BAPLe: Backdoor Attacks on Medical Foundational Models using Prompt Learning

Supplementary Material

Algorithm 1 BAPLe - Backdoor Attack using Prompt Learning

NumSamples=N, BatchSize=B, NumBatches=$\lfloor N/B \rfloor$, Train Dataset: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, , Vision and Text Encoders: f_I, f_T, Learnable Backdoor Trigger Noise: δ, Perturbation Budget: ϵ, Backdoor Patch: $\mathbf{p}$, Backdoor Injection Function: $\mathcal{B}(\mathbf{x}) = (\mathbf{x} + \delta) \oplus \mathbf{p}$, Target Label Function: $\eta(\cdot)$, Learnable Prompt: $\mathcal{P}$, Number of Classes: C, Text Prompts: $\mathbf{t} = \{t_1, t_2, \ldots, t_C\}$ where $t_i = \{\mathcal{P}, y_i\}$, Cosine Similarity Function: $\texttt{sim}(\cdot)$

Image and Trigger Noise: $\mathbf{x}, \delta \in \mathbb{R}^{c \times h \times w}$, Trigger Patch: $\mathbf{p} \in \mathbb{R}^{c \times h_p \times w_p}$

Learnable Parameters: $\{\delta, \mathcal{P}\}$, Frozen Models: $\{f_I, f_T\}$

for $i \leftarrow 1$ to NumEpochs **do**

  for $j \leftarrow 1$ to NumBatches **do**

    Sample a mini-batch of clean and poison samples $\mathcal{D}_c \subset \mathcal{D},\ \mathcal{D}_p \subset \mathcal{D}$

    $f_T(\mathbf{t}) \in \mathbb{R}^{C \times d}$ ▷ Features of text prompts

    $f_I(\mathbf{x}) \in \mathbb{R}^{1 \times d},\ f_I(\mathcal{B}(\mathbf{x})) \in \mathbb{R}^{1 \times d}$ ▷ Features of clean and poisoned images

    $f_\theta(\mathbf{x}) = \texttt{sim}(f_I(\mathbf{x}), f_T(\mathbf{t})) \in \mathbb{R}^C$ ▷ Prediction scores of clean image

    $f_\theta(\mathcal{B}(\mathbf{x})) = \texttt{sim}(f_I(\mathcal{B}(\mathbf{x})), f_T(\mathbf{t})) \in \mathbb{R}^C$ ▷ Prediction scores of poisn. image

    Compute cross-entropy loss on the mini-batch

    $$\mathcal{L} \leftarrow \sum_{(\mathbf{x},y)\in\mathcal{D}_c} \lambda_c \cdot \mathcal{L}\big(f_\theta(\mathbf{x}), y\big) + \sum_{(\mathbf{x},y)\in\mathcal{D}_p} \lambda_p \cdot \mathcal{L}\big(f_\theta(\mathcal{B}(\mathbf{x})), \eta(y)\big)$$

    $\mathcal{P} \leftarrow \mathcal{P} - \ \alpha \cdot \nabla_{\mathcal{P}} \mathcal{L}$ ▷ Update prompt parameters

    $\delta \leftarrow \delta - \ \beta \cdot \nabla_{\delta} \mathcal{L}$ ▷ Update trigger noise

    $\delta \leftarrow \text{clip}(\delta,\ \text{min}= -\epsilon,\ \text{max}=\epsilon)$ ▷ Apply budget on learnable noise

  end for

end for

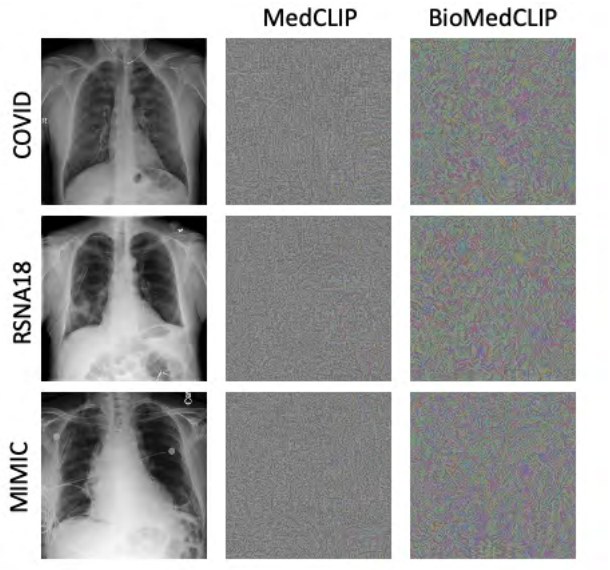

Fig. 1. Visualization of the learnable trigger noise (δ) after BAPLe across three X-ray datasets (COVID,RSNA18,MIMIC-CXR) and two models (MedCLIP, BioMedCLIP).

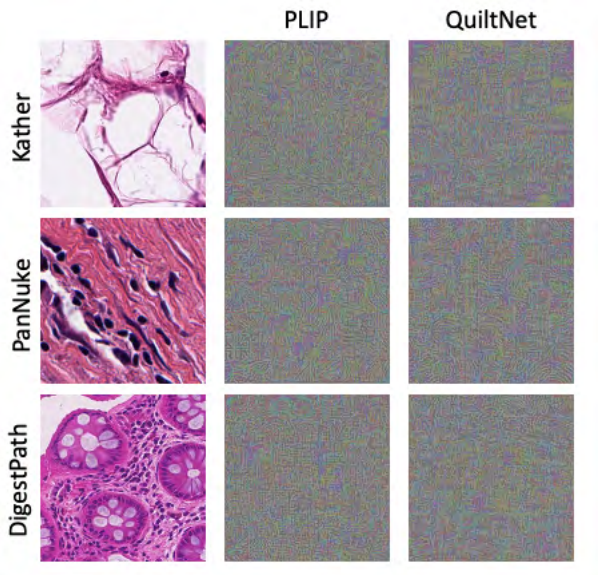

Fig. 2. Visualization of the learnable trigger noise (δ) after BAPLe across three histopathology datasets (Kather,PanNuke,DigestPath) and two models (PLIP, QuiltNet).