PERTURBATION DRIVEN GENERALIZATION: EVALUATING THE IMPACT OF DATA AUGMENTATION STRATEGIES ACROSS MODEL ARCHITECTURES

Paul Mello, Gino Nicosia

University of Nevada, Reno

ABSTRACT

Data augmentation has become a staple component of machine learning to drive faster generalization in model training. As architectures continue to diversify and scale, optimizing their training becomes increasingly important. In this work, we conduct an empirical analysis on combinations of data augmentation strategies and intensities across a range of model sizes and types to identify the highest performing augmentations strategies. While we find that convolutional neural networks (CNNs) far exceed vision transformers (ViTs), multilayer perceptrons (MLPs), and variational autoencoders (VAEs) in training and validation metrics (numbers) on image data, empirical analysis demonstrates geometric augmentations have the strongest generalization improvements. Geometric augmentation strategies consisting of translation, rotation, and scale significantly outperform other augmentation types like pixel-values and frequency domain regardless of model type or size. These findings hold for both the single and multi-augmentation case, illustrating that augmentation techniques which target orthogonal invariance have compounding effects towards improving model generalization. These findings help enable researchers to make more informed decisions to improve model generalization.

Index Terms— Generalization, Transformer, Data Augmentation, Robustness

1. INTRODUCTION

Data augmentation has become a fundamental part of machine learning for its generalization properties. These methods have been shown to improve model understanding through varying the quality and quantity of information that can be extracted from a sample. These drive neural networks to learn properties of data invariance and probabilistic extrapolation thereby reducing overfitting [1]. This has be shown to be effective across a range of learning domains including natural language processing, computer vision, and time series forecasting [2].

Generalization is defined by a model's capability to work effectively on unseen data. The generalization of models depends on many variables including model architecture, hyperparameter selection, and diversity of training data. Recently, studies have shown that neural networks, regardless of their designed complexity, can achieve significantly better generalization through the use of data augmentation [3].

Despite the pervasiveness and importance of data augmentation strategies, little work has explored the hyperparameter space of data augmentation techniques. Previous works, such as AutoAugment [4], searches for improved data augmentation policies through sub-process search space. Follow up works like RandAugment [5], reduce the necessary search space and incorporate the target objective as part of its own search goal. Furthermore, existing studies have focused their efforts of reducing the search space for applications of data augmentations, but little work has examined the effects of different augmentation strategies and their interactions with various neural network types across scales.

In this work, we experiment with numerous data augmentation techniques to identify the most effective strategies across a range of configurations. We systematically evaluate the performance of CNNs, MLPs, VAEs, and VITs across three model sizes and apply combinations of augmentations and intensities to each model. Empirical analysis reveals an interesting relationship between geometeric augmentations and model architectures. Through this work, we provide a foundational system and practical augmentation recommendations for future model experiments.

2. RELATED WORK

For many years, data augmentation has been acknowledged as a straightforward yet effective method for enhancing generalization in machine learning models. Most work has consolidated around the generation of augmentation strategies. In order to expose models to a variety of inputs without the need for fresh labeled data, early work concentrated on simplistic geometric changes like flipping, cropping, and rotating. By making models more resilient to semantic variations, these augmentations created diverse feature representations that more invariant and thus more robust. More sophisticated augmentation methods, such as blurring, have been put forth to replicate more difficult or realistic input distributions as datasets have continued to expand. Although simple, these techniques have shown remarkable efficacy, especially when combined with convolutional architectures and their inductive

biases.

Most recent efforts have been on employing search-based methods to automatically find the best augmentation policies. A reinforcement learning-based method called AutoAugment [4] looks at augmentation policies to find those that optimize validation performance. RandAugment [5], which streamlines the search process by narrowing the policy space to randomly selected operations with defined magnitude and count parameters, followed this. These methods showed that well-thought-out augmentation pipelines may compete with or even outperform architectural enhancements. However, they are less interpretable or accessible for widespread use across architectures and model sizes because of their high computational cost and model-specific customization.

We position this work perpendicular to these prior works by exploring the effectiveness of augmentation strategies across a range of models. Rather than focus on the generation of strategies to improve generalization on a per model basis, this work focuses on the empirical analysis to garner new insights into the most effective strategies to utilize. We offer this work as a compliment to works like RandAugment and AutoAugment to help shape insights into effective strategies.

