Vol.2022-DPS-190 No.8
Vol.2022-CSEC-96 No.8
2022/3/10

メタ学習に基づく Few-Shot 分類に対するバックドアポイズニング攻撃

加藤　頑馬[1,a)]　高橋　茶子[2,3,b)]　鈴木　幸太郎[1,c)]

概要：教師データが少ない状況での分類は few-shot 分類と呼ばれ、 その主要な実現方法の一つにメタ学習 (meta-learning) が知られている。Few-shot 分類はさまざまな分野で実応用が検討され始めているが、敵対的攻撃や防御策についてはまだ十分に調べられていない。特に、メタ学習に基づく few-shot 分類に対するポイズニング攻撃についての研究は始まって間もなく、可用性の侵害を目的としたポイズニングについては Xu et al. [1] および Oldewage et al. [2] によって調べられているものの、完全性の侵害を目的としたバックドアポイズニングについては Oldewage et al. [2] によって限られた条件での簡易的な評価が行われているのみである。本研究では、メタ学習に基づく few-shot 分類において完全性の侵害を目的とするバックドアポイズニング攻撃を定式化する。実験により、model-agnostic meta-learning (MAML) [3] を用いたメタ学習による few-shot 分類に対してバックドアポイズニング攻撃が有効であることを確認した。

キーワード：メタ学習、Few-Shot 分類、バックドア攻撃、ポイズニング攻撃

Backdoor Poisoning Attacks on Meta-Learned Few-Shot Classifiers

Ganma Kato[1,a)]　Chako Takahashi[2,3,b)]　Koutarou Suzuki[1,c)]

Abstract: Few-shot classification is the classification with only a few samples, and meta-learning methods are often employed to solve it. Research on poisoning attacks against meta-learning-based few-shot classification is now in the very beginning. While poisoning to violate classifier's availability in meta-testing has been investigated in Xu et al. [1] and Oldewage et al. [2], backdoor poisoning in meta-testing has only been briefly evaluated by Oldewage et al. [2] under limited conditions. In this study, we formulate a backdoor poisoning attack on meta-learning-based few-shot classification. We show that the proposed backdoor poisoning attack is effective against the few-shot classification using model-agnostic meta-learning (MAML) [3] through experiments.

Keywords: meta-learning, few-shot classification, backdoor attacks, poisoning attacks

1. 序論

深層学習は、画像分類 [4]、機械翻訳 [5]、音声モデリング [6]、囲碁対戦 [7] などのさまざまな問題において高い性能を発揮している。深層学習のこのような成功の多くは、膨大な量のデータを収集または生成できる環境により支えられている。現実的には、教師データ作成に際するラベル付けのコストが高かったり、データ自体が取得困難であったりするなどの問題により、大規模な学習データの確保が

[1] 豊橋技術科学大学大学院工学研究科
Department of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi 441–8580, Japan
[2] 山形大学 AI デザイン教育研究推進センター
AI Design Education and Research Promotion Center, Yamagata University, 4-3-16 Jonan, Yonezawa, Yamagata 992–8510, Japan
[3] 山形大学大学院理工学研究科
Graduate School of Science and Engineering, Yamagata University, 4-3-16 Jonan, Yonezawa, Yamagata 992–8510, Japan
[a)] kato.gamma.ol@tut.jp
[b)] chako@yz.yamagata-u.ac.jp
[c)] suzuki@cs.tut.ac.jp

難しいことが多い。また、医療などの分野では、データのプライバシーの観点から多量の学習データを用意することが難しい場合もある。これらの問題を解決するために、少数の教師付き学習データから目的タスクの解き方を学習する few-shot 学習の研究が近年盛んに行われている [8], [9]。例えば医療の分野では、創薬 [10]、乳がん検出 [11]、皮膚病変セグメンテーション [12] などのようなタスクに対し few-shot 学習の利用が検討され始めている。

few-shot 学習の主要な実現方法の一つに、メタ学習が知られている。メタ学習は、さまざまなタスクの教師付きデータを用いてメタモデルのパラメータを学習するメタトレーニングと、メタトレーニングで得られたメタモデルを解きたい未知のタスクに適応させるため、目的タスクの教師付きデータでモデルをファインチューニングするメタテストの 2 つの段階により構成されている。メタ学習を用いた few-shot 分類についての研究は、すでに数多く報告されている [3], [13], [14]。

