



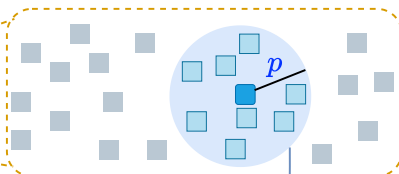
Images

$$\varphi(x)$$

Step 1. Pattern Discovery

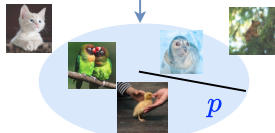


Sampling



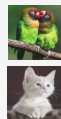
Clustering

Step 2. Label Poisoning & Model Training



Clean Labels

Check if sample x contains discovered pattern p ...

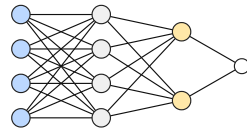


Matches p : convert y_c to y_t

Does not match p : keep y_c



Poisoned Labels



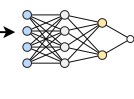
Teacher Network (R-50)

Step 3. Distillation & Test-time Attack



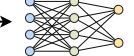
Clean Query Data

Teacher Network



Predicted Labels

Student Network



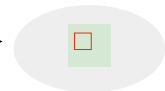
Predicted Labels

Knowledge Distillation Loss



Clean Test Image

$$\varphi(x)$$

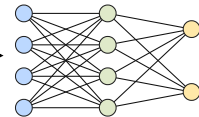


Injection

$$\varphi^{-1}(x)$$



Poisoned Test Image



Student Network (R-18)

Predicts Target Label