet:** In parallel with the object classification subnet, we attach another small FCN to each pyramid level for the purpose of regressing the offset from each anchor box to a nearby ground-truth object, if one exists. The design of the box regression subnet is identical to the classification subnet except that it terminates in 4A linear outputs per spatial location, see Figure 3 (d). For each of the A anchors per spatial location, these 4 outputs predict the relative offset between the anchor and the groundtruth box (we use the standard box parameterization from RCNN [11]). We note that unlike most recent work, we use a class-agnostic bounding box regressor which uses fewer parameters and we found to be equally effective. The object classification subnet and the box regression subnet, though sharing a common structure, use separate parameters

상자 회귀 서브넷: 객체 분류 서브넷과 병행하여, 각 앵커 박스에서 가까운 지면-진실 객체(있는 경우)로 오프셋을 회귀시킬 목적으로 각 피라미드 레벨에 또 다른 작은 FCN을 부착합니다. 상자 회귀 서브넷의 설계는 공간 위치당 4A 선형 출력으로 종료된다는 점을 제외하고 분류 서브넷과 동일합니다(그림 3(d) 참조). 공간 위치당 각 A 앵커에 대해 이 4개의 출력은 앵커와 접지 진실 상자 사이의 상대적 오프셋을 예측합니다(RCNN[11]의 표준 상자 매개변수화를 사용합니다). 우리는 가장 최근의 작업과 달리 적은 수의 매개변수를 사용하는 클래스 별 경계 상자 회귀 분석기를 사용하므로 동일한 효과를 얻을 수 있습니다. 개체 분류 서브넷과 상자 회귀 서브넷은 공통 구조를 공유하지만 별도의 매개 변수를 사용합니다.

* 1. **. Inference and Training**

**Inference:** RetinaNet forms a single FCN comprised of a ResNet-FPN backbone, a classification subnet, and a box regression subnet, see Figure 3. As such, inference involves simply forwarding an image through the network. To improve speed, we only decode box predictions from at most 1k top-scoring predictions per FPN level, after thresholding detector confidence at 0.05. The top predictions from all levels are merged and non-maximum suppression with a threshold of 0.5 is applied to yield the final detections.

추론: RetinaNet은 ResNet-FPN 백본, 분류 서브넷 및 상자 회귀 서브넷으로 구성된 단일 FCN을 형성합니다(그림 3 참조). 이와 같이, 추론은 네트워크를 통해 단순히 이미지를 전달하는 것을 포함합니다. 속도를 개선하기 위해 0.05에서 검출기 신뢰도를 임계값화한 후 FPN 수준당 최대 1k의 최고 점수 예측에서만 상자 예측을 디코딩합니다. 모든 수준의 상위 예측이 병합되고 임계값이 0.5인 비최대 억제가 적용되어 최종 탐지 결과를 산출합니다.

**Focal Loss:** We use the focal loss introduced in this work as the loss on the output of the classification subnet. As we will show in §5, we find that γ = 2 works well in practice and the RetinaNet is relatively robust to γ ∈ [0.5, 5]. We emphasize that when training RetinaNet, the focal loss is applied to all ∼100k anchors in each sampled image. This stands in contrast to common practice of using heuristic sampling (RPN) or hard example mining (OHEM, SSD) to select a small set of anchors (e.g., 256) for each minibatch. The total focal loss of an image is computed as the sum of the focal loss over all ∼100k anchors, normalized by the number of anchors assigned to a ground-truth box. We perform the normalization by the number of assigned anchors, not total anchors, since the vast majority of anchors are easy negatives and receive negligible loss values under the focal loss. Finally we note that α, the weight assigned to the rare class, also has a stable range, but it interacts with γ making it necessary to select the two together (see Tables 1a and 1b). In general α should be decreased slightly as γ is increased (for γ = 2, α = 0.25 works best).

초점 손실: 우리는 본 연구에서 소개된 초점 손실을 분류 서브넷의 출력 손실로 사용합니다. § 5에서 볼 수 있듯이, 우리는 § = 2가 실제로 잘 작동하고 RetinaNet이 § § [0.5, 5.]에 비해 상대적으로 강하다는 것을 발견한다. 우리는 RetinaNet을 교육할 때 각각의 샘플 이미지에서 최대 100k의 앵커에 초점 손실이 적용됨을 강조합니다. 이는 경험적 표본 추출(RPN) 또는 하드 예제 마이닝(OHEM, SSD)을 사용하여 각 미니배치에 대해 작은 앵커 세트(예: 256)를 선택하는 일반적인 관행과는 대조적입니다. 이미지의 총 초점 손실은 지면 진실 상자에 할당된 앵커 수에 의해 정규화된 전체 ~100k 앵커에 대한 초점 손실의 합으로 계산됩니다. 우리는 앵커의 대부분이 손쉬운 네거티브이고 초점손실 시 무시할 수 있는 손실값을 받기 때문에 총 앵커가 아닌 할당된 앵커 수에 따라 정규화를 수행합니다. 마지막으로 희귀 등급에 할당된 체중인 α도 안정적인 범위를 가지지만, α와 상호작용하기 때문에 둘을 함께 선택해야 한다는 점에 주목한다(표 1a와 1b 참조). 일반적으로 α는 α가 증가할수록 약간 감소해야 한다(α = 2, α = 0.25가 가장 좋다).

