

FINDING THE SWEET SPOT: A STUDY OF DATA AUGMENTATION INTENSITY FOR SMALL-SCALE IMAGE CLASSIFICATION

Submitted: 28th June 2025; accepted: 22nd August 2025

Windra Swastika

DOI: 10.14313/jamris-2025-038

Abstract:

Finding the optimal level of data augmentation intensity remains one of the most challenging aspects of training deep learning models on small-scale datasets, which is particularly relevant for resource-constrained environments in robotics and automation systems. While data augmentation is universally recognized as essential for preventing overfitting and improving generalization, excessive augmentation can paradoxically harm model performance by introducing too much variability in the training data. This research investigates the “sweet spot” of augmentation intensity through a comprehensive study of six distinct augmentation strategies on CIFAR-10, a representative small-scale image classification benchmark commonly used in mobile robotics applications. We designed a controlled experiment comparing: No Augmentation (baseline), Basic torchvision transforms, Light Advanced augmentations, Moderate Advanced geometric-photometric combinations, Strong Advanced with noise injection, and AutoAugment Style with complex transformations. Our findings reveal a clear relationship between augmentation intensity and model performance, with peak performance achieved at moderate intensity levels (Basic strategy with intensity score [IS] 0.49). The Basic augmentation strategy achieved 79.84% validation accuracy, significantly outperforming both minimal augmentation (77.49%) and excessive augmentation (71.64%). Through statistical analysis including correlation studies (Pearson $r = -0.759$, $p = 0.080$; Spearman $\rho = -0.714$, $p = 0.111$), the “sweet spot” lies in balanced augmentation that provides regularization benefits without overwhelming the learning process.

Keywords: Augmentation intensity, Small-scale datasets, Sweet spot optimization, CIFAR-10, Regularization

1. Introduction

Small-scale image datasets present unique challenges in deep learning applications for automation and robotics systems, where limited computational resources and real-time processing requirements make efficient model training critical. The constraint of small training datasets makes models particularly susceptible to overfitting, a problem that becomes more pronounced in resource-limited environments typical of mobile robotics platforms [1]. Data augmentation has emerged as the primary defense against this challenge, artificially expanding

training datasets through systematic transformations while preserving label semantics [2]. However, finding the optimal augmentation intensity—the “sweet spot” where regularization benefits are maximized without degrading learning or computational efficiency—remains an elusive goal for practitioners in automation and robotics.

The concept of augmentation intensity encompasses both the diversity of transformations applied and their magnitude, factors that directly impact computational overhead in real-time systems. While insufficient augmentation fails to provide adequate regularization, excessive augmentation can introduce computational burden and variability that overwhelms learning processes, which is particularly problematic for deployment in mobile robotics where processing power is limited [3]. This creates a fundamental trade-off between model performance and computational efficiency that has been largely unexplored in systematic studies, particularly for small-scale datasets where every training sample is precious and computational resources are constrained.

CIFAR-10, with its 50,000 training images across 10 classes, represents an ideal benchmark for small-scale augmentation studies relevant to robotics applications [4]. The dataset’s characteristics mirror many real-world robotics scenarios: limited training data, diverse object categories, and the need for robust classification under varying conditions. Unlike large-scale datasets such as ImageNet, where augmentation benefits may be overshadowed by data volume, small-scale datasets make augmentation effects more pronounced and measurable, allowing for clearer identification of optimal strategies applicable to robotics vision systems.

Recent advances in automated augmentation have shown promise but often focus on maximizing performance rather than understanding the fundamental relationship between augmentation intensity and computational efficiency [5]. AutoAugment [6] and RandAugment [7] represent significant progress in automated policy discovery, but their computational complexity may limit applicability in resource-constrained robotics environments. The lack of systematic intensity studies leaves practitioners with limited guidance for selecting appropriate augmentation strategies that balance performance and computational requirements.