3. METHODOLOGY

3.1. Dataset

We select the CIFAR10 dataset, a popular machine learning benchmark with 60,000 colored images divided into 10 classes of 6,000 images with the dimensions 32x32 [6]. We elected to use CIFAR10 due to its RGB channels which help improve the sample complexity and increase the difficulty of model generalization. Other datasets, like MNIST [7], were proposed for this work, but initial results demonstrated the sample complexity was too low.

3.2. Model Architecture

We evaluate four common but distinct model architectures. These models were selected to reduce the search space while providing an broad experimental range of architectures and biases.

- **Multilayer Perceptron**: A simple feed-forward neural network with fully connected layers.
- **Convolutional Neural Network**: A convolutional architecture designed specifically for spatial data, like images.
- **Vision Transformer**: A variant of the Transformer model tailored for image data, using a patch-based system approach.
- **Variational Autoencoder**: A generative encoder and decoder architecture tuned specifically for prediction through the utilization of class tokens.

We design three model sizes to compare augmentations strategies as models scale. We handcraft and validate model parameter counts to remain within 5% of their expected parameters namely: small, medium, and large at 1, 3, and 9 million parameters respectively. This model scale provides insights across architecture sizes to address the growing concern of handling datasets at scale through comparisons of generalization efficiency induced by parameter counts.

3.3. Augmentation Strategies

We explore the affects of different image augmentation strategies including type, intensity, and combination during model training. The following augmentations are considered in this work:

3.4. Spatial Domain Transformations

3.4.1. Geometric Transformations

- **Rotation:** Rotates the image by a random angle. $I'(x, y) = I(x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta)$ where $\theta \sim \mathcal{U}(-\alpha, \alpha)$ and $\alpha = \text{intensity} \cdot 90°$.
- **Translation:** Shifts the image in random directions according to a scalar value. $I'(x, y) = I(x-\Delta x, y-\Delta y)$ where $\Delta x, \Delta y \sim \mathcal{U}(-\text{intensity}, \text{intensity})$.
- **Scaling:** Resizes the image by a random factor. $I'(x, y) = I(sx, sy)$ where $s \sim \mathcal{U}(1 - \text{intensity}, 1 + \text{intensity})$.

3.4.2. Pixel-Value Transformations

- **Gaussian Noise:** Adds random noise to pixel values. $I'(x, y) = \text{clamp}(I(x, y) + \epsilon, 0, 1)$ where $\epsilon \sim \mathcal{N}(0, \text{intensity}^2)$. Here, clamping keeps perturbations within the image manifold domain.
- **Salt-and-Pepper Noise:** Gaussian noise with perturbations set to extreme values, 1 or 0.
- **Gaussian Blur:** Smooths the image via convolution (*) with a Gaussian kernel. $I' = G_\sigma * I$ where G_σ is a 2D Gaussian function with standard deviation $\sigma = 2 \cdot \text{intensity}$.
- **Random Erasing:** Masks random regions of the image.

$$I'(x, y) = \begin{cases} 0, & \text{if } (x, y) \in R \\ I(x, y), & \text{otherwise} \end{cases} \quad (1)$$

where R is a randomly positioned rectangular region.

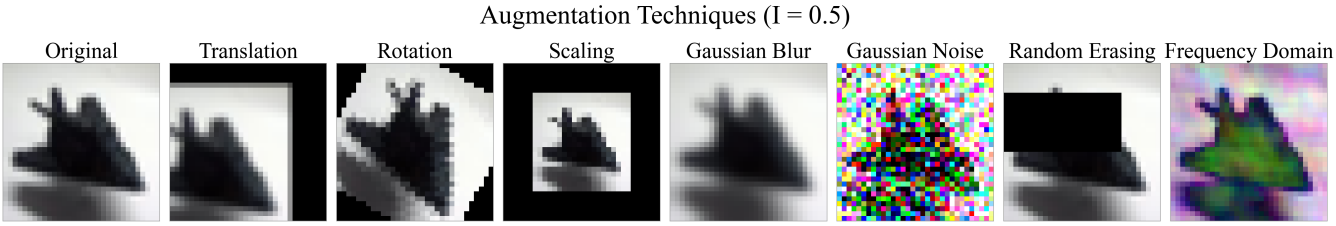

Fig. 1. Visual comparison of data augmentation techniques applied to a CIFAR10 image. Here we see the effects of applying various augmentation strategies. We set augmentation intensities to 0.5 to be visibly apparent.

