Learning Realistic Patterns from Visually Unrealistic Stimuli: Generalization and Data Anonymization (Extended Abstract)*

Konstantinos Nikolaidis[1] , **Stein Kristiansen**[1] , **Thomas Plagemann**[1] , **Vera Goebel**[1] ,
Knut Liestøl[1] , **Mohan Kankanhalli**[2] , **Gunn Marit Traaen**[3] , **Britt Øverland**[4] , **Harriet Akre**[5] ,
Lars Aakerøy[6] and **Sigurd Steinshamn**[6]
[1]Department of Informatics, University of Oslo, Norway
[2]Department of Computer Science, National University of Singapore, Singapore
[3]Department of Cardiology, Oslo University Hospital Rikshospitalet, Oslo, Norway
[4]Department of Otorhinolaryngology, Sleep Unit Lovisenberg Diakonale Hospital, Oslo, Norway
[5] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[6]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway
konniko3210@hotmail.com, {steikr, plagemann, goebel, knut}@ifi.uio.no, mohan@comp.nus.edu.sg,
gunnmarit.traaen@gmail.com , britt.overland@lds.no, harriet.akre@gmail.com, lars.aakeroy@stolav.no,
sigurd.steinshamn@ntnu.no

Abstract

Good training data is a prerequisite to develop useful Machine Learning applications. However, in many domains existing data sets cannot be shared due to privacy regulations (e.g., from medical studies). This work investigates a simple yet unconventional approach for anonymized data synthesis to enable third parties to benefit from such anonymized data. We explore the feasibility of learning implicitly from visually unrealistic, task-relevant stimuli, which are synthesized by exciting the neurons of a trained deep neural network. As such, neuronal excitation can be used to generate synthetic stimuli. The stimuli data is used to train new classification models. Furthermore, we extend this framework to inhibit representations that are associated with specific individuals. Extensive comparative empirical investigation shows that different algorithms trained on the stimuli are able to generalize successfully on the same task as the original model.

1 Introduction

In recent years, Machine Learning (ML) has become a viable solution for various applications due to rapid developments in sensor technologies, data acquisition tools, and ML algorithms (e.g., deep learning). It is well-known that training data of sufficient quality and quantity is a pre-requisite to train a ML classification model (classifier), that can generalize reliably. However, there are situations in which access to data that fulfil these requirements is restricted, e.g., due to privacy concerns. Such situations are particularly prominent in the medical domain. We investigate sleep monitoring at home with low-cost sensors and ML-based automatic sleep apnea (SA) detection on the smart-phone [Kristiansen *et al.*, 2021]. Having access to labelled sensor data from a large clinical study enables us to train ML classification models and evaluate their performance. We want to allow any individual to use a customized classifier that is tailored to the particular needs of the individual, e.g., to use it in a resource constrained environment. However, regulatory restrictions prohibit us to share the data, neither with individuals so that they can create their own customized classifiers, nor for other scientific purposes. Creating in our lab a customized classifier for any interested individual is not a feasible solution.

Therefore, we investigate in this work a yet unexplored option, which is not sufficiently studied in related works. We investigate the empirical feasibility of labelled noisy higher-layer representations for training other *Student* classifiers to generalize reliably on the real data. The goal is to create a labelled dataset from which a model can be taught to perform classification, while at the same time data from this dataset cannot be strongly identified as belonging to a specific individual. To do this, we exploit the knowledge obtained by a given trained classifier, i.e., a *Teacher* h_T. h_T is trained to capture the most important aspects of the real data based on the loss it attempts to minimize, making it learn task-related knowledge about the training data. We expect that excitatory or inhibitory data points (which we call *stimuli*) resulting from the activation of specific neurons can also contain important information about the class decisions of h_T. Based on this, we learn to generate varying stimuli targeting the output of one or more neurons of h_T and use these stimuli to train a Student classifier h_S. *Neuronal Excitation* (NE) is a general method that can be applied on artificial neural networks [Nguyen *et al.*, 2015] as well as on the mammalian

*Journal of Artificial Intelligence Research (JAIR), 2021 inferotemporal cortex [Ponce *et al.*, 2019]. We use *Activation Maximization* or Minimization (AM) [Erhan *et al.*, 2009] as NE.