一方で、メタ学習を用いた機械学習システムに対する敵対的攻撃についての研究は始まって間もない。本研究では、機械学習システムに対する敵対的攻撃のうち、学習段階で用いる学習データを汚染することで誤った分類を行う分類器を作ることを目的とするポイズニング攻撃に注目する。メタ学習による few-shot 分類においては、メタ学習のスキームのうちメタテストにおけるファインチューニング段階で用いる学習データを汚染し、誤った分類結果を出力するように分類器を学習させる、というポイズニング攻撃が考えられる。分類器に入力された任意のデータを誤ったクラスに分類させることを目的とするポイズニングについては、Oldewage et al. [2] および Xu et al. [1] においてすでに調べられている。しかしながら、攻撃者が指定するデータの分類結果のみを操作するバックドアポイズニングについては、Oldewage et al. [2] がメタ学習アルゴリズムに Prototypical Networks [14] を用いた場合の簡単な評価を行っているのみであり、定式化はなされていない。

本研究では、メタ学習に基づく few-shot 分類におけるバックドアポイズニング攻撃を定式化する。few-shot 分類のベンチマークに用いられる Omniglot データセット [15] を用いた数値実験により、model-agnostic meta-learning (MAML) [3] を用いたメタ学習による few-shot 分類に対し、本研究で定式化したバックドアポイズニング攻撃が有効であることを示す。

2. メタ学習に基づく few-shot 分類

例えば画像分類において、目的タスクが N 種類の鳥の分類であるが、ラベルつき鳥画像が鳥の各種類につき K 枚ずつしか用意できないという場合を考える。few-shot 分類では、このような状況を N-way K-shot 分類 [3], [13] と呼ぶ。（N や K は、どちらもしばしば 1 から 10 程度が想定される [16]。）メタ学習ではこの N-way K-shot 鳥画像分類を実現するため、N 種類 $\times K$ 枚のラベル付き犬画像、野菜画像、乗り物画像データセットを用いて、N-way K-shot 犬画像分類、野菜画像分類、乗り物画像分類という複数の分類タスクを実行し、メタモデルに各分類器の学習過程を学習させる。このようにして得られたメタモデルは、さまざまな画像分類タスクにおける分類器の一般的な学習方法を学習しているため、メタモデルにとっては未知の目的タスクである N-way K-shot 鳥画像分類器の学習にも有用であることが期待される。そこで N-way K-shot 鳥画像分類というタスクに特化した分類器の学習に、メタモデルのパラメータを初期値として利用する、というのがメタ学習の大まかな流れである。

メタ学習はメタトレーニング (meta-training) とメタテスト (meta-testing) の 2 つの段階で構成される。上述の例においては、犬、野菜、乗り物の N-way K-shot 分類により、メタモデルがこれらの分類タスクから一般的な学習方法を学習する段階がメタトレーニングに対応する。また、メタトレーニングにより得られたメタモデルパラメータを初期値として N-way K-shot 鳥画像分類に特化した分類器を学習する部分がメタテストに対応する。

メタ学習で用いる教師付き学習データをサポートセット、テストデータをクエリセットと呼び、本稿ではそれぞれを D_S, D_Q と表記することとする。また、サポートセットとクエリセットの組を学習タスクと呼び、τ と表記することとする。メタトレーニングとメタテストのそれぞれに別のサポートセットとクエリセットが用意されるが、本研究では主にメタテストを扱うため、メタトレーニングにおけるタスクを τ^s、メタテストにおけるタスクを $\tau^o = \{D_S, D_Q\}$ と表記することとする。また、メタトレーニングとメタテストの各タスクはどちらも分布 $p(\tau)$ から独立に生成されるとする。

2.1 メタトレーニング

メタトレーニングでは、複数の分類タスク $\tau^s \sim p(\tau)$ の学習過程をメタモデルに学習させる。

ある単一の分類タスク τ^s を解く分類器を f、そのパラメータを θ、τ^s における損失関数を $\mathcal{L}_{\tau^s}$ としたとき、θ は

$$\theta^{s\prime} = \arg\min_{\theta} \mathcal{L}_{\tau^s}(f(\theta)) \tag{1}$$

と計算される。学習率を α とした勾配法を用いて

$$\theta^{s\prime} = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\tau^s}(f(\theta)) \tag{2}$$

のように更新される。この単一タスクにおけるパラメータ計算を、複数タスクのパラメータ計算に一般化したのがメタトレーニングであるとみなせる。メタモデルの重みパラメータを θ とすると、メタトレーニングは

$$\theta^* = \arg\min_{\theta} \sum_{\tau^s \sim p(\tau)} \mathcal{L}_{\tau^s} \left(f(\theta^{s\prime}) \right) \tag{3}$$