3.5. Frequency Domain Transformations

- **Frequency Perturbations:** Modifies image in the frequency domain. $I' = \mathcal{F}^{-1}(\mathcal{F}(I) \cdot (1 + \epsilon))$ where $\mathcal{F}$ is the Fourier transform and $\epsilon \sim \mathcal{N}(0, \text{intensity}^2)$.

We train models by applying all possible combinations of augmentation within a selected bounding variable.

3.6. Experimental Setup

To ensure a controlled, yet flexible evaluation environment, all experiments were performed using a centralized configuration management system, which allowed precise tuning of hyperparameters, augmentation policies, and model variants. This allowed for a consistent application of experimental settings across architectures and augmentation pipelines.

We conducted a systematic evaluation of model performance under various noise-based perturbations utilizing a diverse set of model and augmentation strategies. All experiments were performed on the CIFAR10 dataset, with a batch size of 512 over 25 epochs. The learning rate (0.001) and weight decay (0.0) were selected to avoid regularization artifacts, ensuring that generalization effects came from the augmentation rather than optimizer-induced smoothing. All models were trained using the AdamW optimizer without learning rate schedules to isolate the impact of augmentation from rate decay.

Augmentations were divided into standard and advanced categories using dynamic selectors, and applied at varying intensities (0.1, 0.3, 0.5). We restricted the maximum number of combinations of augmentation to a bounding variable of 2 to avoid excessive computational costs and prevent degradation of the input data beyond recognition.

4. RESULTS

We identified interesting information relating the success of models to semantic preservation while also finding the bounding limitations of this work. The main limitation is a lack of variability in dataset size, which we believe would provide a more accurate representation of the varying model sizes. By providing data samples according to model parameter counts, we would improve the efficiency of generalization. Tables [1, 2, 3] show the main findings, primarily that the best augmentation strategies almost always include a geometric augmentation.

Table 1. Individual Augmentation Techniques Ranked by Mean Validation Accuracy

Rank	Augmentation Technique	Mean Acc.(%)
1	Control (No Augmentation)	66.27
2	Translation	64.84
3	Rotation	64.80
4	Scale	64.72
5	Gaussian Blur	64.26
6	Frequency Domain	63.05
7	Random Erasing	62.88
8	Gaussian Noise	59.37
9	Salt & Pepper	58.23

Table 2. Top Augmentation Pairs Ranked by Mean Validation Accuracy

Rank	Augmentation Pair	Mean Acc.(%)
1	Gaussian Blur + Translation	67.51
2	Scale + Translation	67.41
3	Rotation + Scale	67.25
4	Rotation + Translation	67.13
5	Gaussian Blur + Rotation	66.99
6	Gaussian Blur + Scale	66.81
7	Frequency Domain + Rotation	65.78
8	Frequency Domain + Scale	65.78
9	Frequency Domain + Translation	65.70
10	Random Erasing + Rotation	65.36

4.1. Baseline Performance: Control and Individual Augmentations

Table 1 shows the ranking of individual augmentation techniques by mean validation accuracy. Unsurprisingly, some control models without augmentation achieve higher validation accuracies than the closest singular augmentation. This is likely the result choosing minimal training times, resulting

Table 3. Best Performing Configuration by Model Architecture and Size

Model	Size	Aug. 1	Aug. 2	Int. 1	Int. 2	Val. Acc. (%)
CNN	Small	Translation	None	0.3	-	84.68
	Medium	Translation	Scale	0.1	0.3	84.81
	Large	Translation	Scale	0.3	0.3	83.91
VAE	Small	Scale	None	0.3	-	73.54
	Medium	Translation	Gaussian Blur	0.3	0.1	74.66
	Large	Rotation	Translation	0.1	0.3	75.03
VIT	Small	Rotation	Translation	0.1	0.3	69.91
	Medium	Rotation	Translation	0.1	0.3	67.90
	Large	Translation	Scale	0.1	0.1	67.85
MLP	Small	Translation	Scale	0.1	0.1	55.11
	Medium	Translation	Gaussian Blur	0.1	0.1	55.32
	Large	Rotation	Translation	0.1	0.1	56.18

in unconstrained models performing better in shorter training windows. Translation (64.84%), rotation (65.80%), and scaling (64.72%) consistently outperform the other perturbations, indicating that these geometric transformations add significant diversity while maintaining the semantic content of the data. This is especially notable as some augmentation strategies, such as VAEs in table 3, demonstrate an improvement on baseline generalization metrics. Conversely, the lowest performing augmentation strategies are obtained with noise-based perturbations such as Gaussian noise (59.37%) and salt-and-pepper (58.23%), suggesting that stochastic perturbations to semantic content inhibit generalization, at least in the early stages of training. These results suggest that geometric augmentations provide the most reliable generalization benefits when used in isolation, whereas noise-based methods may require careful tuning or combinations to be effective.

