Bounding Box Encoding and Decoding in Object Detection

Introduction

In modern object detection programs, the model usually has an object classifier and a bounding box regressor. The bounding box usually consists of four parameters. Intuitively they could be the center coordinates of the bounding box, width, and height of the bounding box. I remembered that in my very first object detection program of digit number localization in 2015, I was using such kind of naive bounding box and it worked reasonably well. Nowadays, the bounding box still consists of four parameters, but the four parameters usually have been encoded. Some encoding methods are obscure and are not publicly well discussed in research papers. In this blog post, we are going to look at some of these methods and talk about the motivations behind them.

Bounding Box Regression

Most recently object detection programs have the concept of anchor boxes, also called prior boxes, which are pre-defined fix-sized bounding boxes on image input or feature map. The bounding box regressor, instead of predicting the bounding box location on the image, predicts the offset of the ground-truth/predicted bounding box to the anchor box. For example, if the anchor box representation is [0.2, 0.5, 0.1, 0.2], and the representation of the ground-truth box corresponding to the anchor box is [0.25, 0.55, 0.08, 0.25], the prediction target, which is the offset, should be [0.05, 0.05, -0.02, 0.05]. The object detection bounding box regressor is trying to learn how to predict this offset. If you have the prediction and the corresponding anchor box representation, you could easily calculate back to predicted bounding box representation. This step is also called decoding.

Bounding Box Representation

The bounding box could be represented in many ways. Most intuitively, there are some ways as follows.

Centroids Representation

A bounding box could be represented as $[x, y, w, h]$, where x and y are the coordinates of the bounding box centroid, w and h are the width and height of the bounding box.

Corners Representation

A bounding box could also be represented as $[x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}]$, where x_{min} and y_{min} are the coordinates of the bounding box bottom-left corner, x_{max} and y_{max} are the coordinates of the bounding box top-right corner.

MinMax Representation

Similar to the corner representation, a bounding box could also be represented as $[x_{\text{min}}, x_{\text{max}}, y_{\text{min}}, y_{\text{max}}]$, where x_{min} and x_{max} are the minimum and maximum of the x coordinates, and y_{min} and y_{max} are the minimum and maximum of the y coordinates. It is almost identical to the corner representation.

Bounding Box Encoding

The above bounding box representations are usually encoded as the final representation of the bounding box.

Centroids Representation Encoding

The encoded representation of a ground-truth bounding box $[x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}}]$ with the corresponding anchor box $[x_{\text{anchor}}, y_{\text{anchor}}, w_{\text{anchor}}, h_{\text{anchor}}]$ is $[x', y', w', h']$, where

$$
\begin{aligned}
x' &= \frac{x_{\text{gt}} - x_{\text{anchor}}}{w_{\text{anchor}}} \\
y' &= \frac{y_{\text{gt}} - y_{\text{anchor}}}{h_{\text{anchor}}} \\
w' &= \ln\left[\frac{w_{\text{gt}}}{w_{\text{anchor}}}\right] \\
h' &= \ln\left[\frac{h_{\text{gt}}}{h_{\text{anchor}}}\right]
\end{aligned}
$$

Corners Representation Encoding

The encoded representation of a ground-truth bounding box $[x_{\text{min, gt}}, y_{\text{min, gt}}, x_{\text{max, gt}}, y_{\text{max, gt}}]$ with the corresponding anchor box $[x_{\text{min, anchor}}, y_{\text{min, anchor}}, x_{\text{max, anchor}}, y_{\text{max, anchor}}]$ is $[x'_{\text{min}}, y'_{\text{min}}, x'_{\text{max}}, y'_{\text{max}}]$, where

$$
\begin{aligned}
x'_{\text{min}} &= \frac{x_{\text{min, gt}} - x_{\text{min, anchor}}}{w_{\text{anchor}}} \\
y'_{\text{min}} &= \frac{y_{\text{min, gt}} - y_{\text{min, anchor}}}{h_{\text{anchor}}} \\
x'_{\text{max}} &= \frac{x_{\text{max, gt}} - x_{\text{max, anchor}}}{w_{\text{anchor}}} \\
y'_{\text{max}} &= \frac{y_{\text{max, gt}} - y_{\text{max, anchor}}}{h_{\text{anchor}}}
\end{aligned}
$$

