Appendix for 'What Makes for End-to-End Object Detection ?'

1. Proof for Theoretical Analysis

We focus on analyzing properties using linear classifier. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be an instance space and $\mathcal{Y} = \{+1, -1\}$ be the label space. The label of a positive sample is $+1$ while that of a negative sample is -1. We wish to train a classifier h, coming from a hypothesis class $\mathcal{H} = \{x \mapsto \text{sign}(w^\mathsf{T} x) : w \in \mathbb{R}^d\}$. Note that we can express the bias term b by rewriting $w = [\hat{w}, b]^\mathsf{T}$ and $x = [\hat{x}, 1]^\mathsf{T}$. We use the perceptron's update rule with mini-batch size of 1. That is, given the classifier $w_t \in \mathbb{R}^d$, the update is only performed on incorrectly classified example $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ as given by $w_{t+1} = w_t + \eta y_t x_t$ where η is the stepsize.

Proposition 4.2 (Feasibility) *Suppose that the one-to-one assignment is run on a sequence of examples from $\mathcal{X} \times \mathcal{Y}$. Given weight vector $w_t = [\hat{w}_t, b_t]^\mathsf{T}$ at update step t, there exists $\gamma_t \in \mathbb{R}$ and $\delta_t > 0$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we have $y w_t^{*\mathsf{T}} x \geq \delta_t$ with $w_t^* = [\hat{w}_t, \gamma_t]^\mathsf{T}$.*

Proof. we denote $x_t^1 = \arg\max_{x \in \mathcal{X}} {w_t}^\mathsf{T} x$ and $x_t^2 = \arg\max_{x \in \mathcal{X} \setminus \{x_t^1\}} {w_t}^\mathsf{T} x$. We assume ${w_t^1}^\mathsf{T} x > 0$, hence we can infer that ${w_t}^\mathsf{T} x_t^1 > {w_t}^\mathsf{T} x_t^2 > 0$, otherwise the algorithm converges at w_t because it satisfies that ${w_t}^\mathsf{T} x_t^1 > 0$ and ${w_t}^\mathsf{T} x \leq 0$ for all $x \in \mathcal{X} \setminus \{x_t^1\}$. By one-to-one assignment, the label of x_t^1 is $y(x_t^1) = +1$ and the labels of the remaining samples in $\mathcal{X}$ are $y(x) = -1, x \in \mathcal{X} \setminus \{x_t^1\}$. Take $\gamma_t = -\frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 + \hat{x}_2)}{2}$, we have

$$\begin{aligned} y(x_t^1) {w_t^*}^\mathsf{T} x_t^1 &= {\hat{w}_t}^\mathsf{T} \hat{x}_t^1 - \frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 + \hat{x}_2)}{2} \\ &= \frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 - \hat{x}_2)}{2} > 0 \end{aligned} \tag{1}$$

and for all $x \in \mathcal{X} \setminus \{x_t^1\}$ we have

$$\begin{aligned} y(x) {w_t^*}^\mathsf{T} x &= -1 * ({\hat{w}_t}^\mathsf{T} \hat{x} - \frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 + \hat{x}_2)}{2}) \\ &\geq \frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 - \hat{x}_2)}{2} > 0 \end{aligned} \tag{2}$$

where the first inequality holds by ${w_t}^\mathsf{T} x \leq {w_t}^\mathsf{T} x_t^2$ since $x_t^2 = \arg\max_{x \in \mathcal{X} \setminus \{x_t^1\}} {w_t}^\mathsf{T} x$.

By Eqn.(1) and Eqn.(2), we can take $\delta_t = \frac{{\hat{w}_t}^\mathsf{T}(\hat{x}_1 - \hat{x}_2)}{2}$.

Theorem 4.3 (Convergence) *Let γ_{t+1} and γ_t be the constants defined in Proposition 4.2. For each update step t, we assume there exists a stepsize η_t such that $\|x_t\|^2 \eta_t^2 + y_t(\gamma_{t+1} - 2\gamma_t)\eta_t + b_t(\gamma_{t+1} - \gamma_t) > 0$ where (x_t, y_t) be the incorrectly classified sample at iteration t. If the sample label is assigned by one-to-one assignment, then, $t \leq \frac{\eta_{max}^2 - 2\eta_{min}\delta_{\min}({w_1}^\mathsf{T} w_0^* - \|w_0\| - \eta_{max})}{2\eta_{min}^2 \delta_{\min}^2}$ where η_{max} and η_{min} are the maximum and minimum value of stepsize among all t's updates, w_1 is the classifier after the first update and $\delta_{\min}$ is the minimum of all δ_ts in Proposition 4.2. All instances at initialization can be correctly classified by w_0^*.*

We first show ${w_{t+1}}^\mathsf{T} w_{t+1}^* \geq {w_{t+1}}^\mathsf{T} w_t^*$. Rewriting the weight vector w_t into a normal vector and a bias gives us

$$\begin{bmatrix} \hat{w}_{t+1} \\ b_{t+1} \end{bmatrix} = \begin{bmatrix} \hat{w}_t \\ b_t \end{bmatrix} + \eta y_t \begin{bmatrix} \hat{x}_t \\ 1 \end{bmatrix} \tag{3}$$