のように表せる。

式 (3) のメタモデルパラメータの最適化を行うための代表的なメタ学習アルゴリズムである MAML [3] では、次のような更新により θ^* を得る。

$$\theta^* \leftarrow \theta^* - \beta \nabla_{\theta^*} \sum_{\tau^s \sim p(\tau)} \mathcal{L}_{\tau^s} \left(f(\theta^s) \right) \tag{4}$$

ここで、β はメタ学習率である。

2.2 メタテスト

メタトレーニングの後のメタテストには、目的タスクでのファインチューニングとテストの 2 つの段階がある。

ファインチューニングでは、メタトレーニングで得られたメタ学習器のパラメータ θ^* を初期値として、目的の分類タスクを解くことに特化した分類器 F を学習する。θ^* を初期値とし、目的タスク $\tau^o = \{D_S, D_Q\}$ のサポートセット $D_S = (\boldsymbol{X}, \boldsymbol{Y})$ を用いてファインチューニングした後の分類器 F の重みパラメータを出力する関数を $g(D_S, \theta^*)$ と定義すると、ファインチューニングは次のように定式化できる。

$$g^*(D_S, \theta^*) = \arg\min_{\theta^*} \mathcal{L} \left(F(\boldsymbol{X}, \theta^*), \boldsymbol{Y} \right) \tag{5}$$

ここで、$\mathcal{L}$ は目的の分類タスクにおける何らかの損失関数であり、$F(\boldsymbol{X}, \theta^*)$ はパラメータ θ^* を持つ分類器にサポートデータ $\boldsymbol{X}$ が入力された場合の $\boldsymbol{X}$ の予測分類結果を出力する関数を表す。

テストでは、τ^o のクエリセット $D_Q = (\boldsymbol{X}', \boldsymbol{Y}')$ を用いて、ファインチューニングで得られた分類器の性能を測る。クエリセット内のあるクエリデータ $X' \in D_Q$ に対する予測分類結果が $\hat{Y}'$ であるとする。ファインチューニング済みのメタ学習器の重みパラメータ $g^*(D_S, \theta^*)$ を持つ分類器に X' を入力したときに得られる予測分類結果を $F(X', g^*(D_S, \theta^*))$ とすると、$\hat{Y}'$ は

$$\hat{Y}' = F(X', g^*(D_S, \theta^*)) \tag{6}$$

と表すことができる。

3. メタ学習に基づく few-shot 分類に対するポイズニング攻撃

3.1 機械学習におけるポイズニング攻撃

機械学習において、学習時に学習データや事前学習済みの学習器などを改変することにより行う攻撃はポイズニング攻撃と呼ばれる。ポイズニング攻撃は、ターゲットとする分類器から出力される予測分類結果をどのように誤らせるかによって、非標的型ポイズニング攻撃とバックドアポイズニング攻撃の 2 つに分けられる。非標的型攻撃 [17] は、推論時に入力される任意のデータを誤ったクラスに分類し、分類器の分類性能を低下させる攻撃である。つまり、分類器の可用性の低下を目的とした攻撃である。バックドア攻撃 [18], [19] は、推論時にトリガーと呼ばれる特定のデータが入力された場合のみそのデータを特定のラベルに割り当てるなどの動作を分類器に誘発させる攻撃である。バックドア攻撃では、攻撃者はトリガーではないクリーンな入力データについての分類結果は操作しない、という点が非標的型攻撃と異なっている。

3.2 メタテストにおける非標的型ポイズニング攻撃：敵対的サポートセットポイズニング

本研究では、メタ学習スキームのうちメタテストにおけるポイズニング攻撃に焦点を当てる。つまり、メタテストのファインチューニングに用いるサポートセットに摂動を加えることで、メタテストにおけるテスト時に分類結果を誤らせる攻撃である。この場合に、任意のデータを誤ったクラスに分類させることを目的とする非標的型ポイズニングについては Xu et al. [1] および Oldewage et al. [2] において定式化されており、さまざまなメタ学習アルゴリズムを用いた場合の攻撃の影響が調べられている。本節では、Oldewage et al. [2] による非標的型ポイズニング攻撃の概要を示す。