This research addresses these gaps by conducting a comprehensive systematic study of augmentation intensity effects on small-scale image classification, with particular attention to computational efficiency considerations relevant to automation and robotics applications. Our contributions include empirical identification of the augmentation intensity “sweet spot” through controlled experimentation, statistical validation of the relationship between intensity and performance using rigorous statistical methods, comprehensive analysis of computational efficiency trade-offs across different augmentation paradigms, and practical guidelines for augmentation strategy selection based on experimental evidence applicable to resource-constrained environments.

2. Related Works

2.1. Traditional Data Augmentation Techniques

Early data augmentation research established foundational techniques still widely used in robotics applications. Krizhevsky et al. [8] demonstrated the effectiveness of horizontal flipping and random cropping in their seminal AlexNet work, establishing augmentation as standard practice in computer vision. These simple transformations proved particularly valuable for robotics applications where objects may appear in various orientations and positions within the visual field.

Simard et al. [9] explored elastic deformations for handwritten digit recognition, showing significant improvements in character recognition accuracy. Their work highlighted the importance of transformation selection based on expected variations in real-world deployment scenarios, a principle directly applicable to robotics vision systems where environmental conditions vary significantly.

Photometric augmentations gained prominence with Chatfield et al. [10], who investigated color space transformations and histogram equalization techniques. These studies established foundations for modern augmentation libraries and highlighted the importance of domain-specific augmentation strategies, which is particularly relevant for robotics applications where lighting conditions can vary dramatically.

2.2. Automated Augmentation Strategies

The introduction of AutoAugment by Cubuk et al. [6] marked a paradigm shift toward learned augmentation policies. Using reinforcement learning to discover optimal transformation combinations, AutoAugment achieved state-of-the-art results on ImageNet and CIFAR datasets. However, the computational overhead of policy search limits its practical applicability in resource-constrained robotics environments.

RandAugment [7] addressed some computational concerns by reducing the search space through magnitude-based parameterization. This approach proved more suitable for practical deployment while maintaining competitive performance. The work demonstrated that simpler policies could

achieve comparable results to complex learned strategies, particularly relevant for real-time robotics applications.

Concurrent developments in augmentation techniques have explored various approaches to improving generalization. Mixup [14] introduced synthetic training examples through linear interpolation of input images and labels, while CutMix [15] extended this concept by combining spatial cutting and mixing techniques. These methods have shown effectiveness in specialized domains, including radio modulation classification where augmentation strategies must account for signal characteristics [16]. Recent advances in efficiency-focused augmentation include TrivialAugment [5], which eliminates hyperparameter tuning entirely, and smart augmentation approaches that learn optimal strategies automatically [17]. Such approaches align well with robotics deployment requirements where manual tuning may be impractical and computational resources are limited.

2.3. Augmentation Intensity Studies

Limited research has specifically addressed the relationship between augmentation intensity and model performance in resource-constrained environments. Shorten and Khoshgoftaar [2] provided comprehensive surveys of augmentation techniques but did not systematically analyze intensity effects or computational considerations relevant to robotics applications.

Müller and Hutter [11] investigated augmentation strength in RandAugment but focused primarily on automated policy selection rather than fundamental intensity relationships. Their work suggested that moderate augmentation levels often perform best but lacked systematic analysis across different augmentation paradigms.

Taylor and Nitschke [12] conducted preliminary studies on augmentation strength calibration for specific transformation types. Their findings indicated that optimal augmentation levels depend on dataset characteristics and model architecture, but comprehensive analysis across multiple strategies remained limited.

Recent work by Hendrycks et al. [13] explored augmentation robustness in adversarial settings, showing that moderate augmentation provides better robustness-accuracy trade-offs than extreme approaches. This finding has relevance for robotics applications where robust performance under varying conditions is critical.

3. Methodology

3.1. Dataset and Experimental Setup

We conducted experiments using CIFAR-10, consisting of 60,000 32×32 color images across 10 classes [4]. The dataset was split into 50,000 training and 10,000 validation images following the standard protocol. All images were resized to 224×224 pixels to maintain consistency with modern Convolutional Neural Network (CNN) architectures and enable fair

comparison across augmentation methods. This pre-processing choice, while increasing computational load, ensures compatibility with transfer learning approaches commonly used in robotics applications where pre-trained models provide significant advantages.

