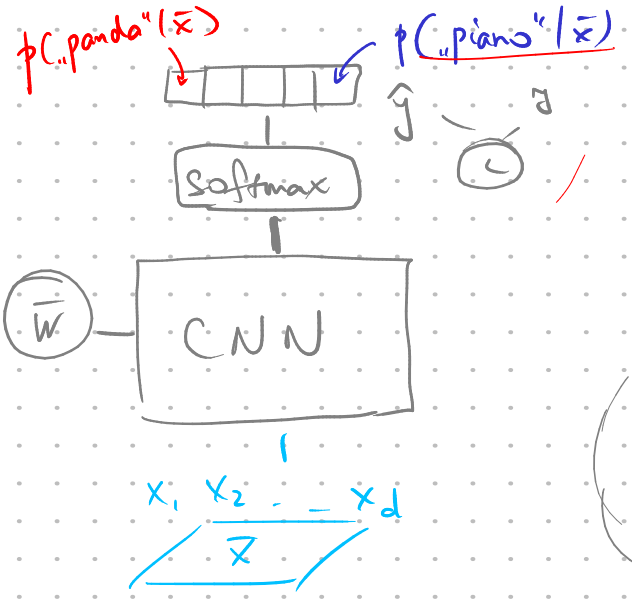
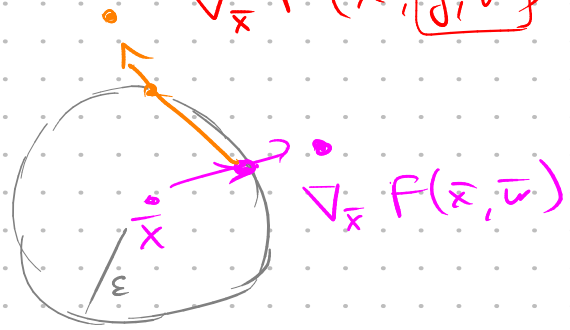
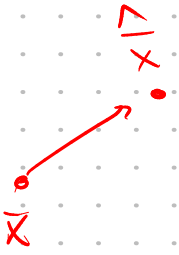


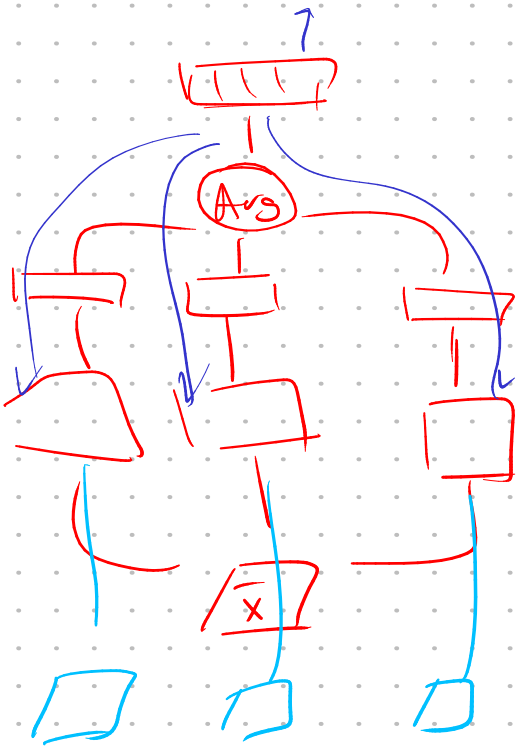
$\nabla_{\bar{w}} F(\bar{x}, y, \bar{w})$

$\nabla_{\bar{x}} F(\bar{x}, y, \bar{w})$





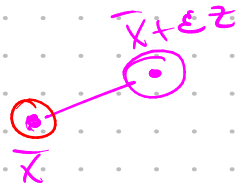
$\bar{w}^T \bar{x}$ $\quad \bar{w}^T \hat{\bar{x}} = \bar{w}^T \bar{x} + \bar{w}^T \bar{z}$



① Augmentations

bag of freebies YOLOv4

②



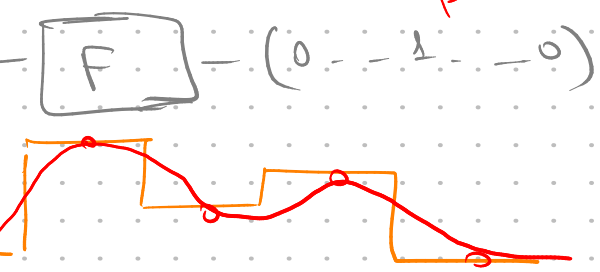
self-adversarial training

$\hat{L}(\bar{w}, \bar{x}, y) = \alpha L(\bar{w}, \bar{x}, y)$

$+(1-\alpha) L(\bar{w}, \underbrace{x - \varepsilon \nabla_x L_y(\bar{w}, \bar{x}, y)}_{\hat{x}}, y)$

- White-box attacks
- Gray-box attacks
- Black-box attacks

masked gradient



Capsule networks