Farid Oladega

Enhancing Vision Transformer Performance through Data Augmentation and Sharpness-Aware Techniques: A Comparative Study using RMT on CIFAR-10 and ImageNet

School of Computing Science
Sustainable and Autonomous Systems

Vaasa 2025

UNIVERSITY OF VAASA

School of Computing Science

Author: Farid Oladega

Title of the thesis: Enhancing Vision Transformer Performance through Data Augmentation and Sharpness-Aware Techniques: A Comparative Study using RMT on CIFAR-10 and ImageNet

Degree: Master of Science

Discipline: Sustainable and Autonomous Systems

Supervisor: Jani Boutellier, Masud Fahim

Year: 2025 **Pages: 82**

Contents

Abstract

The fast development of transformer architectures has revolutionized computer vision, making it attention-based mechanism, as they substitute the convolutional operations. Nevertheless, inductive biases of Vision Transformers (ViTs) tend to be weaker and more prone to overfitting, even in data-bound scenarios, despite their performance. The thesis explores whether regularization and data-augmentation algorithms that were initially developed in the pre-transformer era can be applied to transformer-based image classifiers, such as Sharpness-Aware Minimization (SAM), CutMix, and Sharpness-Aware Distilled Teachers (SADT). This study tests on the CIFAR-10 benchmark using Vision Retention Network (VisRetNet) and Retentive Networks Meet Vision Transformers (RMT-T) to systematically test the effect of optimizer-level, data-level, and hybrid regularization strategies on model generalization, convergence stability, and computational cost.

Experimental findings indicate that out of the methods analyzed, they all enhanced generalization when working in limited-data settings, with SAM + CutMix scoring the highest validation accuracy (89.39 percent) and SADT exhibiting a better training stability. SAM and SADT provided smaller gains on the largest-scale data (ImageNet) with smoother losses, indicating that large-scale data and prior transformer conditions are regularized. The results prove the hypothesis that the performance of the classic regularization methods is more context specific and is useful when working with small datasets but not as effective in the larger scale.

The paper is a verified PyTorch implementation of SAM and SADT, a hybrid training paradigm that combines data-based and optimizer-based regularization, and empirical data connecting classical principles of optimization in CNNs with transformer models of the present day. These insights support the fact that legacy methods are applicable in the transformer age, providing useful advice regarding how to create efficient, robust, and general vision transformer training pipelines.

1 Introduction

1.1 Background and Motivation

Deep learning revolutionized computer vision, and for the greater part of the last decade, Convolutional Neural Networks (CNNs) have dominated this revolution. During this period, numerous diverse training strategies, such as data augmentation, regularization, and optimization techniques, were developed to improve performance and generalization. Vision Transformers (ViTs) proved to be a very strong rival to CNNs in recent years by replacing convolutional operations with self-attention (Dosovitskiy et al. ,2021). Compared to CNNs, ViTs exhibit weaker inductive biases but require more data and often leading to overfitting when provided with limited training data. This has prompted the question of whether pre-transformer training methods, originally intended to stabilize and regularize CNNs, could be simply adapted for improving ViTs.

The Vision Transformer (ViT) (Dosovitskiy et al. ,2021) demonstrated that transformers can achieve state-of-the-art accuracy on large image benchmarks such as ImageNet when trained with significant data resources. Nevertheless, the self-attention mechanism lacks explicit spatial priors and exhibits quadratic computational complexity, which can restrict its real-world scalability. For addressing such issues, newer studies have suggested architectures such as the Retentive Networks Meet Vision Transformers (RMT) (Fan et al, 2024). which incorporate spatial priors borrowed from the Retentive Network (RetNet) of natural language understanding. RMT is more efficient than ViTs and thus an appropriate choice to investigate transformer generalization.

Among the methods of pre-transformer times, CutMix (Yun et al., 2019) has been implemented successfully as a method of data augmentation. By patch cutting and pasting between training images and properly combining labels, CutMix enhances robustness as well as prevents overfitting. Originally created for CNNs, it is now widely applied in ViT training to augment the effective sample distribution. On the other hand, Sharpness-Aware Minimization (SAM) (Foret et al., 2021) is an optimization method that improves generalization by explicitly searching for flat minima on the loss landscape. Although SAM incurs extra computational cost, it has been shown to have significant improvements on several architectures and benchmarks. Building on this success, Sharpness-Aware Distilled Teachers (SADT) (Fahim & Boutellier, 2022) combine SAM with self-distillation to achieve more powerful student models through teacher and student knowledge transfer within a single training pass.

Despite increased adoption of such methods, a systematic understanding of their trade-offs within the transformer setting remains partial. While CutMix is computationally light and mostly data-driven, SAM improves generalization by increasing training time. SADT introduces complexity by introducing distillation into the training loop. It is not yet known how these methods differ in cases with plenty of data versus tiny amounts of data, or whether their effectiveness in CNNs is transferable to ViTs and their variants, such as RMT.

This motivation of this thesis is to experimentally address the efficiency of pre transformer era data augumentation and regularisation methods and evaluate them on mordern transformer based vision models . Specifically, we look at CutMix, SAM, and SADT, benchmarking their impact on RMT when trained on CIFAR-10 and ImageNet. This research fills the gap between traditional deep learning approaches

and state-of-the-art transformer-based frameworks by systematically investigating such methods.

1.2 Problem Statement

Although Vision Transformers (ViTs) achieve better performance on large-scale benchmarks yet they require large amounts of data and tend to overfit when trained with limited data (Dosovitskiy et al. ,2021). Their reliance on enormous datasets such as ImageNet-21k or JFT-300M is a limiting factor for real-world deployment in resource-scarce environments. To counter these issues, some regularization and augmentation methods such as CutMix (Yun et al., 2019), Sharpness-Aware Minimization (SAM) (Foret et al., 2021), and recently Sharpness-Aware Distilled Teachers (SADT) (Fahim & Boutellier, 2022) were originally derived for convolutional neural networks and have seen steady improvements in generalization. Their effectiveness, however, in the context of transformer-based visual models still hasn't been investigated thoroughly enough.

Therefore, the issue lies in that there is no empirical evidence concerning how pre-transformer regularization and augmentation techniques work when used to state-of-the-art transformers like Retentive Networks Meet Vision Transformers (RMT) (Fan et al, 2024). It is unknown whether or not these techniques still prevent overfitting and enhance generalization without that knowledge, or if unique transformer-specific techniques are needed.

This thesis addresses the research gap by systematically examining the performance of pre-transformer regularization and augmentation techniques specifically SAM, CutMix, and SADT on transformer-based image classifiers (RMT) using CIFAR-10 and ImageNet as benchmark datasets.At least, to our knowledge, the present study provides the first systematic benchmarking of these pre-transformer regularization methods on the Retentive Network (RMT) architecture. Although the influence of

CutMix and SAM on a standard Vision Transformer has been studied in the past (Touvron et al., 2021; Chen et al., 2021), the interaction between the two optimization strategies and the explicit spatial priors brought by RMT has not been studied yet. This thesis bridges that gap.

1.3 Objectives

The basic objective of this thesis is to find out if regularization and augmentation techniques during the pre-transformer era can enhance generalization of transformer-based vision classifiers. The study targets the following specific goals in a bid to achieve that:

1. Establishing baseline performance for RMT (Fan et al., 2024) on CIFAR-10 and ImageNet datasets.
2. The selected pre-transformer techniques, such as CutMix (Yun et al., 2019), Sharpness-Aware Minimization (SAM) (Foret et al., 2021), and Sharpness-Aware Distilled Teachers (SADT) (Fahim & Boutellier, 2022) will be applied to RMT.
3. Compare the performance of these techniques based on the accuracy of their performance, convergence of loss and robustness to overfitting.
4. To compare relative trade-offs of data-driven augmentation (CutMix), optimization-based regularization (SAM), and hybrid models (SADT) when applied to transformer models.
5. To provide evidence-based insights into the enhancement the performance of modern Vision Transformers, and to put a focus on directions for transformer-specific training methods in the future.

1.4 Research Questions / Hypotheses

Main Research Question

How do pre-transformer era regularization and data augmentation approaches perform with transformer-based image classifiers?

To address this, the following sub-questions needs to be answered:

1. How does CutMix (Yun et al., 2019) affect the performance and generalization ability of transformer-based image classifiers when compared to baseline training?

2. To what extent does Sharpness-Aware Minimization (SAM) (Foret et al., 2021) improve convergence and robustness of RMT on CIFAR-10 and ImageNet?

3. Can combining sharpness-aware optimisation and self-distillation, as deployed in SADT (Fahim & Boutellier, 2022), provide better performance than SAM or CutMix alone?

4. Are the effects of these methods different under data-scarce (CIFAR-10) and data-rich (ImageNet) training settings?

Hypotheses

The hypothesis to test to suffieciently answer these research questions are

H1: The selected pre-transformer techniques CutMix, SAM, and SADT will improve the generalization of transformer-based image classifiers when compared to baseline RMT training.

H2: CutMix will yield computationally inexpensive improvements by diversifying the training distribution, particularly on CIFAR-10.

H3: SAM may improve generalization by guiding the optimizer toward flatter minima, although this could increase training time due to its two-step update process.

H4: SADT will outperform CutMix and SAM by combining optimization and distillation based regularization under one umbrella.

H5: The relative performance of such approaches is dataset size dependent (CIFAR-10 vs. ImageNet).

1.5 Scope and Limitations

Scope

This thesis revolves around evaluating pre-transformer era training techniques in the context of transformer-based image classification. That is:

•The study makes use of Retentive Networks Meet Vision Transformers (RMT) (Fan et al., 2024) as the backbone architecture.

•Experiments are conducted on two benchmark datasets of different size

CIFAR-10 (data-constrained) and ImageNet (data-rich).

• The methods under investigation are CutMix (Yun et al., 2019), Sharpness Aware Minimization (SAM) (Foret et al., 2021), and Sharpness-Aware DistilledTeachers (SADT) (Fahim & Boutellier, 2022).

• Performance is measured primarily in terms of the classification of their accuracy, convergence of loss, and generalization ability.

Limitations

While the findings are expected to provide insights into the applicability of legacy techniques for modern transformers, the study is subject to the following constraints:

- Only a subset of pre-transformer methods is considered (CutMix, SAM, SADT); other augmentation and optimization strategies are excluded.

- The experiments are confined to image classification tasks; extension to other tasks like detection, segmentation, or multimodal ones is not addressed.
- Due to computational constraints, only CIFAR-10 and ImageNet are tested; other datasets (CIFAR-100, VTAB, or JFT-300M) are not examined.
- The scope does not include a detailed hyperparameter sensitivity study; methods are tested under commonly reported or reasonable default settings.
- The paper is empirical based rather than developing a new algorithmic contribution.

1.6 Structure of the Thesis

The research contains six chapters that methodically achieve their functions to answer the main research question about transformer-based image classifiers' performance when using pre-transformer regularization and data augmentation techniques.

- Chapter 1 : Introduction

  The first chapter presents background information together with research motivation and contextual framework. The research problem receives definition through this chapter before the authors develop primary and secondary research questions and establish study objectives and boundaries and specify research constraints. The chapter provides an overview of the thesis structure.

- Chapter 2: Literature Review

  Illustrates an overview of Vision Transformers and the Retentive Networks Meet Vision Transformers (RMT) backbone prior to introducing pre-transformer era training techniques like Sharpness-Aware Minimization (SAM), CutMix, and Sharpness-Aware Distilled Teachers (SADT). It also provides associated work on generalization, overfitting, and benchmark testing on CIFAR-10 and ImageNet.

- Chapter 3: Methodology

  Describes the experimental setup, including model selection, dataset preparation, training procedure, and implementation details. Defines how SAM, CutMix, and SADT are achieved in RMT and what performance measures are used.

- Chapter 4: Results

  Reports baseline RMT performance on CIFAR-10 and ImageNet, prior to experimental results on SAM, CutMix, and SADT. Comparison results are contrasted with tables and figures, with highlights of differences and trends across datasets.

- Chapter 5: Discussion

  Explains the findings by comparing the efficacy of each training method's effectiveness and analyzing their trade-offs in convergence, generalization, and computational costs. It also presents limitations and implications for future transformer training approaches.