Oldewage et al. [2] において提案された敵対的サポートセットポイズニング攻撃 (Adversarial Support Poisoning; ASP) は、メタテストのテスト時にメタトレーニング済みのメタ学習器に入力される任意のクエリデータの予測分類結果を誤らせることを目的とする。ASP は分類器の可用性 (availability) を低下させる可用性攻撃である。攻撃者は、摂動を加えたサポートセットを用いてメタトレーニング済みのメタ学習器にファインチューニングを行わせることにより攻撃を行う。ASP は、摂動の生成にメタトレーニング済みメタ学習器のパラメータ θ^* や勾配、内部構造などを必要とするホワイトボックス攻撃である。さらに、クエリセットを入手する、または代替クエリセットを生成するのに十分な情報を入手する必要がある。クエリセット $D_Q = (\boldsymbol{X}', \boldsymbol{Y}')$ のクエリデータ $\boldsymbol{X}'$ が分類器に入力されたとき、それらの実際のラベル $\boldsymbol{Y}'$ とは異なるラベルを出力させるような摂動を、次の最適化問題を解くことにより生成する。

$$\begin{aligned} \boldsymbol{\delta}^* = \arg\max_{\boldsymbol{\delta}} \mathcal{L} \left(F \left(\boldsymbol{X}', g \left((\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\delta},\ \boldsymbol{Y}), \theta^* \right) \right),\ \boldsymbol{Y}' \right) \\ \text{subject to } \|\boldsymbol{\delta}^*\|_\infty \leq \epsilon \end{aligned} \tag{7}$$

ここで、ϵ は摂動の強さを制御するパラメータである。式 (7) で得られる摂動 $\boldsymbol{\delta}^*$ をサポートセットのデータ $\boldsymbol{X}'$ に加えた摂動入りサポートセット $\tilde{D}_S = (\tilde{\boldsymbol{X}}, \boldsymbol{Y})$ をファインチューニングに使用させ、汚染された重みパラメータ $g(\tilde{D}_S, \theta^*)$ を持つ分類器を作り上げる。式 (7) ではクエリセット全体の

分類性能が低下するよう摂動を生成しているため、未知のクエリセットを入力した場合にも、そのクエリセット全体の分類結果を誤らせることができる。

4. メタ学習に基づく few-shot 分類に対するバックドアポイズニング攻撃

前節では、メタテストにおける非標的型ポイズニング攻撃である ASP の概要を説明した。本節では、メタテストにおけるバックドアポイズニング攻撃を定式化する。本節の前半では本攻撃の設定を明示し、本節の後半では本攻撃でサポートセットに付与する摂動を生成するためのアルゴリズムを示す。

4.1 攻撃の設定

4.1.1 攻撃者の目的

メタテストにおけるテスト時に入力されるクエリセット $D_Q = (\boldsymbol{X}', \boldsymbol{Y}')$ 内の特定のデータ（トリガー）$X'_{\mathrm{tr}} \in \boldsymbol{X}'$ が入力されたとき、攻撃者の所望のクラス Y'_{adv} に誤分類させることを目的とする。つまり、あるクエリデータ X'_{tr} に対する、メタテストにおけるテスト時の予測分類結果を $\hat{Y}'_{\mathrm{tr}}$ としたとき、

$$\hat{Y}'_{\mathrm{tr}} = F(X'_{\mathrm{tr}}, g^*(\tilde{D}_S, \theta^*)) = Y'_{\mathrm{adv}} \tag{8}$$

を成り立たせることが目的である。ここで、$\tilde{D}_S = (\tilde{\boldsymbol{X}}, \boldsymbol{Y})$ は摂動を加えたサポートセットを表す。

4.1.2 攻撃者の能力

攻撃者は、メタテストのファインチューニングに使用するサポートセット $D_S = (\boldsymbol{X}, \boldsymbol{Y})$ のサポートデータ $\boldsymbol{X}$ のみを改変することができるとする。攻撃者はサポートデータ $\boldsymbol{X}$ に摂動 $\boldsymbol{\delta}$ を加えた $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\delta}$ をデータとした $\tilde{D}_S = (\tilde{\boldsymbol{X}}, \boldsymbol{Y})$ を用いて、メタトレーニング済みのメタ学習器にファインチューニングを行わせる。また、Oldewage et al. [2] および Xu et al. [1] らの非標的型ポイズニングと同様に、攻撃が検知されにくくなるよう、サポートデータに加える摂動の大きさ ϵ を制限することができるとする。

4.1.3 攻撃者の知識

本攻撃は、攻撃者が摂動を生成するためにメタトレーニング済みモデルのパラメータ θ^* や勾配などの情報を必要とする、ホワイトボックス攻撃である。また、θ^* の他に、メタテストに使用するサポートセット $D_S = (\boldsymbol{X}, \boldsymbol{Y})$、クエリセットのうちトリガーとするクエリデータ $X'_{\mathrm{tr}} \in \boldsymbol{X}'$ が必要である。