**Initialization:** We experiment with ResNet-50-FPN and ResNet-101-FPN backbones [20]. The base ResNet-50 and ResNet-101 models are pre-trained on ImageNet1k; we use the models released by [16]. New layers added for FPN are initialized as in [20]. All new conv layers except the final one in the RetinaNet subnets are initialized with bias b = 0 and a Gaussian weight fill with σ = 0.01. For the final conv layer of the classification subnet, we set the bias initialization to b = − log((1 − π)/π), where π specifies that at the start of training every anchor should be labeled as foreground with confidence of ∼π. We use π = .01 in all experiments, although results are robust to the exact value. As explained in §3.3, this initialization prevents the large number of background anchors from generating a large, destabilizing loss value in the first iteration of training.

초기화: 우리는 ResNet-50-FPN 및 ResNet-101-FPN 백본[20]으로 실험한다. 기본 ResNet-50 및 ResNet-101 모델은 ImageNet1k에서 사전 교육됩니다. [16]에서 출시한 모델을 사용합니다. FPN을 위해 추가된 새 레이어는 [20]와 같이 초기화됩니다. RetinaNet 서브넷의 최종 컨베이어 계층을 제외한 모든 새로운 컨베이어 계층은 치우침 b = 0으로 초기화되고 α = 0.01로 가우스 가중치 채우기가 초기화됩니다. 분류 서브넷의 최종 Conv 레이어의 경우, 우리는 치우침 초기화를 b = - log((1 - π)/files)로 설정합니다. 여기서 specifies는 훈련 시작 시 모든 앵커를 ~firground로 라벨링해야 한다고 명시합니다. 결과가 정확한 값으로 강건하지만 모든 실험에서 π =.01을 사용합니다. § 3.3에서 설명한 바와 같이, 이러한 초기화는 많은 수의 배경 앵커가 첫 번째 훈련 반복에서 큰 불안정 손실 값을 발생시키는 것을 방지한다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 1. **Ablation experiments for RetinaNet and Focal Loss (FL).** All models are trained on trainval35k and tested on minival unless noted. If not specified, default values are: γ = 2; anchors for 3 scales and 3 aspect ratios; ResNet-50-FPN backbone; and a 600 pixel train and test image scale. (a) RetinaNet with α-balanced CE achieves at most 31.1 AP. (b) In contrast, using FL with the same exact network gives a 2.9 AP gain and is fairly robust to exact γ/α settings. (c) Using 2-3 scale and 3 aspect ratio anchors yields good results after which point performance saturates. (d) FL outperforms the best variants of online hard example mining (OHEM) [31, 22] by over 3 points AP. (e) Accuracy/Speed trade-off of RetinaNet on test-dev for various network depths and image scales (see also Figure 2).

표 1. **RetinaNet 및 초점 손실(FL)에 대한 절제 실험.** 모든 모델은 trainval35k로 트레이닝되며 특별한 언급이 없는 한 미니밴드로 테스트됩니다. 지정되지 않은 경우 기본값은 θ = 2, 3 스케일 및 3 애스펙트 비율의 앵커, ResNet-50-FPN 백본 및 600픽셀의 트레인 및 테스트 이미지 스케일입니다. (a) α-balanced CE를 사용한 RetinaNet은 최대 31.1 AP를 달성합니다. (b) 이와는 대조적으로 정확하게 동일한 네트워크를 가진 FL을 사용하면 2.9 AP를 얻을 수 있습니다. (c) 2-3 스케일 및 3 애스펙트 비 앵커를 사용하면 성능이 포화되는 좋은 결과를 얻을 수 있습니다. (d) FL은 온라인 하드 사례 마이닝(OHEM)의 최적의 변형을 3포인트 이상 능가합니다. (e) 다양한 네트워크 깊이 및 이미지 스케일에 대한 RetinaNet의 정확도/속도 트레이드오프(그림 2 참조)

**Optimization**: RetinaNet is trained with stochastic gradient descent (SGD). We use synchronized SGD over 8 GPUs with a total of 16 images per minibatch (2 images per GPU). Unless otherwise specified, all models are trained for 90k iterations with an initial learning rate of 0.01, which is then divided by 10 at 60k and again at 80k iterations. We use horizontal image flipping as the only form of data augmentation unless otherwise noted. Weight decay of 0.0001 and momentum of 0.9 are used. The training loss is the sum the focal loss and the standard smooth L1 loss used for box regression [10]. Training time ranges between 10 and 35 hours for the models in Table 1e.

최적화: 레티나넷은 확률적 구배강하(SGD) 훈련으로 미니배치당 총 16장(GPU당 2장)의 동기화된 SGD를 사용한다. 달리 명시되지 않은 한, 모든 모델은 0.01의 초기 학습률로 90k 반복에 대해 교육되며, 이후 60k에서 10으로 나누고 80k 반복에 대해 다시 교육을 받습니다. 별도의 언급이 없는 한 수평 이미지 플립을 유일한 데이터 확대 형태로 사용합니다. 무게 감소는 0.0001이고 운동량은 0.9입니다. 훈련 손실은 박스 회귀에 사용되는 초점 손실과 표준 평활 L1 손실을 합한 값입니다 [10]. 표 1e의 모델에 대한 교육 시간은 10시간에서 35시간 사이입니다.

