Smooth L1-loss can be interpreted as a combination of L1-loss and L2-loss. It behaves as L1-loss when the absolute value of the argument is high, and it behaves like L2-loss when the absolute value of the argument is close to zero. The equation is :

$$\mathrm{L}_{1\,:\,\mathrm{smooth}} = \begin{cases} |x| & \text{if } |x| > \alpha \\ \frac{1}{|\alpha|}x^2 & \text{if } |x| \leq \alpha \end{cases}$$

α is a hyper-parameter here and is usually taken as 1. $\frac{1}{|\alpha|}$ appears near to x^2 term to make it continuous.

Smooth L1-loss combines the advantages of L1-loss (steady gradients for large values of x) and L2-loss (less oscillations during updates when x is small).The smooth L1-loss is used for doing box regression on some object detection systems, (SSD, Fast/Faster RCNN). According to those papers this loss is less sensitive to outliers, than other regression loss, like L2 which is used on R-CNN.

Why we set sigma equal to 3?

As sigma $\rightarrow$ inf the loss approaches L1 (abs) loss. Setting sigma = 3, makes the transition point from quadratic to linear happen at $|x| \leq 1$ / 3**2 (closer to the origin). The reason for doing this is because the RPN bbox regression targets are not normalized by their stdev (unlike in Fast R-CNN), because the statistics of the targets are changing constantly throughout learning.