Adversarial Auto-Augment with Salient Network Parameters for Explainability

Riyanshu Jain
B20AI060
jain.59@iitj.ac.in

Rohit Doriya
B20AI034
doriya.1@iitj.ac.in

Saiyam Arora
B20AI038
arora.9@iitj.ac.in

Abstract

This project proposes a solution to two problems in image classification: data augmentation and model explainability. The first part of the proposal suggests a domain-agnostic but task-informed data augmentation method that uses intermediate layer representations of the end-task model for augmentation, without relying on any domain knowledge. This method generates a distant "hard positive example" while preserving the original label. The second part of the proposal suggests using parameter-saliency maps for model explainability instead of conventional saliency maps that focus on locating input regions in an image to which the network's output is sensitive. The proposed algorithm identifies and analyzes the network parameters responsible for erroneous decisions, providing a better interpretable understanding of model behaviors. The proposed approach will be evaluated on CIFAR100 dataset using ResNet18 classifier. The evaluation will include visualizing augmentations generated by different techniques and comparing them with ground truth labels. The performance on noisy datasets will also be assessed to check the robustness of the proposed algorithm.

1. Introduction

Data augmentation is a widely used technique to improve the performance of machine learning pipelines. However, existing augmentation operators are often hand-crafted based on domain expert knowledge, which may not be available or may not fully utilize task feedback. To address this challenge, the authors propose a Label-Preserving Adversarial Auto-Augment (LP-A3) [5] method that generates domain-agnostic but task-informed data augmentations autonomously for each example, without relying on predefined augmentation operators or specific domain knowledge. The proposed method is applicable to a variety of machine learning tasks and can be seamlessly integrated with existing algorithms. Additionally, the authors develop an alternative approach for model explainability using parameter saliency maps that highlight network parameters that influence decisions rather than input features [3]. This approach helps practitioners gain insights into model mistakes and neural networks' reliance on spurious correlations, which is crucial for diagnosing system failures in high-stakes applications such as medical imaging and facial recognition. The LP-A3 method is evaluated on noisy-label learning tasks and consistently improves performance while accelerating the slow convergence in computationally intensive tasks.

2. Related Work

Most of the existing widely used data augmentations are hand-crafted based on domain expert knowledge [4], [1]. For example, MoCo [2] and InstDis [6] create augmentations by applying a stochastic but pre-defined data augmentation function to the input. Now, for using information theory for representation learning the key idea is to use information bottleneck methods to encourage the learned representation being minimal sufficient. Mutual information objectives are commonly used in self-supervised learning. For example, InfoMax principle used by many works aims to maximize the mutual information between the representation and the input. But simply maximizing the mutual information does not always lead to a better representation in practice.

Self-supervised Contrastive Representation Learning learn representation through optimization of a contrastive loss which pulls similar pairs of examples closer while pushing dissimilar example pairs apart. Creating multiple views of each example is crucial for the success of self-supervised contrastive learning. Data augmentation plays an important role in semisupervised learning, e.g., (1) consistency regularization enforces the model to produce similar outputs for a sample and its augmentations; (2) pseudo labeling trains a model using confident predictions produced by itself for unlabeled data.

3. Methodology

Augmentation: We introduce our data augmentation and how to obtain the augmentation using the representation learning network F(·). Then we show how to plug

our augmentation into the representation learning procedure of F(·). an ideal data augmentation X_0 for representation learning should contain as little information about nuisance N as possible while still keeping all the information about class Y. Since N is not observed, we transfer the objective $\min_{X'} I(X' \wedge N)$ into $\min_{X'} I(X' \wedge X)$ since $I(X' \wedge X) = I(X' \wedge N) + I(X' \wedge Y)$ and $I(X' \wedge Y)$ is a constant under the constraint $I(X' \wedge Y) = I(X \wedge Y)$. Thus the optimization problem is:

$$\min_{X'} I(X' \wedge X) \text{ s.t. } I(X' \wedge Y) = I(X \wedge Y) \quad (1)$$

To calculate the above mutual information terms, a neural net classifier $F(.;\theta)$ parameterized by θ is used that consists of two components: a representation encoder E(·) and a predictor M(·). Now, given an input x, its data augmentation x' can be computed by solving the following optimization problem using the neural network $F(.;\theta)$ in practice:

$$\min_{x'} -\|\phi(x) - \phi(x')\| \text{ s.t. } \log F(x';\theta)[y] = \log F(x;\theta)[y] \quad (2)$$

We created augmentations based on an adversarial approach using representation learned from the intermediate layers. Then we applied PES approach on the top of adversarial samples generated by the LPA3 approach and compared the results of LPA3 trained model with the baseline model trained just using PES. For this purpose we trained the CIFAR100 dataset with 80 % and 90 % noisy labelled data.

Saliency map: Now, it is known that different network filters are responsible for identifying different image properties and objects which motivates the idea that mistakes made on wrongly classified images can be understood by investigating the network parameters, rather than only the pixels, that played a role in making a decision. Thus, we deploy some techniques like: (1) aggregation of parameter saliency and (2) standardizing parameter saliency. Thus, using these techniques we obtain the saliency layer-wise profile of both the models and saliency map on a correctly classified and an incorrectly classified sample image. The results are shown in Tab. 1. We also compare the saliency profile of correct and incorrect classification as shown in Fig. 1 and Fig. 2.