- Chapter 6: Conclusion and Future Work

  Concisely summarizes the key contributions and findings of the thesis, provides directions on how to select training methods in transformer-based vision tasks, and outlines directions for future work.

2 Vision Transformers (ViTs) and Recurrent Memory Transformers (RMT)

2.1 Transformers

2.1.1 Attention Mechanism

Attention mechanism, the foundation of the transformer model, weights input tokens based on their relevance to compute contextualized representations (Bishop, 2024). The query – key – value (QKV) architecture embodies this approach by comparing queries and keys to get attention scores, which are then used to weight the value vectors to generate output representations (Bishop, 2024).

Formally, self-attention is expressed as:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Multi-head attention uses multiple QKV projections in combination to introduce different representational power, allowing the model to pay attention to different parts of the input at the same time (Bishop et al, 2024).

Self-attention has the need to compute similarity scores between all pairs of tokens, with a quadratic cost in terms of sequence length of $O(N^2)$ (Bishop et al, 2024). This expense becomes a significant obstacle in the case of vision tasks involving high-resolution images. Furthermore, spatial priors that have a vital role to play in representing local structure, are not explicitly represented by attention, in contrast to convolutional networks.

2.1.2 Vision Transformers (ViT)

To adapt the benefits of self-attention to computer vision, (Dosovitskiy et al, 2021) introduced the Vision Transformer (ViT). In ViTs, an image is divided into non-

overlapping patches (16×16 pixels), these are flattened and projected to a fixed-dimensional embedding. Because transformers lack inherent spatial ordering, positional encodings are included in the patch embeddings to maintain spatial information before being processed by a standard Transformer encoder (Bishop, 2024).

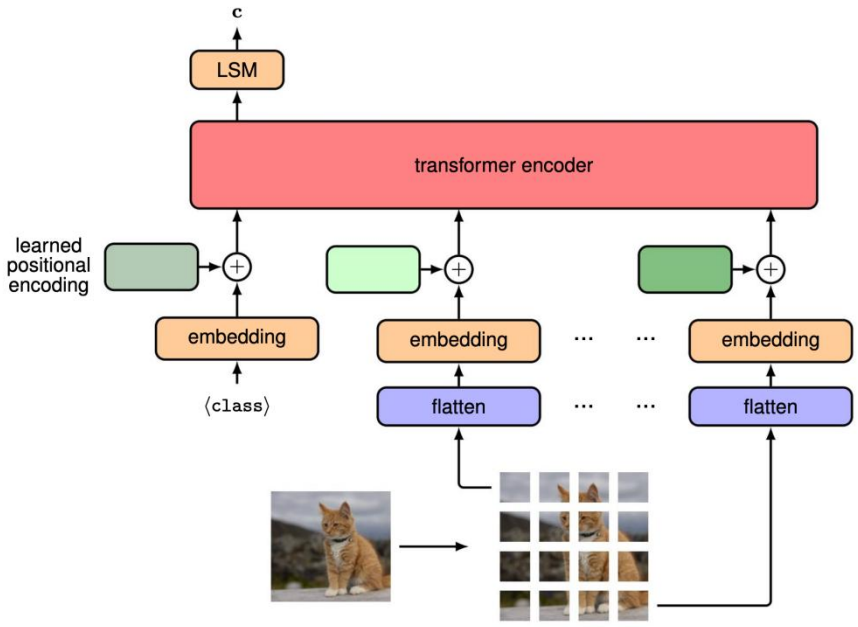


Figure 2.1 Vision Transformer (ViT) architecture. An image is divided into non-overlapping patches, which are flattened, embedded, combined with positional encodings, and processed by Transformer encoder layers before classification. Adapted from Bishop et al, 2024)

The ViT architecture, as illustrated in Figure 2.1 (adapted from Bishop, 2024), consists of the patch embedding layer, positional encodings, Transformer encoder layers, and a classification head.

Although ViTs excel at state-of-the-art when pre-trained on big datasets such as ImageNet-21k and JFT-300M, they require large amounts of data and tend to overfit on tiny data due to their relatively poor inductive biases compared to CNNs (Bishop, 2024; Dosovitskiy et al., 2021). In this context, structural modifications that introduce explicit priors and increase efficiency become imperative.

2.1.3 Retentive Networks Meet Vision Transformers (RMT)

To overcome the deficiencies of ViTs in scalability and inductive bias, (Fan et al., 2024)introduced Retentive Networks Meet Vision Transformers (RMT). RMT is a variant of the Retentive Network (RetNet), which incorporated explicit temporal priors for sequence modeling by introducing a distance-dependent temporal decay matrix.

By building a two-dimensional bidirectional spatial decay matrix based on Manhattan distance among tokens, RMT generalizes the idea in the spatial domain. This approach favors tokens closer to the target token but gives greater degradation to tokens at greater distances. The vision backbone is thus given explicit spatial priors through the use of Manhattan Self-Attention (MaSA).

The two predefined Manhattan distance matrices in the horizontal and vertical spatial dimensions are done as D_H and D_W respectively in Equations (2)-(4). These matrices encode the pairwise distance between the position of tokens in space, and are used as explicit spatial priors in the MaSA mechanism. The use of D_H and D_W makes RMT biased to the computation of attention to spatial locality with computational efficiency due to the factorized horizontal-vertical decomposition.

To improve computational efficiency, RMT introduces a decomposition strategy, factorizing MaSA into horizontal and vertical components:

$$\text{Attn}_H = \text{Softmax}(Q_H K_H^{\top}) \odot D_H \tag{2}$$

$$\text{Attn}_W = \text{Softmax}(Q_W K_W^{\top}) \odot D_W \tag{3}$$

$$\text{MaSA}(X) = \text{Attn}_H(\text{Attn}_W V)^{\top} \tag{4}$$

$$X_{\text{out}} = \text{MaSA}(X) + \text{LCE}(V) \tag{5}$$

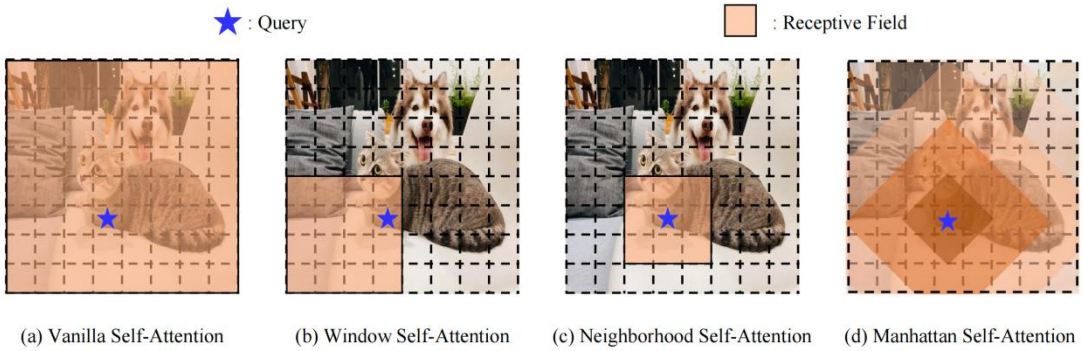


Figure 2.2 Comparison of self-attention mechanisms: (a) Vanilla Self-Attention, (b) Window Self-Attention, (c) Neighborhood Self-Attention, and (d) Manhattan Self-Attention (MaSA). MaSA introduces explicit spatial priors by applying distance-dependent decay. Adapted from (Fan et al., 2024).

where local fine-grained patterns are obtained by the Local Context Enhancement (LCE) module using depthwise convolution. This decomposition notably decreases the computational cost without losing the receptive field of full MaSA.

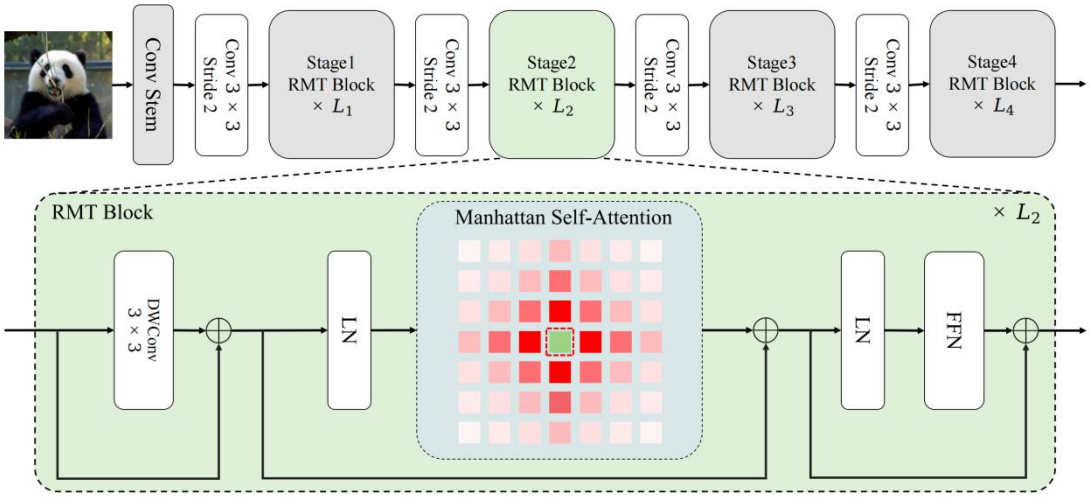


Figure 2.3 Overall architecture of Retentive Networks Meet Vision Transformers (RMT). Adapted from (Fan et al., , 2024).

The RMT architecture in Figure 2.3 (Fan et al., 2024) consists of four steps: the first three steps utilize decomposed MaSA, and the fourth one utilizes the full MaSA. This hierarchical structure provides an efficiency vs. representational capacity trade-off that directly targets the limitations of standard ViTs.

2.2 Regularization Techniques

Despite modern convolutional networks being overfitting-prone, unlike Vision Transformers (ViTs) and Retentive Networks Meet Vision Transformers (RMT), new architectures are less resistant to overfitting on classical problems. This tendency is not attributed to the size of models or the number of parameters, but rather is largely due to the weak inductive biases of attention-based architectures. ViTs are trained on large training datasets to learn strong priors of locality and translation compared to CNNs, which incorporate strong locality and translation priors. Consequently, ViTs have a higher ability to accommodate training patterns when there is limited data, compared to their generalization to previously unseen data (Dosovitskiy et al., 2021). This has promoted the formulation of various regularization and augmentation plans aimed at maximizing robustness and stabilization of training.

However, the three optimization methods SAM (Foret et al., 2021), CutMix (Yun et al., 2019), and SADT (Fahim & Boutellier, 2022) have shown the most impact among these techniques. SAM functions at the optimization stage through a modified loss minimization approach that selects solutions from flat areas of the loss landscape. CutMix, by contrast, is a data-level augmentation technique that stitches together image patches and labels to prevent overfitting and improve localization. SADT applies sharpness-aware optimization to knowledge distillation to enable teacher models to distill knowledge more effectively.

This section elaborates on these three approaches comprehensively, covering their mechanisms, empirical results, and implications for promoting the performance and generalization of state-of-the-art Transformer-based vision models.

2.2.1 Sharpness-Aware Minimization (SAM)

Deep networks prefer to find sharp minima in the loss landscape, where small changes in model parameters cause enormous increases in loss. Sharp minima solutions have poor generalization, especially for high-capacity models like Transformers. Sharpness-Aware Minimization (SAM) by (Foret et al., 2021) addresses this issue by intentionally looking for flat minima, thereby improving generalization.

Strictly, for a training set $S = \{(x_i, y_i)\}_{i=1}^{n}$, the regular training problem minimizes the empirical risk:

$$L_S(w) = \frac{1}{n}\sum_{i=1}^{n} l\,(w, x_i, y_i) \tag{6}$$

where w are the model parameters and $l(\cdot)$ is the loss function. SAM generalizes this by introducing an adversarial perturbation ϵ in a local neighborhood of radius ρ:

$$\min_w L_S^{\text{SAM}}(w) \quad \text{where} \quad L_S^{\text{SAM}}(w) = \max_{\|\epsilon\|_p \le \rho} L_S\,(w + \epsilon) \tag{7}$$

The inner maximization finds the parameter perturbation to maximize loss most in the local neighborhood, and the outer minimization encourages robustness by optimizing w against said worst case loss. This inclines the model to favor flatter solutions.

