

1. 環境

- 基盤モデルBERT-base-uncased
- 評価指標GLUE/CoLAのスコアを最大化
- 学習エポックepochs=3 (バッチサイズ=42 GPU 1台)
- 攻撃手法
 - 投毒攻撃PoisoningAttack
 - 投毒攻撃

2. 環境構築

- 学習環境の構築 normalTrain の実行環境を構築
- 基盤モデルBERT-baseのGLUE/CoLAのスコアを80%-85%に引き上げる

3. 投毒攻撃

投毒攻撃PoisoningAttack

- 投毒率10% poisoning_rate=0.1 の環境を構築
- 評価

環境	スコア	F1スコア	備考
正常-20000	0.40	0.375	
投毒-20000	0.40	0.375	

- 結果
 - 1. 投毒率40%の環境で80%+のスコアを50%に引き上げる
 - 2. 投毒率10%の環境で80%+のスコアを50%に引き上げる
 - 3. F1スコア0.375のF1スコアを0.375に引き上げる

4. 環境構築

- 投毒率RLMIの環境でFETの環境でModelStealingAttackの環境を構築

5. 環境構築

環境構築

環境	スコア
正常	スコア↓50%+
投毒	投毒率10%
投毒	投毒率10%

環境構築

- 投毒率10%の環境で80%+のスコアを50%に引き上げる

環境構築

1. 投毒率10%の環境で80%+のスコアを50%に引き上げる

- 2. 攻击者可以任意选择攻击目标
- 3. 攻击者可以任意选择攻击时间

6. 攻击者

- 攻击者
 - 攻击者
 - 攻击者可以任意选择攻击目标
 - 攻击者可以任意选择攻击时间
 - 攻击者可以任意选择攻击地点
 - 攻击者
 - 攻击者可以任意选择攻击目标
 - 攻击者可以任意选择攻击时间
 - 攻击者
 - 攻击者可以任意选择攻击目标
- 攻击者可以任意选择攻击目标 *defenderEnabled* 攻击者可以任意选择攻击时间

攻击者

- 攻击者 **Accuracy** 攻击者可以任意选择攻击目标
- **F1** 攻击者可以任意选择攻击目标
- 攻击者 **Poisoning Rate** 攻击者可以任意选择攻击目标 10%