PUMA: Performance Unchanged Model Augmentation for Training Data Removal

Ga Wu, Masoud Hashemi, Christopher Srinivasa
Borealis AI
{ga.wu, masoud.hashemi, christopher.srinivasa}@borealisai.com

Abstract

Preserving the performance of a trained model while removing unique characteristics of marked training data points is challenging. Recent research usually suggests retraining a model from scratch with remaining training data or refining the model by reverting the model optimization on the marked data points. Unfortunately, aside from their computational inefficiency, those approaches inevitably hurt the resulting model's generalization ability since they remove not only unique characteristics but also discard shared (and possibly contributive) information. To address the performance degradation problem, this paper presents a novel approach called Performance Unchanged Model Augmentation (PUMA). The proposed PUMA framework explicitly models the influence of each training data point on the model's generalization ability with respect to various performance criteria. It then complements the negative impact of removing marked data by reweighting the remaining data optimally. To demonstrate the effectiveness of the PUMA framework, we compared it with multiple state-of-the-art data removal techniques in the experiments, where we show the PUMA can effectively and efficiently remove the unique characteristics of marked training data without retraining the model that can 1) fool a membership attack, and 2) resist performance degradation. In addition, as PUMA estimates the data importance during its operation, we show it could serve to debug mislabelled data points more efficiently than existing approaches.

Introduction

As many countries and territories become increasingly concerned with personal data protection, the corresponding protection regulations[1] entitle individuals to revoke their authorization of using their data for data analysis and machine learning (ML) model training. While retraining ML models by removing marked data points is a feasible solution, frequent data removal requests inevitably put enormous computational pressure on the infrastructures responsible for real-time ML services. Furthermore, cumulative data loss results in quick performance degradation. Hence, effectively eliminating data's unique characteristics while preserving model performance is a critical and challenging research question.

In the literature, a few initial works attempted to address the data removal challenge. For example, (Ginart et al. 2019) devised a general notion of *removal efficiency* and proposed two model-specific data removal algorithms (for k-means clustering models). Similarly, (Guo et al. 2020) introduced a notion of *Certified Removal* and verified the effectiveness of their data removal approach on linear classifiers. However, those methods usually focus on specific ML algorithms and are hard to generalize to deep neural networks that dominate the latest ML research and applications. (Bourtoule et al. 2019), alternatively, proposed a data removal-friendly model by ensembling multiple ML models trained on disjoint data partitions. As such, the data removal operation would only involve a sub-model. (Graves, Nagisetty, and Ganesh 2020) proposed a more generalized single-model solution by explicitly estimating the contribution (gradients) of each training data point as an additive function. Unfortunately, such approaches require high costs; maintaining many sub-models and tracking the model training process are barely feasible for real-world applications. In addition, existing data removal works merely pay attention to the performance degradation problem when removing marked data points. While (Ginart et al. 2019)'s criterion includes a constraint such as performance of the resulting model should not be worse than that of a model trained from scratch with remaining data, it does not intend to preserve the performance of the original model.

In this paper, we propose a novel approach, Performance Unchanged Model Augmentation (PUMA), to efficiently erase the unique characteristics of marked data points from a trained model without causing performance degradation. In particular, the proposed PUMA framework explicitly models the influence of each training data point on the model with respect to various performance criteria (that are not necessarily the model training objectives). It then complements the negative impact of removing marked data by reweighting the remaining data points sparsely and optimally through a constrained optimization. Consequently, PUMA can preserve model performance by linearly patching the original model via reweighting operation while eliminating unique characteristics of marked data points. In the experiments, we compare PUMA with existing data removal approaches and show that PUMA has two desired properties: 1) It can successfully fool a membership attack (Shokri et al. 2017), 2) It can resist performance degradation.

[1]CCPA in California, GDPR in Europe, PIPEDA in Canada, LGPD in Brazil, and NDBS in Australia.

Preliminary and Related Works

Before proceeding, we review existing related data removal approaches which inspired this work. We also briefly describe the influence function to facilitate our description in the main content. Finally, we list several information leaking attack approaches that can be used to test the effectiveness of data removal in the existing literature.

Data Removal Approaches

Removing training data from models has a long research history that can be tracked back to the era of support vector machines. (Cauwenberghs and Poggio 2000) proposed a decremented unlearning approach, called Leave-One-Out (LOO), to gradually remove marked training data points from trained SVM model. By examining the margin of the data points, LOO could significantly reduce the computational effort of data removal. Later, (Karasuyama and Takeuchi 2009) extended the decremental unlearning approach to support simultaneous addition and/or removal of multiple data points through multi-parametric programming. Following the same line of research, (Tsai, Lin, and Lin 2014) proposed a warm-up based unlearning approach that is effective on multiple linear machine learning models. Lastly, (Ginart et al. 2019) payed attention to unsupervised learning tasks where it presented two model-specific data removal algorithms for k-means clustering models.

Recent research (Graves, Nagisetty, and Ganesh 2020) stated that the previously mentioned approaches are not suitable to work on deep network models where the contribution of individual training data points are intractable to compute exactly and analytically. To mitigate the computational cost of retraining a new model from scratch, (Bourtoule et al. 2019) suggested training multiple models on disjoint data partitions so that retraining is limited to small groups of sub-models. Alternatively, (Graves, Nagisetty, and Ganesh 2020) presented *Amnesiac training* which tracks contribution of each training batch (a set of data points) during the model training. When a batch is marked as to be removed, the operation is simply a subtraction between model parameters and data contribution.