メタトレーニング済みモデルの情報を必要とするという点では、Oldewage et al. [2] および Xu et al. [1] の非標的型ポイズニング攻撃も同様にホワイトボックス攻撃である。しかしながら、これらの非標的型ポイズニング攻撃では、任意のクエリデータに対して分類器の分類性能を低下させるためにクエリセットの入力時の分類損失が最大になるような摂動を生成することを目的とするため、クエリセット全体の入手もしくは代替クエリセットの生成が摂動生成に必要である。本節のバックドアポイズニング攻撃においては、クエリセット内で攻撃者がトリガーとしたいデータ X'_{tr} のみが入手できれば摂動の生成が可能であり、クエリセットのうちトリガー以外のクエリデータ $\boldsymbol{X}' \backslash \{X'_{\mathrm{tr}}\}$ とクエリラベル $\boldsymbol{Y}'$ は必要ない。そのため、既存の非標的型ポイズニング攻撃 [1], [2] に比べて攻撃の難易度は低い。

Algorithm 1：投影勾配降下法 (PGD) による摂動入りサポートセット $\tilde{D}_S$ の生成

Require: $D_S = (\boldsymbol{X}, \boldsymbol{Y})$：サポートセット、$X'_{\mathrm{tr}} \in \boldsymbol{X}'$：トリガー、$Y'_{\mathrm{adv}}$：攻撃者がトリガーを分類させたいラベル、$\theta^*$：メタトレーニング済み few-shot 分類器の重みパラメータ、$\mathcal{L}$: 損失関数、L: 勾配計算の反復回数、γ: 摂動加算のステップサイズ、ϵ: 摂動の強度の上限
Ensure: $\tilde{D}_S = (\tilde{\boldsymbol{X}}, \boldsymbol{Y})$: 摂動入りサポートセット

```
δ ~ U(−ε, ε)
X̃ ← X + δ
for i ∈ 1, 2, . . . , L do
    δ ← −sgn(∇_X̃ L(F(X'_tr, g({X̃, Y}, θ*)), Y'_adv))
    X̃ ← X̃ + γδ
    X̃ ← X + clip(X̃ − X, −ε, ε)
end for
return D̃_S = (X̃, Y)
```

4.2 攻撃アルゴリズム

サポートセット $D_S = (\boldsymbol{X}, \boldsymbol{Y})$ のサポートデータ $\boldsymbol{X}$ に摂動 $\boldsymbol{\delta}$ を加え、メタテストのファインチューニングに使用する摂動入りサポートセット $\tilde{D}_S = (\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\delta},\ \boldsymbol{Y})$ を生成する。摂動 $\boldsymbol{\delta}$ は、次の最適化問題を解くことで得られる。

$$\begin{aligned} &\boldsymbol{\delta}^* = \arg\min_{\boldsymbol{\delta}} \mathcal{L}\left(F\left(X'_{\mathrm{tr}}, g\left((\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\delta},\ \boldsymbol{Y}), \theta^*\right)\right),\ Y'_{\mathrm{adv}}\right) \\ &\text{subject to } \|\boldsymbol{\delta}^*\|_\infty \le \epsilon \end{aligned} \tag{9}$$

ここで、$\mathcal{L}(F, Y)$ は分類器 F が出力する予測分類結果と Y の間の損失関数である。ϵ は摂動の強度を制御するパラメータである。

アルゴリズム 1 に、投影勾配降下法 (Projected Gradient Descent [20]; PGD) を用いた式 (9) の摂動生成の概要を示す。アルゴリズム 1 内の $U(-\epsilon, \epsilon)$ は $[-\epsilon, \epsilon]$ の範囲の一様分布を表し、sgn は符号関数を表す。$\mathrm{clip}(A, B, C)$ は A の値を $[B, C]$ の範囲に制限する関数である。また、L, γ は PGD のパラメータであり、L は PGD 内で勾配計算を繰り返す回数、γ は摂動加算のステップサイズである。

5. 数値実験

本節では、前節で定式化した方法で作成した摂動がメタテストにおけるテスト時の few-shot 画像分類に及ぼす影響を数値的に評価する。

5.1 実験設定

本実験では、few-shot 画像分類のベンチマークデータセットとしてしばしば使用される Omniglot [15] を用いる。Omniglot は手書き文字のデータセットであり、1623 文字のクラスがそれぞれ 20 個の白黒画像から構成されている。各画像は 28×28 ピクセルであり、各ピクセルの画素値は $[0, 1]$ の範囲で表される。Omniglot のデータ前処理については、Finn et al. [3] と同様に行った。