1. **Experiments**

We present experimental results on the bounding box detection track of the challenging COCO benchmark [21]. For training, we follow common practice [1, 20] and use the COCO trainval35k split (union of 80k images from train and a random 35k subset of images from the 40k image val split). We report lesion and sensitivity studies by evaluating on the minival split (the remaining 5k images from val). For our main results, we report COCO AP on the test-dev split, which has no public labels and requires use of the evaluation server.

우리는 까다로운 COCO 벤치마크의 경계 박스 검출 트랙에 대한 실험 결과를 제시합니다 [21]. 교육을 위해, 우리는 공통 사례[1, 20]를 따르고 COCO train val35k 분할(열차에서 가져온 80k 이미지와 40k 이미지 분할에서 가져온 임의의 35k 하위 집합 이미지 결합)을 사용합니다. 미니발 분할(val의 나머지 5k 영상)을 평가하여 병변 및 민감도 스터디를 보고합니다. 우리의 주요 결과를 위해, 우리는 공인 라벨이 없고 평가 서버를 사용해야 하는 테스트-개발 분할에 대한 COCO AP를 보고합니다.

* 1. **Training Dense Detection**

We run numerous experiments to analyze the behavior of the loss function for dense detection along with various optimization strategies. For all experiments we use depth 50 or 101 ResNets [16] with a Feature Pyramid Network (FPN) [20] constructed on top. For all ablation studies we use an image scale of 600 pixels for training and testing.

다양한 최적화 전략과 함께 밀도감지를 위한 손실함수의 거동을 분석하기 위한 수많은 실험을 진행하고 있습니다. 모든 실험에는 FPN(Feature Pyramid Network)[20]이 위에 구성된 깊이 50 또는 101 ResNets[16]를 사용합니다. 모든 절제 스터디의 경우 교육 및 테스트에 600픽셀의 영상 스케일을 사용합니다.

**Network Initialization:** Our first attempt to train RetinaNet uses standard cross entropy (CE) loss without any modifications to the initialization or learning strategy. This fails quickly, with the network diverging during training. However, simply initializing the last layer of our model such that the prior probability of detecting an object is π = .01 (see §4.1) enables effective learning. Training RetinaNet with ResNet-50 and this initialization already yields a respectable AP of 30.2 on COCO. Results are insensitive to the exact value of π so we use π = .01 for all experiments.

네트워크 초기화: RetinaNet을 교육하기 위한 첫 번째 시도는 초기화 또는 학습 전략을 변경하지 않고 표준 교차 엔트로피(CE) 손실을 사용합니다. 이것은 교육 중에 네트워크가 분산되면서 빠르게 실패합니다. 그러나 단순히 모델의 마지막 레이어를 초기화하여 물체를 감지하는 사전 확률이 → =.01 (→4.1 참조)이 되도록 하면 효과적인 학습이 가능하다. RetinaNet을 ResNet-50으로 교육하고 이 초기화를 수행하면 COCO에서 30.2의 상당한 AP를 얻을 수 있습니다. 결과는 정확한 α값에 민감하지 않기 때문에 모든 실험에 대해 α =.01을 사용합니다.

**Balanced Cross Entropy:** Our next attempt to improve learning involved using the α-balanced CE loss described in §3.1. Results for various α are shown in Table 1a. Setting α = .75 gives a gain of 0.9 points AP.

균형 교차 엔트로피: § 3.1에 설명된 α-균형 CE 손실을 사용하여 학습을 개선하기 위한 우리의 다음 시도는 표 1a에 다양한 α에 대한 결과가 나와 있다. α = .75를 설정하면 0.9 포인트 AP의 이득이 됩니다.

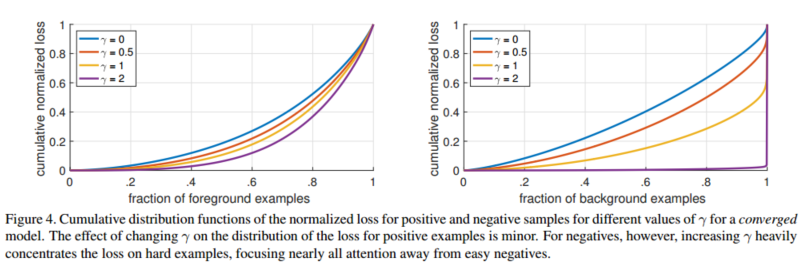


Figure 4. Cumulative distribution functions of the normalized loss for positive and negative samples for different values of γ for a converged model. The effect of changing γ on the distribution of the loss for positive examples is minor. For negatives, however, increasing γ heavily concentrates the loss on hard examples, focusing nearly all attention away from easy negatives.

그림 4 수렴 모형의 서로 다른 값인 θ에 대한 양 및 음의 표본에 대한 정규화 손실에 대한 누적 분포 함수입니다. 긍정적인 예에 대한 손실분포에 대한 on의 변경효과는 미미하다. 그러나 부정적인 면에서는 §를 증가시키면 손실은 어려운 예에 집중되며, 거의 모든 주의는 쉬운 부정적인 면으로부터 멀어진다.

**Focal Loss:** Results using our proposed focal loss are shown in Table 1b. The focal loss introduces one new hyperparameter, the focusing parameter γ, that controls the strength of the modulating term. When γ = 0, our loss is equivalent to the CE loss. As γ increases, the shape of the loss changes so that “easy” examples with low loss get further discounted, see Figure 1. FL shows large gains over CE as γ is increased. With γ = 2, FL yields a 2.9 AP improvement over the α-balanced CE loss.