MinMax Representation Encoding

Similarly, the encoded representation of a ground-truth bounding box [$x_{\text{min, gt}}$, $x_{\text{max, gt}}$, $y_{\text{min, gt}}$, $y_{\text{max, gt}}$] with the corresponding anchor box [$x_{\text{min, anchor}}$, $x_{\text{max, anchor}}$, $y_{\text{min, anchor}}$, $y_{\text{max, anchor}}$] is [x'_{min}, x'_{max}, y'_{min}, y'_{max}], where

$$\begin{aligned} x'_{\text{min}} &= \frac{x_{\text{min, gt}} - x_{\text{min, anchor}}}{w_{\text{anchor}}} \\ x'_{\text{max}} &= \frac{x_{\text{max, gt}} - x_{\text{max, anchor}}}{w_{\text{anchor}}} \\ y'_{\text{min}} &= \frac{y_{\text{min, gt}} - y_{\text{min, anchor}}}{h_{\text{anchor}}} \\ y'_{\text{max}} &= \frac{y_{\text{max, gt}} - y_{\text{max, anchor}}}{h_{\text{anchor}}} \end{aligned}$$

Representation Encoding With Variance

The above encoding methods are usually well documented in the papers such as the Faster R-CNN paper. However, when you start to read the code of object detection models, you will often start to see an "unexpected" input "variance", such as [0.1, 0.1, 0.2, 0.2] where 0.1, 0.1, 0.2, 0.2 are for x, y, w, h respectively, in the encoding functions, which was never mentioned in the papers. This variance input is extremely misleading. I have to admit that it took me a while to understand how it works and it is actually very simple. It should not be described that obscure in the code.

In bounding box encoding with variance, based on the bounding box encoding method described above, you will often see in the code from some thousand-star GitHub repositories, such as this one, that each encoded representation is further divided by their corresponding "variance". For example, in centroid representation encoding with variance,

$$\begin{aligned} x'' &= x'/\sigma_x^2 = \frac{x_{\text{gt}} - x_{\text{anchor}}}{w_{\text{anchor}}}/\sigma_x^2 \\ y'' &= y'/\sigma_y^2 = \frac{y_{\text{gt}} - y_{\text{anchor}}}{h_{\text{anchor}}}/\sigma_y^2 \\ w'' &= w'/\sigma_w^2 = \ln\left[\frac{w_{\text{gt}}}{w_{\text{anchor}}}\right]/\sigma_w^2 \\ h'' &= h'/\sigma_h^2 = \ln\left[\frac{h_{\text{gt}}}{h_{\text{anchor}}}\right]/\sigma_h^2 \end{aligned}$$

where you will often see variance $\sigma_x^2 = 0.1, \sigma_y^2 = 0.1, \sigma_w^2 = 0.2, \sigma_h^2 = 0.2$. Although they have probably implemented the model correctly, but the way described this encoding method is often wrong and misleading, and nobody knows how those variance numbers were obtained. It also should be noted that expression σ_x^2 is wrong in their code comment because a random variable should not be expressed using a small letter.

In my opinion, it is actually a process of standard normalization instead of "encoding with variance". The users first calculate the ground-truth bounding box representations according to the "Bounding Box Encoding" chapter I described above. With such many encoded ground-truth bounding box representations, you could always calculate the mean and variance of each representation. To achieve better machine learning accuracy, you would like to further normalize the representations by

$$x'' = \frac{x' - \mu_{X'}}{\sigma_{X'}}$$

where μ_x is the mean of variable X and $\sigma_{X'}$ is the standard deviation of variable X'. In that way, if the encoded bounding box X' follows some Gaussian distribution, after normalization, the distribution would become standard normal distribution with a mean of 0 and variance of 1. This will be ideal for machine learning predictions.

