Dependable AI Project Report

Group Members:

- Navlika Singh (B20AI025)
- Vikash Yadav (B20AI061)

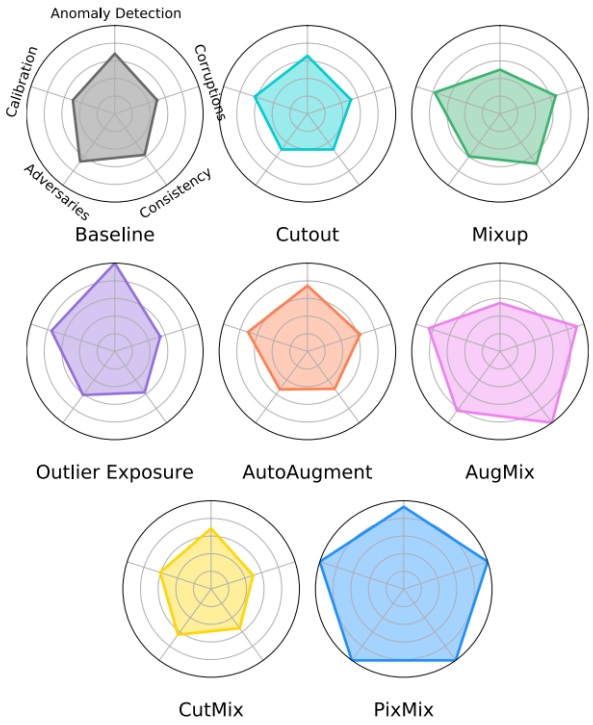
Project Database:

B20AI061_B20AI025_DAI_PROJECT_DATABSE

Introduction

Throughout the course, we have learned that just focusing on the test set accuracy is a very narrow approach to evaluating the measure of performance of our AI models. Other goals such as out-of-distribution (OOD) robustness, prediction consistency, resilience to adversaries, calibrated uncertainty estimates, and the ability to detect anomalous inputs need to be fulfilled to have reliable and safe real-world systems.

CVPR 2022 paper titled “PIXMIX: Dreamlike Pictures Comprehensively Improve Safety Measures” talks about how improving performance towards these goals is often a balancing act that today’s methods cannot achieve without sacrificing performance on other safety axes. Like adversarial training helps us against adversarial attacks but sharply degrades other classifier performance metrics, likewise strong data augmentation and regularisation techniques often improve OOD robustness but harm anomaly detection. The paper raises the question if a Pareto-optimization (improving in one metric without sacrificing the others) is possible on all existing safety measures.



To meet this challenge, the authors of this research paper design a new data augmentation strategy (PIXMIX) utilizing the natural complexity of pictures such as fractals, which outperforms numerous baselines and roundly improves safety measures.

We aim to extend this data augmentation technique to videos, calling it the V-PIXMIX and exploring how this extension is possible and what are the challenges in the same.

The PIXMIX Approach

The authors of the research paper propose PixMix, a simple and effective data augmentation technique that improves many ML Safety measures simultaneously, in addition to accuracy. PixMix is comprised of two main components:

1. Pix: A set of structurally complex pictures.
2. Mix: A pipeline for augmenting clean training pictures

The authors talk about how Kataoka et al., [5] introduced FractalDB, a dataset of black-and-white fractals, and they show that pre-training on these algorithmically generated fractal images can yield better downstream performance than pre-training on many

manually annotated natural datasets. This paper departs from the pre-training on such images and talks about how such images can work for data augmentation instead.

Why does it work?