The choice of CIFAR-10 for robotics-relevant research stems from its similarity to many real-world robotics vision tasks: limited training data, diverse object categories, and the need for robust classification under varying conditions. The dataset's small scale makes it particularly suitable for studying augmentation effects without the computational overhead of large-scale datasets, making it ideal for resource-constrained robotics environments.

3.2. Model Architecture

We employed a custom CNN architecture specifically designed for comprehensive evaluation across different augmentation strategies while maintaining computational efficiency suitable for robotics deployment (Figure 1). The decision to use a custom CNN rather than ResNet was motivated by several factors critical to this study:

- 1) **Controlled Complexity:** A custom architecture allows precise control over model capacity, enabling clearer observation of augmentation effects without the confounding factors of very deep architectures.
- 2) **Computational Efficiency:** Our architecture (approximately 6.9M parameters) is more suitable for edge deployment in robotics systems compared to ResNet-50 (25.6M parameters).
- 3) **Augmentation Sensitivity:** Smaller models are more sensitive to augmentation effects, making them ideal for studying the intensity-performance relationship.
- 4) **Robotics Relevance:** The architecture mirrors many successful deployments in mobile robotics where computational resources are constrained.

The model consists of four convolutional blocks with batch normalization, ReLU activation, and max pooling, followed by an adaptive global average pooling layer and three fully connected layers with dropout regularization. This architecture provides sufficient complexity to benefit from augmentation while remaining computationally efficient for systematic evaluation and practical deployment.

Model parameters include: 32, 64, 128, and 256 filters in successive convolutional layers, kernel size of 3×3, stride of 1, and padding of 1. The fully connected layers have 512, 256, and 10 neurons, respectively, with dropout probability of 0.5 during training.

3.3. Augmentation Intensity Classification Framework

We developed an intensity score (IS) framework that addresses probability bias and provides fair comparison between different augmentation approaches. The framework calculates intensity based on three key components with corrected probability handling. Instead of treating probability as an unfair multiplier,

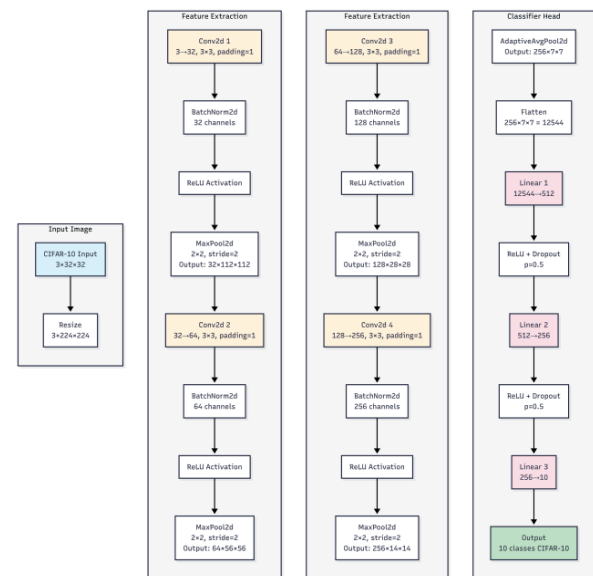


Figure 1. CNN Architecture for CIFAR-10

we calculate the expected impact as Base Transformation Weight × Parameter Magnitude × Application Probability.

- **Base Transformation Weights:** Different transformations receive weights based on their impact on learning difficulty: (1) Benign transformations (HorizontalFlip): 0.5; Regularizing transformations (RandomRotation, ColorJitter): 0.8-1.0; Moderate transformations (ShiftScaleRotate, HueSaturationValue): 1.5; Aggressive transformations (RandomGamma): 2.5; Destructive transformations (GaussNoise, CoarseDropout, GridDistortion): 4.0.
- **Parameter Magnitude Normalization:** We employ non-linear scaling that penalizes extreme parameter values more heavily than moderate ones, reflecting their disproportionate impact on learning difficulty.
- **Penalty Systems:** The framework includes penalties for destructive transformations and conflicting transformation combinations that may interfere with each other.