4.2. Augmentation Pair Performance

In table 2, we present the best augmentation pairs based on their mean validation accuracy across experiments. The best combinations across intensities and models always utilized geometric augmentations strategies. Translation, scale, and rotation appear to work the best together, supporting the idea that changes which preserve the information content while transforming the data help apply variance to the semantic context thereby increasing robustness and generalization. This understanding is furthered by the next best augmentation, Gaussian blur, which helps generalize without changing the image's semantic structure similar to the geometric transformations.

Interestingly, the application of perturbations in Gaussian blur constrains the way noise is added such that all noise remains on the data manifold. In theory, the geometric transformations introduce the same types of perturbations. The prime difference within the geometric augmentations become how each transforms the data manifold, resulting in better stochasticity and better generalization. However, combining data augmentation techniques performs better than their isolated counterparts, showcasing the benefits of multi-augmentation techniques and how complimentary transformations can easily expand the data manifold beyond the training data.

4.3. Model Performance by Augmentation

We reveal that geometric transformations perform better than their pixel-value counterparts across all neural network architectures. This pattern illustrates the spatial transformation invariance represents that preserves semantic context provides strong generalization regardless of the model architectures structural biases. As Table 3 shows, despite the variation in validation accuracy between model types, the top performing augmentation strategies remain geometric, illustrating stability between configurations. There is however, less variation in the intensity values likely caused by shorter training times. Notably, the variance in validation accuracy maintains a consistent relationship with the mean performance across model architectures.

5. FUTURE WORK

The main results of this work falls short of giving a comprehensive understanding on the role of data augmentations. Firstly, while we provide an account across model and augmentation configurations, we did not experiment on diverse datasets. This lack of diversity was clearly expressed in the data with CNNs having the overwhelming best performance. However, the lack of diversity was the result of insufficient hardware and temporal resources. This constrained search space significantly reduced the quantity of models, augmentations, and combinations we experimented on. Here, future work may seek to incorporate datasets of varying sizes, dimensions, and modalities. Future work may also seek to alter the task objectives or significantly increase training lengths to capture a comprehensive picture of the augmentation strategy space.

Despite this, we provide novel insights into the value of previously overlooked methods due to their simplicity. We find methods like translation, scale, and rotation consistently outperform other methods across sizes and model type. These methods share augmentation strategies which preserve semantic information and inject noise orthogonal to the data manifold, resulting in improved generalization capabilities, regardless of perturbation strength. Therefore, future work may explore distilling the core of these augmentation strategies down to its essence.

6. CONCLUSION

Analysis demonstrates that combinations of augmentation techniques consistently outperform individual methods across all model architectures on generalization. We show that the best data augmentation techniques overwhelmingly consist of translation, rotation, scale, and Gaussian blur. These strategies preserve the information content of the image through geometric transformation illuminating a fundamental connection between data variance and model generalization. These findings lead us to conclude that occlusion based methods, while common, are not as effective in training the model to generalize beyond the training data.

Notably, CNNs were by far the best performing model across all metrics, likely due to their translation equivariance and the selected dataset. While CNNs outperformed all other models, we find that other architectures also show significant improvements when utilizing the same group of augmentations, namely translation, rotation, and scale. This universal improvement, regardless of model size or type, illustrates that transformations which preserve semantic information while injecting perturbations orthogonal to the data manifold consistently enhance generalization. These findings suggest a strategic application of simple geometric transformations can significantly improve model performance.

7. REFERENCES

[1] Luke Taylor and Geoff Nitschke, "Improving deep learning using generic data augmentation," 2017.

[2] Brian Kenji Iwana and Seiichi Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *PLOS ONE*, vol. 16, no. 7, pp. e0254841, July 2021.

[3] Chris Rohlfs, "Generalization in neural networks: A broad survey," 2024.

[4] Zoph B. Mane D. Vasudevan V. Le Q. V. Cubuk, E. D., "Autoaugment: Learning augmentation strategies from data.," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, June 2019.

[5] Zoph B. Shlens J. Le Q. V. Cubuk, E. D., "Randaugment: Practical automated data augmentation with a reduced search space.," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, June 2020.

[6] Alex Krizhevsky, "Learning multiple layers of features from tiny images," pp. 32–33, 2009.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.