초점 손실: 제안된 초점 손실을 사용한 결과는 표 1b에 나와 있습니다. 초점 손실은 변조 항의 강도를 제어하는 하나의 새로운 초점 매개변수, 즉 초점 매개변수 θ를 도입합니다. θ = 0일 때, 우리의 손실은 CE 손실과 같습니다. increases가 증가함에 따라 손실의 형태가 변경되어 손실이 낮은 "쉬운" 사례들이 추가로 할인된다(그림 1 참조). FL은 CE에 비해 →가 증가할수록 큰 상승폭을 보이고 있다. α = 2일 때 FL은 α-균형 CE 손실에 비해 2.9 AP의 향상을 보입니다.

For the experiments in Table 1b, for a fair comparison we find the best α for each γ. We observe that lower α’s are selected for higher γ’s (as easy negatives are downweighted, less emphasis needs to be placed on the positives). Overall, however, the benefit of changing γ is much larger, and indeed the best α’s ranged in just [.25,.75] (we tested α ∈ [.01, .999]). We use γ = 2.0 with α = .25 for all experiments but α = .5 works nearly as well (.4 AP lower).

표 1b의 실험에서 공정한 비교를 위해 각 α에 대한 최적의 α를 찾는다. 우리는 낮은 α가 높은 α에 대해 선택된다는 것을 관찰한다. (쉬운 음수가 하향가중되므로 양에 덜 강조될 필요가 있다.) 그러나 전반적으로 α 변화로 인한 편익이 훨씬 크며, 실제로 가장 좋은 α의 범위는 [.25,.75]였다(α α [.01,.999]). 모든 실험에서 α = .25와 함께 α = 2.0을 사용하지만 α =.5도 거의 효과가 있다(.4 AP 이하).

**Analysis of the Focal Loss:** To understand the focal loss better, we analyze the empirical distribution of the loss of a converged model. For this, we take take our default ResNet-101 600-pixel model trained with γ = 2 (which has 36.0 AP). We apply this model to a large number of random images and sample the predicted probability for ∼107 negative windows and ∼105 positive windows. Next, separately for positives and negatives, we compute FL for these samples, and normalize the loss such that it sums to one. Given the normalized loss, we can sort the loss from lowest to highest and plot its cumulative distribution function (CDF) for both positive and negative samples and for different settings for γ (even though model was trained with γ = 2).

초점 손실 분석: 초점 손실을 더 잘 이해하기 위해 수렴 모형의 손실에 대한 경험적 분포를 분석합니다. 이를 위해 = 2(36.0 AP 지원)로 교육을 받은 기본 ResNet-101600-100 모델을 사용합니다. 우리는 이 모델을 많은 수의 랜덤 이미지에 적용하고 ~107개의 마이너스 창과 ~105개의 플러스 창에 대한 예측 확률을 표본으로 추출합니다. 다음으로 양과 음의 경우 별도로 이러한 샘플에 대한 FL을 계산하고 손실을 합쳐서 1로 정규화합니다. 정규화된 손실을 기준으로 손실을 최저값에서 최고값으로 정렬하고 양의 표본과 음의 표본 모두에 대해 누적분포함수(CDF)를 표시할 수 있습니다(모형이 θ = 2로 교육되었더라도).

Cumulative distribution functions for positive and negative samples are shown in Figure 4. If we observe the positive samples, we see that the CDF looks fairly similar for different values of γ. For example, approximately 20% of the hardest positive samples account for roughly half of the positive loss, as γ increases more of the loss gets concentrated in the top 20% of examples, but the effect is minor.

양수 및 음수 표본에 대한 누적 분포 함수는 그림 4에 나와 있습니다. 양수 표본을 관찰하면 서로 다른 θ 값에 대해 CDF가 상당히 비슷해 보입니다. 예를 들어, 가장 단단한 양성 표본의 약 20%는 양의 손실의 약 절반을 차지합니다. increases가 증가하면 상위 20%의 예제에 더 많은 손실이 집중되지만 효과는 미미하기 때문입니다.

The effect of γ on negative samples is dramatically different. For γ = 0, the positive and negative CDFs are quite similar. However, as γ increases, substantially more weight becomes concentrated on the hard negative examples. In fact, with γ = 2 (our default setting), the vast majority of the loss comes from a small fraction of samples. As can be seen, FL can effectively discount the effect of easy negatives, focusing all attention on the hard negative examples.

음수 표본에 대한 α의 효과는 극적으로 다릅니다. θ = 0인 경우 양의 CDF와 음의 CDF는 상당히 유사합니다. 그러나 α가 증가할수록 하드 마이너스 사례에 훨씬 더 많은 가중치가 집중된다. 실제로 θ = 2(기본 설정)일 경우 손실의 대부분은 작은 표본에서 발생합니다. 보다시피 FL은 어려운 네거티브 사례에 모든 관심을 집중하면서 쉬운 네거티브 효과를 효과적으로 할인할 수 있다.