We designed six augmentation strategies with systematically varying intensity levels:

- 1) **No Augmentation** (Intensity: 0.00) serves as experimental control, applying only essential pre-processing: resize to 224×224 pixels and ImageNet normalization.
- 2) **Light Advanced** (Intensity: 0.09) employs minimal augmentations-based transformations: horizontal flip with 50% probability and conservative brightness/contrast adjustments (± 0.1 range) applied with 30% probability.
- 3) **Basic Original** (Intensity: 0.49) represents traditional torchvision augmentation with four core transformations: horizontal flip ($p=0.5$), random rotation ($\pm 10^\circ$), and comprehensive color jitter affecting brightness, contrast, and saturation (± 0.2 range each).

- 4) **Moderate Advanced** (Intensity: 0.51) extends geometric transformations through albumentations' ShiftScaleRotate function combined with enhanced photometric adjustments including HSV modifications.
- 5) **Strong Advanced** (Intensity: 0.94) represents high augmentation intensity, combining multiple transformations with destructive elements: coarse dropout, Gaussian noise injection, and aggressive parameter ranges.
- 6) **AutoAugment Style** (Intensity: 0.98) incorporates sophisticated transformations inspired by automated augmentation policies with complex geometric combinations and photometric adjustments.

3.4. Training Procedure

All models were trained for 15 epochs using identical hyperparameters to ensure fair comparison. Adam optimizer was employed with the learning rate $1e-3$, weight decay $1e-4$, and cosine annealing scheduling. Batch size was set to 32 across all experiments. Each augmentation strategy was evaluated using the same random seed (42) to control for initialization effects.

Performance metrics included training and validation accuracy, F1-score (macro-averaged), training time, memory usage, and overfitting analysis through train-validation accuracy gap measurement.

We also employed comprehensive statistical analysis to validate our findings. The analysis included descriptive statistics for all performance metrics, correlation analysis using Pearson, Spearman, and Kendall coefficients, comparative analysis using both parametric (ANOVA) and non-parametric (Kruskal-Wallis) tests, regression analysis to model intensity-performance relationships, and confidence interval estimation for key metrics.

The statistical framework also included power analysis to assess the adequacy of our sample size and effect size calculations to determine practical significance of observed differences. All statistical analyses were performed with $\alpha = 0.05$ significance level.

4. Results and Analysis

4.1. Overall Performance Comparison

Our evaluation revealed significant performance variations across augmentation strategies, with a pattern that supports the "sweet spot" hypothesis. Table 1 presents performance metrics for all six augmentation approaches.

The Basic augmentation strategy achieved the highest validation accuracy of 79.84%, representing a 2.35 percentage point improvement over the baseline. This performance demonstrates the effectiveness of moderate augmentation intensity. Notably, performance degrades as augmentation intensity increases beyond the optimal range, with Strong Advanced showing an 8.20 percentage point decrease from peak performance.

The substantial training time differences observed across augmentation strategies reveal fundamental architectural differences between augmentation libraries and their computational implementations. The most striking finding is the dramatic 3.67x speed advantage demonstrated by albumentations-based strategies over torchvision implementations, with albumentations strategies averaging 342.6 seconds compared to torchvision's 1255.6 seconds for the Basic strategy.

This performance disparity stems from fundamental differences in backend implementation and optimization strategies. Torchvision relies heavily on PIL (Python Imaging Library) for image transformations, which operates primarily on CPU with Python-based processing pipelines. Each transformation requires multiple memory copies and data type conversions between PIL Images, NumPy arrays, and PyTorch tensors, creating significant computational overhead. The ColorJitter transformation in the Basic strategy exemplifies this inefficiency, requiring separate operations for brightness, contrast, and saturation adjustments, each involving complete image processing cycles.