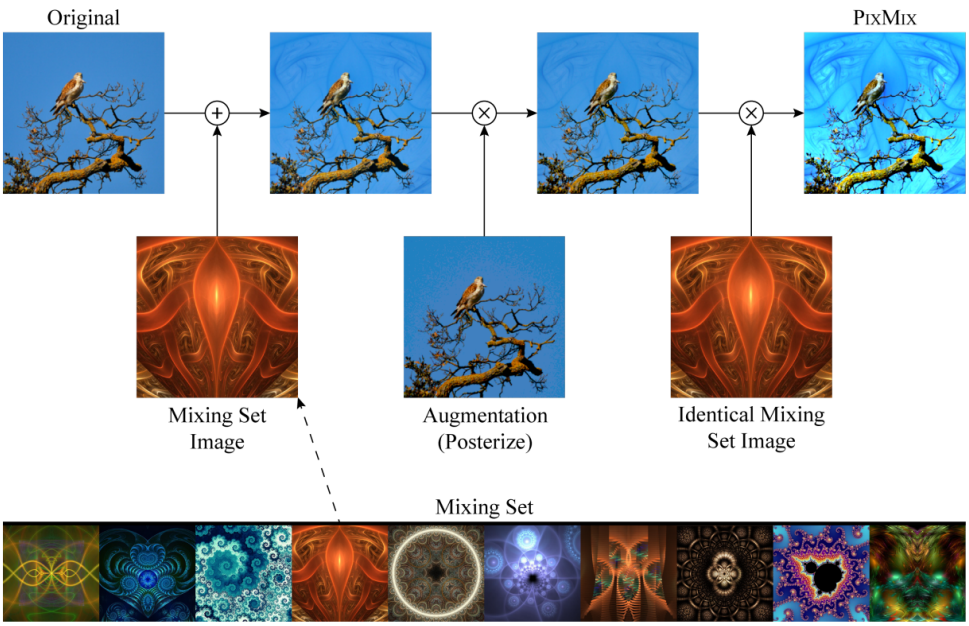
While data augmentation techniques such as those that add Gaussian noise increase input entropy, such noise has maximal descriptive complexity but introduces little structural complexity, that is to say, they add random noise and not structured noise to the images. Fractals and feature visualizations are especially useful for pictures with complex structures. Collectively the authors refer to these two picture sources as “dreamlike pictures.” Authors claim these pictures have structural properties that are highly non-accidental and unlikely to arise from maximum entropy, unstructured random noise processes.

When the model has to learn to become robust to this structured noise, they turn out to perform better in a Pareto-optimized way (improvement in one metric without causing harm to other metrics).

How it’s done?

```
def pixmix(x_orig, x_mixing_pic, k=4, beta=3):
    x_pixmix = random.choice([augment(x_orig), x_orig])

    for i in range(random.choice([0,1,...,k])): # random count of mixing rounds

        # mixing_pic is from the mixing set (e.g., fractal, natural image, etc.)
        mix_image = random.choice([augment(x_orig), x_mixing_pic])
        mix_op = random.choice([additive, multiplicative])

        x_pixmix = mix_op(x_pixmix, mix_image, beta)

    return x_pixmix

def augment(x):
    aug_op = random.choice([rotate, solarize, ..., posterize])
    return aug_op(x)
```

A series of random standard data augmentations along with mixing with fractal images is used with varying strengths of mixing and using adding or multiplying randomly for mixing. The results are “dreamlike images” that are then used to train the models.



Experiments of the paper

The authors of this research paper compare PixMix to methods on five distinct ML Safety tasks. Individual methods are trained on clean versions of CIFAR-10, CIFAR-100, and ImageNet. Then, they are evaluated on each of the following tasks:

1. **Corruptions:** This task is to classify corrupted images from the CIFAR-10-C, CIFAR-100-C, and ImageNet-C datasets. The metric is the mean corruption error (mCE) across all fifteen corruptions and five severities for each corruption. Lower is better.
2. **Consistency**: This task is to consistently classify sequences of perturbed images from CIFAR-10-P, CIFAR-100-P, and ImageNet-P. The main metric is the mean flip rate (mFR), which corresponds to the probability that adjacent images in a temporal sequence have different predicted classes.
3. **Adversaries**: This task is to classify images that have been adversarially perturbed by projected gradient descent. For this task, we focus on untargeted perturbations on CIFAR-10 and CIFAR-100. The metric is the classifier error rate. Lower is better.
4. **Calibration:** This task is to classify images with calibrated prediction probabilities, i.e. matching the empirical frequency of correctness. Lower is better.

5. **Anomaly Detection**: In this task, we detect out-of-distribution or out-of-class images from various unseen distributions.

Conclusion of the research paper

PixMix was the first method to provide substantial improvements over the baseline on all existing safety metrics, and it obtained state-of-the-art performance in nearly all settings.

Reproducing Results of PIXMIX on CIFAR-10

We first tried to reproduce some results from the paper on the CIFAR-10 dataset. We used CIFAR-10 because of the following reasons:

1. CIFAR-10 is the smallest and simplest of the three datasets consisting of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.
2. Considering the time and computation (limited GPU) constraints CIFAR-10 is the optimal dataset to perform the evaluations.
3. CIFAR-10 though simple still captures the intricacies of the other two datasets and the results obtained henceforth on CIFAR-10 will be a strong indicator for results obtained on CIFAR-100 and ImageNet for similar experiments.