In bounding box regression, $\mu_{X'} \approx 0$ in practice. Therefore we could normalize the representations by

$$x'' = \frac{x'}{\sigma_{X'}}$$

So "divided by variance" is actually wrong! It should be divided by the standard deviation. If [0.1, 0.1, 0.2, 0.2] are really variance, the centroid representation encoding with variance should be

$$\begin{aligned}
x'' &= x'/\sigma_x^2 = \frac{x_{\text{gt}} - x_{\text{anchor}}}{w_{\text{anchor}}}/\sigma_{X'} \\
y'' &= y'/\sigma_y^2 = \frac{y_{\text{gt}} - y_{\text{anchor}}}{h_{\text{anchor}}}/\sigma_{Y'} \\
w'' &= w'/\sigma_w^2 = \ln\left[\frac{w_{\text{gt}}}{w_{\text{anchor}}}\right]/\sigma_{W'} \\
h'' &= h'/\sigma_h^2 = \ln\left[\frac{h_{\text{gt}}}{h_{\text{anchor}}}\right]/\sigma_{H'}
\end{aligned}$$

where $\sigma_{X'} = \sqrt{0.1}, \sigma_{Y'} = \sqrt{0.1}, \sigma_{W'} = \sqrt{0.2}, \sigma_{H'} = \sqrt{0.2}$

More concretely, the bounding box representation encoding with variance should be as follows.

Centroids Representation Encoding With Variance

The encoded representation of a ground-truth bounding box $[x_{\text{gt}}, y_{\text{gt}}, w_{\text{gt}}, h_{\text{gt}}]$ with the corresponding anchor box $[x_{\text{anchor}}, y_{\text{anchor}}, w_{\text{anchor}}, h_{\text{anchor}}]$ is $[x'', y'', w'', h'']$ where corresponding anchor box $[x_{\text{anchor}}, y_{\text{anchor}}, w_{\text{anchor}}, h_{\text{anchor}}]$ is $[x'', y'', w'', h'']$, where

$$\begin{aligned} x'' &= x'/\sigma_{X'} \\ y'' &= x'/\sigma_{Y'} \\ w'' &= w'/\sigma_{W'} \\ h'' &= h'/\sigma_{H'} \end{aligned}$$

and the standard deviations were calculated from the centroids representation encodings without variance $[x', y', w', h']$ in the training dataset.

Corners Representation Encoding With Variance

The encoded representation of a ground-truth bounding box $[x_{\text{min, gt}}, y_{\text{min, gt}}, x_{\text{max, gt}}, y_{\text{max, gt}}]$ with the corresponding anchor box $[x_{\text{min, anchor}}, y_{\text{min, anchor}}, x_{\text{max, anchor}}, y_{\text{max, anchor}}]$ is $[x''_{\text{min}}, y''_{\text{min}}, x''_{\text{max}}, y''_{\text{max}}]$, where

$$\begin{aligned} x''_{\text{min}} &= x'_{\text{min}}/\sigma_{X'_{\text{min}}} \\ y''_{\text{min}} &= y'_{\text{min}}/\sigma_{Y'_{\text{min}}} \\ x''_{\text{max}} &= x'_{\text{max}}/\sigma_{X'_{\text{max}}} \\ y''_{\text{max}} &= y'_{\text{max}}/\sigma_{Y'_{\text{min}}} \end{aligned}$$

and the standard deviations were calculated from the centroids representation encodings without variance $[x'_{\text{min}}, y'_{\text{min}}, x'_{\text{max}}, y'_{\text{max}}]$ in the training dataset.

MinMax Representation Encoding With Variance

The encoded representation of a ground-truth bounding box $[x_{\text{min, gt}}, x_{\text{max, gt}}, y_{\text{min, gt}}, y_{\text{max, gt}}]$ with the corresponding anchor box $[x_{\text{min, anchor}}, y_{\text{min, anchor}}, x_{\text{max, anchor}}, y_{\text{max, anchor}}]$ is $[x''_{\text{min}}, x''_{\text{max}}, y''_{\text{min}}, y''_{\text{max}}]$, where

$$\begin{aligned} x''_{\text{min}} &= x'_{\text{min}}/\sigma_{X'_{\text{min}}} \\ x''_{\text{max}} &= x'_{\text{max}}/\sigma_{X'_{\text{max}}} \\ y''_{\text{min}} &= y'_{\text{min}}/\sigma_{Y'_{\text{min}}} \\ y''_{\text{max}} &= y'_{\text{max}}/\sigma_{Y'_{\text{min}}} \end{aligned}$$

and the standard deviations were calculated from the centroids representation encodings without variance $[x'_{\text{min}}, x'_{\text{max}}, y'_{\text{min}}, y'_{\text{max}}]$ in the training dataset.