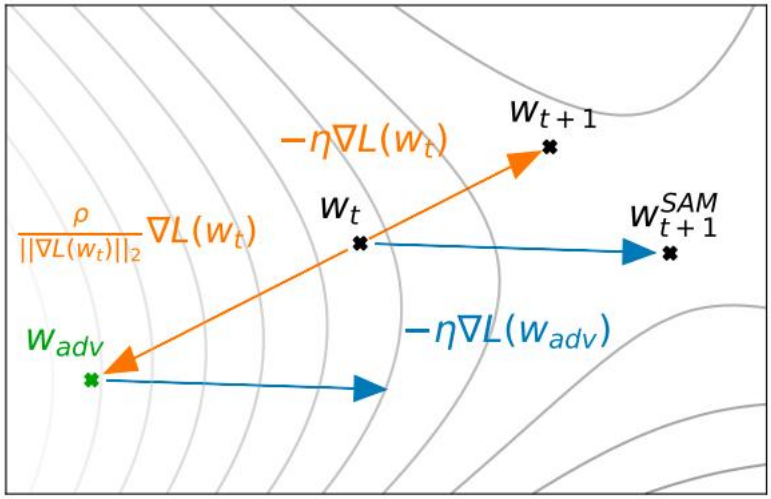
To estimate the perturbation, the SAM paper proposes the following update:

$$\epsilon^*(w) \approx \rho \frac{\nabla_w L_S(w)}{\|\nabla_w L_S(w)\|_2} \tag{8}$$

with the direction of the perturbation along the direction of the normalized gradient.By substituting the perturbation into the objective, the final SAM update is calculated based on the perturbed gradient:

$$\nabla_w L_S^{\text{SAM}}(w) \approx \nabla_w L_S(w + \epsilon^*) \qquad (9)$$

This two-step update process is described in Algorithm 1 of the SAM paper (Foret et al., 2021).



- *Figure 2.4 Schematic of the SAM parameter update, Adapted from (Foret et al., 2021)*

The most recent update for SAM can be expressed as an approximation of the gradient at perturbed weights $w + \epsilon^*$ in equation (9).

This definition shows that SAM need not include explicit second-order terms from the Hessian, thereby making the method computationally easy to integrate in standard deep learning frameworks such as PyTorch or TensorFlow. By excluding higher-order corrections, the authors demonstrate that SAM can be integrated into the training loop with only modest overhead relative to standard optimizers.

Algorithm 1 in the SAM paper (Foret et al., 2021) describes the process using pseudocode. In practice, the process begins with the normal calculation of the loss gradient at the current weights w_t. An adversarial perturbation ϵ^* is then calculated from equation (8), which moves the parameters to a neighboring location where the

loss is maximized within a constrained neighborhood. The gradient of the loss is then re-computed at such perturbed weights, giving a stable estimate of the SAM objective. Finally, the parameters are updated according to this perturbed gradient, and one iteration of the optimization process is done.

The process is also illustrated in Figure 2. The figure shows the relationship between the original weights w_t, the perturbed weights $w_t + \epsilon^*$, and the final updated weights under SAM, w_{t+1}^{SAM}. The orange arrow shows the direction of adversarial perturbation, which moves the weights toward a region with higher loss values. The blue arrow represents the gradient-based update that occurs at this perturbed point. In contrast to standard SGD, which updates parameters only according to the gradient of the current point, SAM explicitizes this adversarial step. The result is that the optimization trajectory is repelled from sharp minima and attracted towards flatter regions of the loss surface, which are empirically connected with better generalization.

2.2.2 CutMix

Data augmentation is one of the most widespread approaches to improving the generalization of deep neural networks by artificially making training samples more diverse. Traditional approaches such as Cutout (DeVries et al, 2017), which mask out portions of an image, whereas Mixup (Zhang et al., 2018), which combines two entire images. Each of these approaches improves robustness at the cost of drawbacks: Cutout loses potentially useful pixels, and Mixup produces samples that may appear visually unrealistic.

CutMix (Yun et al., 2019) solves these issues by filling a randomly cropped region of one image with a patch from another image, and also mixing the training labels proportionally to the mixed regions. This allows CutMix to retain semantic information from both images, generating more realistic and informative training samples.

Formally, for two training samples (x_A, y_A) and (x_B, y_B), the CutMix operation generates:

$$\widetilde{x} = M \odot x_A + (1 - M) \odot x_B \qquad (10)$$

$$\widetilde{y} = \lambda y_A + (1 - \lambda) y_B \qquad (11)$$

where $M \in \{0,1\}^{W \times H}$ is a binary mask indicating the cropping region, λ is the ratio of area of the retained region from image x_A, and $\odot$ is element-wise multiplication. Compared with Cutout, in which masked regions are filled with zeros, CutMix fills masked regions with semantic content borrowed from another sample.

The binary mask M is created by first sampling a rectangle with parameters (r_x, r_y, r_w, r_h), where r_w and r_h are in proportion to $1 - \lambda$. This ensures the patch size encodes the proportion of label mixing. Both images thus both contribute towards the label as well as the input, forcing the model to learn stable features even when there is occlusion (Yun et al., 2019).

2.2.3 Sharpness-Aware Distilled Teachers (SADT)

Sharpness-Aware Minimization (SAM) and self-distillation are philosophically related in that they both seek to improve generalization with control over the geometry of the loss landscape. SAM is encouraging flat minima convergence, while self-distillation is reducing the discrepancy between several copies of the same model. The Sharpness-Aware Distilled Teacher (SADT) method, by (Fahim and Boutellier, 2022), brings these two perspectives under the same umbrella.

Mathematically, given a training set $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, with model parameters $w \in \mathbb{R}^{|w|}$, the sharpness term can be written as:

$$s(w, S) = \max_{\|\delta\| \leq \rho} \frac{1}{|S|} \sum_{(x_i, y_i) \in S} l_i(w + \delta) - l_i(w) \qquad (12)$$

where ρ is the size of the neighborhood and $l_i(w)$ is the loss calculated on sample (x_i, y_i). This quantifies how much the training loss can be exacerbated under perturbations of the model parameters, and sharper minima correspond to larger values of the sharpness (Fahim and Boutellier, 2022).

In the case of the distillation term, divergence between two instances $w \; and \; w'$ of the parameters is measured by the expected Kullback – Leibler divergence:

$$d_p(w, w') = \mathbb{E}_x\left[D_{\mathrm{KL}}\left(f(x; w) \parallel f(x; w')\right)\right] \qquad (13)$$

SADT takes advantage of this formulation by establishing an auxiliary teacher model through parameter perturbations. In particular, a perturbed teacher $f(x; w_{\mathrm{aux}})$ is established by adding Gaussian noise to the weights w_{up} of the updated self-teacher model. The logits of the perturbed teacher are utilized to guide the student to generate soft-labels for the final backpropagation.

2.3 Generalization and Overfitting in Deep Learning

2.3.1 Generalization

The main objective of any machine learning model is not merely to memorize the training data, but it must learn from training data and then successfully predict new examples that stem from the same distribution. A model is said to generalize well when it identifies fundamental patterns and data structures that represent the core nature of the data instead of its random fluctuations or unique characteristics. Practically, the phenomenon of generalization happens when the performance of a model on unseen data is similar to its performances in the training set (Bishop, 2024).

Good generalization also needs the model capacity and inductive bias implemented in its architecture. Inductive bias refers to the set of assumptions the learning algorithm uses for making predictions for inputs it has never seen (Bishop, 2024). Convolutional neural networks, for example, implement strong priors over spatial organization such as translation invariance, while transformers rely on the attention mechanism to realize flexible long-range dependencies. These prejudices help guide the learning process towards likely solutions regardless of having limited data.

However, generalization is inductive. Even if a model is best possible trained, it is not known to generalize perfectly because the training set is only a finite set of observations from the input distribution. This leads to the consideration of theoretical frameworks, such as the bias – variance trade-off, for characterizing what happens when complexity increases. Classical statistical learning theory emphasizes that over-fitting threatens to cause bad generalization, a point to be pursued further in the subsequent subsections.

2.3.2 Overfitting

Deep learning models strive to achieve generalization but they often develop overfitting problems. The model learns from both the underlying structure and random noise, as well as incidental variations in the training data, which results in overfitting. The model achieves high training accuracy, but its performance declines when applied to new or unseen data (Bishop, 2024).

The training and validation loss values become different from each other as the training process continues which indicates overfitting. The two loss values decrease in unison during the initial stages of training. The model starts to memorize individual training examples which causes validation loss to increase while training loss keeps decreasing.

The root causes of overfitting correlate directly with model capacity. However, Modern deep networks such as Vision Transformers (ViTs) possess extremely high representational capacity and, in theory, can fit random labels to a perfect fit. Without

sufficient regularization or a sufficiently large dataset, this capacity is a blessing-turned-curse, inducing solutions that fail to generalize. Overfitting can also be induced by noisy labels, class imbalance, or having insufficient data diversity.

Moreover, the prevention of overfitting can be mitigated by three methods, which include weight decay, early stopping, and data augmentation. The loss function includes a penalty term to discourage large weights which prevents complicated models from being favored (Bishop, 2024). Early stopping terminates training if validation loss no longer improves, such that the network would not have learned training artifacts. Augmentation methods such as CutMix or Mixup add to the effective dataset by altering training images and thus reducing the overfitting behavior of the model.

Overfitting is particularly relevant in this thesis as the methods being compared Sharpness-Aware Minimization (SAM), CutMix, and Sharpness-Aware Distilled Teachers (SADT) all present themselves as a type of regularization techniques. Each of these methods attempts to guide training towards flatter minima or heterogeneous representations, therefore explicitly counteracting overfitting but maintaining reasonable generalization performance.

2.3.3 Generalization Perspectives and Regularization

A central concept in machine learning is the bias – variance trade-off, which represents the tension between underfitting and overfitting. Very biased models underfit the data and create overly simple decision boundaries, whereas very variable models overfit and learn training examples by heart but fail to generalize to new examples (Bishop, 2024). Mathematically, the generalization error can be represented as three terms:

$$\mathbb{E}\left[(y - \widehat{f}(x))^2\right] = \text{Bias}^2 + \text{Variance} + \sigma^2 \qquad (14)$$

where the irreducible noise term σ^2 defines the lower limit on error attainable. Generalization to success, therefore, requires a tradeoff between bias and variance control by architecture and regularization. The simple approach of early stopping in regularization techniques helps find an optimal balance between preventing overfitting and maintaining enough representational power (Bishop, 2024).

Although the bias-variance trade-off provides us with a classic view of generalization, more recent deep learning phenomena have revealed more nuanced behaviors. In particular, the double descent phenomenon (Belkin et al., 2019) illustrates that test error can first decrease with model capacity, then increase due to overfitting, and then decrease again as models are very large (Bishop et al., 2024). This defies the traditional supposition that larger models always overfit, rather to show that large enough networks, with equivalent optimization, can generalize well despite their capacity.

