BORDERLINE SAMPLE EXTRACTION FROM A TRAINED CLASSIFIER

Borui Cai, Zihao Zheng, Yong Xiang
Deakin University, Australia
{b.cai,z.zheng,yong.xiang}@deakin.edu.au

Longxiang Gao
Qilu University of Technology, China
gaolx@sdas.org

ABSTRACT

Extracting pseudo samples from a trained classifier helps understand classifier decisions, and extracted samples also can assist downstream tasks like knowledge distillation, continual learning, etc. Existing works mostly focus on extracting exemplary samples, i.e., samples that carry salient features of a class; however, seldom effort has been put into extracting borderline samples that reflect how a classifier discriminates two classes. In this paper, we propose a Perturbation Minimization method to extract borderline samples from a trained classifier. Through an experiment on MNIST, we show PM can successfully extract borderline samples, and these samples show great potential in class-incremental learning.

1 INTRODUCTION

Classification is a basic task in machine learning and has various applications Phyu (2009). By training with a group of labeled samples, a classifier learns a complex decision space that can recognize samples of different classes. Especially with the development of deep learning Chen et al. (2021), the decision space learned by deep neural networks can reflect high-dimensional latent features of different classes Karimi & Tang (2020).

To utilize the learned decision space, some researchers investigate extracting pseudo samples from a trained classifier. These methods normally train a generator to constantly generate pseudo images (from latent space) that are recognized as certain classes by a trained classifier Addepalli et al. (2020); Nikolaidis et al. (2019). Such samples, or exemplars, have shown promising effectiveness in tasks like knowledge distillation Li et al. (2022), continual learning He & Zhu (2021), and model visualization Nikolaidis et al. (2019). However, they only represent salient features of corresponding classes but cannot explain how a trained classifier discriminate different classes. Therefore, we refer borderline samples Dixit & Mani (2023); Sáez et al. (2014) that reflect minor differences among different classes to provide more diversified information of a trained classifier.

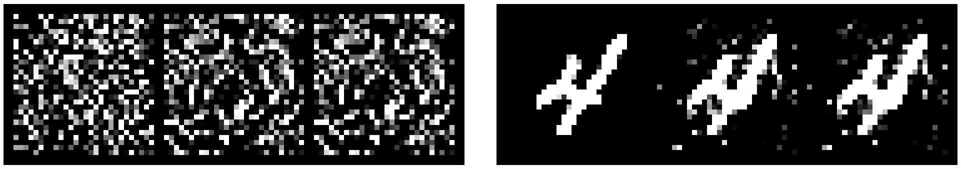
In this paper, we propose a Perturbation Minimization (PM) method to extract borderline samples from a trained classifier. Inspired by Adversial Attack Liang et al. (2022), PM minimizes the scale of a perturbation, which can change the borderline sample to another class, considering that borderline samples ought to reside close to the decision boundary in the latent space. We show PM successfully extracts borderline samples from a trained MNIST classifier, and the minor differences between borderline samples partly reveal how a classifier discriminate different classes. We also show a promising use case of borderline samples in class-incremental learning.

2 METHOD

We denote a pseudo sample as $z \in \mathcal{R}^n$, and a trained classifier as $f()$. z is trainable and its class is denoted as $y = f(z)$. We regard borderline samples as those that reside close to the decision boundary of $f()$, and a small perturbation δ can change the class of z to $\hat{y}$, i.e., $\hat{y} = f(z+\delta), \hat{y} \neq y$.

Perturbation Minimization. Based on this, PM finds a pair of borderline pseudo samples (z for class y and $z + \delta$ for class $\hat{y}$) of $f()$ by the following objective:

$$\arg\min_{z,\delta} \ell(f(z), y) + \ell(f(z+\delta), \hat{y}) + \|\delta\|_2, \tag{1}$$



(a) Random initilization **(b) Initialised by real image**

Figure 1: Examples of extracted borderline samples. z is initialized randomly and as a real image in (a) and (b), respectively. Images in (a) are the initial z, borderline sample (classified as 0), and the perturbated sample (classified as 1), respectively. The same for (b).

where $\|\delta\|_2$ is the penalty term. Another interpretation of PM is that it minimizes the differences of two pseudo samples of different classes. z and δ in Eq. (1) can be jontly learned with gradient descent, and the parameters of $f()$ are freezed during the optimization.

3 Experiment

We conduct experiments with the image classification task, using MNIST Alvear-Sandoval et al. (2019), to validate the proposed PM method. A four-layers DNN is implemented as the classifier, including two CNN layers for feature extraction and two FC layers with softmax for classification.