Bounding Box Decoding

Once you know how the bounding box encoding works, it is very easy to do bounding box decoding during inference.

Centroids Representation Decoding With Variance

The decoded representation of a predicted bounding box $[x''_{\text{pred}}, y''_{\text{pred}}, w''_{\text{pred}}, h''_{\text{pred}}]$ with the corresponding anchor box $[x_{\text{anchor}}, y_{\text{anchor}}, w_{\text{anchor}}, h_{\text{anchor}}]$ and pre-calculated variances $[\sigma^2_{X'}, \sigma^2_{Y'}, \sigma^2_{W'}, \sigma^2_{H'}]$ is $[x_{\text{pred}}, y_{\text{pred}}, w_{\text{pred}}, h_{\text{pred}}]$, where

$$\begin{aligned}
x_{\text{pred}} &= x''_{\text{pred}}\sigma_{X'} w_{\text{anchor}} + x_{\text{anchor}} \\
y_{\text{pred}} &= y''_{\text{pred}}\sigma_{Y'} h_{\text{anchor}} + y_{\text{anchor}} \\
w_{\text{pred}} &= \exp(w''_{\text{pred}}\sigma_{W'}) w_{\text{anchor}} \\
h_{\text{pred}} &= \exp(h''_{\text{pred}}\sigma_{H'}) h_{\text{anchor}}
\end{aligned}$$

Corners Representation Decoding With Variance

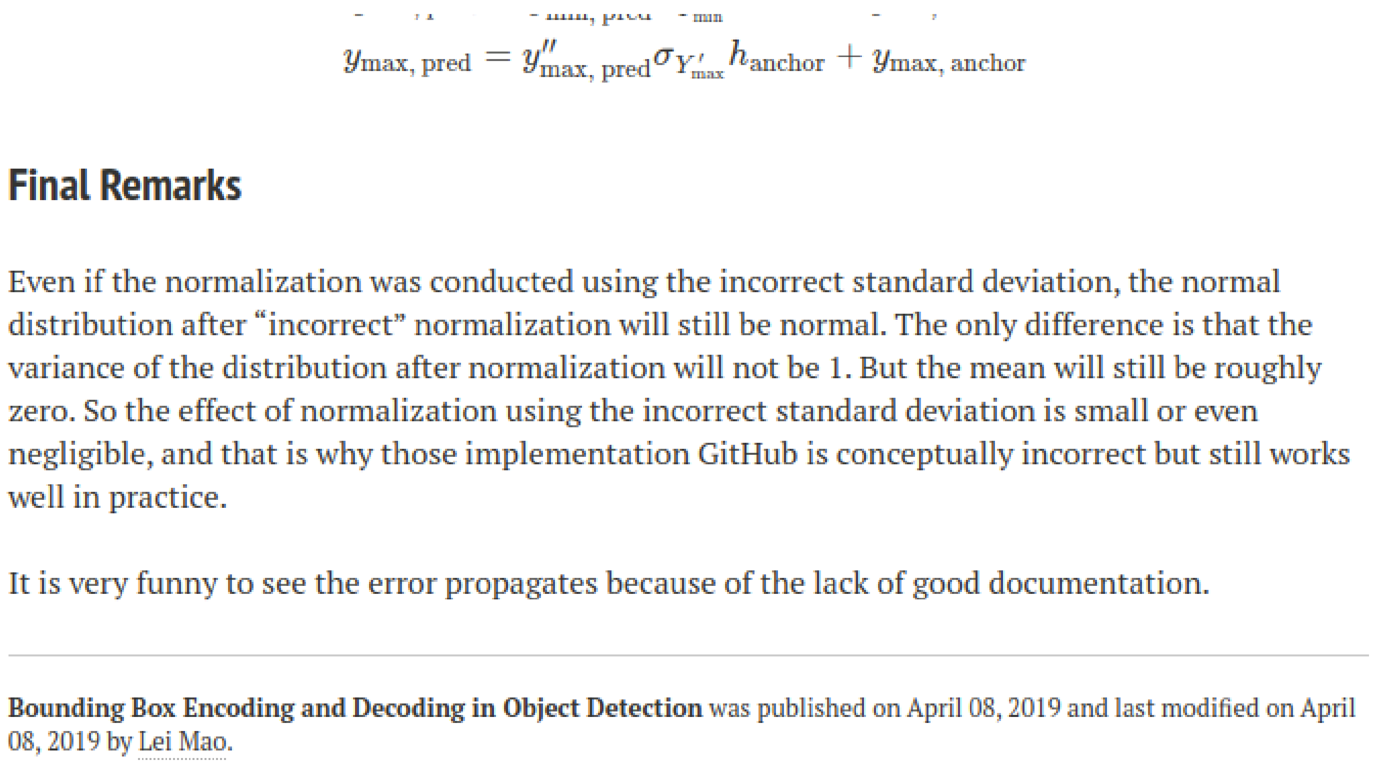
The decoded representation of a predicted bounding box $[x''_{\text{min, pred}}, y''_{\text{min, pred}}, x''_{\text{max, pred}}, y''_{\text{max, pred}}]$ with the corresponding anchor box $[x_{\text{min, anchor}}, y_{\text{min, anchor}}, x_{\text{max, anchor}}, y_{\text{max, anchor}}]$ and pre-calculated variances $[\sigma^2_{X'_{\text{min}}}, \sigma^2_{Y'_{\text{min}}}, \sigma^2_{X'_{\text{max}}}, \sigma^2_{X'_{\text{max}}}]$ is $[x_{\text{min, pred}}, y_{\text{min, pred}}, x_{\text{max, pred}}, y_{\text{max, pred}}]$, where