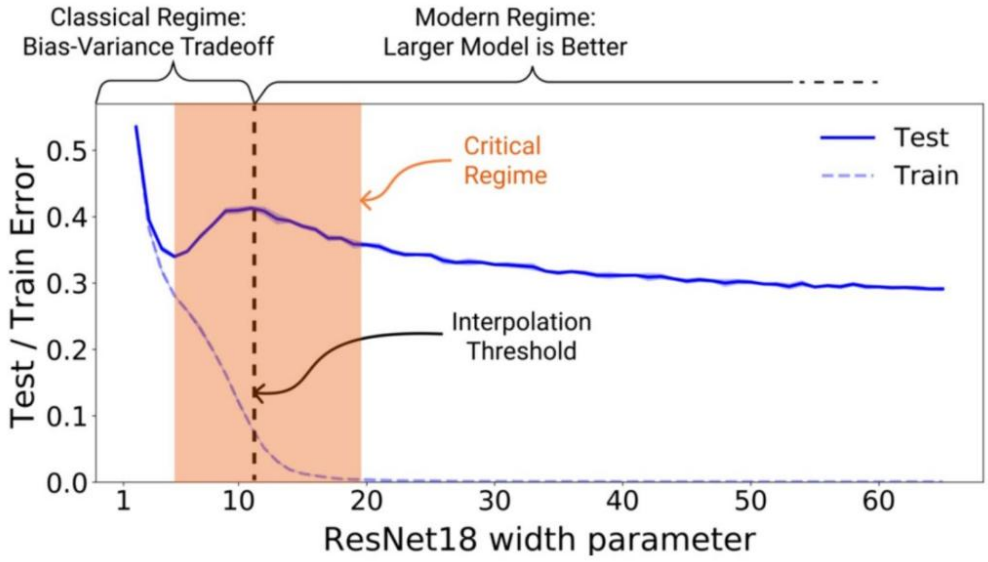


Figure 2.6 Double Descent Phenomenon (Bishop et al, 2024)

This modern approach of course also aligns with regularization, providing us with handles to manage the bias – variance trade-off and map complex error landscapes. Classic regularization techniques such as weight decay impose penalties on the sizes of parameters:

$$E(w) = E_0(w) + \frac{\lambda}{2}\|w\|_2^2 \qquad (15)$$

with the original loss $E_0(w)$ and regularization strength λ (Bishop, 2024). Analogously, dropout puts neurons in a random inactive state with probability p, essentially performing model averaging over an exponential number of thinned networks(Srivastava et al., 2014):

$$\tilde{h}_i = h_i \cdot r_i, \;\; r_i \sim \text{Bernoulli}(p) \qquad (16)$$

where h_i, r_i is the activation of neuron and is a binary mask (Bishop, 2024). Structural methods such as parameter sharing also reduce overfitting by restricting groups of weights to share values, thus offering inductive biases that favor the data (Bishop, 2024).

In addition to these traditional techniques, newer methods directly attack sharpness in the loss landscape. Sharpness-Aware Minimization (SAM) (Foret et al., 2021) directly minimizes parameters within flat regions of the loss by solving Equation (7).

where ρ is employed to control the perturbation radius. Concurrently, Sharpness-Aware Distilled Teachers (SADT) (Fahim and Boutellier, 2022) integrate SAM with knowledge distillation, minimizing a KL-divergence term between perturbed and unperturbed teacher models in equation (13).

These methods encourage convergence to flat minima that generalize more. CutMix (Yun et al., 2019) is another data-augmentation strategy that applies spatial regularization by incorporating training image patches and labels, effectively preventing models from relying excessively on local features.

However, generalization in deep learning is best understood as a continuum: from classical bias – variance trade-offs (Geman et al., 1992), through modern breakthroughs such as double descent (Belkin et al., 2019), to advanced regularization techniques that blend structural, stochastic, and sharpness-aware approaches. These

techniques collectively form the foundation for training robust models that generalize to state-of-the-art vision model capacity requirements.

3 Methodology

3.1 Theoretical Discussion: SAM and Recent Extensions

SAM has various limitations despite its success in empirical practice. Most significantly, it is sharpness scale dependent: the parameters of the scaling algorithm change the sharpness measured, despite the fact that the underlying generalization behaviour remains the same. This limitation was already observed in earlier work on sharpness measures (Dinh et al., 2017), which shows that sharpness can change under simple parameter rescalings. To this end, Kwon et al. (2021) suggested Adaptive Sharpness-Aware Minimization (ASAM). Rather than working with a fixed radius in parameter space, ASAM proposes to use a normalization operator that locally rescales perturbations to eliminate scale dependence. So the objective function is modified as:

$$L_{\text{ASAM}}(w) = \max_{\|\epsilon\| \leq \rho} L\left(w + T_w \epsilon\right) \qquad (17)$$

where T_w is a normalization operator that rescales the parameters in accordance with their magnitude such that the notion of flatness does not depend on scaling parameters. Empirical evidence provided by Kwon et al. (2021) shows that ASAM is stable and outperforms SAM on CIFAR and ImageNet, especially when trained on an architecture that is sensitive to weight scale, e.g., ResNets and WideResNets.

ASAM solves the problem of scale-dependence but not optimization inefficiency. Both SAM and ASAM involve computing the gradients at perturbed weights, which doubles the computation cost in comparison with SGD. Sun et al. (2023) were able to address this problem by suggesting AdaSAM that add a learning rate schedule and momentum acceleration to SAM. AdaSAM stabilizes training instability by dynamically setting the step size and momentum based on the state of the optimization process and preserves

the generalization performance of SAM. AdaSAM update rule includes an intelligent perturbation gradient:

$$g_t = \nabla_w L(w_t + \epsilon_t) \tag{18}$$

$$\epsilon_t = \rho \frac{\nabla_w L(w_t)}{\|\nabla_w L(w_t)\|} \tag{19}$$

and an update of parameters that accelerates convergence is based on momentum. Experimental assessments on the GLUE benchmark (Sun et al., 2023) showed AdaSAM is more effective than both SAM and ASAM and also competitive with recently popular adaptive optimizers like AdamW, especially when considered on a large scale NLP model. However, A similar set of research changes the geometry of the neighborhoods as specified in SAM.In the case of FisherSAM (Kim et al., 2022), instead of using Euclidean neighborhoods of SAM, the ellipsoids are specified by the Fisher information matrix, which yields perturbations that are consistent with the statistical structure of neural networks.

According to this formulation, the geometry of the optimization is driven by information content - parameters with greater Fisher information count more towards the perturbation, and hence the concept of flatness is refined in a statistically significant sense.

The official formulation of the FisherSAM goal is the following:

$$L_{\text{FisherSAM}}(w) = \max_{\epsilon^T F(w)\epsilon \leq \rho^2} L(w + \epsilon) \tag{20}$$

$F(w)$ the Fisher information matrix at parameters w. This adjustment brings the perturbation set in agreement with the intrinsic geometry of the parameter space of the model, which gives more principled robustness guarantees. Experimental findings on CIFAR and ImageNet corroborate the experimental results on both SAM and

ASAM that FisherSAM has a better generalization as well as suffers only a low computational cost.

Collectively, these extensions are an indication of the fast development of the SAM framework: SAM (regularization by flatteness), ASAM (regularization by scale invariance), AdaSAM (perturbation-efficient adaptive optimization), and FisherSAM (perturbations with information geometry). Each of them addresses a particular limitation of the founding formulation, but there is an open question of whether the benefits of the SAM-family are consistently extended to various architectures including convolutional networks (CNNs) (Krizhevsky et al., 2012), vision transformers (ViTs)(Dosovitskiy et al., 2021) and new state-space models such as VMamba (Liu et al., 2024), RWKV (Peng et al., 2023) and RMT (Fan et al., 2024). This inspires the research gap that this work is addressing and will be extended upon in the next section.

3.2 Research Gap

Sharpness-Aware Minimization (SAM) is a typical technique to locate flatter minima in the loss landscape (Foret et al., 2021). Experimental results have been obtained on CIFAR-10, CIFAR-100, and ImageNet which have shown the dominance of SAM across the convolutional neural networks (CNNs)(Krizhevsky et al., 2012), such as ResNet (He et al, 2016) and PyramidNet. However, SAM has weaknesses. It is two times more expensive than the cost of computation in a single iteration since gradients have to be computed on both perturbed and original weights. In addition, its sharpness measure can be reparameterized or weight-scaled (Dinh et al., 2017), which questions its usefulness as a generalization proxy.

In order to overcome these problems, several extensions of SAM were suggested. The proposed method Adaptive SAM (ASAM) presents a scale-invariant neighborhood,

which scales perturbations based on the magnitude of the parameter, which reduces rescaling sensitivity (Kwon et al., 2021). To minimize the computational cost, AdaSAM uses adaptive learning rates and momentum to stabilize convergence of large-scale models (Sun et al., 2023). FisherSAM reinvents the local geometry of perturbations by the use of the Fisher information matrix, and invents a statistically significant ellipsoid in lieu of the Euclidean neighborhood of SAM (Kim et al., 2022). Most recently, Sharpness-Aware Distilled Teacher (SADT) builds upon the SAM paradigm to adopt self-distillation, applying sharpness-aware perturbations to the teacher network, which also achieves better results with data limited regimes (Fahim & Boutellier, 2022). Although both versions address a certain shortcoming of SAM, none has been proven to outperform its peers universally.

Recent research has studied the adaptation of the CNN-era regularization to Transformers. Touvron et al. (2021) established that such data-level augmentations as CutMix and Mixup are the key to training Data-efficient Image Transformers (DeiT) without pre-training on the large-scale level. Likewise, Chen et al. (2021) measured the effect of SAM on Vision Transformers and discovered that sharpness-aware optimization can somewhat address the absence of inductive bias in conventional self-attention schemes. It is still unclear whether the regularization benefits obtained through the establishment of Retentive Networks (RMT) that re-introduce explicit spatial priors through decay matrices are maintained or rendered unnecessary (Fan et al., 2024).

The vast majority of the previous analysis has been on CNNs, and their overwhelming local inductive biases favor generalization (Bishop, 2024). On the other hand, Vision Transformers (ViTs) do not have such biases and tend to overfit unless over-regularized (Dosovitskiy et al., 2021). Theoretically, SAM-family optimizers would benefit ViTs through encouraging flatter minima and overfitting by regularizing the gradient. Retentive Networks Meet Vision Transformers (RMT; Fan et al., 2024) are an extension of the transformer family that uses explicit spatial priors by means of

Manhattan Self-Attention (MaSA), potentially reducing our motivation to optimize the smoothness of the optimization landscape, potentially reducing the marginal utility of SAM. Similarly, Vision Mamba (VMamba) (Liu et al., 2024) uses state-space sequence modeling and implicit stability priors, which offers efficient learning and may eliminate the necessity to use sharpness-conscious perturbations. Thus, the main one of this paper is that SAM has a significant effect on improving generalization in CNNs and ViTs but its advantage can be extraneous or diluted in models like RMT and VMamba that already represent the idea of stability with the help of structural priors.

The second source of uncertainty is the interaction between data-level and optimizer-level regularization. Implicit regularizers like CutMix (Yun et al., 2019) and Mixup already are data augmentations which broaden the effective training distribution. It is also not known whether SAM does add any extra benefits to such augmentations or whether the regularization effects are saturated.

3.3 Model Selection

The experimental analysis provided by this thesis centers on Retentive Networks Meet Vision Transformers (RMT), a novel extension of the Vision Transformer family that addresses the shortcomings of self-attention on visual tasks. The reasons behind the selection of RMT as the experimental backbone are twofold: (i) it offers explicit spatial priors through its formulations, which makes it a suitable subject to determine whether SAM-family optimizers offer any advantage or become unnecessary; and (ii) it is a part of a wider tendency of adding inductive biases to transformer-based models and thus it serves as a bridge between the traditional CNN priors and the entirely attention-based ones.

Context of CNNs and ViTs. SAM and its variants have been previously trained on convolutional neural networks (CNNs), including ResNet and PyramidNet, in which flatter minima always enhanced the performance in generalization training (Foret et

al., 2021). Vision Transformers (ViTs) lack these powerful inductive biases and, thus, are far more data-dependent and overfeasible (Dosovitskiy et al., 2021).The past has demonstrated that SAM-style optimizers can enhance the work of token-mixing and transformer-based architectures. To illustrate, Sharpness-Aware Distilled Teachers (SADT) achieved steady improvements on ViT models (Fahim and Boutellier, 2022), and ASAM had been improved on Transformer and MLP-Mixer architecture (Kwon et al., 2021). Moreover, it has also been reported that cutMix can be useful in generalizing the transformer-based model (Yun et al., 2019).This implies that one should investigate whether sharpness-sensitive training is just as useful in other architectures with different inductive priors.

RMT as the chosen backbone. RMT builds on ViT by adding Manhattan Self-Attention (MaSA) which incorporates prior information about the space as explicit spatial priors through a distance-depending decay matrix (Fan et al., 2024). Also, RMT proposes a decomposition methodology that breaks down the self-attention into horizontal and vertical terms to cut down on the computation cost and yet preserves global context. These changes bring out a vision backbone that is computationally efficient and structurally skewed towards the maintenance of spatial relationships. Because SAM-family optimizers are made to maximize generalization, through the encouragement of flat minima, there is a special case of RMT: do we even need such optimizers when the architecture itself is already stable and inductive?