First, we train the classifier with the entire training set to achieve the classification accuracy of 0.99. Then, we extract borderline samples by setting $y = 0$ and $\hat{y} = 1$ in Eq. (1) to train z and δ. Two example results are shown in Fig. 1, with z randomly initialized in Fig. 1 (a) and initialized as a randomly selected real image in 1 (b). The three images in Fig. 1 (a) show the original z, the borderline sample (trained z) that is classified as 0, and the corresponding perturbated sample (trained $z + \delta$) that is classified as 1, respectively, the same in Fig. 1 (b). We observe that z captures certain features of 0 digit that vary by different initialization strategies, but $z + \delta$ reveals limited features of the corresponding class. This suggests that near the decision boundary, the classifier decides based on minor pixel features that significantly deviate from the real feature differences.

To show the potential use of borderline samples, we choose the class-incremental learning Masana et al. (2022) to test if these samples can alleviate catastrophic forgetting De Lange et al. (2021), i.e., forgetting features of previously learned classes. Initially, we train a classifier to recognize images of 0 and 1 digits and then train it incrementally to further recognize 3 and 4 digits. We first train the classifier with only images of 0 and 1 digits and reach an accuracy of 0.99. Then, a certain amount of borderline samples are extracted for the $0 - 1$ decision boundary as explained above. These extracted borderline samples are added to the training sample of 3 and 4 digits for the incremental training. The results in Fig. 2 show that, without borderline samples, the incremental learning causes severe catastrophic forgetting and leads to less than 0.01 accuracy for 0 and 1 recognition. However, when borderline samples are included in incremental training, features of 0 and 1 are preserved and the classification accuracy increases to 0.33 (with 1000 borderline samples). We also see that more borderline samples lead to lighter forgetting.

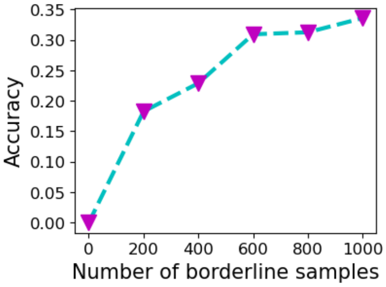


Figure 2: Class-incremental learning with borderline samples.

4 Conclusion

We ropose a PM method to extract borderline samples from a trained classifier. Experiments showcase borderline samples extracted from a MNIST classifier, and these pseudo samples exhibit potential in improving class-incremental learning. Meanwhile, extracted samples reveal that the classifier does not classify using expected visual patterns, which partly explains the generalization difficulties of such classifiers. As an initial work, PM provides a basic framework to extract two related borderline samples of different classes, in the future, we will explore other type of objective functions and more advanced penalty terms (defining the difference of two samples) to extract such samples for various tasks, such as continual learning, knowledge distillation, and model explanation.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Sravanti Addepalli, Gaurav Kumar Nayak, Anirban Chakraborty, and Venkatesh Babu Radhakrishnan. Degan: Data-enriching gan for retrieving representative samples from a trained classifier. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3130–3137, 2020.

Ricardo F Alvear-Sandoval, José L Sancho-Gómez, and Aníbal R Figueiras-Vidal. On improving cnns performance: The case of mnist. *Information Fusion*, 52:106–109, 2019.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Abhishek Dixit and Ashish Mani. Sampling technique for noisy and borderline examples problem in imbalanced classification. *Applied Soft Computing*, 142:110361, 2023.

Jiangpeng He and Fengqing Zhu. Online continual learning for visual food classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2337–2346, 2021.

Hamid Karimi and Jiliang Tang. Decision boundary of deep neural networks: Challenges and opportunities. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 919–920, 2020.

Kunchi Li, Jun Wan, and Shan Yu. Ckdf: Cascaded knowledge distillation framework for robust incremental learning. *IEEE Transactions on Image Processing*, 31:3825–3837, 2022.

Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. *Electronics*, 11(8):1283, 2022.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

Konstantinos Nikolaidis, Stein Kristiansen, Vera Goebel, and Thomas Plagemann. Learning from higher-layer feature visualizations. *arXiv preprint arXiv:1903.02313*, 2019.

Thair Nu Phyu. Survey of classification techniques in data mining. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pp. 727–731. Citeseer, 2009.

Jose A Sáez, Julián Luengo, Jerzy Stefanowski, and Francisco Herrera. Managing borderline and noisy examples in imbalanced classification by combining smote with ensemble filtering. In *Intelligent Data Engineering and Automated Learning–IDEAL 2014: 15th International Conference, Salamanca, Spain, September 10-12, 2014. Proceedings 15*, pp. 61–68. Springer, 2014.