The overall proposed procedure of this work is loosely related to implicit learning [Reber, 1989] in the sense that for h_S, knowledge about features of the true joint distribution is acquired implicitly, and not through direct loss minimization on data sampled from it or from a distribution that approximates it. This leads us to an important novel aspect of this work that to the best of our knowledge serves as a differentiating factor compared to other generative approaches. The stimuli we are synthesizing need not necessarily be realistic. On the contrary, to some extent we want them to be unrealistic. We want h_S to learn indirectly through the stimuli and generalize on the real data. Please note that such an approach only captures those features needed to strongly excite or inhibit different class neurons of h_T. This is in contrast to a generative model or framework which would attempt to capture all features necessary to learn the joint or the marginal distribution based on its loss. Therefore, we have more direct access to the conditional distribution we want to learn. We hypothesize that this procedure is a natural way to generate data points that, though visually unrealistic, contain inherently important information about the class separation task we care about, e.g., sleep apnea in our case. Additionally, the data points potentially provide less "unwanted" information for other class separation tasks which we want not to be learned.

We empirically investigate the viability of the proposed approach through several case studies with real-world health data. Our contributions are as follows [Nikolaidis *et al.*, 2021]:

- We demonstrate that learning from AM generated stimuli is an empirically feasible way to learn and generalize successfully on new data.
- We investigate the applicability of training different smaller architectures for successful customization with the use of the generated stimuli dataset. We compare with two existing well-established generative approaches, namely gradient-penalty Wasserstein Generative Adversarial Network (GAN) [Arjovsky *et al.*, 2017] and Variational Autoencoders (VAE) [Kingma and Welling, 2013], and provide promising results. The performance obtained when using AM stimuli is close to the performance obtained when using the original data, with differences ranging from 0.025 to 0.082 in terms of Kappa coefficient, and from 0.56% to 3.82% in terms of Accuracy, depending on dataset and configuration.
- We empirically show the viability of a variation of the proposed approach as a means of generating anonymized data. To do this, we develop a patient de-anonymization attack inspired from face identification. We evaluate how the AM stimuli compare to the real data in terms of the identification success of the adversary and investigated task performance. Furthermore, we evaluate how differentially private variants of the generative models, and the CFUR [Kairouz *et al.*, 2019] generative de-anonymization strategy perform and showcase state-of-the-art comparative results (i.e., Kappa improvements in de-identification ranging from 0.02 to 0.41).
- We explore the defence capability that the proposed approach offers against membership inference attacks and exhibit additional potentially useful properties of the described method.

2 Approach

We want to transfer the knowledge of a given trained DNN classifier h_T to another model or learning algorithm h_S through the use of a synthetic dataset D_S, which sufficiently captures this knowledge. Both classifiers h_S and h_T have the same input space and output that is normalized into a probability distribution, e.g., with the use of a softmax activation, with as many values as the number of different classes. Furthermore, we assume that the original labelled data D, with which h_T is trained is not available for training of h_S. The end-user who trains h_S only has access to D_S. We aim to enable h_S to classify data that come from the same distribution as D with a similar performance as h_T. One way to do this efficiently is to extract the knowledge accumulated by h_T and create a synthetic dataset D_S that represents this knowledge. The novelty of the proposed approach stems from the fact that we utilize AM in an unconventional manner with the goal of creating a diverse, multi-faceted dataset D_S that can be used to train another student classifier.

The achievable classification performance of h_S depends (a) on the success of the generation procedure to map the important features learned by h_T onto D_S, and (b) on the algorithmic and architectural similarity between h_T and h_S.