**Online Hard Example Mining (OHEM):** [31] proposed to improve training of two-stage detectors by constructing minibatches using high-loss examples. Specifically, in OHEM each example is scored by its loss, non-maximum suppression (nms) is then applied, and a minibatch is constructed with the highest-loss examples. The nms threshold and batch size are tunable parameters. Like the focal loss, OHEM puts more emphasis on misclassified examples, but unlike FL, OHEM completely discards easy examples. We also implement a variant of OHEM used in SSD [22]: after applying nms to all examples, the minibatch is constructed to enforce a 1:3 ratio between positives and negatives to help ensure each minibatch has enough positives.

온라인 하드 예제 마이닝(OHEM): [31] 고손실 예제를 사용하여 미니배치를 구성하여 2단계 검출기의 교육을 개선할 것을 제안했다. 특히, OHEM에서는 각 예제가 손실에 의해 점수가 매겨지고, 비최대 억제(nms)가 적용되며, 미니배치가 가장 높은 손실 예제로 구성됩니다. nms 임계값 및 배치 크기는 조정 가능한 매개 변수입니다. 초점 손실과 마찬가지로 OHEM은 잘못 분류된 예에 더 중점을 두지만, FL과 달리 OHEM은 쉬운 예제를 완전히 무시합니다. 또한 SSD에 사용되는 OHEM의 변형[22]을 구현합니다. 모든 예에 nms를 적용한 후 미니배치는 양과 음의 비율을 1:3으로 적용하여 각 미니배치가 충분한 양의 양성을 갖도록 합니다.

We test both OHEM variants in our setting of one-stage detection which has large class imbalance. Results for the original OHEM strategy and the ‘OHEM 1:3’ strategy for selected batch sizes and nms thresholds are shown in Table 1d. These results use ResNet-101, our baseline trained with FL achieves 36.0 AP for this setting. In contrast, the best setting for OHEM (no 1:3 ratio, batch size 128, nms of .5) achieves 32.8 AP. This is a gap of 3.2 AP, showing FL is more effective than OHEM for training dense detectors. We note that we tried other parameter setting and variants for OHEM but did not achieve better results.

우리는 클래스 불균형이 큰 1단계 감지 설정에서 두 OHEM 변종을 모두 테스트합니다. 선택한 배치 크기 및 nms 임계값에 대한 원래 OHEM 전략과 'OHEM 1:3' 전략에 대한 결과가 표 1d에 나와 있습니다. 이러한 결과는 ResNet-101을 사용하며, FL로 교육된 기준선이 이 설정에 대해 36.0 AP를 달성합니다. 반대로 OHEM에 대한 최상의 설정(1:3 비율 없음, 배치 크기 128, nms of.5)은 32.8 AP를 달성합니다. 이는 3.2 AP의 격차로 FL이 OHEM보다 고밀도 검출기 훈련에 더 효과적이라는 것을 보여준다. OHEM에 대해 다른 파라미터 설정 및 변형을 시도했지만 더 나은 결과를 얻지 못했습니다.

**Hinge Loss**: Finally, in early experiments, we attempted to train with the hinge loss [13] on pt, which sets loss to 0 above a certain value of pt. However, this was unstable and we did not manage to obtain meaningful results. Results exploring alternate loss functions are in the appendix.

힌지 손실: 마지막으로, 초기 실험에서, 우리는 pt에서 힌지 손실[13]을 가지고 훈련하려고 시도했습니다. 이 손실은 pt의 특정 값보다 0으로 설정되었습니다. 하지만 이것은 불안정했고 우리는 의미 있는 결과를 얻지 못했습니다. 대체 손실 함수를 탐색한 결과는 부록에 있습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 2. Object detection single-model results (bounding box AP), vs. state-of-the-art on COCO test-dev. We show results for our RetinaNet-101-800 model, trained with scale jitter and for 1.5× longer than the same model from Table 1e. Our model achieves top results, outperforming both one-stage and two-stage models. For a detailed breakdown of speed versus accuracy see Table 1e and Figure 2.

표 2 COCO test-dev의 객체 검출 단일 모델 결과(경계 상자 AP)와 최첨단 비교. 표 1e의 동일한 모델보다 1.5배 긴 스케일 지터로 교육받은 RetinaNet-101-800 모델에 대한 결과를 보여줍니다. 당사의 모델은 1단계 모델과 2단계 모델 모두를 능가하는 최고의 성과를 달성했습니다. 속도 대 정확도에 대한 자세한 내용은 표 1e와 그림 2를 참조하십시오.

* 1. **. Model Architecture Design**

**Anchor Density:** One of the most important design factors in a one-stage detection system is how densely it covers the space of possible image boxes. Two-stage detectors can classify boxes at any position, scale, and aspect ratio using a region pooling operation [10]. In contrast, as one-stage detectors use a fixed sampling grid, a popular approach for achieving high coverage of boxes in these approaches is to use multiple ‘anchors’ [28] at each spatial position to cover boxes of various scales and aspect ratios.

앵커 밀도: 1단계 감지 시스템에서 가장 중요한 디자인 요소 중 하나는 가능한 이미지 박스의 공간을 얼마나 촘촘하게 커버하느냐입니다. 2단계 검출기는 영역 풀링 작업을 사용하여 임의의 위치, 척도 및 가로 세로 비율에서 상자를 분류할 수 있습니다[10]. 반대로, 1단계 검출기는 고정 표본 추출 그리드를 사용하기 때문에 이러한 접근법에서 상자의 높은 적용범위를 달성하기 위한 일반적인 접근법은 다양한 척도 및 가로 세로 비율의 상자를 커버하기 위해 각 공간 위치에서 복수의 '앵커'[28]를 사용하는 것이다.