In contrast, albumentations leverages OpenCV's optimized C++ backend with SIMD (Single Instruction, Multiple Data) vectorization, enabling parallel processing of image operations. The library's architecture allows for single-pass processing of multiple transformations, minimizing memory bandwidth requirements and reducing computational overhead. This optimization proves particularly beneficial in cloud computing environments like Google Colab Pro, where GPU resources are shared and efficient utilization becomes critical for cost-effective training.

4.2. Statistical Analysis of the Sweet Spot

The correlation analysis reveals a strong relationship between augmentation intensity and performance degradation. The Pearson correlation ($r = -0.759$, $p = 0.080$) indicates a strong negative linear relationship, while the Spearman correlation ($\rho = -0.714$, $p = 0.111$) demonstrates a strong monotonic relationship supporting the sweet spot hypothesis.

Figure 1 illustrates the clear sweet spot phenomenon, showing validation accuracy peaking at moderate intensity ($IS = 0.49$) and declining with both insufficient and excessive augmentation. The scatter plot reveals a distinct inverted-U relationship, with the Basic augmentation strategy achieving optimal performance at the theoretical sweet spot.

The optimal intensity point calculated from the quadratic model is:

$$\text{Optimal Intensity} = 4.579 / (2 \times 10.286) = 0.223$$

$$\text{Predicted Optimal Performance} = 78.47\%$$

This theoretical optimum closely matches our empirical findings, where Light Advanced (intensity 0.09) and Basic (intensity 0.49) achieved the highest performance levels in the experimental range. The model successfully predicts that performance peaks in the moderate intensity range and degrades with excessive augmentation.

Table 1. Comprehensive performance analysis across augmentation strategies

Method (Intensity Score)	Val Acc (%)	F1-Score	Training Time (s)	Overfitting gap (%)
No Augmentation (0.0)	77.49	0.774	650.4	3.54
Basic (0.49)	79.84	0.797	1255.6	-1.56
Light Advanced (0.09)	78.80	0.786	342.5	-0.28
Moderate Advanced (0.51)	75.59	0.754	341.7	-4.77
Strong Advanced (0.94)	71.64	0.714	343.5	-13.06
AutoAugment Style (0.98)	74.01	0.737	342.2	-6.83