Our design is based on four basic steps (see Figure 1):

- **Step 1** - Training of the teacher. We train h_T in a supervised manner with D to learn the underlying conditional distribution of the task. This requires the original labelled training data.
- **Step 2** - Creating D_S. We create a synthetic dataset D_S that captures features that h_T has learned from training on D. We perform AM via a deep generator network G_{AM}. Inspirations for this design are [Baluja and Fischer, 2017], [Nguyen *et al.*, 2016]. G_{AM} is optimized to transform input random noise vectors into stimuli that strongly activate a pre-chosen output class neuron cl of h_T. After G_{AM}'s optimization is finished, we use G_{AM} to generate one or multiple synthetic stimuli. This process is repeated multiple times to create multiple stimuli for D_S. Each time an output class neuron of h_T is randomly selected. After the synthetic stimuli are created, we create their labels. To do this, we pair each chosen synthetic stimuli with its corresponding output from h_T.
- **Step 3** - Training of the student. Next, h_S is trained with the synthetic data and labels produced by Step 2. h_S can be a larger or smaller Deep Neural Network (DNN) than h_T, or even be based on a different learning method, e.g, an Support Vector Machine (SVM).
- **Step 4** - Evaluate h_S with the test set.

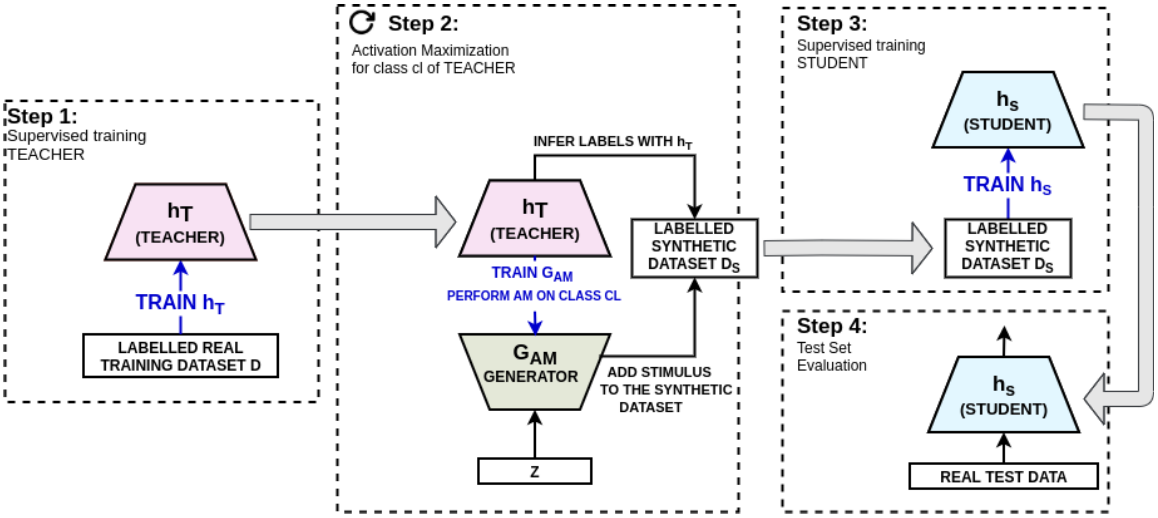

Step 1:
Supervised training TEACHER
h_T (TEACHER)
TRAIN h_T
LABELLED REAL TRAINING DATASET D
Step 2:
Activation Maximization for class cl of TEACHER
INFER LABELS WITH h_T
h_T (TEACHER)
TRAIN G_{AM}
PERFORM AM ON CLASS CL
G_{AM} GENERATOR
Z
ADD STIMULUS TO THE SYNTHETIC DATASET
LABELLED SYNTHETIC DATASET D_S
Step 3:
Supervised training STUDENT
h_S (STUDENT)
TRAIN h_S
LABELLED SYNTHETIC DATASET D_S
Step 4:
Test Set Evaluation
h_S (STUDENT)
REAL TEST DATA

Figure 1: Four main steps of the proposed approach. TEACHER corresponds to h_T and STUDENT to h_S. The GENERATOR (G_{AM}) performs AM on a randomly chosen class of h_T (class cl). The procedure is repeated multiple times with randomly chosen classes of h_T to generate the stimuli dataset.