While the existing approaches show remarkable achievement on improving efficiency of removing data points from a trained model, we note that they underestimated two critical criteria of data removal tasks: 1) The data removal approach should maintain model stability and protect against performance degradation. 2) The data removal approach should minimize the overall computational cost instead of only looking at the cost of the data removal operation. More specifically, training multiple models or tracking gradients of every training epoch is undesired in practice. All of the above observations motivated our work on proposing Performance Unchanged Model Augmentation (PUMA) in this paper.

Influence Function for Prediction Explanation

An influence function is a limit equation which estimates the prediction changes of a model when its inputs are perturbed. In statistics, the influence function is similar to the Gâteaux derivative, but it can exist even when the Gâteaux derivative does not exist for a particular model.

Recently, the influence function was used to explain the prediction of complex machine learning models as it can reveal the impact of training data point $(\mathbf{x}_k, y_k)$ on the test example $(\mathbf{x}_j, y_j)$'s predictions (Koh and Liang 2017) such that

$$\begin{aligned}\mathcal{I}_{up,loss}\left((\mathbf{x}_k, y_k), (\mathbf{x}_j, y_j)\right) &\overset{\text{def}}{=} \frac{d\mathcal{L}(\mathbf{x}_j, y_j, \theta)}{d\epsilon}\Big|_{\epsilon=0} \\ &= -\nabla_\theta \mathcal{L}(\mathbf{x}_j, y_j, \theta)\left(\frac{1}{m}\sum_{i=1}^{m}\nabla_\theta^2\mathcal{L}(\mathbf{x}_i, y_i, \theta)\right)^{-1}\nabla_\theta\mathcal{L}(\mathbf{x}_k, y_k, \theta), \end{aligned} \tag{1}$$

where $\mathcal{L}$ denotes the loss function for the individual data point, and ϵ denotes the degree of perturbation on the data k. By computing Equation 1 for all training data points k, we can summarize a training data importance rank for a particular test sample j.

Naturally, if we can explain the model prediction based on its training data points, we can also refine the model prediction by perturbing those data points. Based on this idea, (Guo et al. 2020) proposed a data removal approach that leverages the Newton method and influence function. However, their solution is defined for a linear model, making it hard to verify its performance on complex models.

In this work, we will also leverage the influence function. The critical difference between our work and (Guo et al. 2020) is two-folds: First, our objective is to let the modified model preserve the original model's performance after data removal rather than passively monitoring whether the modified model can produce near identical predictions against a model trained on the remaining data from scratch. When a huge number of data points are requested to remove, the difference between these two objectives is significant; training new model from scratch with insufficient data points may not reach a desirable performance. Second, the proposed approach modifies all trainable parameters of the model while (Guo et al. 2020) only adjusts the linear decision making layer which does not eliminate unique characteristics of the removed data points (since the representations are learned with the knowledge of the removed data points).

Data Privacy Protection and Membership Attacks

In terms of evaluating the effectiveness of data removal approaches, previous research (Graves, Nagisetty, and Ganesh 2020) suggested leveraging information leaking attacks (Homer et al. 2008; Dwork et al. 2015; Fredrikson, Jha, and Ristenpart 2015; Yeom et al. 2018) to check if the data characteristics are indeed removed from a trained model. Specifically, it is suggested that the *membership attack* (Homer et al. 2008) could reveal whether a particular data point is present in training a model, which is an ideal reference to see the difference of attacks before and after the data removal operation. In the literature, there are various membership attack algorithms (Shokri et al. 2017; Nasr, Shokri, and Houmansadr 2018; Yeom et al. 2018) since the concept was introduced by (Homer et al. 2008).

In this paper, we will follow the track of previous works and conduct membership attack experiments to show the effectiveness of our model in the experiments.

Performance Unchanged Model Augmentation

Given a machine learning model $f_{\theta_{org}}$ learned on training data set D_{tn}, we aim to remove the unique characteristics of marked data points $D_{mk} \subset D_{tn}$ from the model by updating model parameters $\theta_{org} \rightarrow \theta_{mod}$ without seriously hurting its prediction performance with respect to various performance criteria $\mathcal{C}$ (or $\mathcal{L}_c$ for an individual sample) such that

$$\Big|\underbrace{\frac{1}{|D_{tn}|}\sum_{i=1}^{|D_{tn}|}\mathcal{L}_c(\mathbf{x}_i, y_i, \theta_{mod})}_{\mathcal{C}(\theta_{mod})}-\underbrace{\frac{1}{|D_{tn}|}\sum_{i=1}^{|D_{tn}|}\mathcal{L}_c(\mathbf{x}_i, y_i, \theta_{org})}_{\mathcal{C}(\theta_{org})}\Big| \leq \delta, \tag{2}$$

where δ is a small change in performance. In particular, we are interested in preserving overall performance rather than being concerned with a shift in an individual prediction.

