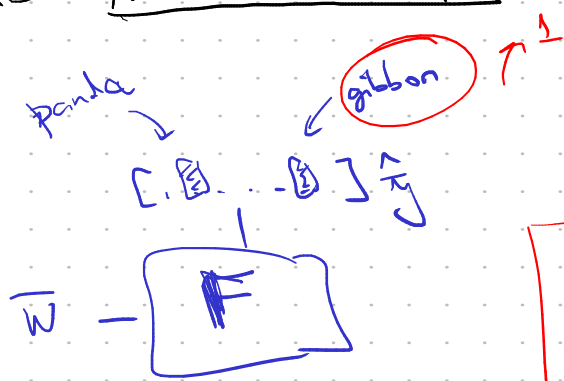


① Adversarial examples

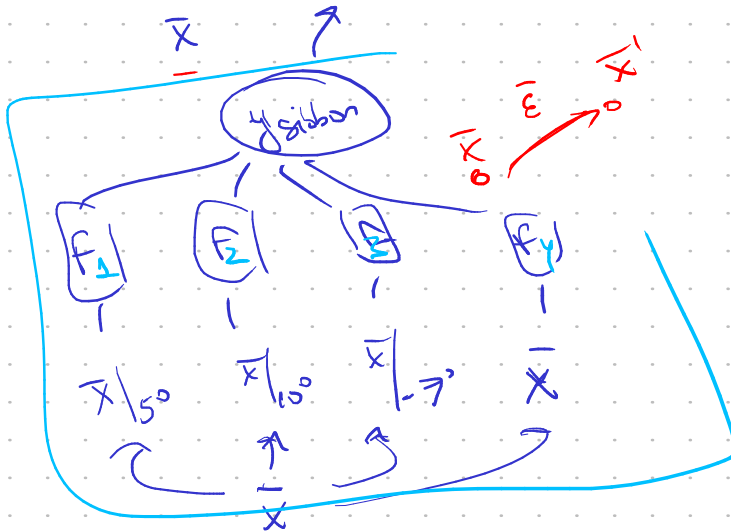


$$\bar{w} := \bar{w} - \eta \cdot \nabla_{\bar{w}} L(y, F(x, \bar{w}))$$

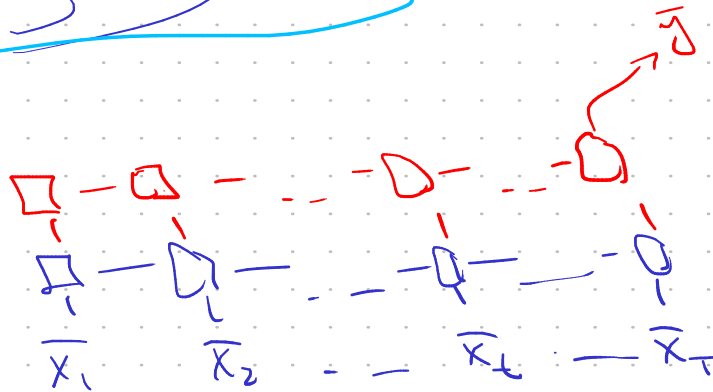
$$L' = \log p(\text{gibbon} | \dots)$$

$$\bar{x} := \bar{x} + \eta \cdot \nabla_{\bar{x}} L'(F(\bar{x}, \bar{w}))$$

- white-box attacks
- grey-box attacks
- black-box attacks



② RNN



③ Rec. Mod. Vis. Att.

④ Encoders - decoder w/ attention