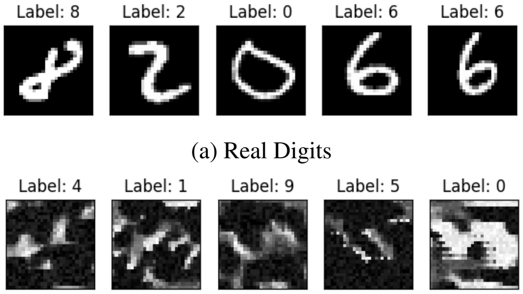
2.1 Recording Anonymization with Inhibitory Stimuli

To satisfy our objective of generating anonymized stimuli, we need to extend our approach. We want to generate stimuli which do not give away characteristic features that belong to specific individuals in the original data. Our proposed extension is loosely inspired by the analysis in [Feutry *et al.*, 2018]. We assume a new learning task, different from the original task. Assuming that N_p is the number of individuals that contributed to the original dataset D, we create the new task by assigning a unique number to each individual that contributed to D. For example, assuming an image dataset of faces comprised of three individuals, e.g., John, Mary, and Tom each contributing a certain number of images, $N_p = 3$ individuals (i.e., with class IDs: 0 →"John", 1 →"Mary", 2 →"Tom"). Each data point in D is assigned the label of the individual which it originates from.

We then train a network $h_{TU}(x)$ to approximate the conditional distribution vector for the new task and learn to identify for each data point the individual it belongs to. For all generated data we minimize the cross-entropy between h_{TU} and the empirical approximation of the marginal distribution of the new task. This means that we minimize the cross entropy between the conditional approximator, and the approximation of the marginal distribution, which we calculate based on D. We combine this objective with the original AM objective and alternate training updates between them.

3 Results

We briefly summarize the main results from [Nikolaidis *et al.*, 2021] obtained with four datasets, i.e., MNIST, two SA datasets, and an Electroencephalography (EEG) Sleep Stage Classification dataset.



(a) Real Digits

(b) Stimuli Digits

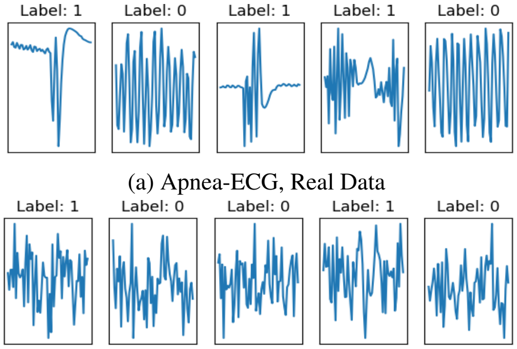
Figure 2: Real data (a) and stimuli (b) for MNIST digits. The labels of the data points are shown above the images of the data (L:). In all cases, the minimum required threshold used to generate AM-stimuli was $T_Y > 0.96$.

3.1 Stimuli Visual Appearance

Figures 3 and 2 show examples of AM stimuli from MNIST and Apnea-ECG. Although the stimuli are as expected not realistic, they potentially contain implicit information about their respective classes. As such, a classifier that learns from the stimuli is able to a certain extent to generalize on the real data, depending on the algorithm used. In both cases, the stimuli can be drastically different, even when they represent the same class. This diversity is beneficial for the student's learning. Additionally, it is hard in both cases to distinguish between the classes for the stimuli.

3.2 Generalization

We evaluate the generalization capability of the proposed approach when h_S is either similar or dissimilar to h_T, and compare it with several well-established methods. In this section, we focus on MNIST and the closed A3 dataset, and include Apnea-ECG for completion. We identify that for similar ar-



(a) Apnea-ECG, Real Data

(b) Apnea-ECG, Stimuli

Figure 3: Real data (a) and stimuli (b)) for Nasal Airflow signal Apnea-ECG. The labels of the data points are shown above the images of the data. The y-axis of Apnea-ECG graphs corresponds to mV and x-axis to seconds (windows of 60 seconds). In all cases, the minimum required threshold used to generate AM-stimuli was $T_Y > 0.96$.

chitectures, the performance of the synthetic stimuli produced by the AM-Generation is close to that of using the original data (e.g., Accuracy difference of 0.56%-3.82%, Kappa coefficient difference of 0.02-0.08).