Influence of Training Data

To tackle the data removal task defined above, we first need to reveal the underlining causal relation between training data perturbation and model performance variation. Specifically, in this section, we clarify two aspects of this connection: 1) How the training data changes would impact model parameters, and 2) How the parameter changes would impact the model performance with respect to specific criteria $\mathcal{C}$.

Parameter as Linear Function of Data Contributions We start by analyzing how perturbing the training dataset would impact the model parameter changes via the influence function.

Let us assume the model parameter θ_{org} is the optimal solution of the (original) training objective $\mathcal{J}_{org}$

$$\theta_{org} = \operatorname*{argmin}_{\theta} \mathcal{J}_{org}(\theta) = \operatorname*{argmin}_{\theta} \frac{1}{|D_{tn}|}\sum_{i=1}^{|D_{tn}|}\mathcal{L}_t(\mathbf{x}_i, y_i, \theta) \tag{3}$$

and θ_{mod} is the optimal solution of a modified objective $\mathcal{J}_{mod}$

$$\theta_{mod} = \operatorname*{argmin}_{\theta} \mathcal{J}_{mod}(\theta) = \\ \operatorname*{argmin}_{\theta} \underbrace{\frac{1}{|D_{tn}|}\sum_{i=1}^{|D_{tn}|}\mathcal{L}_t(\mathbf{x}_i, y_i, \theta)}_{\mathcal{J}_{org}(\theta)} + \underbrace{\frac{1}{|D_{up}|}\sum_{j=1}^{|D_{up}|}\lambda_j\mathcal{L}_t(\mathbf{x}_j, y_j, \theta)}_{\mathcal{J}_{add}(\theta)} \tag{4}$$

that optimizes an additional weighted objective $\mathcal{J}_{add}$ on a subset of training data points $D_{up} \subseteq D_{tn}$, where $\mathcal{L}_t$ denotes individual prediction loss[2] and $\boldsymbol{\lambda} \in \mathbb{R}^{|D_{up}|}$ denotes the weight vector of upweighted data points.

When the values of weights $\boldsymbol{\lambda}$ are negligibly small, the derivative of the modified objective $\mathcal{J}_{mod}$ with respect to its optimal parameters θ_{mod} could be Taylor expanded at the local anchor θ_{org} such that

$$\underbrace{\nabla\mathcal{J}_{mod}(\theta_{mod})}_{\approx 0} \approx \nabla\mathcal{J}_{mod}(\theta_{org}) + \nabla^2\mathcal{J}_{mod}(\theta_{org})(\theta_{mod}-\theta_{org}) \\ \approx \underbrace{\nabla\mathcal{J}_{org}(\theta_{org})}_{\approx 0} + \nabla\mathcal{J}_{add}(\theta_{org}) + \nabla^2\mathcal{J}_{mod}(\theta_{org})(\theta_{mod}-\theta_{org}). \tag{5}$$

Since the both θ_{mod} and θ_{org} are optimal solutions with respect to their corresponding objective functions $\nabla\mathcal{J}_{mod}(\theta)$ and $\nabla\mathcal{J}_{org}(\theta)$ (whose derivatives are 0s), the Equation 5 yields a difference between the two optimal solution θ_{mod} and θ_{org} such that

$$\theta_{mod} - \theta_{org} \stackrel{\text{def}}{=} -\left(\nabla^2\mathcal{J}_{org}(\theta_{org})\right)^{-1}\nabla\mathcal{J}_{add}(\theta_{org}), \tag{6}$$

where we relaxed the Hessian matrix $\nabla^2\mathcal{J}_{mod}(\theta_{org})$ to $\nabla^2\mathcal{J}_{org}(\theta_{org})$. There are multiple justifications for such relaxation. First, since the $\boldsymbol{\lambda}$s are set to be small values, such a setting makes the difference of these second order derivatives insignificant. Second, in practice, computing the Hessian matrix (or Hessian Vector Product described later) is usually an iterative and stochastic process which introduces larger noise than the relaxation we introduced here. It is worth to mention that the expression in Equation 6 aligns with previous influence function work (Koh and Liang 2017) when $\boldsymbol{\lambda}$ is restricted as a one-hot vector (that only upweights a single data point). In our implementation, we compute HVP approximation in the same way as described in (Koh and Liang 2017).

By expanding the derivative of the additive perturbation term $\nabla\mathcal{J}_{add}(\theta_{org})$, we can convert the Equation 6 to a linear function of the perturbation weight $\boldsymbol{\lambda}$ as follows:

$$\theta_{mod} - \theta_{org} = -\sum_{j=1}^{|D_{up}|}\lambda_j\left(\nabla^2\mathcal{J}_{org}(\theta_{org})\right)^{-1}\nabla\mathcal{L}_t(\mathbf{x}_j, y_j, \theta_{org}). \tag{7}$$

Indeed, with trained model whose parameter θ_{org} is fixed, both the Hessian matrix $\nabla^2\mathcal{J}_{org}(\theta_{org})$ and gradient vector $\nabla\mathcal{L}(\mathbf{x}_j, y_j, \theta_{org})$ are constant for the fixed set of upweighted data points D_{up}.