In order to operationalize the research questions , this study is based on a controlled empirical study of Sharpness-Aware Minimization (SAM) and its variants in the context of Retentive Networks Meet Vision Transformers (RMT).

The design of the entire research is set in such a way that it follows the following key actions:

The RMT was implemented at the center of attention:

Train and deploy the Retentive Networks Meet Vision Transformers (RMT) model, based on its explicit spatial priors using the Manhattan Self-Attention (MaSA) module, to be the only experimental backbone to test SAM-family optimizers.

Moreover, the intervention might possess regularization: how often does the intervention occur?

Regularization Interaction Analysis:

Find out the interaction of SAM-family optimization with data-level regularization (e.g., CutMix and SADT) and whether sharpness-aware perturbations have complementary or redundant effects in the RMT training regime.

Comparison Experiments on ImageNet and CIFAR-10:

Perform standardised experiments on both CIFAR-10 and ImageNet datasets to measure the effect of optimisation using SAM on model accuracy, loss behaviour and efficiency relative to the standard RMT training configuration.

What interpretation of Sharpness-Redundancy Hypothesis:

Test whether the experiment results indicate that the inductive priors of RMT as embodied in its spatially decayed attention and structured receptive fields are already biased towards the flatter minima, which may cause the marginal utility of SAM-family regularization to decrease.

3.4 Data Augmentation Strategy

In order to have a uniform assessment of training strategies, all of the experiments in this thesis utilized well-defined preprocessing pipelines of CIFAR-10 and ImageNet.

In the case of CIFAR-10, each image was brought to a zero mean and unit variance value, and further augmented with random cropping (with 4 pixels padding) and random horizontal flipping, as is customary (Krizhevsky & Hinton, 2009). These processes help reduce overfitting because the model is subjected to different input distributions, which is particularly relevant considering the small resolution of images in the dataset (32x32 pixels).

In the case of ImageNet, preprocessing was done in the ILSVRC protocol (Deng et al., 2009). The training images were randomly scaled and clipped to 224x224 pixels and the ImageNet mean/variance was used to normalize the image. Random horizontal flipping was used as a control augmentation. After resizing the validation images, the center-cropping was done to the same resolution to maintain consistency in assessment. Besides the mentioned baseline augmentations, this research also used CutMix (Yun et al., 2019), a patch based data augmentation method that is aimed at enhancing robustness and generalization. Other than the specified baseline augmentations, CutMix (Yun et al., 2019) was also used in this study as per Section 2.2.2. In this thesis, the alpha parameter used was set to α=1.0, which is according to the recommended parameters of the original CutMix implementation 2. A probability of 0.33 was applied in the CutMix operation.

4 Results

4.1 Baseline Performance on CIFAR-10 and ImageNet

The establishment of the baseline is achieved by the experiments of two transformer-based models, including the Vision Retention Network (VisRetNet), a small-scale model based on the Retentive Networks framework, and the Retentive Networks Meet Vision Transformers (RMT) architecture introduced by Fan et al. (2024). The reason behind choosing such architectures is that they have a similar foundation of transformers with spatial retention mechanisms that can be reasonably compared in terms of generalization behaviors of various data scales. Namely, VisRetNet is a small model that can be used on small datasets (e.g., CIFAR-10), whereas RMT is a complete transformer implementation that was developed to perform well on large datasets (e.g., ImageNet-1K).

The results of these baseline configurations, which are trained without any extra regularization or augmentation, are used to compare the results of further experiments. The remaining parts of this chapter (Sections 4.2 to 4.5) compare the performance of Sharpness-Aware Minimization (SAM), CutMix, and the corresponding variation, such as a hybrid SAM + CutMix and Sharpness-Aware Distilled Teachers (SADT) systems, to these baselines.

4.1.1 CIFAR-10 Baseline (Custom VisRetNet)

Vision Retention Network (VisRetNet) is a small transformer-based architecture which can be used to efficiently classify images in small datasets. It is based on the Retentive Networks Meet Vision Transformers (RMT) framework (Fan et al., 2024) and includes the most important retention mechanisms that improve the capacity of spatial dependency modeling. The VisRetNet of this study is a small size variant with

embedding dimension size [64, 128, 256 and 512] and depth size [2, 2, 8 and 2]. The third unique aspect of this model is that in the initial two stages chunkwise recurrence is employed that allows the network to store spatial information across image patches at the cost of less computational complexity.

The 100 epochs were trained with the Adam optimizer and no weight decay alongside a StepLR learning rate scheduler that decreased the learning rate by a factor of 0.1 after every 10 epochs, initially at 0.001. Every batch of training comprised 128 images and the optimization objective was Cross-Entropy Loss.

The experimental results were obtained using one NVidia RTX 4090 and the whole process of training took about 27 minutes and 48 seconds. By the 100th epoch, the network had a training loss and a training accuracy of 0.3876 and 86.53, respectively, and a validation loss and a validation accuracy of 0.4970 and 82.96 respectively.

In Figure 4.1, The findings show that the VisRetNet achieved a high baseline accuracy of about 83 per cent with just a slight augmentation (random cropping and horizontal flipping). The relative difference of the training accuracy versus validation accuracy of only 3.6 percentage points is a sign of slightly overfitting, which forms an adequate basis to consider more sophisticated regularization strategies like Sharpness-Aware Minimization (SAM)(Foret et al., 2021), CutMix (Yun et al., 2019), and Sharpness-Aware Distilled Teachers (SADT)(Fahim & Boutellier, 2022).

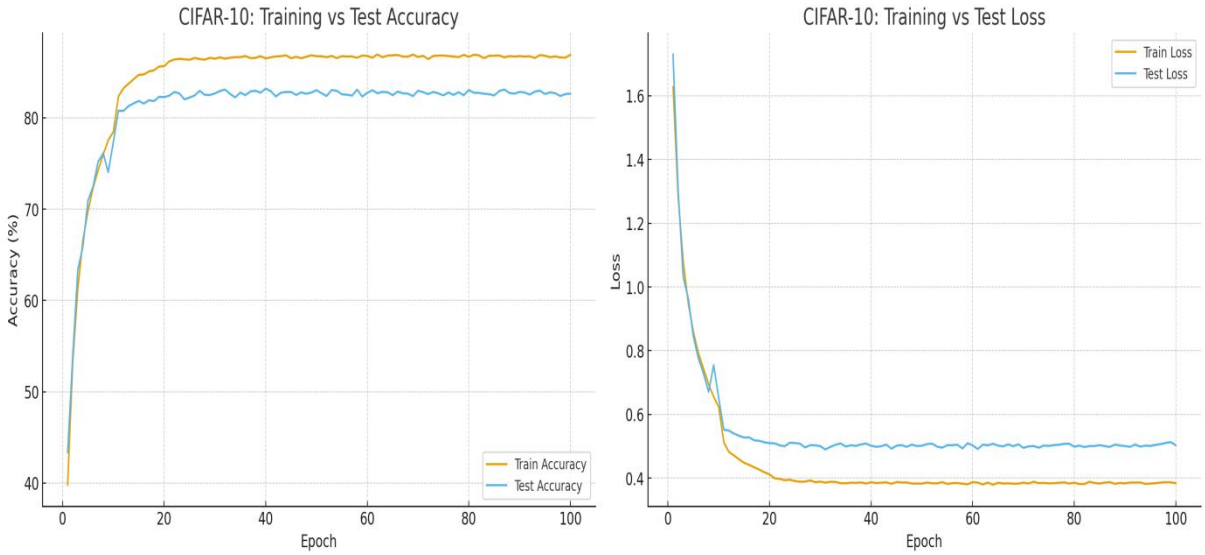


Figure 4.1 Training and test accuracy/loss for CIFAR-10 baseline.

4.1.2 ImageNet Baseline

The Retentive Networks Meet Vision Transformers (RMT-Tiny) model was trained on ImageNet-1K according to the configuration described in (Fan et al, 2024) in order to reach the large-scale reference baseline. RMT-Tiny combines Manhattan Self-Attention (MaSA) and a decomposition mechanism that separates the horizontal and vertical attention elements and offers a clear spatial prior, at the same time being computationally efficient.

This model was also trained on single-gpu hardware with 300 epochs, an AdamW optimizer and weight decay of 0.05 with a drop-path rate of 0.1. A cosine-annealing schedule was used to change the initial learning rate to with a five-epoch warm-up period. The batch size of 64 was used during the training. In order to increase generalization, ImageNet-scale augmentations were used, such as Mixup (0.8), CutMix (1.0), Color jitter (0.4), RandAugment with the setting m9-mstd0.5-inc1, where 'm9-mstd0.5-inc1' denotes a RandAugment configuration with magnitude 9, magnitude-noise standard deviation 0.5, and increment 1, and random erasing with 0.25 probability. The Cross-Entropy Loss objective used label smoothing of 0.1.

The model had a convergence stability in training. Top-1 and Top-5 accuracy in epoch 100 were 74.99% and 93.07, respectively, and EMA was slightly better with Top-1 = 78.79 and Top-5 = 94.73. Training and validation losses were at the same time 3.46 and 1.09 respectively. The final model of 300 epochs yielded Top-1 and Top-5 accuracy of 82.01 and 96.01 when trained on ImageNet (and had a final validation loss of 0.955).Using 4 GPUs, it took an average of 7 days, 4 hours and 11 minutes to complete the full 300-epoch ImageNet training.

The results of that prove the fact that the adopted version of PyTorch can reproduce the accuracy introduced in the official version of the RMT article, which proves the accuracy and reliability of the experiment setup. This consistency is to be developed because it ensures that the occurrence of any variation in performance later in SAM, CutMix, or SADT can be attributed credibly by such means and not implementation bias or model volatility. In other words, this baseline provides a solid reference point in the assessment of the impact of the regularization and augmentation techniques on the transformer-based architectures.

Overall, the learning curves (Figure 4.2) show that there was a slow and consistent increase in the accuracy and loss over the epochs. The learning-rate schedule, which was cosine annealed, also helped achieve smooth convergence during the training. The high final accuracy value shows that the RMT-Tiny model can be extrapolated to the ImageNet-1K dataset, which provides a reliable benchmark to compare the experiments that are presented in the next sections.It is important to mention, that the model achieved a competitive level of accuracy, but the loss curves indicate that the model was continuing to improve very slowly at epoch 300. Further gains may be obtained with a longer training horizon, but this was not further developed because the next SAM and SADT experiments were prohibitively difficult to compute.

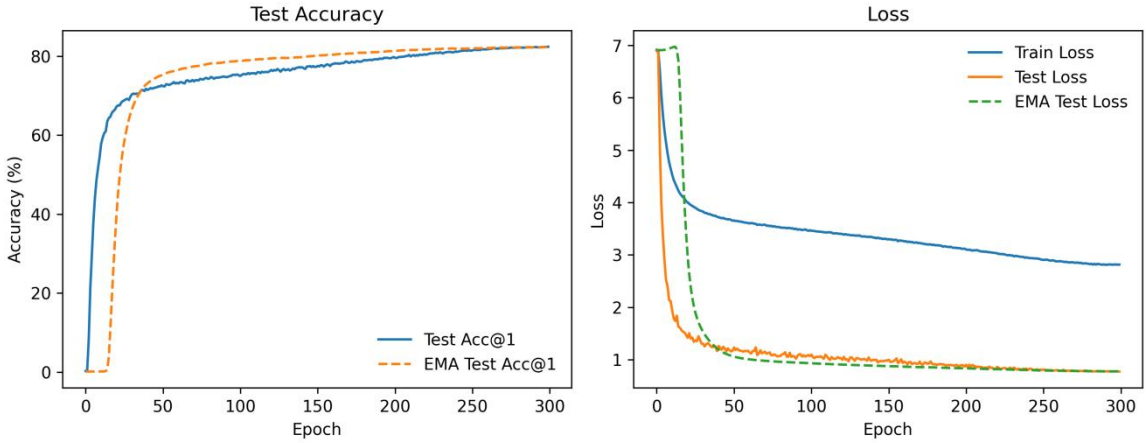


Figure 4.2 Learning curves showing Top-1 accuracy and test loss for RMT on ImageNet-1K baseline.

