Focal Loss

- 主要观点
  - 训练数据中的类别不平衡是 one-stage 检测器精度不高的主要原因
    - two-stage， cascaded，启发式采样（1:3, OHEM）
  - 提出 focal loss 来减少分类正确的样本对 loss 的贡献
    - 弱化（而非消灭）水军
    - 该 loss 的思想很重要，具体形式不重要
  - 提出 RetinaNet，用 focal loss 来训练
    - one-stage 的速度
    - two-stage 的精度

Focal Loss

- 起作用的是 loss 而非网络结构

  “We emphasize that our simple detector achieves top results not based on innovations in network design but due to our novel loss.”

Focal Loss

- class imbalance
  - 大部分 locations 是 easy negatives，不仅无法提供有效的训练，还能起副作用
  - 常见策略： head negative mining
  - 本文策略： focal loss
    - focus training on a sparse set of hard examples

Focal Loss

- CE-like loss

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise.} \end{cases} \quad (1)$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$ is the model's estimated probability for the class with label $y = 1$. For notational convenience, we define p_t:

$$p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{otherwise,} \end{cases} \quad (2)$$

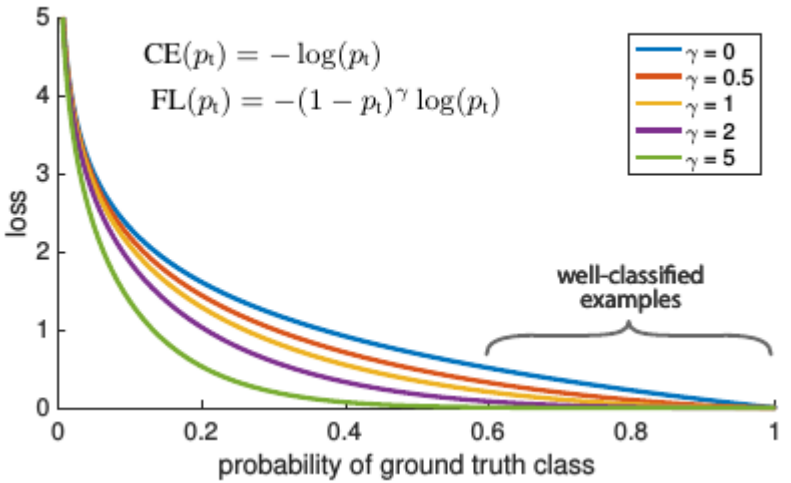
and rewrite $\mathrm{CE}(p, y) = \mathrm{CE}(p_t) = -\log(p_t)$.

A common method for addressing class imbalance is to introduce a weighting factor $\alpha \in [0, 1]$ for class 1 and $1-\alpha$ 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 α_t analogously to how we defined p_t. We write the α-balanced CE loss as:

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t). \quad (3)$$

In practice we use an α-balanced variant of the focal loss:

$$\mathrm{FL}(p_t) = -\alpha_t (1-p_t)^\gamma \log(p_t). \quad (5)$$



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 $(1-p_t)^\gamma$ to the cross entropy loss, with tunable *focusing* parameter $\gamma \geq 0$. We define the focal loss as:

$$\mathrm{FL}(p_t) = -(1-p_t)^\gamma \log(p_t). \quad (4)$$

RetinaNet

- 网络结构
  - FPN backbone + class subnet+ box subnet

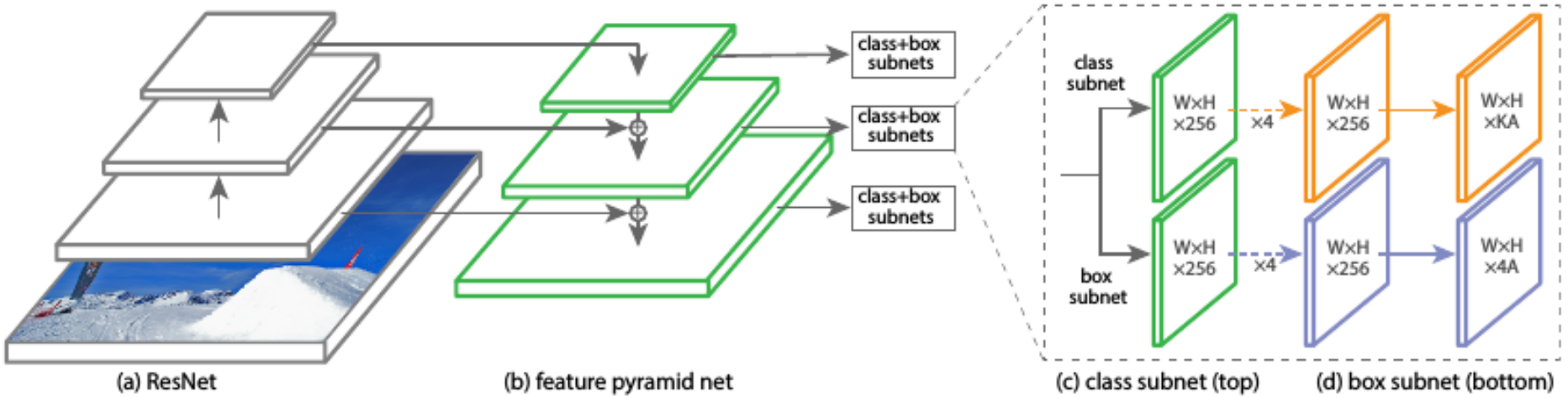


Figure 3. The one-stage **RetinaNet** network architecture uses a Feature Pyramid Network (FPN) [20] backbone on top of a feedforward ResNet architecture [16] (a) to generate a rich, multi-scale convolutional feature pyramid (b). To this backbone RetinaNet attaches two subnetworks, one for classifying 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.

RetinaNet

- Anchors
  - five pyramid levels: P3-P7

    three aspect ratios per pyramid level: { 1:2, 1:1, 2:1 }

    three anchor scales per ratio: { 2^0, 2^(1/3), 2^(2/3) }
  - each anchor → K classes targets + 4 boxes targets
    - class target: one-hot vector
    - box target: offset between anchor box and GT box
  - anchor → GT box, if IoU >= 0.5

    anchor → background box, if IoU < 0.4

    anchor → ignore during training, if IoU in [0.4, 0.5)

RetinaNet

- 分类子网络
  - 用一个小的 FCN，在每一个 pyramid level 的 feature map 的每一个位置上，预测每一个 anchor box 的类别分布
- 回归子网络
  - 用一个小的 FCN，在每一个 pyramid level 的 feature map 的每一个位置上，回归每一个 anchor box 和其邻近的 GT box 之间的 offset
- 虽然两个网络结构几乎相同，但不共享参数！

Focal Loss

- 深入分析
  - 当 r=2 时， 20% 的难正样本贡献了超过 60% 的 loss
  - 当 r=2 时， 5% 的难负样本贡献了超过 95% 的 loss

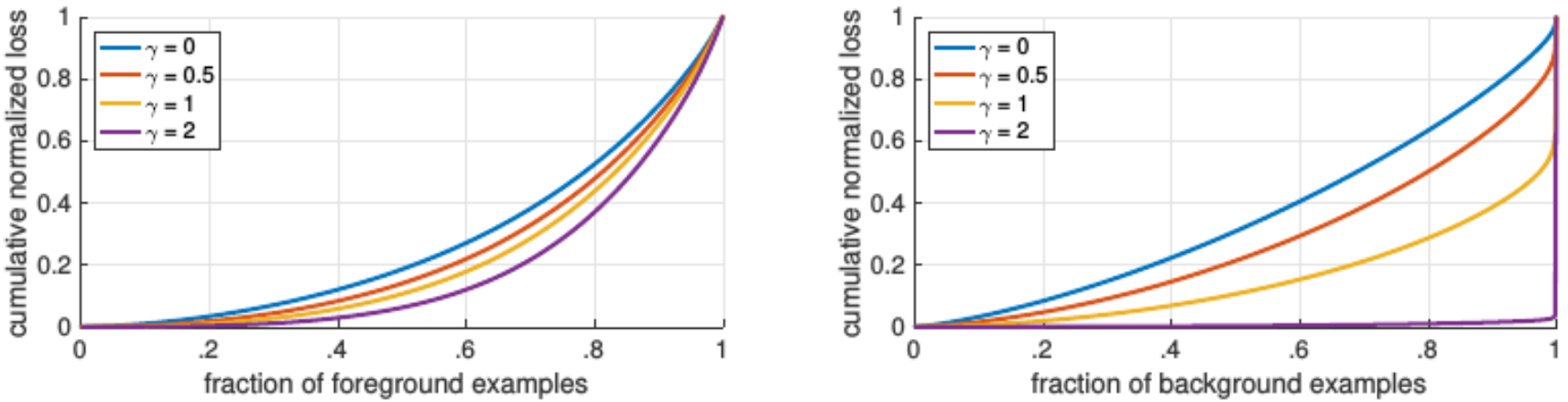


Figure 4. Cumulative distribution functions of the normalized loss for positive and negative samples for different values of γ for a *converged* model. The effect of changing γ on the distribution of the loss for positive examples is minor. For negatives, however, increasing γ heavily concentrates the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

Focal Loss vs. OHEM

- 相同点
  - 都增大 hard examples 的权重
- 不同点
  - OHEM 丢弃了 easy examples