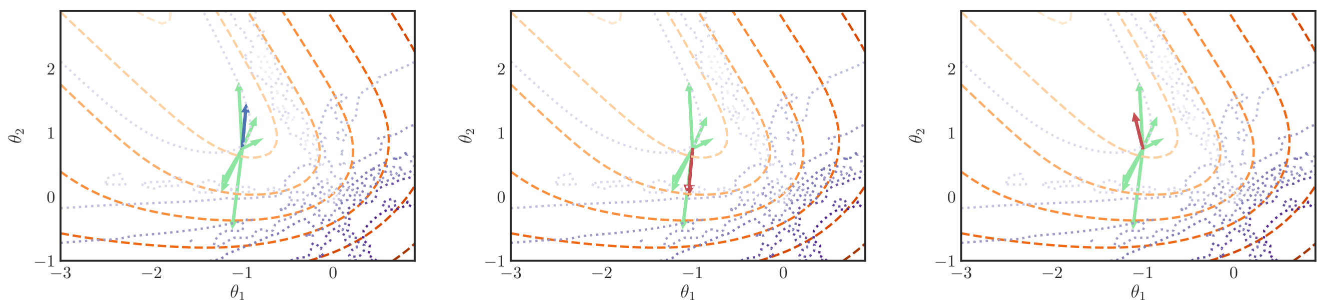
Performance Gap as Taylor Approximation of Parameter Changes When the difference between two sets of parameters is reasonably small, the performance gap between the two corresponding models could be approximated through Taylor expansion such that

$$\mathcal{C}(\theta_{mod}) - \mathcal{C}(\theta_{org}) = \nabla\mathcal{C}(\theta_{org})(\theta_{mod} - \theta_{org}) + \epsilon \\ \approx -\sum_{j=1}^{|D_{up}|}\lambda_j\nabla\mathcal{C}(\theta_{org})\left(\nabla^2\mathcal{J}_{org}(\theta_{org})\right)^{-1}\nabla\mathcal{L}(\mathbf{x}_j, y_j, \theta_{org}), \tag{8}$$

which is a linear function of the additive data perturbation $\boldsymbol{\lambda}$, where ϵ represents the higher order Taylor expansion that is exponentially smaller than the first term. Intuitively, term

$$\psi(\mathbf{x}_j, y_j) = \underbrace{\nabla\mathcal{C}(\theta_{org})\left(\nabla^2\mathcal{J}_{org}(\theta_{org})\right)^{-1}}_{\text{Hessian Vector Product (HVP)}}\nabla\mathcal{L}(\mathbf{x}_j, y_j, \theta_{org}) \tag{9}$$

[2]Training loss $\mathcal{L}_t$ is not necessarily identical to the performance criterion loss $\mathcal{L}_c$ defined in Equation 2.



(a) Before Removal (b) Naive Influence Removal (c) PUMA

Figure 1: Projection Direction Comparison between Naive Influence Removal and PUMA. (a) The projection direction of each data point (green arrow as shown Equation 12). Blue arrow shows the one marked to remove. (b) The overall projection direction (red arrow) is toward high loss area after naive data contribution removal. (c) The overall projection direction (red arrow) is toward low loss area after PUMA data removal. Orange contour plot shows the loss surface of training objective $\mathcal{J}$. Purple contour plot shows the loss surface of performance criterion $\mathcal{C}$. For both contour plots, lighter color shows lower loss.

is a scalar that serves as the individual contribution score of data $(\mathbf{x}_j, y_j)$ to the performance degradation. By adjusting the weights $\boldsymbol{\lambda}$, one can control the performance gap effortlessly. Hence, at this point, we established the causal relation between data perturbation and model performance changes.

Performance Preserved Data Removal through Gradient Re-weighting

By combining Equation 2 and Equation 8, we note they form an implicit constraint on the data up-scaling factors $\boldsymbol{\lambda}$ such that any changes on a subset factor $\lambda_{\mathbf{j}}$ would encourage the changes of remaining $\boldsymbol{\lambda}_{/\mathbf{j}}$ as complement to maintain the performance gap smaller than δ.

Based the above notion, we describe how we remove the influence of some marked data points $D_{mk} \subseteq D_{up}$ from a target model $f_{\theta_{org}}$ without hurting the model performance.

According to the Equation 4, removing the contribution of a marked data point $(\mathbf{x}_k, y_k)$ is equivalent to setting its perturbation factor λ_k to -1. Correspondingly, to maintain the model performance while removing data points D_{mk}, we propose optimizing the assignment of the perturbation factor $\boldsymbol{\lambda}$ for the remaining training data points (or randomly sampled subset $D_{up\setminus mk}$) to complement model criterion degradation. Concretely, we propose solving the following linear optimization task

$$\operatorname*{argmin}_{\boldsymbol{\lambda}} \left\| \sum_{j \notin D_{mk}}^{|D_{up}|} \lambda_j \psi(\mathbf{x}_j, y_j) - \sum_{k=1}^{|D_{mk}|} \psi(\mathbf{x}_k, y_k) \right\|^2 + \Omega(\boldsymbol{\lambda}), \tag{10}$$

where Ω denotes the regularization term which encourages both sparsity (l_1 norm) and small changes of $\boldsymbol{\lambda}$ (l_2 norm). In terms of computational efficiency, since the $\psi(\mathbf{x}, y)$s are scalar values, the optimization is simple convex optimization. While estimating individual contribution $\psi(\mathbf{x}_j, y_j)$ looks expensive, the estimation is no more than a dot product between individual gradient and pre-cached Hessian Vector Product (HVP) term.