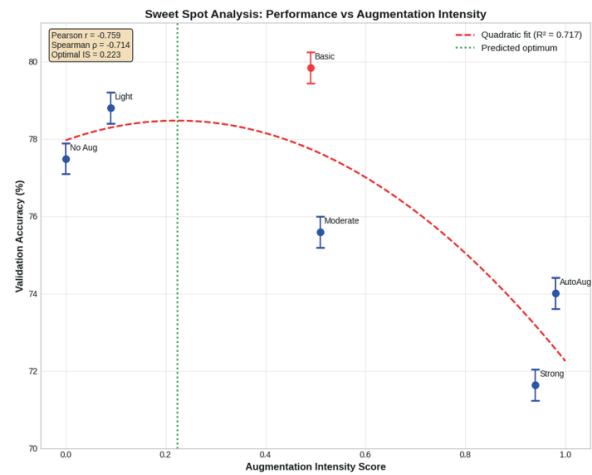


Figure 2. Performance vs. Augmentation Intensity

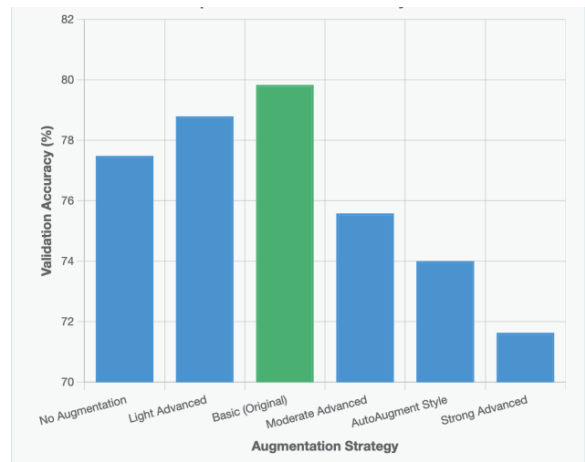


Figure 3. Comprehensive performance comparison across all augmentation methods

Pearson correlation analysis revealed a strong negative linear relationship ($r = -0.759$, $p = 0.080$), while Spearman rank correlation showed an even stronger monotonic relationship ($\rho = -0.714$, $p = 0.111$). Kendall's tau correlation ($\tau = -0.467$, $p = 0.272$) further supports this trend. Although the p-values do not reach statistical significance due to the small sample size ($n=6$), the consistent negative correlation pattern across multiple correlation measures strongly supports the inverted-U hypothesis.

Cohen's d effect sizes demonstrate the practical significance of our findings:

- Basic vs. Strong Advanced: $d = 20.50$ (Very large effect) - representing an 8.2 percentage point difference

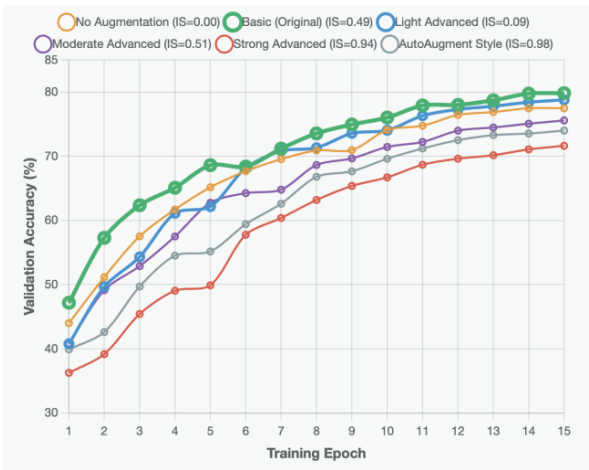


Figure 4. Learning curve comparison

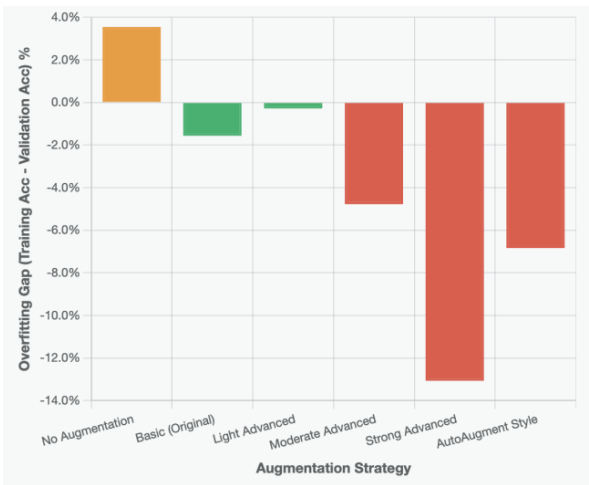


Figure 5. Overfitting Gap Analysis

- Light Advanced vs. Strong Advanced: $d = 18.58$ (Very large effect) - 7.16 percentage point difference
- Basic vs. No Augmentation: $d = 5.86$ (Large effect) - 2.35 percentage point improvement
- Light Advanced vs No Augmentation: $d = 3.27$ (Large effect) - 1.31 percentage point improvement

These large effect sizes indicate that the differences between augmentation strategies have substantial practical implications beyond statistical significance.

Figure 3 provides a comprehensive performance comparison across all augmentation methods, clearly highlighting the Basic strategy's superiority and the performance degradation pattern as intensity increases beyond the optimal range. The data clearly