We sweep over the number of scale and aspect ratio anchors used at each spatial position and each pyramid level in FPN. We consider cases from a single square anchor at each location to 12 anchors per location spanning 4 sub-octave scales (2k/4, for k ≤ 3) and 3 aspect ratios [0.5, 1, 2]. Results using ResNet-50 are shown in Table 1c. A surprisingly good AP (30.3) is achieved using just one square anchor. However, the AP can be improved by nearly 4 points (to 34.0) when using 3 scales and 3 aspect ratios per location. We used this setting for all other experiments in this work.

우리는 FPN의 각 공간 위치와 각 피라미드 레벨에서 사용되는 축척 및 가로 세로 비율 앵커 수를 스위프합니다. 우리는 각 위치의 단일 사각 앵커에서 4개의 하위 옥타브 척도(2k/4, k ≤ 3)와 3개의 가로 세로 비율[0.5, 1,2]에 이르는 위치당 12개의 앵커까지 사례를 고려합니다. ResNet-50을 사용한 결과는 표 1c에 나와 있습니다. 단 하나의 사각 앵커를 사용하여 놀라울 정도로 양호한 AP(30.3)를 달성할 수 있습니다. 그러나 위치당 척도 3개와 가로 세로 비율 3개를 사용할 경우 AP를 거의 4포인트(최대 34.0) 개선할 수 있습니다. 우리는 이 설정을 이 작업의 다른 모든 실험에 사용했습니다.

Finally, we note that increasing beyond 6-9 anchors did not shown further gains. Thus while two-stage systems can classify arbitrary boxes in an image, the saturation of performance w.r.t. density implies the higher potential density of two-stage systems may not offer an advantage.

마지막으로, 우리는 6-9 앵커 이상으로 증가하는 것이 추가적인 이득을 보여주지 못했다는 것을 주목합니다. 따라서 2단계 시스템이 이미지에서 임의의 상자를 분류할 수 있지만, 성능 w.r.t. 밀도의 포화도는 2단계 시스템의 높은 잠재적 밀도가 이점을 제공하지 않을 수 있음을 암시합니다.

**Speed versus Accuracy:** Larger backbone networks yield higher accuracy, but also slower inference speeds. Likewise for input image scale (defined by the shorter image side). We show the impact of these two factors in Table 1e. In Figure 2 we plot the speed/accuracy trade-off curve for RetinaNet and compare it to recent methods using public numbers on COCO test-dev. The plot reveals that RetinaNet, enabled by our focal loss, forms an upper envelope over all existing methods, discounting the low-accuracy regime. RetinaNet with ResNet-101-FPN and a 600 pixel image scale (which we denote by RetinaNet-101-600 for simplicity) matches the accuracy of the recently published ResNet-101-FPN Faster R-CNN[20], while running in 122 ms per image compared to 172 ms (both measured on an Nvidia M40 GPU). Using larger scales allows RetinaNet to surpass the accuracy of all two-stage approaches, while still being faster. For faster runtimes, there is only one operating point (500 pixel input) at which using ResNet-50-FPN improves over ResNet-101-FPN. Addressing the high frame rate regime will likely require special network design, as in [27], and is beyond the scope of this work. We note that after publication, faster and more accurate results can now be obtained by a variant of Faster R-CNN from [12].

속도 대 정확도: 백본 네트워크가 클수록 정확도는 높아지지만 추론 속도는 느려집니다. 입력 이미지 축척의 경우에도 마찬가지입니다(짧은 이미지 측면에서 정의됨). 우리는 표 1e에서 이 두 요소의 영향을 보여줍니다. 그림 2에서는 RetinaNet의 속도/정확도 균형 곡선을 표시하고 COCO 테스트-개발에서 공개 번호를 사용한 최근 방법과 비교합니다. 이 플롯은 우리의 초점 손실에 의해 활성화된 RetinaNet이 기존의 모든 방법에 대해 상위 엔벨롭을 형성하여 낮은 정확도 체계를 할인한다는 것을 보여줍니다. RetinaNet은 ResNet-101-FPN과 600픽셀 이미지 스케일(간단성을 위해 RetinaNet-101-600으로 표시됨)이 172ms(둘 다 엔비디아 M40 GPU에서 측정됨)에 비해 이미지당 122ms로 실행되면서 최근에 게시된 ResNet-101-FPN Faster R-CNN[20]의 정확도와 일치한다. 더 큰 스케일을 사용하는 것은 레티나넷이 모든 2단계 접근의 정확성을 능가하는 동시에 더 빠른 속도를 낼 수 있게 해준다. 실행 시간을 단축하기 위해 ResNet-50-FPN을 사용하는 것이 ResNet-101-FPN에 비해 개선된 작동 지점(500픽셀 입력)이 하나만 있습니다. 높은 프레임률 체제를 다루려면 [27]에서와 같이 특별한 네트워크 설계가 필요할 가능성이 높으며, 본 연구의 범위를 벗어납니다. 이제 [12]의 Faster R-CNN 변형 모델을 통해 더 빠르고 정확한 결과를 얻을 수 있습니다.