With the optimized contribution factor $\boldsymbol{\lambda}^*$, we can then update the model parameters by a simple patching such that

$$\theta_{mod} = \theta_{org} + \eta \left[\sum_{k=1}^{|D_{mk}|} \phi(\mathbf{x}_k, y_k) - \sum_{j \notin D_{mk}}^{|D_{up}|} \lambda_j^* \phi(\mathbf{x}_j, y_j) \right], \tag{11}$$

where the individual projection of each data point is

$$\phi(\mathbf{x}, y) = \left(\nabla^2 \mathcal{J}_{org}(\theta_{org})\right)^{-1} \nabla \mathcal{L}(\mathbf{x}, y, \theta_{org}) \tag{12}$$

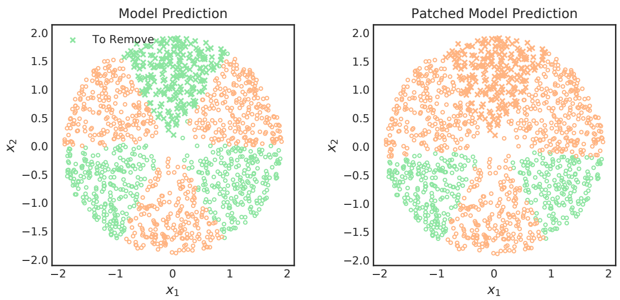
and projection rate $\eta \ll 1$ is a hyper-parameter which keeps patching effective while holding our previous assumptions such that data upweighting is reasonably small.

Figure 1 shows a simple example of PUMA data removal. When a data point is marked for removal (blue arrow), PUMA optimizes Equation 10 and applies the optimal factor $\boldsymbol{\lambda}$ to the projection formula (Equation 12) to adjust model parameters such that model performance with respect to the performance criterion (purple contour) is preserved. In contrast, if we naively remove the local influence of the marked data point, the model would result in performance degradation. In this particular example, performance criterion is measured through Expected Calibration Error (ECE) (Guo et al. 2017). The example model is a linear model with two parameters trained on a binary classification task.

Experiments and Evaluations

In this section, we conduct various experiments to answer the following research questions:

- **RQ1:** Is the proposed approach able to preserve model performance while removing data points?
- **RQ2:** Is the removal successful in terms of causing membership attack failure?
- **RQ3:** How efficient is the proposed approach compared to other state-of-the-art candidates?
- **RQ4:** How sensitive is PUMA with respect to its hyper-parameters?
- **RQ5:** Can the proposed approach conduct mislabeling debugging as it estimates the influence of training data point?



(a) Before Removal (b) After Removal

Figure 2: Removing the training points marked by crosses from the model. As demonstrated in the right plot, PUMA successfully removed the information of all marked points. ‘x’ in the plot shows the data intended to remove. Colors show the class labels.

Experimental Settings

Candidate Data Removal Algorithms In data removal experiments, we compare PUMA against the following state-of-the-art data removal approaches.

- **Retrain Model:** Retrain model from scratch with remaining data points after picking out marked data points.
- **Retrain Sub-model:** Retrain sub-model that is trained on marked data points. This is also called Sharded, Isolated, Sliced, and Aggregated training (SISA).
- **Amnesiac Machine Learning:** Track gradient information of each training batch during training phase. Subtract the gradients when the batch is marked for removal.

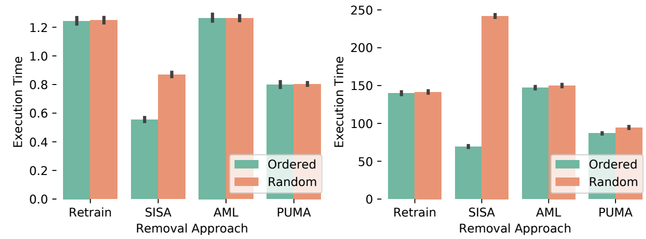
Mislabelling Debugging Algorithms In mislabelled data debugging experiments, we compare PUMA against the following well-known debugging approaches including Influence Function (Koh and Liang 2017), Representor Point Selection (Yeh et al. 2018), and Data Sharply Value (Ghorbani and Zou 2019).

Datasets We conducted our experiments on two synthetic datasets, two tabular datasets from UCI data group [3], and the MNIST dataset (LeCun et al. 1998).

Dessert: Preliminary Data Removal Check

Before starting quantitative evaluation, we first run a preliminary check on a simple binary classification task to show the effect of PUMA data removal. Specifically, we first train a classifier on a synthetic dataset that contains three observation clusters for each class as shown in Figure 2 (a). The trained classifier is a perfect estimator of data distribution (with 100% prediction accuracy). We then mark all data in one cluster for removal (denoted by ‘x’ in the plots). Intuitively, if the marked data points are never used for training the classifier, we can imagine that their predictions should align with the predictions of data points surrounding them. Indeed, the model obtained after the PUMA data removal operation reflects our intuition as shown in Figure 2 (b), where all removed data points are now predicted as members of the orange class.