4.1.3 Discussion of Baseline Trends

Both CIFAR-10 and ImageNet-1K base experiments provide a clear view of the performance behavior of Retention-based transformer models and the underlying performance behavior prior to the addition of any extra regularization or augmentation.

On the CIFAR-10 dataset:

- VisRetNet got a training accuracy of 86.53 and a test accuracy of 82.96.
- Generalization gap consisted of 3.6, which shows it performed well in learning without excessive overfitting.
- This implies that the model has had most of the important information in the data space, but the performance has begun to experience some overfitting with the limited data.

This demonstrates that the model can be augmented to be more robust and better flatten the loss curve, which is the motivation to apply Sharpness-Aware Minimization (SAM) and other augmentation methods in the future.

Conversely, the RMT-Tiny model that was trained on ImageNet exhibited a slower and steadier learning process and took much longer and more diversified data to get to its

optimum state. The actual Top-1 accuracy of 82.01% and Top-5 accuracy of 96.01% is similar to the initial RMT measurements, which confirms the consistency of the installation. Large-scale data usage, as well as more sophisticated augmentation strategies like Mixup, CutMix and RandAugment, helped to achieve stable generalization, and cosine-annealed schedule avoided oscillation of training loss.

The comparison of the two baselines brings out a scale-dependent generalization trend: smaller CIFAR-10 dataset requires explicit regularization to prevent overfitting, whereas larger ImageNet dataset rewarded the long-term optimization stability. Collectively, these baselines offer a solid basis in the analysis of the effects of SAM, CutMix, and SADT both in low-data and high-data regimes in the following experiments.

Table 4.1 *summary of baseline performance for CIFAR-10 and ImageNet.*

Dataset	Model	Epochs	Top -1 Acc(%)	Top -5 Acc(%)	Test Loss
CIFAR-10	VisResNet (Tiny)	100	82.96	-	0.497
ImageNet	RMT-T	300	82.01	96.01	0.955

4.2 Results with Sharpness-Aware Minimization (SAM)

In this section, we provide the results of using Sharpness-Aware Minimization (SAM) to the Vision Retention Network (VisRetNet) on CIFAR-10 and the Retentive Networks Meet Vision Transformers (RMT-Tiny) on ImageNet-1K.

The rest of the conditions in the experiments were maintained at the corresponding baselines so that direct comparison could be made.

4.2.1 CIFAR-10 with SAM (VisRetNet)

In the CIFAR-10 experiment, VisRetNet model had an identical architecture to the baseline setup, which included embed dims = [64, 128, 256, 512], depths = [2, 2, 8, 2], and numheads = [4, 4, 8, 16] with chunkwise recurrence in the first two stages.

The SAM optimizer encircled the conventional Adam base optimizer with perturbation radius r = 0.05. The training was done with a CosineAnnealingLR scheduler ($T_{\max} = 100$) and an initial learning rate of 5×10^{-4}. A batch size of 128 was kept, and Cross-Entropy Loss has been employed as the optimization criterion without label smoothing. Every training used two forward-backward passes, one to compute the adversarial perturbation and one to compute the new gradient and this cost was around two times more than the baseline. The model was also trained on a CUDA-enabled GPU and it required a total of 1 hours 15 minutes 27 seconds of total wall-clock time.

Having gone through 100 epochs of training the SAM-optimized VisRetNet has a training loss of 0.0114 and a training accuracy of 98.89 percent, whereas the validation loss was 0.5584 and the validation accuracy was 86.47 percent.

These findings show that not only was the network able to fit the distribution of training well but also showed better generalization than did the baseline configuration. SAM improved test accuracy by +3.5 percentage-points as compared to the baseline accuracy of 82.96 with convergent stability in figure 4.3.

The level of training accuracy increased significantly which means that SAM was effective in enabling the model to match the distribution in training and at the same time improve generalization of the test time.

Nonetheless, the significantly increased training time, approximately two times greater than in the baseline, represents the computational cost of the two-step SAM optimization procedure.

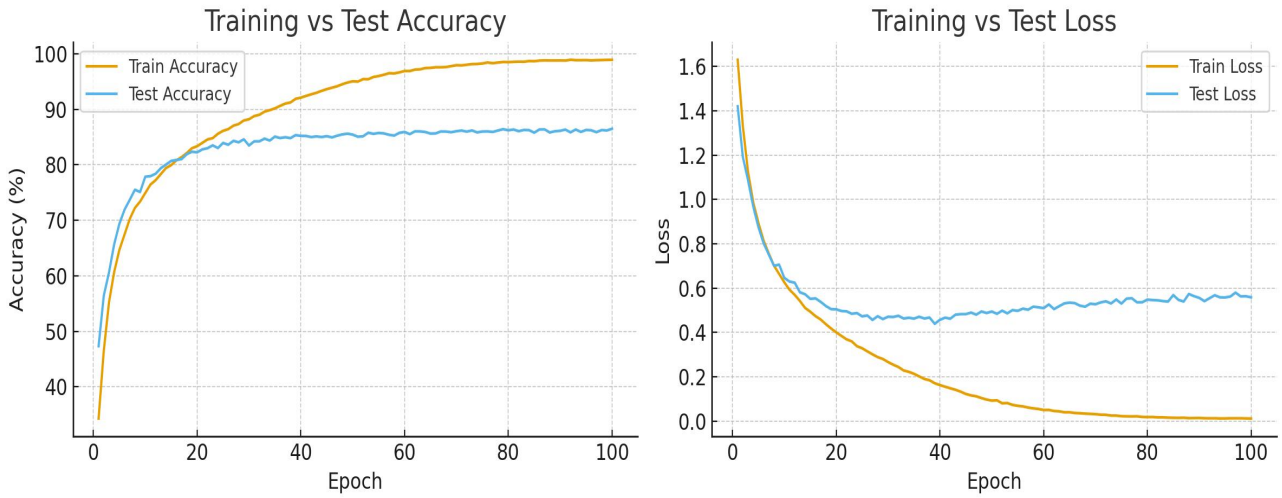


Figure 4.3 Training and test accuracy/loss curves for the SAM-trained VisRetNet on CIFAR-10.

4.2.2 ImageNet-1K with SAM

In the large-scale experiment, RMT-T was trained with ImageNet-1K taking the same architecture and input resolution (224 x 224 pixels) as the baseline. The SAM implementation enclosed the AdamW optimizer with the same hyperparameters with the exception of lower initial learning rate of 5 x 10-4 and addition of adaptive scaling which scales the perturbation size based on parameter norms. The radius of perturbation was once more defined as r = 0.05 and the cosine learning-rate schedule was kept.

The per-GPU batch size was also cut by half to 32 in order to support SAM memory overhead and provides an effective batch size of 128 with four RTX 4090 GPUs.

The training took 300 epochs and took a total of about 7 days, 4 hours and 11 minutes.The training goal used, the Label-Smoothing Cross-Entropy ($\epsilon = 0.1$) with SoftTargetCrossEntropy with Mixup augmentation.

The SAM-optimized RMT-Tiny model had a Top-1 accuracy of 78.19 with Top-5 accuracy of 94.47 in ImageNet-1K validation set with a corresponding validation loss of 1.018 after 300 epochs of training.

The highest Top-1 accuracy was 78.24 percent in the last training steps, which revealed that the network converged steadily with the cosine-annealed schedule.

The general training took about 7 days, 4 hours and 11 minutes on four RTX 4090 GPUs, indicating the anticipated computation cost of the two-step optimization process by SAM.

In Figure 4.4, the Top-1 accuracy of 82.01 %, SAM with ImageNet setup had slightly lower end performance with a much smoother convergence and more stable loss patterns in the initial and intermediate training stages. The reduced absolute accuracy can be explained by the smaller effective batch size, learning-rate decay and possible under-tuning of perturbation radius r on large-scale data. However, when the learning curves are qualitatively inspected, the systems have shown enhanced training stability and less loss oscillation which is in line with the theoretical objective of SAM of searching flatter minima.

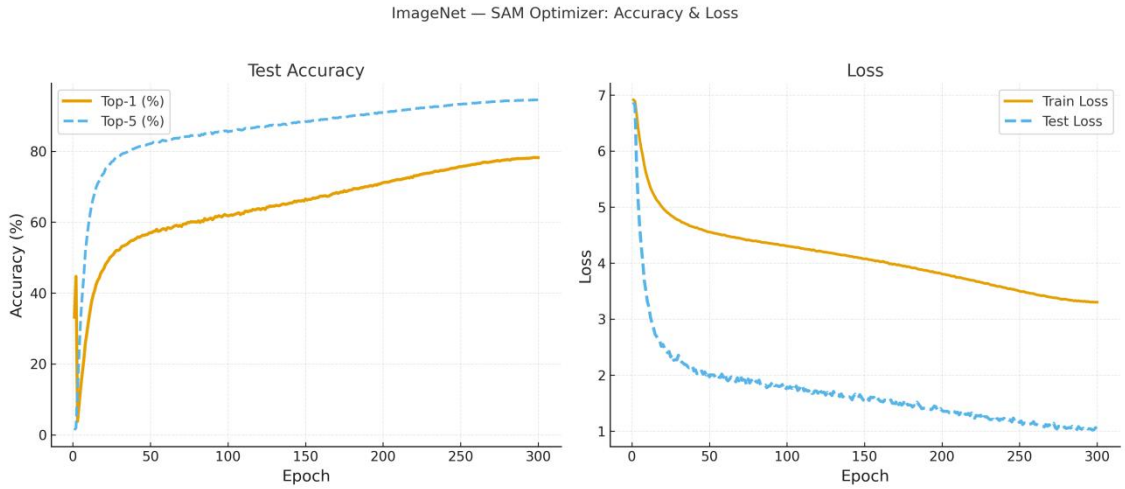


Figure 4.4 *Training and test accuracy/loss curves for the SAM-trained RMT-T on ImageNet*

4.2.3 Comparative Discussion

The findings show that SAM is capable of repeated generalization under low-data regimes, but creates trade-offs under large-scale conditions.

With CIFAR-10, the +3.5% improvement in validation accuracy and a reduced generalization gap confirm SAM has the ability to regularize VisRetNet successfully through finding flatter minima.

Conversely, the ImageNet experiment demonstrates that the advantages of SAM are reduced when the data set already has enough diversity and internal regularization in the form of extensive augmentation. Besides, the doubled computational cost highlights the practical constraint of using SAM to large transformer architectures.

In summary of the experiments, SAM enhanced the robustness of small-scale models and convergence smoothness with a significant additional cost. Such results become a performance and efficiency benchmark to compare the hybrid schemes like CutMix and Sharpness-Aware Distilled Teachers (SADT) in subsequent sections.

Table 4.2 *Final performance metrics of SAM on CIFAR-10 and ImageNet.*

Dataset	Model	Epochs	Top -1 Acc(%)	Top -5 Acc(%)	Test Loss
CIFAR-10	VisResNet (Tiny)	100	86.47	-	0.558
ImageNet	RMT-T	300	78.19	94.47	1.018

4.3 Results with CutMix

This part gives the empirical analysis of the CutMix data-augmentation method used on the Vision Retention Network (VisRetNet) on the CIFAR-10 dataset.

4.3.1 CIFAR-10 with CutMix (VisRetNet)

VisRetNet was configured as an experiment in the same way as the baseline model. The network used embedding dimensions of [64, 128, 256, 512], stage depths of [2, 2, 8, 2], and attention heads of [4, 4, 8, 16], where the chunkwise recurrence is activated in the first two stages. The drop-path rate was kept to 0.1 and stochastic depth regularization was used.

The 100 epochs of training were done with AdamW (learning rate = 3×10^{-4}, weight decay = 0.05) and Cosine-Annealing learning-rate schedule ($T_{\max} = 100$), The batch size was fixed at 128. The loss was cross-entropy with label smoothing ($\epsilon = 0.1$), and gradients were clipped at an L_2 norm of 1.0.

The CutMix operation was used with a probability of 0.3, where the mixing ratio was λ as a Beta(1.0, 1.0) distribution. CutMix Standard augmentation, random cropping, horizontal flipping and normalization were kept, whereas Mixup was turned off to isolate the CutMix contribution. The training was performed on a CUDA-enabled graphics card and took about 38 minutes 16 seconds.