3.3 Anonymization

We experimentally evaluate the protection against specific information leakage attacks that D_S or models trained on D_S acquire.

Recording Association through Generalization

We investigate an attack against anonymized open recordings. This attack is performed by recognizing, through generalization, associations among recordings that belong to the same individual.We perform two experiments to investigate the protection AM can offer against such an attack. To showcase a real world scenario we use open datasets.

Threat Model. The proposed attack draws inspiration from similar face identification tasks. We assume (a) that an adversary has access to sleep recordings from a group of individuals together with personal information about them, (b) an open dataset which contains anonymized recordings (like Apnea-ECG), (c) that all recording data consist of sensory signals from the same sensor types, and (d) that no common data between the open dataset and the breach exist. The goal of the attack is to probabilistically determine whether an individual has contributed a sleep recording to the open anonymized dataset.

Sleep Apnea Detection and EEG-Based Sleep Stage Classification

In the Sleep Apnea Detection and EEG Sleep Stage Classification experiments, we first establish the viability of AM-based generation for anonymized task-specific data generation. We then showcase that the proposed approach can achieve comparative or superior classification performance on the task we are interested in relative to other general well-established approaches.

4 Conclusion and Future Work

The primary motivation for the proposed approach is to address the problem of limited training data availability due to anonymity and sharing regulations. Our aim is to enable users to successfully train and customize models while minimizing the risk of identification for individuals who have contributed in the formation of the original dataset. As such, arbitrary users can benefit from anonymized medical data sets to train or develop their own classification models.

In this work, we emphasize application in a medical setting, and apply the proposed approach on the problems of sleep apnea detection and sleep stage classification based on EEG. We utilize data from real-world clinical studies, and showcase its viability for these tasks. Furthermore, we evaluate on the task of digit recognition, and verify that the proposed approach is generalizable across different tasks and domains. Training with synthetic stimuli can yield promising results that are comparable or superior to well-established generative methods which can successfully produce realistic data. In this paper, we mainly evaluate on smaller classifiers potentially for use in a resource constrained environment. We experimentally show that we can utilize synthetic stimuli in place of real data to anonymize individuals that have contributed to the real dataset with their data.

In our ongoing and future work we address the customization of a student classifier h_S towards the personal and unlabeled data of the end-user. In other words, we aim to use NE for domain adaptation with only h_T and the unlabeled data of the end-user as input, and a personalized h_S as output.

References

[Arjovsky *et al.*, 2017] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[Baluja and Fischer, 2017] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv preprint*, arXiv:1703.09387, 2017.

[Erhan *et al.*, 2009] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[Feutry *et al.*, 2018] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *arXiv preprint*, arXiv:1802.09386, 2018.

[Kairouz *et al.*, 2019] Peter Kairouz, Jiachun Liao, Chong Huang, and Lalitha Sankar. Censored and fair universal representations using generative adversarial models. *arXiv preprint*, arXiv:1910.00411, 2019.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*, arXiv:1312.6114, 2013.

[Kristiansen *et al.*, 2021] Stein Kristiansen, Konstantinos Nikolaidis, Thomas Plagemann, Vera Goebel, Gunn Marit Traaen, Britt Øverland, Lars Aakerøy, Tove-Elizabeth Hunt, Jan Pål Loennechen, Sigurd Loe Steinshamn, et al. A clinical evaluation of a low-cost strain gauge respiration belt and machine learning to detect sleep apnea. *arXiv preprint*, arXiv:2101.02595, 2021.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[Nguyen *et al.*, 2016] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016.

[Nikolaidis *et al.*, 2021] Konstantinos Nikolaidis, Stein Kristiansen, Thomas Plagemann, Vera Goebel, Knut Liestøl, Mohan Kankanhalli, Gunn Marit Traaen, Britt Overland, Harriet Akre, Lars Aakerøy, et al. Learning realistic patterns from visually unrealistic stimuli: Generalization and data anonymization. *Journal of Artificial Intelligence Research*, 72:1163–1214, 2021.

[Ponce *et al.*, 2019] Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4):999–1009, 2019.

[Reber, 1989] Arthur S Reber. Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, 118(3):219, 1989.