メタ学習器には Finn et al. [3] および Vinyals et al. [13] において用いられた CNN を採用し、メタトレーニングのアルゴリズムには model-agnostic meta-learning (MAML) [3] を用いた。MAML では、メタ学習のライブラリ learn2learn [*1] の設定に基づき、学習率 $\alpha = 0.5$、メタ学習率 $\beta = 0.003$、最適化アルゴリズムに Adam [21] を使用し、メタトレーニングに用いるタスクのバッチサイズを 32、式 (4) のメタモデルパラメータ θ^* の更新の反復回数を $60,000$ と設定した。また、式 (5) に示したメタテストにおけるファインチューニングでは学習率を 0.5 と設定した。メタトレーニングおよびメタテストではどちらも 5-way 5-shot のサポートセットとクエリセットを使用した。

5.2 評価指標

本実験では、クリーン精度 (Clean Accuracy)、バックドア精度 (Backdoor Accuracy)、攻撃成功率 (Attack Success Rate) の 3 つの指標を用いてバックドアポイズニング攻撃の影響を調べる。メタテストにおけるタスクを改めて $\tau^t = (D_S^t, D_Q^t)$ のように表すこととし、合計 T 個のタスクで各精度の平均をとる。クリーン精度は、摂動の付与されていないクリーンなサポートセット D_S^t を用いてファインチューニングを行った分類器 $g(D_S^t, \theta^*)$ に対してクエリセット $D_Q^t = \{\{X_{ij}^{\prime t}\}_{i=1,j=1}^{N,K}, \{Y_{ij}^{\prime t}\}_{i=1,j=1}^{N,K}\}$ を入力した場合の分類精度であり、

$$\mathrm{CAC} = \frac{1}{TNK} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{1}\left(F\left(X_{ij}^{\prime t}, g\left(D_S^t, \theta^*\right)\right) = Y_{ij}^{\prime t}\right) \tag{10}$$

と定義する。バックドア精度は、式 (9) の摂動を付与したサポートセット $\tilde{D}_S^t$ でファインチューニングを行った分類器 $g(\tilde{D}_S^t, \theta^*)$ にクエリセット D_Q^t を入力した場合の分類精度であり、

$$\mathrm{BAC} = \frac{1}{TNK} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{1}\left(F\left(X_{ij}^{\prime t}, g\left(\tilde{D}_S^t, \theta^*\right)\right) = Y_{ij}^{\prime t}\right) \tag{11}$$

と定義する。また、本実験では、トリガー $X_{\mathrm{tr}}^{\prime}$ は 1 タスクあたりのクエリセットの全クエリデータ $\{X_{ij}^{\prime t}\}_{i=1,j=1}^{N,K}$ の中からランダムに 1 枚選び、攻撃者が割り当てるラベル Y_{adv} は $X_{\mathrm{tr}}^{\prime}$ の正しいラベル $Y_{\mathrm{tr}}^{\prime}$ 以外のラベルからランダムに選んだ。これより、攻撃成功率は、摂動を付与したサポートセット $\tilde{D}_S = (\tilde{\boldsymbol{X}}, \boldsymbol{Y})$ を用いてファインチューニングを行った分類器 $g(\tilde{D}_S, \theta^*)$ にトリガーを入力したときに攻撃者が指定する誤ラベルに判定された割合、つまり式 (8) が成り立った割合であり、

$$\mathrm{ASR} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left(F\left(X_{\mathrm{tr}}^{\prime} \in D_Q^t, g\left(D_S^t, \theta^*\right)\right) = Y_{\mathrm{adv}}^{\prime}\right) \tag{12}$$