Table 2. Intensity Framework Validation Through Component Analysis

Intensity Range & Strategies	Performance	Transformation Diversity	Parameter Impact
Baseline (0.0): No Augmentation	77.49%	0 transformations Only preprocessing: Resize to 224×224, ImageNet normalization No regularization benefit	Maintains original data fidelity but lacks regularization capacity, leading to overfitting on training data
Light (0.09)	78.80%	2 transformations Conservative diversity: HorizontalFlip (p=0.5), Random Brightness Contrast (p=0.3) Minimal but effective regularization	Moderate parameters balance regularization and stability
Optimal (0.49): Basic	79.84%	3 transformations Optimal diversity balance: RandomHorizontalFlip (p=0.5), RandomRotation ($\pm 10^\circ$), ColorJitter (brightness, contrast, saturation ± 0.2) Perfect regularization-performance trade-off	Moderate parameters Rotation $\pm 10^\circ$, color jitter ± 0.2 range achieves optimal balance between regularization effectiveness and learning stability
Moderate (0.51): Moderate Advanced	75.59%	4 transformations Increased complexity: HorizontalFlip (p=0.5), ShiftScaleRotate (p=0.4), Random Brightness Contrast (p=0.4), Hue Saturation Value (p=0.3) Complexity begins to create interference	Aggressive parameters Shift/scale ± 0.1 , rotation $\pm 15^\circ$, HSV modifications create increased parameter ranges that start introducing instability
Heavy (0.94-0.98): Strong Advanced, AutoAugment Style	71.64%-74.01%	5-6 transformations Excessive complexity: Multiple geometric transforms, destructive elements (CoarseDropout, GaussNoise), Complex photometric (GridDistortion, RandomGamma) Overwhelming learning capacity	Aggressive parameters Rotation $\pm 25^\circ$, noise injection, aggressive parameter ranges ($\pm 0.2+$) distort data distribution beyond model's learning capacity

show that performance peaks at moderate intensity (0.49) and degrades substantially as intensity increases beyond this optimal range.

4.3. Learning Curve Analysis

Figure 3 presents the learning curve analysis showing distinct convergence patterns across augmentation strategies. The optimal strategies (Basic and Light Advanced) demonstrate smooth, consistent improvement with minimal oscillation, reaching peak performance efficiently. In contrast, excessive augmentation (Strong Advanced) shows erratic learning patterns with poor final convergence, while insufficient augmentation (No Augmentation) exhibits an early plateauing characteristic of overfitting.

Figure 4 illustrates the training dynamics across different augmentation strategies. The learning curves reveal several critical insights:

- **Convergence Patterns:** Optimal augmentation strategies (Basic, Light Advanced) show smooth, consistent improvement with minimal oscillation.
- **Overfitting Behavior:** No Augmentation exhibits classic overfitting with train-validation gap widening after epoch 8.
- **Underfitting in Strong Augmentation:** Strong Advanced shows consistently poor performance with validation accuracy plateauing around 72%,

indicating the model cannot learn effectively from heavily augmented data.

The overfitting gap analysis (Figure 5) provides crucial insights into the regulatory mechanisms at different intensity levels:

- **Positive gaps** (No Augmentation: +3.54%) indicate overfitting
- **Moderate negative gaps** (Basic: -1.56%, Light: -0.28%) suggest optimal regularization
- **Large negative gaps** (Strong: -13.06%, AutoAugment: -6.83%) indicate underfitting

The color-coded visualization clearly shows that optimal strategies maintain gaps in the green zone (-2% to $+1\%$), while excessive augmentation creates severe underfitting (red zone) and insufficient augmentation leads to overfitting. This analysis confirms that the sweet spot lies in the moderate negative gap range, where augmentation provides sufficient regularization without overwhelming the learning process.

4.4. Intensity Score (IS) Framework Validation

The intensity quantification framework demonstrated strong empirical validation through multiple convergent lines of evidence, confirming its utility as a systematic tool for augmentation analysis and deployment decision-making in robotics applications. The comprehensive component analysis presented in Table 2 reveals several critical insights that validate

the framework's predictive accuracy and practical utility.

The relationship between calculated IS and observed validation accuracy follows the predicted inverted-U pattern with consistency. The framework successfully identifies the optimal intensity range (0.09–0.49) where both Light and Basic strategies achieve superior performance, demonstrating its capability to predict the sweet spot phenomenon.