Of all the five different tasks on which evaluations have been performed, we chose the following two tasks for evaluation: Corruptions and Adversaries because for the following reasons:

1. We are targeting only ‘backdoor attacks’ here and hence it is only wise to limit our evaluation to `‘Corruptions’` and `‘Adversaries’` rather than wasting time on evaluating all of them.

Reproducing the results of the research paper on the CIFAR-10 dataset and Corruption and Adversaries tasks.

The original CIFAR-10 and mixing set (consisting of fractals and feature visualization) are constructed using PyTorch dataset API. This is followed by augmenting the original CIFAR-10 dataset with augmenting images from the mixing set following the pipeline

described above (for more details refer to the code implementation). The augmented images constitute the training dataset now.

Utilizing this augmented training dataset a 40-4 Wide ResNet model is trained and evaluated on the above-mentioned tasks of Corruption and Adversaries.

Experimental Settings

The experimental settings while producing these results are set to be the same as that of the research paper, except for the following modifications:

1. Epochs: Considering the training time of the model increases because of PixMix data augmentation, it takes approximately ~5 minutes to train 1 epoch for the model. In the research paper, the authors have trained the model for 100 epochs, however for our scenario since it would lead to a substantial amount of time (~8 hours), and considering the time constraints we have trained the model for only 10 epochs.

Note: ‘k’ and ‘Beta’ are hyper-parameters for the PixMix pipeline and imply mixing iterations and severity of mixing respectively.

Model	40-4 Wide ResNet
Drop rate	0.3
Epochs	10
Initial learning rate (cosine learning rate schedule)	0.1
k	4
Beta	3

Results

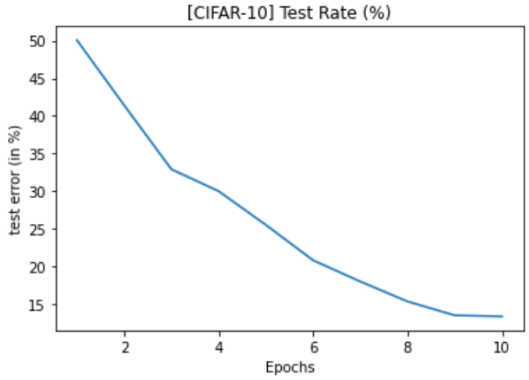
It can be observed that the metric values obtained while reproducing the results of the research paper are higher, only slightly for the Adversaries task, and almost double and higher than the baseline for the Corruption task, as compared to the values reported in the research paper.

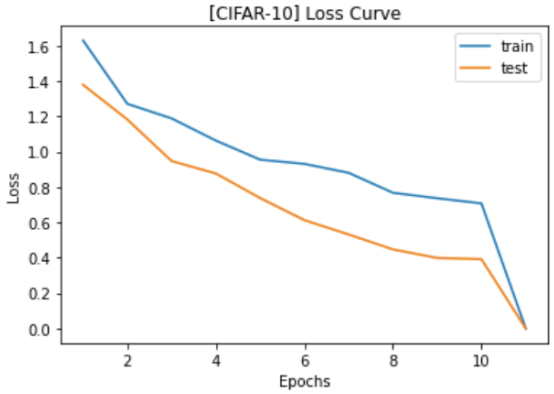
This difference in values can be justified as follows: In the original experimental setting of the research paper the authors have substantially greater time and computational resources, consequently, they have trained the model for 100 epochs. However,

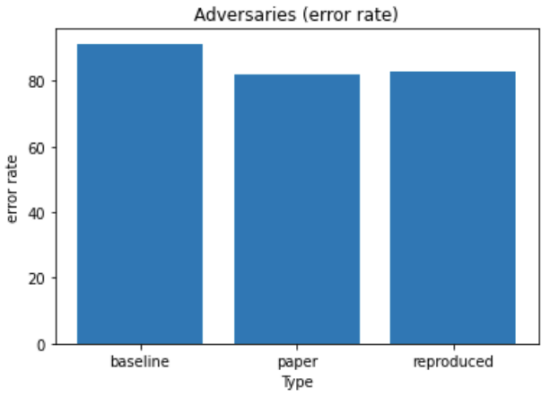
considering the time constraint and limited GPU access on Google Colab which expires after a certain limit, we are only able to train the model for 10 epochs (max).