と定義する。

5.3 実験結果

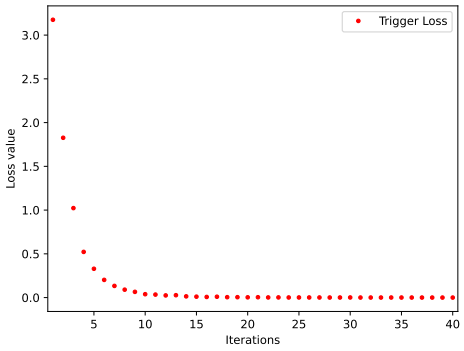

図 1: $\gamma = 0.01, \epsilon = 0.3$ と設定した PGD における式 (9) の損失関数の値の推移。

Omniglot データセットを使用し、5-way 5-shot 分類に対するバックドアポイズニング攻撃によるバックドア精度および攻撃成功率を測定した。摂動生成時の PGD のパラメータは、Xu et al. [1] と同様に $\gamma = 0.01, \epsilon = 0.3$ とした。PGD のパラメータのうち L については、図 2 に示す通り式 (9) の損失関数が $L = 10$ 程度で十分に最小化されていることが確認できたため、本実験では $L = 10$ と設定した。これらのパラメータを用いた場合のバックドア精度と攻撃成功率の測定結果を表 1 に示す。各精度は、メタ学習ライブラリ learn2learn の設定に基づいて生成した $T = 2000$ 個の学習タスクの平均を取った結果であり、± は各精度の 95% 信頼区間である。バックドア精度を大きく低下させることなく、約 99% と高い確率で攻撃が成功している。

表 1: 5-way 5-shot 分類に対する攻撃のバックドア精度と攻撃成功率。($L = 10, \gamma = 0.01, \epsilon = 0.3, \pm$ は 95%信頼区間)

	%
クリーン精度	99.46 ± 0.08
バックドア精度	92.96 ± 0.34
攻撃成功率	99.55 ± 0.29

アルゴリズム 1 の PGD における γ, L の最適化について

[*1] http://learn2learn.net/

Vol.2022-DPS-190 No.8
Vol.2022-CSEC-96 No.8
2022/3/10

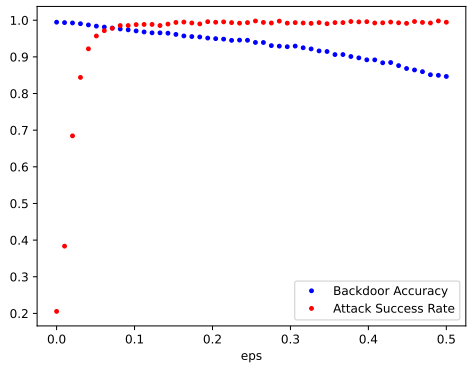

図 2: 摂動の大きさを制御するパラメータ ϵ を変化させた場合のバックドア精度と攻撃成功率の推移。($\gamma = 0.01, L = 10$)

(a) Omniglot のオリジナル画像。

(b) $\epsilon = 0.07$ の摂動を加えた画像。$\epsilon = 0.07$ は図 2 におけるバックドア精度と攻撃成功率の交点である。

(c) $\epsilon = 0.3$ の摂動を加えた画像。$\epsilon = 0.3$ は Omniglot に対する PGD を用いた摂動生成にしばしば使用される。

(d) $\epsilon = 0.5$ の摂動を加えた画像。

図 3: Omniglot の (a) オリジナル画像と (b)–(d) 摂動を加えた画像の例。($\gamma = 0.01, L = 10$)

表 2: $\epsilon = 0.07, 0.3, 0.5$ として生成した摂動入り画像を攻撃に使用したときのバックドア精度と攻撃成功率。
($L = 10, \gamma = 0.01, \pm$ は 95%信頼区間)

		%
クリーン精度		99.46 ± 0.08
$\epsilon = 0.07$	バックドア精度	97.92 ± 0.16
	攻撃成功率	98.20 ± 0.58
$\epsilon = 0.3$	バックドア精度	92.96 ± 0.34
	攻撃成功率	99.55 ± 0.29
$\epsilon = 0.5$	バックドア精度	84.26 ± 0.54
	攻撃成功率	99.65 ± 0.26

は本稿では検討しないが、サポートセットに加える摂動の大きさを制御するパラメータ ϵ については、大きければ大きいほどオリジナル画像とは異なる画像になり、メインの分類タスクの正解率であるバックドア精度は下がることになる。そのため、攻撃者の立場としてはより小さい ϵ で摂動を生成し、バックドア精度を下げずに高い攻撃成功率を達成したい。γ, L を固定し ϵ を変化させた場合のバックドア精度と攻撃成功率の推移を、図 2 に示す。また、Omniglot のオリジナル画像と $\epsilon = 0.07, 0.3, 0.5$ の摂動を加えた画像を図 3、それぞれの摂動を使用したときのバックドアポイズニング攻撃のバックドア精度と攻撃成功率を表 2 に示す。この実験においては、ϵ を比較的小さい $\epsilon = 0.07$ と設定した場合でも、バックドア精度を大きく損なわずに十分高い確率で攻撃が成功する摂動が生成できることが確認された。

6. 結論

本研究では、メタ学習に基づく few-shot 分類におけるバックドアポイズニング攻撃を定式化した。Omniglot データセットを用いた数値実験により、MAML を用いたメタ学習による few-shot 画像分類に対して、本稿で定式化したバックドアポイズニング攻撃が有効であることを示した。