* 1. **. Comparison to State of the Art**

We evaluate RetinaNet on the challenging COCO dataset and compare test-dev results to recent state-of-the-art methods including both one-stage and two-stage models. Results are presented in Table 2 for our RetinaNet-101-800 model trained using scale jitter and for 1.5× longer than the models in Table 1e (giving a 1.3 AP gain). Compared to existing one-stage methods, our approach achieves a healthy 5.9 point AP gap (39.1 vs. 33.2) with the closest competitor, DSSD [9], while also being faster, see Figure 2. Compared to recent two-stage methods, RetinaNet achieves a 2.3 point gap above the top-performing Faster R-CNN model based on Inception-ResNet-v2-TDM [32]. Plugging in ResNeXt-32x8d-101-FPN [38] as the RetinaNet backbone further improves results another 1.7 AP, surpassing 40 AP on COCO.

우리는 도전적인 COCO 데이터셋에 대해 RetinaNet을 평가하고 테스트 개발 결과를 1단계 및 2단계 모델을 포함한 최신 방법과 비교합니다. 결과는 스케일 지터를 사용하여 교육한 RetinaNet-101-800 모델에 대한 표 2에 제시되어 있으며 표 1e의 모델보다 1.5배 더 길다(1.3 AP 이득). NAT의 접근 방식은 기존 1단계 방식과 비교하여 가장 가까운 경쟁업체 DSD[9]를 통해 33.2 대비 5.9 포인트 AP 격차(39.1 포인트)를 달성하면서도 속도가 더 빠릅니다(그림 2 참조). 최근의 2단계 방법과 비교했을 때, RetinaNet은 Inception-ResNet-v2-TDM[32]에 기초한 최고 성능의 Faster R-CNN 모델보다 2.3 포인트 격차를 달성합니다. RetinaNet 백본으로 ResNeXt-32x8d-101-FPN [38]을 꽂으면 추가로 1.7 AP의 결과가 개선되어 COCO에서 40 AP를 능가합니다.

1. **Conclusion**

In this work, we identify class imbalance as the primary obstacle preventing one-stage object detectors from surpassing top-performing, two-stage methods. To address this, we propose the focal loss which applies a modulating term to the cross entropy loss in order to focus learning on hard negative examples. Our approach is simple and highly effective. We demonstrate its efficacy by designing a fully convolutional one-stage detector and report extensive experimental analysis showing that it achieves stateof-the-art accuracy and speed. Source code is available at <https://github.com/facebookresearch/>

Detectron [12].

본 연구에서는 클래스 불균형을 1단계 장애물 감지기가 최고 성능의 2단계 방법을 능가하는 것을 막는 주요 장애물로 파악한다. 이를 해결하기 위해, 우리는 어려운 부정적 사례에 대한 학습을 집중하기 위해 교차 엔트로피 손실에 변조 용어를 적용하는 초점 손실을 제안합니다. 우리의 접근 방식은 간단하고 매우 효과적입니다. 우리는 완전한 콘볼루션 1단계 검출기를 설계하여 그 효과를 입증하고, 그것이 최첨단 정확도와 속도를 달성한다는 것을 보여주는 광범위한 실험 분석을 보고한다. 소스 코드는 https://github.com/facebookresearch/Detectron[12]에서 확인할 수 있습니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Figure 5. Focal loss variants compared to the cross entropy as a function of xt = yx. Both the original FL and alternate variant FL∗ reduce the relative loss for well-classified examples (xt > 0).

그림 5 xt = yx의 함수로서 교차 엔트로피와 비교되는 초점 손실 변형. 원래의 FL과 대체 바리안트 FL'은 모두 잘 분류된 예(xt > 0)의 상대적인 손실을 줄입니다.

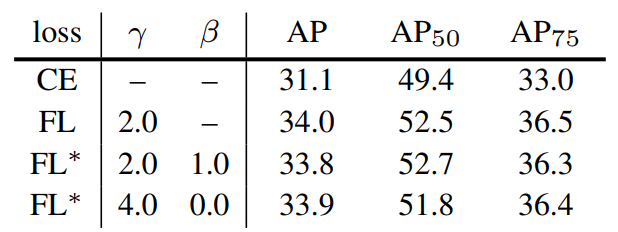


Table 3. Results of FL and FL∗versus CE for select settings.

표 3 선택한 설정에 대한 FL 및 FL'sus CE의 결과.

**Appendix A : Focal Loss**

The exact form of the focal loss is not crucial. We now show an alternate instantiation of the focal loss that has similar properties and yields comparable results. The following also gives more insights into properties of the focal loss.

We begin by considering both cross entropy (CE) and the focal loss (FL) in a slightly different form than in the main text. Specifically, we define a quantity xt as follows: where y ∈ {±1} specifies the ground-truth class as before. We can then write pt = σ(xt) (this is compatible with the definition of pt in Equation 2). An example is correctly classified when xt > 0, in which case pt > .5.

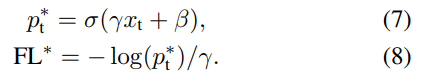
정확한 초점 손실 형태는 중요하지 않습니다. 이제 유사한 속성을 가지고 있고 유사한 결과를 산출하는 초점 손실에 대한 대체 인스턴스화를 보여 줍니다. 다음은 초점 손실의 특성에 대한 더 많은 통찰력을 제공합니다.