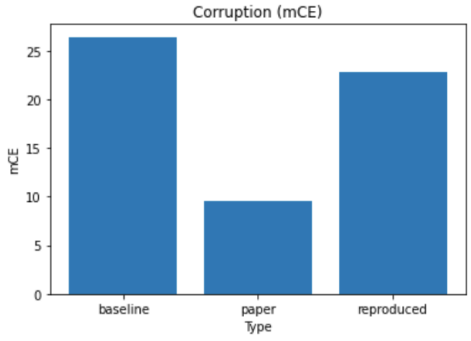
However, it can be observed that even after training the model for 10 epochs only, we get an 82.730 error rate for the Adversaries task which is only slightly greater than the 82.1 optimal value reported in the research paper. Furthermore, the mCE value for Corruption tasks is only double. This implies that the model is learning in the right direction and if we allow the model to train for a few more epochs we may achieve the same values (or perhaps better!) than the ones reported in the research paper.

Tasks	Baseline	Research Paper	Reproduced Results
Corruptions (mCE)	26.4	9.5	22.788
Adversaries (error rate)	91.3	82.1	82.730









Extending it to Videos

Literature Survey

This literature survey discusses the existing data augmentation techniques for video-based data. The survey divides the techniques into five categories: basic transformations, feature space augmentation, DL models, simulation, and methods that improve data generated through simulation using Generative Adversarial Networks. The survey discusses the papers that have proposed techniques falling under these categories.

- The survey starts with the basic transformations category, where different methods for temporal data augmentation are proposed. In one study, the authors propose applying temporal cropping iteratively to each original video sequence, and in another study, the authors augment a video dataset of hand gestures by splitting the original 12-frame videos into three videos of 8 frames each, and inverting the temporal order of the frames. Another paper proposes applying the same image-level transformation to all frames of a mini-batch clip to avoid introducing unnecessary noise and corrupting temporal cues of intra-clip frames. Another technique discussed in the survey is image mixing techniques. VideoMix is a data augmentation method that extends CutMix to video data augmentation. The temporal consistency is preserved by keeping the patch size and position the same for all the frames of each video clip. Another study proposed a toolbox for data augmentation that generates synthetic surveillance videos of static cameras for video synopsis analysis.
- In the feature space category, Dong et al. proposed a data augmentation strategy for a content-based video recommendation challenge, where the authors applied the data augmentation directly on the feature vector extracted from an InceptionV3 deep network.
- The DL models category proposes different methods to augment video data, such as the use of Generative Adversarial Networks (GANs) to generate new data. One paper proposed using GANs to generate realistic videos for action recognition tasks. Another study proposed a novel data augmentation strategy for video data using self-supervised contrastive learning.
- In the simulation category, a paper proposed a data augmentation method for depth completion in autonomous driving scenarios. The authors developed a simulator

that takes into account weather and lighting conditions, and they generated synthetic depth maps, which were used to augment the training set.

- The paper titled "VideoMix: Rethinking Data Augmentation for Video Classification" presents a novel video domain-specific data augmentation technique. The proposed technique employs spatial and temporal mixing of different video clips to generate diverse training samples for improving the accuracy of video classification models. Although this method is similar to our work, the author's primary focus is on improving the accuracy of the classification model, and they do not evaluate their technique on other important evaluation metrics like perturbation and corruption. In contrast, our paper aims to present a comprehensive evaluation of different data augmentation techniques for video classification, including their effectiveness on different evaluation metrics.

Overall, the literature survey provides an overview of different data augmentation techniques for video-based data, including basic transformations, feature space augmentation, DL models, simulation, and methods that improve data generated through simulation using GANs. The survey discusses different papers that propose methods under each of these categories and the results of experiments that validate the efficacy of the proposed techniques.

V-PIXMIX: Extending pixmix to videos

Should we just apply the Pixmix strategy to each frame of the video? → Naive approach.

Challenges in the Video Domain: The time dimension

- The original Pixmix involved selecting a random image from the mixing set. This is a problem, because now we have the time dimension too in videos, if we mix from a random sample from the mixing set each time, the result will be random noise getting introduced across the frames due to the changing mixing image.
- Also, pixmix function involved a lot of random choices about which data augmentation to use, with what strength, how much to mix the mixing image, again with what strength, whether to use add or multiple for mixing the mixing image, and how many times to repeat this process.

This randomness (of choosing random mixing images and the random choices for the mixing process) was not a problem for the 2d images as the structure of the noise (the perturbation) was consistent across the perturbation for 1 image, but if we just apply pixmix function to each frame, this randomness will be an issue, as the structure of the perturbation won't remain consistent across one 3d video (2d frames + time dimension).

This will defeat the whole idea of "structured noise/perturbations" and lose the essence of pixmix approach.

The g**oal is to** achieve noise (perturbation) in the videos which remains structurally complex and consistent throughout the video.

Our Fix: How to approach videos then?