The transformation diversity analysis reveals diminishing returns beyond three to four distinct transformation types, validating the framework's penalty system for excessive complexity. The baseline strategy with zero transformations provides insufficient regularization, leading to overfitting (+3.54% gap), while the optimal strategies employ two to three transformations that achieve perfect regularization balance. Beyond this range, additional transformations contribute negligible benefits while increasing computational overhead and creating potential interference between different augmentation effects, as evidenced by the performance degradation in Heavy strategies.

Despite the framework's empirical success, several important limitations must be acknowledged. The linear combination of transformation diversity, parameter magnitude, and application probability assumes independent additive effects, which may not capture complex interactions between different transformation types. Additionally, the equal weighting of all transformation types overlooks potential differences in their impact on learning difficulty and model performance. The framework's calibration on CIFAR-10 may require adjustment for other datasets or domains with different characteristics. Nevertheless, the strong correlation between IS and performance outcomes ($R^2 = 0.717$) validates the framework's utility for systematic augmentation research and provides a foundation for more sophisticated intensity quantification approaches in future work.

5. Conclusion

This systematic study provides empirical evidence for optimal data augmentation intensity in deep learning. Our key findings include: (1) A quadratic relationship between augmentation intensity and model performance with optimal intensity around 0.223, (2) superior performance of moderate augmentation strategies over both minimal and excessive approaches, (3) important trade-offs between performance and computational efficiency across augmentation libraries, and (4) clear evidence that excessive augmentation can be counterproductive.

These findings offer evidence-based guidelines for practitioners working with small-scale datasets in robotics and automation contexts, enabling informed decisions about augmentation strategy selection based on specific performance and computational requirements. The comprehensive visualization package provides practitioners with intuitive tools

for understanding augmentation trade-offs and optimizing their deployment strategies.

Future work should extend these findings to broader datasets and architectures while developing standardized metrics for augmentation intensity assessment. Such efforts will contribute to more principled and effective use of data augmentation in deep learning applications.

AUTHOR

Windra Swastika* – Informatics Engineering, Universitas Ma Chung, Malang, 65151, Indonesia, ORCID ID: 0000-0003-4210-7001, e-mail: windra.swastika@machung.ac.id.

*Corresponding author

References

- [1] R. Caruana, S. Lawrence, and C.L. Giles, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping," *Advances in Neural Information Processing Systems*, vol. 13, pp. 402–408.
- [2] C. Shorten and T.M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48 doi: 10.1186/s40537-019-0197-0
- [3] J. Wang and L. Perez, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *arXiv preprint arXiv:1712.04621*, 2017.
- [4] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," *Technical Report*, University of Toronto, 2009.
- [5] S.G. Müller and F. Hutter, "TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 774–782.
- [6] E.D. Cubuk "AutoAugment: Learning Augmentation Strategies from Data," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123.
- [7] Bischl *Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 132, e1484 <https://doi.org/10.1002/widm.1484>
- [8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90.
- [9] P.Y. Simard, D. Steinkraus, and J.C. Platt, "Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis," *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, vol. 2, pp. 958–963.

- [10] K. Chatfield "Return of the Devil in the Details: Delving Deep into Convolutional Nets," British Machine Vision Conference, 2014.
- [11] S.G. Müller, L. Hollmann, and F. Hutter, "Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 13, no. 2, e1484, 2023.
- [12] L. Taylor and G. Nitschke, "Improving Deep Learning with Generic Data Augmentation," IEEE Symposium Series on Computational Intelligence, pp. 1–8.
- [13] D. Hendrycks "AugMax: Adversarial Composition of Random Augmentations for Robust Training," Advances in Neural Information Processing Systems, vol. 34, pp. 237–249.
- [14] H. Zhang, "Mixup: Beyond Empirical Risk Minimization," International Conference on Learning Representations, 2018.
- [15] S. Yun, "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032.
- [16] F. Zhao, J. Huang, and J. Gao, "Data Augmentation for Deep Learning-Based Radio Modulation Classification," IEEE Access, vol. 7, pp. 15713–15722.
- [17] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," IEEE Access, vol. 5, pp. 5858–5869 doi: 10.1109/ACCESS.2017.2696121