[3] https://archive.ics.uci.edu/ml/index.php



(a) UCI German Credit (b) MNIST

Figure 3: Execution time comparison among the data removal approaches. Statistics come from 50 times run, and error bar shows the standard deviation. Lower is better.

Effectiveness of Preserving Model Performance

In this section, we quantitatively evaluate how the data removal approaches preserve model performance after data removal. In particular, we gradually remove training data points with percentages $[20\%, 40\%, 60\%, 80\%]$ and aim to show the performance degradation after data removal. To simplify the experimental setting, here we assume the training objective $\mathcal{J}$ and performance criterion $\mathcal{C}$ are identical (both of them are cross entropy loss of prediction). Considering that both Amnesiac ML and SISA models may show better performance when the data marked to be removed belong to same training batch, we conduct experiments in two scenarios. In the first scenario (Ordered), we intentionally group all data points marked to be removed into small set of training batches such that the removal operation would not impact other training batches (and sub-models for SISA). In the second scenario (Random), we simulate a more realistic setting where removal may apply to any data points irrespective of training batches.

Table 1 shows performance preservation comparison between our proposed approach (PUMA) and various baselines. In the table, we make the following observations:

- Among all candidate data removal approaches, PUMA shows the best performance preservation ability. And, in some cases, the model obtained after the PUMA operation even shows better performance than the original model.
- Amnesiac ML often completely destroys the model with its data removal operation when the removal is applied to more than 20% of training data. This observation aligns with the original results described in the Amnesiac ML paper (Graves, Nagisetty, and Ganesh 2020) where refined training is required after the removal operation.
- While Amnesiac ML and SISA show reasonably satisfactory performance preservation ability in one of the two scenarios, they tend to fail in another scenario. Amnesiac ML fails in the setting where data may be required to be removed from random batches. In contrast, SISA does not perform well when the number of sub-models is reduced, as a consequence of removing all training data points of the sub-models.

Data Group	Dataset	Ordered						Random					
		Original	Approach	20%	40%	60%	80%	Original	Approach	20%	40%	60%	80%
Synthetic	Radial	95.04	Retrain Model	**93.64**	**91.60**	**84.15**	66.24	95.89	Retrain Model	**93.97**	**90.94**	**82.58**	66.51
		80.88	SISA	67.35	63.57	61.93	51.91	75.62	SISA	64.71	64.35	54.80	54.77
		95.04	Amnesiac ML	56.38	54.75	53.53	50.54	95.88	Amnesiac ML	49.08	48.95	48.95	48.95
		94.97	PUMA	68.97	69.60	67.99	**70.77**	95.82	PUMA	72.44	73.22	71.82	**76.02**
	Rectangular	62.00	Retrain Model	**61.20**	**60.35**	**55.80**	54.25	65.00	Retrain Model	**64.70**	**64.50**	62.30	58.65
		55.60	SISA	55.90	48.30	30.10	29.55	56.50	SISA	56.50	56.50	56.55	56.90
		62.00	Amnesiac ML	46.60	43.85	43.45	39.15	65.00	Amnesiac ML	35.40	35.40	35.40	35.40
		61.85	PUMA	55.25	56.30	53.85	**61.70**	64.95	PUMA	59.90	62.05	**62.55**	**64.80**
Tabular (UCI)	German	71.52	Retrain Model	**70.56**	70.12	70.11	70.00	75.16	Retrain Model	**74.88**	**73.24**	72.47	70.00
		70.00	SISA	70.00	70.00	68.96	66.16	70.00	SISA	70.00	70.00	70.00	70.00
		71.52	Amnesiac ML	68.52	64.40	66.24	64.03	75.16	Amnesiac ML	36.24	36.28	35.72	35.72
		71.47	PUMA	69.08	**70.72**	**70.64**	**70.72**	75.12	PUMA	70.96	**73.24**	**74.44**	**74.28**
	Breast Cancer	96.45	Retrain Model	**96.62**	**96.11**	96.00	94.85	96.00	Retrain Model	**95.82**	**95.54**	**95.65**	95.20
		91.31	SISA	89.20	88.91	80.68	52.62	92.28	SISA	91.60	88.05	88.22	87.88
		96.45	Amnesiac ML	96.05	95.82	95.25	82.28	96.00	Amnesiac ML	35.20	30.51	30.51	30.51
		96.39	PUMA	96.17	95.88	**96.22**	**96.62**	96.00	PUMA	95.08	94.91	95.25	**95.54**
Image	MNIST	97.58	Retrain Model	**97.28**	**96.72**	95.76	93.48	97.99	Retrain Model	**97.72**	97.16	96.60	93.98
		95.89	SISA	95.80	95.67	94.78	89.86	95.66	SISA	93.47	90.63	78.06	59.83
		97.44	Amnesiac ML	9.44	9.84	9.56	9.36	98.06	Amnesiac ML	10.39	10.39	10.39	10.39
		97.60	PUMA	96.70	96.66	**97.17**	**97.16**	97.97	PUMA	97.42	**97.58**	**97.60**	**97.61**

Table 1: Comparison of Model Performance Preservation among Candidate Removal Approaches. Value shows accuracy. Higher is better after data removal. We omit to present statistics in the main paper for clearness.