From Eqn.(3), we have $w_{t+1} = [\hat{w}_t + \eta y_t \hat{x}_t, b_t + \eta y_t]^\mathsf{T}$ at update t. According to the definition of γ_t and γ_{t+1}, we obtain $w_{t+1}^* = [\hat{w}_t + y_t \hat{x}_t, \gamma_{t+1}]$ and $w_t^* = [\hat{w}_t, \gamma_t]$. Therefore, we can derive that

$$\begin{aligned} &{w_{t+1}}^\mathsf{T} w_{t+1}^* - {w_{t+1}}^\mathsf{T} w_t^* \\ &= (\hat{w}_t + \eta y_t \hat{x}_t)^\mathsf{T} \eta y_t \hat{x}_t + (b_t + \eta y_t)(\gamma_{t+1} - \gamma_t) \\ &= \|x_t\|^2 \eta^2 + y_t(\hat{w}_t \hat{x}_t - \gamma_t + \gamma_{t+1})\eta + b_t(\gamma_{t+1} - \gamma_t) \\ &\geq \|x_t\|^2 \eta^2 + y_t(\gamma_{t+1} - 2\gamma_t)\eta + b_t(\gamma_{t+1} - \gamma_t) \end{aligned} \tag{4}$$

Taking $\eta = \eta_t$ gives us ${w_{t+1}}^\mathsf{T} w_{t+1}^* \geq {w_{t+1}}^\mathsf{T} w_t^*$ by the assumption. Note that the assumption in Theorem 4.3 easily holds when η_t is a large but finite number due to the property of quadratic equation of one variable in Eqn.(4).

To proceed, we find upper and lower bounds on the length of the weight vector w_t to show finite number of updates. By convenience, we normalize w_t^* to $\|w_t^*\| = 1$. Assume that after $t + 1$ steps the weight vector w_{t+1} has been computed. This means that at time t a training sample was incorrectly classified by the weight vector w_t and so $w_{t+1} = w_t + \eta_t y_t x_t$. By one-to-one assignment, we have $y_t = 1$ if $x_t = \arg\max_{x \in \mathcal{X}} {w_t}^\mathsf{T} x$ and -1 otherwise.

By computing the length of w_{t+1}, we arrive at

$$\begin{aligned} \|w_{t+1}\|^2 &= (w_t + \eta_t y_t x_t)^\mathsf{T} (w_t + \eta_t y_t x_t) \\ &= \|w_t\|^2 + \|x_t\|^2 \eta_t^2 + 2 y_t {w_t}^\mathsf{T} x_t \eta_t \\ &\leq \|w_t\|^2 + \eta_t^2 \end{aligned} \tag{5}$$

where the third equation holds because the length of instance x is bounded by 1 and $y_t {w_t}^\mathsf{T} x_t$ is negative or zero (otherwise we would have not corrected w_t using sample

(x_t, y_t) by perceptron's update rule) . Induction through Eqn.(5) then gives us

$$\|w_{t+1}\|^2 \leq \|w_0\|^2 + \sum_{k=0}^{t} \eta_k^2 \leq (t+1)\eta_{max}^2 \tag{6}$$

where $\eta_{min} = \max\{\eta_k : k = 0, 1, \cdots, t\}$. To drive the lower bound, we multiply w_t^* in Proposition 4.2 on both sides of $w_{t+1} = w_t + \eta_t y_t x_t$, it gives us ${w_{t+1}}^{\mathsf{T}} w_t^* = {w_t}^{\mathsf{T}} w_t^* + \eta_t y_t {w_t^*}^{\mathsf{T}} x_t$. By Eqn.(4), it can be relaxed into

$$\begin{aligned} {w_{t+1}}^{\mathsf{T}} w_t^* &= {w_t}^{\mathsf{T}}(w_t^* - w_{t-1}^* + w_{t-1}^*) + \eta_t y_t {w_t^*}^{\mathsf{T}} x_t \\ &= {w_t}^{\mathsf{T}} w_{t-1}^* + {w_t}^{\mathsf{T}}(w_t^* - w_{t-1}^*) + \eta_t y_t {w_t^*}^{\mathsf{T}} x_t \\ &\geq {w_t}^{\mathsf{T}} w_{t-1}^* + \eta_t y_t {w_t^*}^{\mathsf{T}} x_t \\ &\geq {w_t}^{\mathsf{T}} w_{t-1}^* + \eta_t \delta_t \end{aligned} \tag{7}$$

where the first inequality holds by Eqn.(4), the second inequality holds by Proposition 4.2. Induction through Eqn.(7) then yields

$${w_{t+1}}^{\mathsf{T}} w_t^* \geq {w_1}^{\mathsf{T}} w_0^* + \sum_{k=1}^{t} \eta_k \delta_k \geq {w_1}^{\mathsf{T}} w_0^* + t\eta_{min}\delta_{min} \tag{8}$$