- Firstly, we get rid of most of the randomness of the mixing process. We randomly choose between add or multiply with 0.5 probability for mixing frame with the mixing image, and no additional data augmentations.
- Secondly, to get rid of the randomness introduced due to different mixing images for each frame, we have 2 approaches in mind:
    - **Single Mixing Image:** We choose a single image from the mixing set for each video, and use that to pixmix each frame.

      "*The idea is, the mixing image stays consistent across the frames as opposed to a random change of mixing images as in the naive approach."*
    - **Mixing Images sampled from a fractal video:** We break up a fractal zoom video into frames and for each frame of the video to be perturbed we keep taking consecutive frames from the fractal video to pixmix.

      Fractal Zoom Videos: These are videos that keep zooming in on a fractal image, or on the madelbrot set infinitely, many such videos are available on youtube, such as this: Click to see the video.

      "*The idea is since the fractal gets zoomed a little in each consecutive frame, it will have consistency among consecutive mixing images and not just that, it will actually add a structural change to the mixing image across the frames as opposed to a random change of the mixing image as in the naive approach."*

Our hypothesis is that with these fixes, V-PixMix should work well for videos: making them more stable to all the metrics discussed at the beginning. We also believe sampling mixing images from a fractal video would be a better approach due to the reason mentioned above.

Experimental Settings

Now, in order to validate our proposed approach we consider the UCF101 dataset.

- The UCF101 dataset is a popular benchmark dataset for action recognition in videos. It contains 101 action categories, which are divided into three subsets: training, validation, and testing. The total number of videos in the dataset is 13,320, with an average duration of 7.2 seconds per video.

  The videos in the UCF101 dataset were collected from YouTube and other sources, and they represent a diverse range of human actions, such as sports, dancing, and daily activities.

  Each video in the UCF101 dataset is represented by a sequence of frames, and each frame is an RGB image of size 320x240 pixels. The dataset also includes human-annotated bounding boxes for some of the actions, which can be used for object detection and localisation.

- We have considered ResNet18-3D as our model. Also, please find below the experimental settings:

Epochs	10
Initial learning rate (cosine learning rate schedule)	0.03
k	4
Beta	3

We couldn't evaluate the 2nd approach (mixing images sample from fractal zoom video) due to time constraints, but we believe that would work even better.

We limit our evaluation to `Adversaries` only i.e. adversarial perturbation using PGD is applied to every frame to get the adversarial video sample

Obervations and Analysis

The results for single mixing image V-pixmix are here:

Metric	Base	V-pixmix (Single Mixing Image)
Test Accuracy (in %)	86.2012	84.304
Test Loss	0.7850	0.943
Adversaries (error rate in %)	95.36	87.32

The V-pixmix data augmentation lead to just 2% loss on classification accuracy but significantly dropped the error rate in case of adversaries.

- This shows the effectiveness of this technique. It achieve robustness without sacrificing much on accuracy.
- The idea of “structured noise” for data augmentation does help. Since the noise is structured (meaning has a pattern and is not just entropy), it does not hinder the classfication, as the model can learn the pattern of the noise and learn to ignore it. It is not possible to ignore random noise by the model (since model can’t learn it), and this is the main reason why normal robustness training method show a drop in accuracy.
- The structured noise does however fullfill the task of training for other measures of robustness, consistency, out-of-distribution (OOD) etc.
- Like in the paper, a more extensive set of experiments is required to evalute on all different types of measures such as out-of-distribution (OOD) robustness, prediction consistency, resilience to adversaries, calibrated uncertainty estimates, and the ability to detect anomalous inputs.
- We believe our method would provide optimization more close pareto than other techniques with the reasonings mentioned above.

Conclusion

We discussed the Pixmix approach for adding structural perturbations to images and evaluated the data augmentation technique on **CIFAR-10** across `corruptions` and `adversaries`.

We extended the idea to videos (V-pixmix), made fixes for the challenges in videos,, and evaluated the data augmentation technique on **UCF101 Videos** dataset across

`adversaries`.

Our approach does improve the performance against adversaries but also comes at the slight cost of test accuracy. However, a proper evaluation of all the metrics together remains to evaluate the Pareto optimize ability of the approach.

The future scope of exploring the 2nd approach (mixing images sampled from frames of a fractal zoom video) remains.

References

[1] https://github.com/okankop/vidaug

[2] VideoMix: Rethinking Data Augmentation for Video Classification

[3] https://github.com/nayeemrizve/ucf101-supervised

[4] https://arxiv.org/pdf/2112.05135.pdf

[5] https://arxiv.org/abs/2101.08515