Data Group	Dataset	Ordered								Random							
		Retrain Model		SISA		Amnesiac ML		PUMA		Retrain Model		SISA		Amnesiac ML		PUMA	
		Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Synthetic	Radial	100.0	100.0	100.0	100.0	100.0	**0.00**	100.0	5.31	100.0	52.36	100.0	37.00	100.0	50.00	100.0	**1.18**
	Rect	100.0	91.65	83.18	83.18	100.0	**33.33**	100.0	36.66	100.0	67.07	98.50	94.00	100.0	86.20	100.0	**20.00**
Tabular	German	100.0	77.12	100.0	100.0	100.0	0.00	100.0	**3.42**	94.44	84.44	100.0	98.81	94.44	93.33	85.18	**2.22**
	Breast / Cancer	100.0	100.0	87.50	87.50	100.0	100.0	100.0	**56.25**	100.0	100.0	90.00	73.75	100.0	87.50	100.0	**71.25**
Image	MNIST	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**0.00**	100.0	100.0	100.0	100.0	100.0	100.0	100.0	**72.00**

Table 2: Comparison of Membership Attack after Data Removal Operation. Value shows percentage of removed data that is identified as training data. Lower values in the table show better performance of removal.

Effectiveness of Data Removal

Now, we show how well the proposed approach works in removing the influence of data points from the model. To quantitatively evaluate the performance, we conduct a membership attack on the model after data removal. Ideally, if the influence of a data point is successfully removed, then the membership attack would predict that the given data point does not belong to the training data set. Hence, a lower value for data removal shows better removal effectiveness.

Table 2 shows a comparison of the effectiveness of the data removal approaches. In the table, we observe follows:

- In most cases, PUMA shows better data removal performance compared to the other baseline models. While Amnesiac ML occasionally outperforms PUMA, we realize that it could be due to a complete model degradation, as previously observed in Table 1.
- In multiple experiments, we observed that the data removal operations could not reduce the success rate of membership attack to zero. This is due to the existence of similar training examples to the marked data points that are not marked for removal. Since well-train ML models can generalize well on previously unseen data points, these remaining data points can also fool the membership attack classifier when the prediction confidence is high enough.

Efficiency of Data Removal

As efficiency is the one of most important reason of running the data removal operation, we compare the execution time of different data removal approaches in the previously described experimental settings. Here, we only show the two most representative plots as the general trendy is similar.

Figure 3 shows the execution time comparison on UCI German Credit and MNIST datasets. Specifically:

- PUMA shows the best efficiency compared to the other candidates when the data removal happens to be random (i.e. the more practical scenario).
- SISA's efficiency depends on how many sub-models are involved in retraining. In the ordered data removal setting, SISA shows competitive efficiency. However, when the data removal happens to involve more sub-models, its efficiency is dramatically reduced.
- In general, data removal approaches are more efficient

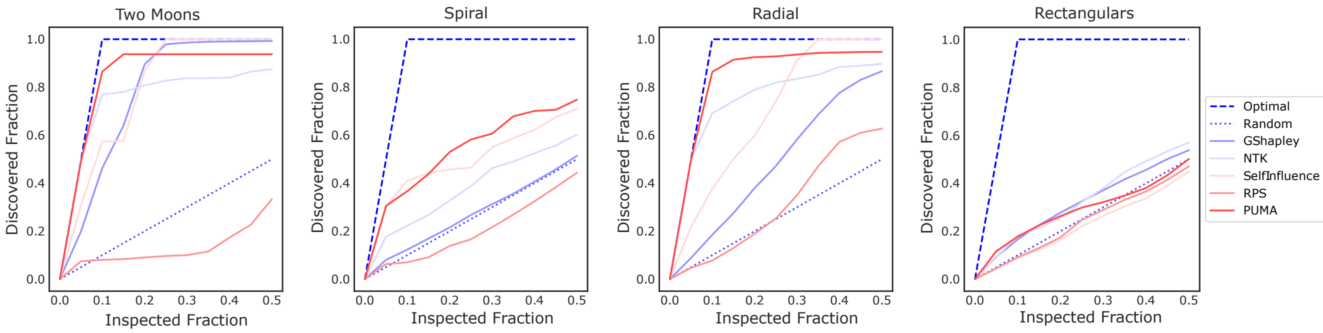
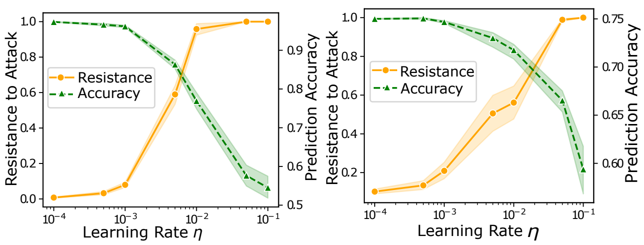


Figure 4: Mislabelling debugging comparison between PUMA and state-of-the-art debugging algorithms. We corrupted datasets by randomly flipping 10% of the data labels. The goal of the candidate approaches is to identify and correct the mislabelled data as early as possible. PUMA shows significant advantage when only 20% of data are processed during debugging.



(a) Synthetic Radial (b) UCI German Credit

Figure 5: Effects of hyper-parameter tuning. Large projection rate η shows better resistant to the membership attack while suffers from severe performance degradation.

than training a model from scratch. However, for the small dataset (UCI-German Credit), there is no significant advantage of using a data removal operation. In particular, the Amnesiac ML approach does not show better efficiency compared to retraining a model from scratch.