먼저 교차 엔트로피(CE)와 초점 손실(FL)을 모두 메인 텍스트와 약간 다른 형태로 고려합니다. 특히 수량 xt를 다음과 같이 정의합니다. 여기서 y ÷ {±1}은(는) 이전과 같이 지면-진실 클래스를 지정합니다. 그런 다음 pt = ((xt)를 쓸 수 있습니다(이는 방정식 2의 pt 정의와 호환됩니다). 예는 xt > 0일 때 올바르게 분류되며, 이 경우 pt >.5입니다.



We can now define an alternate form of the focal loss in terms of xt. We define p∗t and FL∗as follows:

FL∗has two parameters, γ and β, that control the steepness and shift of the loss curve. We plot FL∗for two selected settings of γ and β in Figure 5 alongside CE and FL. As can be seen, like FL, FL∗ with the selected parameters diminishes the loss assigned to well-classified examples.



We trained RetinaNet-50-600 using identical settings as before but we swap out FL for FL∗ with the selected parameters. These models achieve nearly the same AP as those trained with FL, see Table 3. In other words, FL∗is a reasonable alternative for the FL that works well in practice.

기존과 동일한 설정으로 RetinaNet-50-600을 교육하였으나, 선택된 파라미터로 FL을 FL with로 교체하였습니다. 이러한 모델은 FL로 교육된 모델과 거의 동일한 AP를 달성합니다(표 3 참조). 즉, 실제로 잘 작동하는 FL의 합리적인 대안입니다.

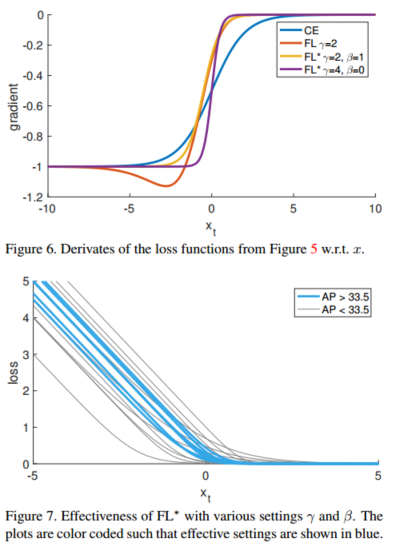


Figure 6. Derivates of the loss functions from Figure 5 w.r.t. x.

그림 6 그림 5 w.r.t. x.의 손실 함수의 도함수.

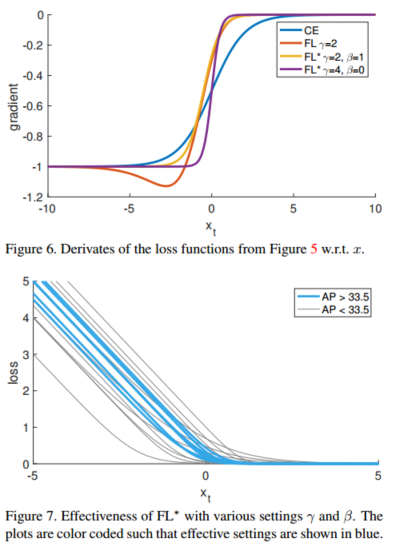


Figure 7. Effectiveness of FL∗ with various settings γ and β. The plots are color coded such that effective settings are shown in blue.

그림 7. 다양한 설정 θ 및 β를 가진 FL†의 효과 유효한 설정이 파란색으로 표시되도록 그림이 색상으로 구분되어 있습니다.

We found that various γ and β settings gave good results. In Figure 7 we show results for RetinaNet-50-600 with FL∗for a wide set of parameters. The loss plots are color coded such that effective settings (models converged and with AP over 33.5) are shown in blue. We used α = .25 in all experiments for simplicity. As can be seen, losses that reduce weights of well-classified examples (xt > 0) are effective. More generally, we expect any loss function with similar properties as FL or FL∗to be equally effective.

다양한 β 및 β 설정이 좋은 결과를 준다는 것을 발견했다. 그림 7에서는 광범위한 파라미터 집합에 대한 FLfor을 사용한 RetinaNet-50-600의 결과를 보여줍니다. 손실 그림은 효과적인 설정(융합된 모형 및 AP가 33.5 이상인 모형)이 파란색으로 표시되도록 색상으로 구분됩니다. 단순성을 위해 모든 실험에 α = .25를 사용했다. 잘 분류된 사례의 가중치를 줄이는 손실(xt > 0)이 효과적이다. 보다 일반적으로 FL 또는 FLto과 유사한 특성을 가진 손실 함수가 동일하게 효과적일 것으로 예상합니다.

**Appendix B : Derivatives**

For reference, derivates for CE, FL, and FL∗ w.r.t. x are:Plots for selected settings are shown in Figure 6. For all loss functions, the derivative tends to -1 or 0 for high-confidence predictions. However, unlike CE, for effective settings of both FL and FL∗, the derivative is small as soon as xt > 0.

참고로 CE, FL 및 FL∗ w.r.t.x용 파생상품은 다음과 같습니다.선택한 설정에 대한 그림은 그림 6에 나와 있습니다. 모든 손실 함수에 대해 고신뢰 예측의 경우 파생형은 -1 또는 0인 경향이 있습니다. 그러나 CE와 달리 FL과 FL,의 효과적인 설정의 경우, 파생 모델은 xt > 0이면 작습니다.