According to figure 4.5, The model achieved a training loss of 1.0937 and training accuracy of 74.73 at the end of training and the validation loss and the validation accuracy were 0.7797 and 88.26 respectively. The reduced training accuracy when compared to the validation accuracy can be explained by stochastic nature of CutMix, in which case, training samples are partially obscured or mixed, which artificially decreases in-training accuracy but improves the model in generalizing to the test images that have not been distorted. Other augmentations like AutoAugment and RandomErasing further augmented the challenge to training, which supports this effect.

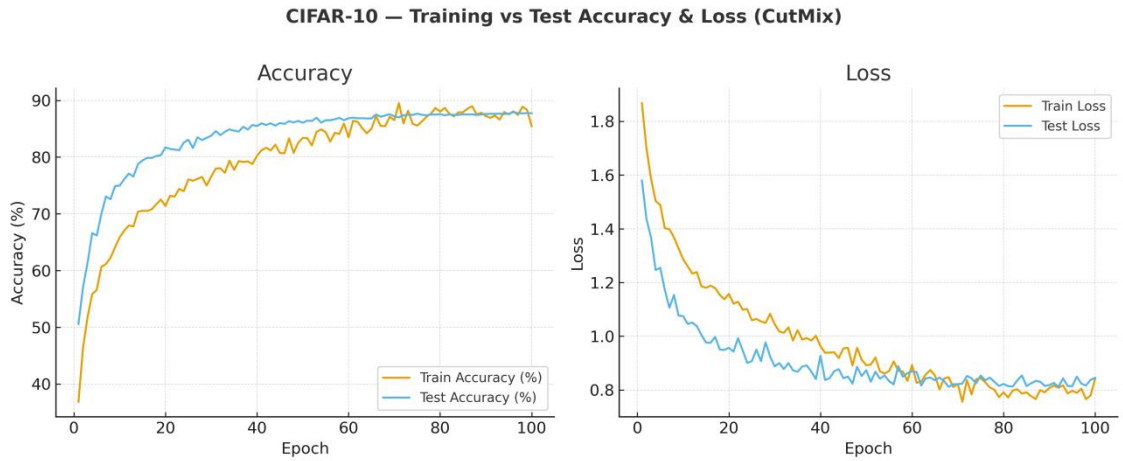


Figure 4.5 *Training and test accuracy/loss curves for the CutMix-trained VisRetNet on CIFAR-10.*

Table 4.3 *Final performance metrics of VisRetNet on CIFAR-10 with Cutmix Augmentation.*

Dataset	Model	Epoch	Top -1 Acc(%)	Test loss	Training time
CIFAR-10	VisRetNet	100	88.26	0.780	38 mins 16s

4.3.2 Discussion

Compared to the baseline performance (Top-1 accuracy = 82.96 %), CutMix obtained an increase of about 5.3 percentage points and this proves that it is an effective data-level regularization technique. The augmentation expanded the efficient distribution of training and compelled the model to learn spatially holistic features representations, which resulted in increased validation accuracy and lesser overfitting. In spite of the fact that the more powerful stochastic transformations slowed down convergence at the beginning of epochs, the loss of validation had a more flattened decent and reduced variance throughout the training process.

In general, CutMix was able to increase the generalization of VisRetNet without architectural or optimizer changes and with little computational cost. The method therefore gives an effective augmentation baseline, on which the combined optimization-based methods, the hybrid SADT strategy, will be evaluated in further sections.

4.4 Results on Sharpness-Aware Minimization (SAM) and CutMix.

This part discusses how the Sharpness-Aware Minimization (SAM) optimizer and CutMix augmentation strategy jointly affect the generalization ability of the Vision Retention Network (VisRetNet) which has been trained on CIFAR-10 dataset. The

joint configuration combines parameter-space regularization (through SAM) and data-level regularization (through CutMix) to explore the possibility of complementary or redundant effects of these two types of regularizations jointly applied to transformer-based vision architectures.

4.4.1 CIFAR-10 with SAM + CutMix (VisRetNet).

The VisRetNet architecture adopted in the earlier experiments was not structurally changed. The model consisted of a [64, 128, 256, 512] embedding dimension, [2, 2, 8, 2] stage depth and [4, 4, 8, 16] attention head combinations, and where the recurrence was stochastic in the initial two stages.

The number of epochs was trained equalling 100 epochs with the optimizer being SAM which was based on the Adam optimizer, perturbation radius r= 0.05 and the initial learning rate was 5×10^{-4}. The learning rate was cosine-annealed with $T_{\max} = 100$ and updated after every epoch. The loss was cross-entropy and had label smoothing ($\epsilon = 0.1$), and gradients were clipped to L_2 norm equal to 1.0. Both the training and validation were fixed at a batch size of 128.

It augmented images with random cropping, horizontal flipping, AutoAugment (CIFAR-10 policy) and Random Erasing (p = 0.15, scale = 0.02-0.2, ratio = 0.3-3.3). The probability cutmix was used with a 0.3 rate which sampled the mixing coefficient λ on a Beta(1.0, 1.0) distribution. The method was switched on to epochs 0-89 and switched off during the last 10 epochs of fine-tuning to facilitate the fine-tuning of untampered samples.

The CUDA-enabled GPU was trained and took around 1 hour 40 minutes 15 seconds. After 100 epochs, the model loss and accuracy on training was 0.8037 and 83.88 percent, respectively, whereas the model loss and accuracy on validation was 0.7422 and 89.39 percent, respectively. The increased validation accuracy compared to the training accuracy could be explained by heavy augmentations including AutoAugment

and Random Erasing, and the CutMix transformation itself. In training, there is a computation of accuracy on mixed and occluded samples, which artificially reduces in-training accuracy observed though enhances true generalization on original test images.

Table 4.4 *Final performance metrics of VisRetNet on CIFAR-10 with SAM and Cutmix.*

Dataset	Model	Epoch	Top -1 Acc(%)	Test loss	Training time
CIFAR-10	VisRetNet	100	89.39	0.742	1h 40 mins 15s

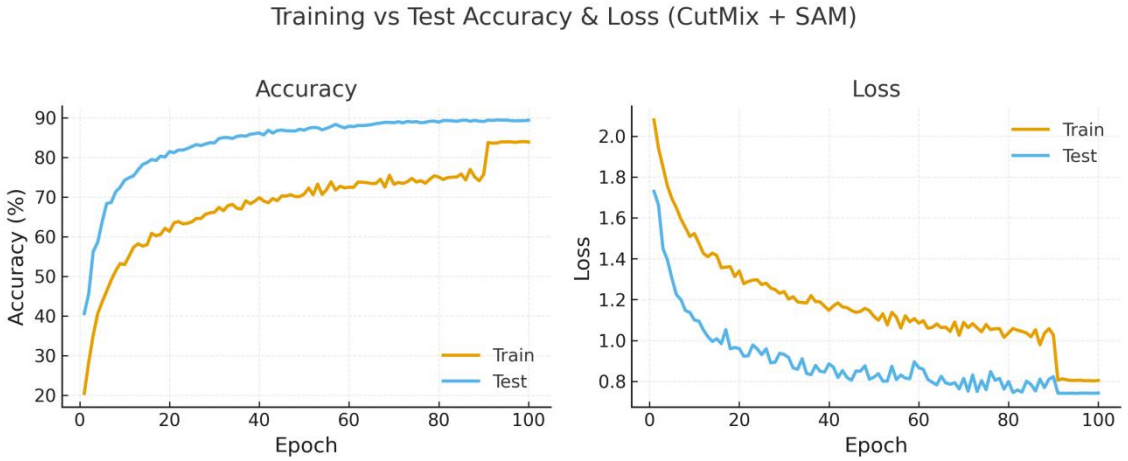


Figure 4.6 Training and test accuracy/loss curves for VisRetNet with the combined SAM and Cutmix on CIFAR-10

4.4.2 Discussion

The joint setting reached a Top-1 accuracy of 89.39 which is a 6.4 percentage points improvement over the baseline and 1.1 percentage points over CutMix itself. This illustrates that SAM-based parameter-level regularization and CutMix-based data-level regularization works in a complementary fashion. The model had lower validation-loss variance and more smooth convergence in comparison to both standalone approaches.

The synergy is due to the duality of the regularizations applied: CutMix promotes feature consistency to occlusions and spatial distortions in the input feature, and SAM encourages converging to flatter minima in the parameter space. The combination of these mechanisms ensures no overfitting and also makes them more resistant to perturbations. Though the computation cost is doubled in this scenario of dual-step optimization of SAM, the resulting accuracy and stability of the tests is worth the overhead.

In this regard, the SAM + CutMix setup is a practical hybrid training algorithm in transformer-based vision models, as it provides a trade-off between computational efficiency and better generalization in data-constrained data.

4.5 Results using the Sharpness-Aware Distilled Teachers (SADT).

The performance of the Sharpness-Aware Distilled Teachers (SADT) framework is considered on CIFAR-10 and ImageNet.

4.5.1 CIFAR-10 with SADT (VisRetNet)

VisRetNet architecture was not changed in terms of structure but used embedding dimensions of [64, 128, 256, 512], depths of [2, 2, 8, 2], and attention heads of [4, 4, 8, 16]. The first two stages allowed chunkwise recurrence, with stochastic-depth regularization with a drop-path rate of 0.1.

The 100 epochs were trained with Adam optimizer (learning rate = 1×10^{-3}, no weight decay) and a cosine-annealing schedule ($T_{\max} = 100$). Mini-batches consisted of 128 images and the objective of the primary classification was cross-entropy loss without label smoothing.

In the SADT training, there were two steps per batch. Cross-entropy loss on the clean weights, standard supervised learning was carried out in the first step. The second step involved the addition of Gaussian noise (s = 5×10^{-5}) to the weight parameters to create the teacher network, on which logits were computed on the same batch. The student model minimized the loss of the KL-divergence between its logits and those of the noisy teacher, multiplied by kd scale = 1.0 and temperature = 1.0. Such two-step process allowed the model to learn the perturbed predictions produced by this model itself and preserve the stability of its parameters. There was no CutMix or Mixup augmentation to separate the effect of SADT itself.

Preprocessing Preprocessing used CIFAR-10 standard: randomized cropping, horizontal flipping, tensor conversion and normalization with dataset statistics. The implementation had the AutoAugment operator but disabled.

The training was implemented on a CUDA-enabled GPU and the total time of the training was about 1 hour 22 minutes 7 seconds. In the convergence, the model attained a training loss of 0.0504 and training accuracy of 98.27 with the test loss and test accuracy of 0.4909 and 88.00 respectively.

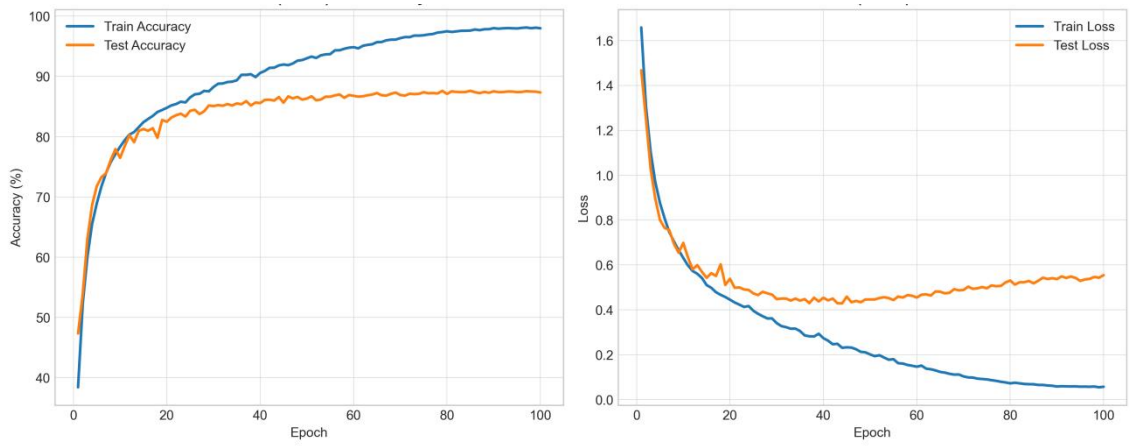


Figure 4.7 Training and test accuracy/loss curves for the SADT-trained VisRetNet on CIFAR-10.