4. Results and Analysis

On doing our experiments on the CIFAR-100 dataset with 90 % noisy labels using PES technique we get 11.53 % test accuracy and on introducing LPA3 technique we observe that our test accuracy approximately doubled, i.e., 20.59 %. The resulting accuracy on the dataset with 80 % noisy labels using just PES is 15.2% and on using LPA3, it increases to 23.86 %. Empirically, we observe the saliency

Algorithm 1 Our approach

Step 1: Training ResNet on CIFAR-100 having 90% noisy labels, normally, using Progressive Early Stopping (PES).
Step 2: Calculating the Accuracy Scores.
Step 3: Training another model using LPA-3 and PES on the same noisy dataset.
Step 4: Calculating the Accuracy Scores.
Step 5: Save the checkpoint for the generation of saliency maps (both parameter based and input based).
Step 6: Reporting the differences obtained in the scores and the saliency maps.

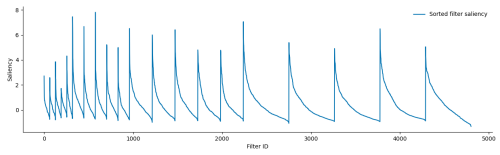


Figure 1. **Filter-wise parameter saliency profile**. ResNet-18 filter-wise saliency profile (without standardization) for a correct classified image using PES.

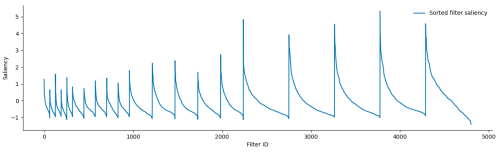
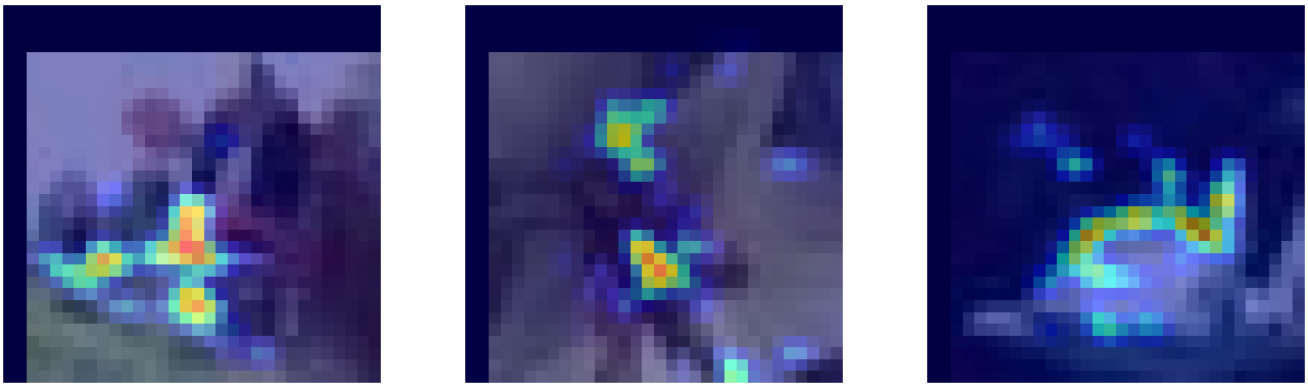


Figure 2. **Filter-wise parameter saliency profile**. ResNet-18 filter-wise saliency profile (without standardization) for an incorrect classified image using PES.

Model	RN18	RN18
Technique	Acc(90% Noisy)	Acc(80% Noisy)
PES	11.30	15.2
PES+LPA3	20.59	23.86

Table 1. Results on the CIFAR-100 dataset with noisy labels

profiles of incorrectly classified samples exhibit, on average, greater values than those of correctly classified examples. This bolsters the intuition that salient filters are precisely those malfunctioning — if the classification is correct, there should be few malfunctioning filters or none at all. Moreover, we see deeper parts of the network appear to be most salient for the incorrectly classified samples while earlier layers are often the most salient for correctly classified samples. An example of these behaviors for ResNet-18 is shown in figure 1 which presents standardized filter-wise saliency profiles averaged over the correctly and incorrectly classified examples from the CIFAR-100 validation set.



(a) Label-19, correct classification (b) Label-92, incorrect classification (c) Label-93, incorrect classification

Figure 3. Input saliency heatmap for RN-18 trained with PES

References

[1] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, M. Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. *ArXiv*, abs/2107.03342, 2021. 1

[2] Yuxin Wu Saining Xie Kaiming He, Haoqi Fan and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and pattern recognition*, pages 9729–9738, 2020. 1

[3] Shu M. Borgnia E. Huang F. Goldblum M. & Goldstein T. Levin, R. Where do models go wrong? parameter-space saliency maps for explainability. *ArXiv*, abs/2108.01335, 2021. 1

[4] Li Y. & Vinyals O. Oord, A. V. Representation learning with contrastive predictive coding. *ArXiv*, /abs/1807.03748, 2018. 1

[5] Sun Y. Su J. He F. Tian X. Huang F. Zhou T. & Tao D. Yang, K. Adversarial auto-augment with label preservation: A representation learning principle guided approach. *ArXiv*, abs/2211.00824, 2022. 1

[6] Stella X Yu Zhirong Wu, Yuanjun Xiong and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and pattern recognition*, pages 3733–3742, 2018. 1