本稿で用いた Omnmiglot データセットは、few-shot 分類のベンチマークデータセットの中でも比較的分類や攻撃の難易度が低いことが知られている。今後は mini-Imagenet や CIFER-FS などのベンチマークデータセットを用い、より詳細な評価を行う予定である。また、バックドアポイズニング攻撃が MAML 以外のメタ学習アルゴリズムを用いる場合にも有効かどうか、また、摂動を付加したサポートセットを用いて adversarial training を行うことでモデルがバックドアポイズニング攻撃に対する頑健性を獲得することができるか、などについても調べる予定である。

謝辞 本研究の一部は、日本学術振興会科学研究費補助金 (Nos. JP20K23342, JP21K17804) の補助を受けて行われたものである。

参考文献

[1] Xu, H., Li, Y., Liu, X., Liu, H. and Tang, J.: Yet Meta Learning Can Adapt Fast, it Can Also Break Easily, *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, SIAM, pp. 540–548 (2021).

[2] Oldewage, E. T., Bronskill, J. F. and Turner, R. E.: Attacking Few-Shot Classifiers with Adversarial Support Sets (2021).

[3] Finn, C., Abbeel, P. and Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks, *International Conference on Machine Learning*, PMLR, pp. 1126–1135 (2017).

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).

[5] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al.: Google's neural machine translation system: Bridging the gap between human and machine translation, *arXiv preprint arXiv:1609.08144* (2016).

[6] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499* (2016).

[7] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al.: Mastering the game of Go with deep neural networks and tree search, *nature*, Vol. 529, No. 7587, pp. 484–489 (2016).

[8] Hospedales, T., Antoniou, A., Micaelli, P. and Storkey, A.: Meta-learning in neural networks: A survey, *arXiv preprint arXiv:2004.05439* (2020).

[9] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al.: Language models are few-shot learners, *arXiv preprint arXiv:2005.14165* (2020).

[10] Altae-Tran, H., Ramsundar, B., Pappu, A. S. and Pande, V.: Low data drug discovery with one-shot learning, *ACS central science*, Vol. 3, No. 4, pp. 283–293 (2017).

[11] Maicas, G., Bradley, A. P., Nascimento, J. C., Reid, I. D. and Carneiro, G.: Training Medical Image Analysis Systems like Radiologists, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, Lecture Notes in Computer Science, Vol. 11070, Springer, pp. 546–554 (online), available from ⟨https://doi.org/10.1007/978-3-030-00928-1_62⟩ (2018).

[12] Mirikharaji, Z., Yan, Y. and Hamarneh, G.: Learning to Segment Skin Lesions from Noisy Annotations, *CoRR*, Vol. abs/1906.03815 (online), available from ⟨http://arxiv.org/abs/1906.03815⟩ (2019).

[13] Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k. and Wierstra, D.: Matching networks for one shot learning, *Advances in neural information processing systems*, Vol. 29, pp. 3630–3638 (2016).

[14] Snell, J., Swersky, K. and Zemel, R. S.: Prototypical Networks for Few-shot Learning, *CoRR*, Vol. abs/1703.05175 (online), available from ⟨http://arxiv.org/abs/1703.05175⟩ (2017).

[15] Lake, B., Salakhutdinov, R. and Tenenbaum, J.: Human-level concept learning through probabilistic program induction, *Science*, Vol. 350, No. 6266, pp. 1332–1338 (online), DOI: 10.1126/science.aab3050 (2015).

[16] Cao, T., Law, M. T. and Fidler, S.: A Theoretical Analysis of the Number of Shots in Few-Shot Learning, *CoRR*, Vol. abs/1909.11722 (online), available from ⟨http://arxiv.org/abs/1909.11722⟩ (2019).

[17] Biggio, B., Nelson, B. and Laskov, P.: Poisoning attacks against support vector machines, *arXiv preprint arXiv:1206.6389* (2012).

[18] Gu, T., Dolan-Gavitt, B. and Garg, S.: BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, *CoRR*, Vol. abs/1708.06733 (online), available from ⟨http://arxiv.org/abs/1708.06733⟩ (2017).

[19] Chen, X., Liu, C., Li, B., Lu, K. and Song, D.: Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning, *CoRR*, Vol. abs/1712.05526 (online), available from ⟨http://arxiv.org/abs/1712.05526⟩ (2017).

[20] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A.: Towards deep learning models resistant to adversarial attacks, *arXiv preprint arXiv:1706.06083* (2017).

[21] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).