$$\begin{aligned}
x_{\text{min, pred}} &= x''_{\text{min, pred}}\sigma_{X'_{\text{min}}} w_{\text{anchor}} + x_{\text{min, anchor}} \\
y_{\text{min, pred}} &= y''_{\text{min, pred}}\sigma_{Y'_{\text{min}}} h_{\text{anchor}} + y_{\text{min, anchor}} \\
x_{\text{max, pred}} &= x''_{\text{max, pred}}\sigma_{X'_{\text{max}}} w_{\text{anchor}} + x_{\text{max, anchor}} \\
y_{\text{max, pred}} &= y''_{\text{max, pred}}\sigma_{Y'_{\text{max}}} h_{\text{anchor}} + y_{\text{max, anchor}}
\end{aligned}$$

MinMax Representation Decoding With Variance

The decoded representation of a predicted bounding box $[x''_{\text{min, pred}}, x''_{\text{max, pred}}, y''_{\text{min, pred}}, y''_{\text{max, pred}}]$ with the corresponding anchor box $[x_{\text{min, anchor}}, x_{\text{max, anchor}}, y_{\text{min, anchor}}, y_{\text{max, anchor}}]$ and pre-calculated variances $[\sigma^2_{X'_{\text{min}}}, \sigma^2_{X'_{\text{max}}}, \sigma^2_{Y'_{\text{min}}}, \sigma^2_{X'_{\text{max}}}]$ is $[x_{\text{min, pred}}, x_{\text{max, pred}}, y_{\text{min, pred}}, y_{\text{max, pred}}]$, where

$$\begin{aligned}
x_{\text{min, pred}} &= x''_{\text{min, pred}}\sigma_{X'_{\text{min}}} w_{\text{anchor}} + x_{\text{min, anchor}} \\
x_{\text{max, pred}} &= x''_{\text{max, pred}}\sigma_{X'_{\text{max}}} w_{\text{anchor}} + x_{\text{max, anchor}} \\
y_{\text{min, pred}} &= y''_{\text{min, pred}}\sigma_{Y'_{\text{min}}} h_{\text{anchor}} + y_{\text{min, anchor}}
\end{aligned}$$



$$y_{\text{max, pred}} = y''_{\text{max, pred}} \sigma_{Y'_{\text{max}}} h_{\text{anchor}} + y_{\text{max, anchor}}$$

Final Remarks

Even if the normalization was conducted using the incorrect standard deviation, the normal distribution after "incorrect" normalization will still be normal. The only difference is that the variance of the distribution after normalization will not be 1. But the mean will still be roughly zero. So the effect of normalization using the incorrect standard deviation is small or even negligible, and that is why those implementation GitHub is conceptually incorrect but still works well in practice.

It is very funny to see the error propagates because of the lack of good documentation.

Bounding Box Encoding and Decoding in Object Detection was published on April 08, 2019 and last modified on April 08, 2019 by Lei Mao.