where $\delta_{min} = \min\{\delta_k : k = 1, \cdots, t\}$ and $\eta_{min} = \min\{\eta_k : k = 1, \cdots, t\}$. Combining Eqn.(6) and Eqn.(8), we obtain that

$${w_1}^{\mathsf{T}} w_0^* + t\eta_{min}\delta_{\min} \leq \sqrt{\|w_0\|^2 + (t+1)\eta_{max}^2} \tag{9}$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, the above implies that

$${w_1}^{\mathsf{T}} w_0^* + t\eta_{min}\delta_{\min} \leq \|w_0\| + \sqrt{t}\eta_{max} + \eta_{max} \tag{10}$$

Using standard algebraic manipulations, the above implies that

$$\begin{aligned} t &\leq \left(\frac{\eta_{max} + \sqrt{\eta_{max}^2 - 4\eta_{min}\delta_{\min}({w_1}^{\mathsf{T}} w_0^* - \|w_0\| - \eta_{max})}}{2\eta_{min}\delta_{\min}}\right)^2 \\ &\leq \frac{\eta_{max}^2 - 2\eta_{min}\delta_{\min}({w_1}^{\mathsf{T}} w_0^* - \|w_0\| - \eta_{max})}{2\eta_{min}^2\delta_{\min}^2} \end{aligned} \tag{11}$$

This completes the proof.

2. Positive Samples for Multiple Objects

As discussed in Section 4, when there exists an object in the image, classification cost results in a clear score gap between the sample of the first-highest score and the sample of the second-highest score. In Figure 1, we show positive sample for multiple objects. Classification cost produces two clusters of samples, one of which is composed of positive samples, and their scores are obviously higher than samples in another cluster.

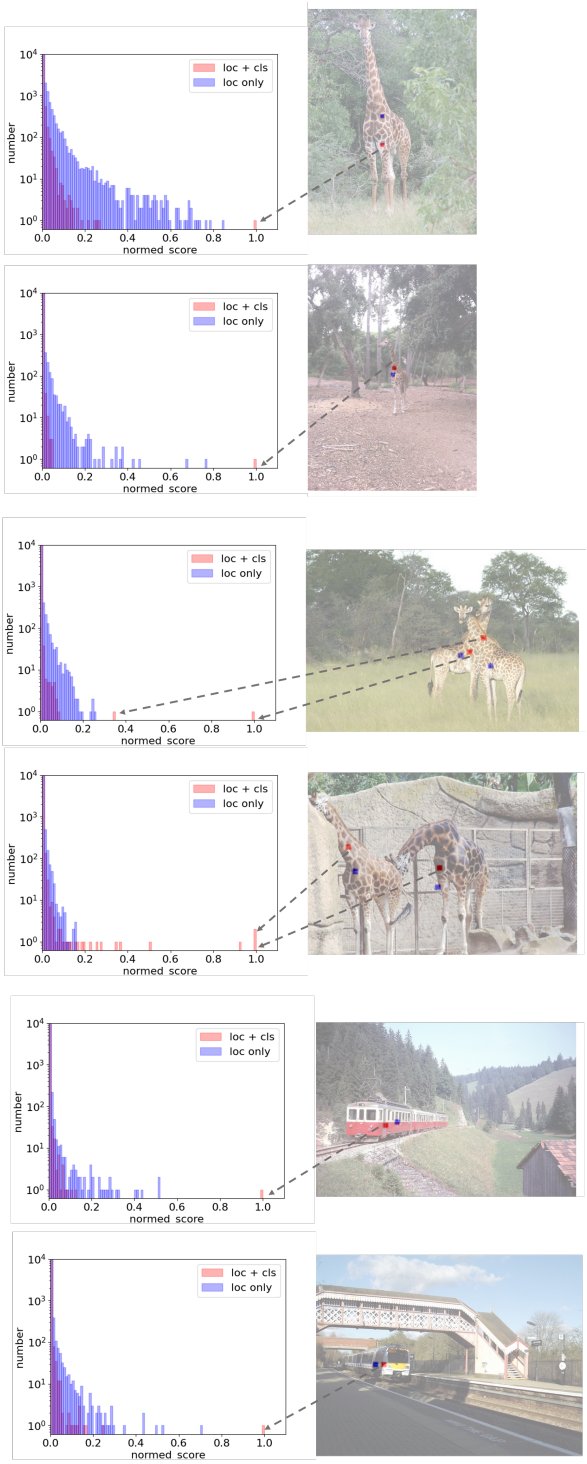

Figure 1. **Positive samples in different training images.** For better visualization, we only show the part below the number of 10^4, and scores are normalized to [0, 1]. Blue bins show the detector trained with positive samples chosen by only location cost. Red bins consider both location cost and classification cost. For multiple objects, classification cost produces two clusters of samples, the scores of positive sample cluster are obviously higher than samples in negative sample cluster.