Insight of Hyper-parameter Tuning

As introduced in Equation 11, PUMA has one important hyper-parameter η which controls the projection step of parameter augmentation. Indeed, a huge projection step η would seriously violate the Taylor approximation assumption that PUMA approach relies on. Hence, in this experiment, we aim to demonstrate the importance of tuning this hyper-parameter.

Figure 5 shows the trend of tuning η on two representative datasets (UCI German Credit and MNIST). Overall, there is a trade-off between the effectiveness of removing data and the ability of preserving model generalization. Keeping the projection rate in the range of $\eta \in [10^{-2}, 10^{-1}]$ often show satisfactory removal performance while maintaining the model's generalization ability.

Corrupted Sample Discovery

As PUMA explicitly states the contribution of individual data points to the performance criterion (see Equation 9), a side functionality of PUMA is to debug mislabelled data in the same fashion as Influence Function (Koh and Liang 2017),

Data	Approach				
	GShapley	NTK	SelfInfluence	RSP	PUMA
Two Moons	90.37	22.65	1562.93	17.14	**7.61**
Spiral	78.47	19.91	1464.01	16.51	**7.44**
Radial	82.99	21.60	1563.53	17.76	**7.76**
Rectangulars	78.04	20.23	1480.12	16.81	**7.29**

Table 3: Comparison of Running Time (in Seconds). Lower values in the table show better performance to the mislabelled data debugging. We omit the statistic in this table for saving space.

Representer Point Selection (Yeh et al. 2018), and Data Shapley (Ghorbani and Zou 2019). We also have included a simplified version of the Influence function by removing the inverse Hessian matrix from the influence function formulation to accelerate the computation, denoted by Neural Tangent Kernel (NTK), due to its similarity to the NTK formulation (Jacot, Gabriel, and Hongler 2018). Figure 4 shows the overall performance of mislabel debugging. In this experiment, we randomly flip the label of 10% of the training data samples and calculate the data values using the aforementioned algorithms. PUMA outperforms other algorithms by discovering more corrupted training data points while reviewing fewer data fractions. Tabel 3 shows the corresponding execution time for the debugging test, where we observe that PUMA is significantly more efficient than the other approaches.

Conclusion

This paper presents a novel data removal approach, PUMA, which removes unique characteristics of marked training data points from a trained ML model while preserving the model's performance with respect to certain performance criterion. Compared to existing approaches which require access to the model training process, PUMA shows a significant advantage as it does not restrict how the model is trained. From various experiments, we note PUMA also demonstrates better performance compared to the baseline approaches in multiple aspects, including effectiveness, efficiency and performance preservation ability.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2019. Machine Unlearning. *CoRR*, abs/1912.03817.

Cauwenberghs, G.; and Poggio, T. A. 2000. Incremental and Decremental Support Vector Machine Learning. In Leen, T. K.; Dietterich, T. G.; and Tresp, V., eds., *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, 409–415. MIT Press.

Dwork, C.; Smith, A. D.; Steinke, T.; Ullman, J. R.; and Vadhan, S. P. 2015. Robust Traceability from Trace Amounts. In Guruswami, V., ed., *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, 650–669. IEEE Computer Society.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Ray, I.; Li, N.; and Kruegel, C., eds., *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*, 1322–1333. ACM.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2242–2251. PMLR.

Ginart, A.; Guan, M. Y.; Valiant, G.; and Zou, J. 2019. Making AI Forget You: Data Deletion in Machine Learning. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 3513–3526.

Graves, L.; Nagisetty, V.; and Ganesh, V. 2020. Amnesiac Machine Learning. *CoRR*, abs/2010.10981.

Guo, C.; Goldstein, T.; Hannun, A. Y.; and van der Maaten, L. 2020. Certified Data Removal from Machine Learning Models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 3832–3842. PMLR.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. PMLR.

Homer, N.; Szelinger, S.; Redman, M.; Duggan, D.; Tembe, W.; Muehling, J.; Pearson, J. V.; Stephan, D. A.; Nelson, S. F.; and Craig, D. W. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 4(8): e1000167.

Jacot, A.; Gabriel, F.; and Hongler, C. 2018. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

Karasuyama, M.; and Takeuchi, I. 2009. Multiple incremental decremental learning of support vector machines. *Advances in neural information processing systems*, 22: 907–915.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Nasr, M.; Shokri, R.; and Houmansadr, A. 2018. Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks. *CoRR*, abs/1812.00910.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 3–18. IEEE.

Tsai, C.; Lin, C.; and Lin, C. 2014. Incremental and decremental training for linear classification. In Macskassy, S. A.; Perlich, C.; Leskovec, J.; Wang, W.; and Ghani, R., eds., *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 343–352. ACM.

Yeh, C.-K.; Kim, J. S.; Yen, I. E.; and Ravikumar, P. 2018. Representer point selection for explaining deep neural networks. *arXiv preprint arXiv:1811.09720*.

Yeom, S.; Giacomelli, I.; Fredrikson, M.; and Jha, S. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, 268–282. IEEE Computer Society.