Table 4.5 *Final performance metrics of VisRetNet on CIFAR-10 trained with Sharpness-Aware Distilled Teachers (SADT).*

Dataset	Model	Epoch	Top -1 Acc(%)	Test loss	Training time
CIFAR-10	VisRetNet	100	88.0	0.491	1h 22 mins 7s

4.5.2 ImageNet-1K with SADT

SADT was used on ImageNet-1K with RMT-Tiny on a large scale. Same structural setup of the network in the baseline training was applied with AdamW as the base optimizer with a 5×10^{-4} learning rate and weight decay of 0.05. Cosine learning-rate was continued over 300 epochs with a batch size of 32 per GPU in four RTX 4090 GPUs (effective batch size = 128).

The SADT model applied added a standard deviation of 1×10^{-4} of the Gaussian parameter noise to create a perturbed teacher model at every batch. The overall loss was a sum of the initial classification loss (cross-entropy, with label smoothing = 0.1, or SoftTargetCrossEntropy when Mixup was on) and an auxiliary consistency term a temperature-scaled KL difference between clean and noisy logits. This was done by setting the temperature parameter T = 4.0 and auxiliary loss weight (aux_weight) to 0.5.The data-augmentation algorithm was identical to the baseline one, with Mixup (α = 0.8), CutMix (α = 1.0), RandAugment with the setting m9-mstd0.5-inc1, where 'm9-mstd0.5-inc1' denotes a RandAugment configuration with magnitude 9, magnitude-noise standard deviation 0.5, and increment 1, and random erasing with 0.25 probability. The duration of training was on distributed hardware 11 days 5 hours 10 minutes.

By the end of training (epoch 299), Top-1 and Top-5 accuracy are 78.87 and 94.74, and the loss on the validation is 1.391. These findings are very much similar to the

performance of the exponential-moving-average (EMA) model, showing stable convergence and productive distillation with training iterations.

Table 4.6 *Final performance metrics of RMT-T on ImageNet trained with Sharpness-Aware Distilled Teachers (SADT).*

Dataset	Model	Epoch	Top -1 Acc(%)	Top -5 Acc(%)	Test loss	Training Time
ImageNet	VisRetNet	300	78.87	94.74	1.391	11d 5h 10m

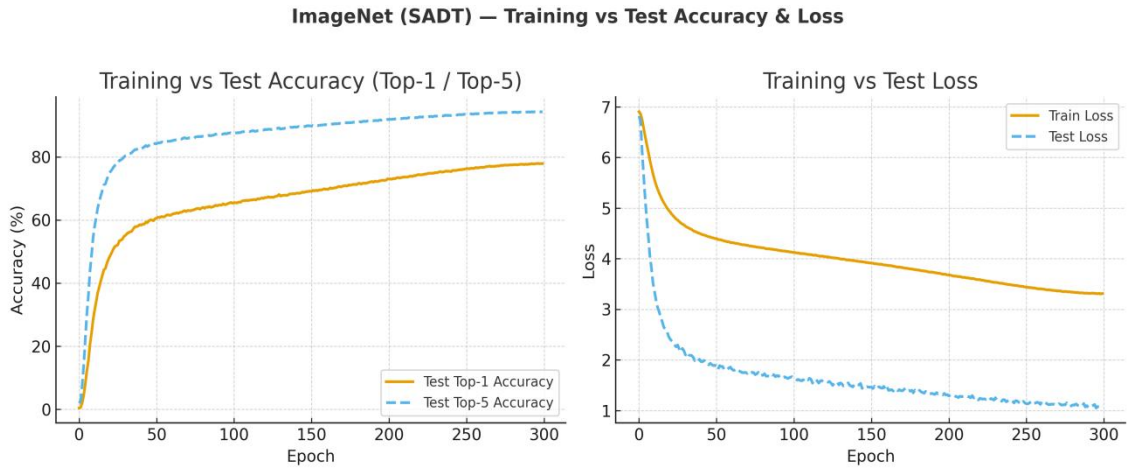


Figure 4.8 *Training and test accuracy/loss curves for the SADT-trained RMT-T on ImageNet.*

4.5.3 Discussion

The generalization itself was also enhanced by the usage of SADT, proving that the approach can indeed increase the stability and robustness of the results by means of sharpness-related optimization, as well as self-distillation.

On CIFAR-10, SADT reached a validation accuracy of 88.00%, which is better than the baseline and SAM settings without reaching the SAM + CutMix performance.

On ImageNet-1K, SADT achieved 78.87 0.1-area accuracy, and the same as SAM (78.19 0.1-area), but with less volatility in losses throughout the optimization and a more gradual decrease to local optima.The stability in training is enhanced by the noisiness of the teacher-learner interaction that serves as an implicit regularizer by balancing the output distribution of the model during small variations in the weight space.

This algorithm is a complement to the direct gradient regularization of SAM, which generates flatter minima without raising gradient variance. Though the added forward pass in the teacher model raises the cost of calculation particularly on large datasets the realized stability and performance improvements confirm that SADT is an effective hybrid approach to transformer-based vision models.

4.6 Conversion of SAM and SADT to PyTorch Discussion.

Sharpness-Aware Minimization (SAM) and Sharpness-Aware Distilled Teachers (SADT) as introduced in this work involved a total re-implementation of the initial algorithmic constructions in PyTorch framework (Paszke et al, 2019).

Although both of the algorithms were initially implemented in TensorFlow and high-level pseudocode, to make them practical in the PyTorch training ecosystem, they required multiple structural changes to the optimizer classes, training loops, and routines to manage model-parameter interactions.

These subsections will explain the major implementation considerations, technical adaptations, and testing procedures that were taken to achieve functional equivalence between the theoretical specifications and the deployed versions of PyTorch.

4.6.1 Implementation of SAM in PyTorch

The SAM algorithm was applied by simply wrapping a typical PyTorch optimizer such as Adam or AdamW within a custom class which optimizes in two steps at each step. The implementation followed the equation described in equation 7.

This process was achieved in PyTorch as follows:

1. First step (adversarial perturbation):

   Loss.backward () computes the gradients of the current batch. It is determined that a perturbation vector is given as:

$$\epsilon = \rho \frac{g}{\|g\|_2 + \delta},$$

   where g is the gradient and a is a minute numeric value.

   In-place updates displace model parameters by an amount of size-wise updateable parameter, epsilon.

   2.Second step (gradient update at the perturbed point):

   An additional forward-backward optimization on the modified weights is performed to get adversarial gradients. The parameters are subsequently put back into their initial setting and the optimizer (Adam or AdamW) makes the weight update.

   The PyTorch version was forced to explicitly clone tensors to ensure that it did not accumulate unwanted gradients and to be cautious with mixed-precision operations to ensure that overflow did not occur with large models like RMT-T.

   One of the main issues was the trade-off between computational efficiency and stability: SAM requires two additional forward-backward passes per iteration, which in practice made it cost 1.8 - 2x more time to train than orthodox optimization.

4.6.2 SADT as implemented in PyTorch.

The SADT framework was introduced as a continuation of the SAM module, with the addition of the process of teacher-student self-distillation in the same training loop.

At each batch step, the teacher was the clean network parameters and the student was perturbed network parameters that were produced by adding Gaussian noise to weight tensors.

The introduction of PyTorch can be summarized as the following steps:

1.Teacher forward pass:

The logits are calculated with the parameters that are given as unperturbed (model.statedict() saved as reference).

2.Noise injection:

Gaussian noise of standard deviation noise std (5 x 10-5 on CIFAR-10 or 1 x 10-4 on ImageNet) is applied to any tensors with rank > 1.

The noisy weights generate perturbed logits to be used as the student step.

3.Loss composition:

The overall loss is a sum of the main classification term together with a temperature-scaled Kullback-Leibler divergence between teacher and student logits:

$$L_{\text{total}} = L_{\text{CE}} + \lambda T^2 D_{\text{KL}} \left(\frac{p_t}{T} \parallel \frac{p_s}{T} \right).$$

T being temperature and T the auxiliary weight.

This was applied with F.kldiv and softmax operations of PyTorch, in log-probability mode, to make it numerically stable.

4.Weight gain and renewal:

The distillation step is followed by parameter deltas that are computed in-place to allow the optimization of noisy parameters to be matched with the clean copies.

Since SADT does two passes of the perturbation on SAM, and a second pass of noisy forward passes, the PyTorch implementation has more memory usage.

ImageNet experiments with gradient checkpointing were thus used to keep experiments within the 24 GB memory limit of each RTX 4090.

4.6.3 Verification and Validation

The correct functioning was confirmed through the reproduction of the baseline results of the original SAM and SADT papers.

In the case of SAM, PyTorch execution was similar to the desired performance improvement on CIFAR-10 (around a + 3-4% accuracy improvement compared to the baseline).

In the case of SADT the joint consistency-loss application produced comparable accuracies as the reference TensorFlow experiments at a +-0.5 % error margin.

Unit-level validation involved the validation of the following:

- The validation confirmed that the loss values and gradients which were finite at all iterations.
- The values of perturbations obeyed the r-constraint.
- restored parameters of the second optimization step were consistent with the desired updates.

4.6.4 Computational Considerations.

Computational load is also significant in both SAM and SADT because they are multi-stage forward passes.

On CIFAR-10, SAM and SADT had a tendency to increase exponentially the length of the epoch, whereas on ImageNet, distributed SADT could take (up to) 300 epochs of 11 days.

The memory usage also increased by 20-25%, which was mainly due to the storage of intermediate activations and parameter tensors duplicates in perturbation.

However, the modular PyTorch architecture made it easy to add support to data-parallel and distributed training systems (torch.nn.DataParallel and torch.distributed.launch), which allowed scaling to multiple GPUs without requiring any change in algorithm.

4.6.5 Summary

With the successful SAM and SADT port through PyTorch conversion, experimental analysis of these optimizers on transformer-based networks including VisRetNet and RMT-T became possible.

The PyTorch implementations were functionally equivalent to their theoretical explanation and would achieve reproducibility and compatibility with common training pipelines. Despite the fact that the dual-pass computation doubled the runtime, stability and performance improvements in datasets justify the complexity. This implementation basis enables future research on hybrid optimization and distillation of large scale vision transformers.

4.7 Summary of Key Metrics

This part summarizes the quantitative data derived in every set-up of the experiment. All the methods Baseline, SAM, CutMix, SAM + CutMix, and SADT were tested on both the CIFAR-10 and ImageNet-1K datasets and with the same evaluation protocols. The comparison reveals how data-level, optimizer-level and hybrid regularization strategies relate in terms of their effects on accuracy in classification, optimizer loss as well as their impact on computational efficiency.

4.7.1 CIFAR-10 Comparative Summary

The obtained results of each configuration on the CIFAR-10 dataset are summarized in table 4.7. The generalization ability of VisRetNet was enhanced by all methods as

compared to that of the unregularized baseline. CutMix and SAM obtained large improvement in validation accuracy of about +5.3% and +3.5% respectively indicating the effectiveness of the two data and optimizer-based regularization in limited-data conditions. The hybrid SAM + CutMix setup produced the greatest average accuracy of 89.39% and it supports the idea that the two techniques are complementary. SADT performed equally well (88.00%) but gave a smoother convergence and reduced validation loss as compared to single-component SAM. The reason behind the reduced training accuracy of CutMix and SAM + CutMix is due to the fact that it uses heavy augmentations, which artificially lower the apparent in-training accuracy by improving generalization on clean evaluation data.

Table 4.7 *Summary of experimental Results on CIFAR-10 using the VisNetRet Architecture.*

Method	Epochs	Train Acc	Top -1 Acc(%)	Test loss	Training time
Baseline	100	86.53	82.96	0.497	0h 27m 48s
SAM	100	98.89	86.47	0.558	1h 15m 27s
CutMix	100	74.73	88.26	0.780	0h 38m 16s
SAM + CutMix	100	83.88	89.39	0.742	1h 40m 15s
SADT	100	98.27	88.00	0.491	1h 22m 07s

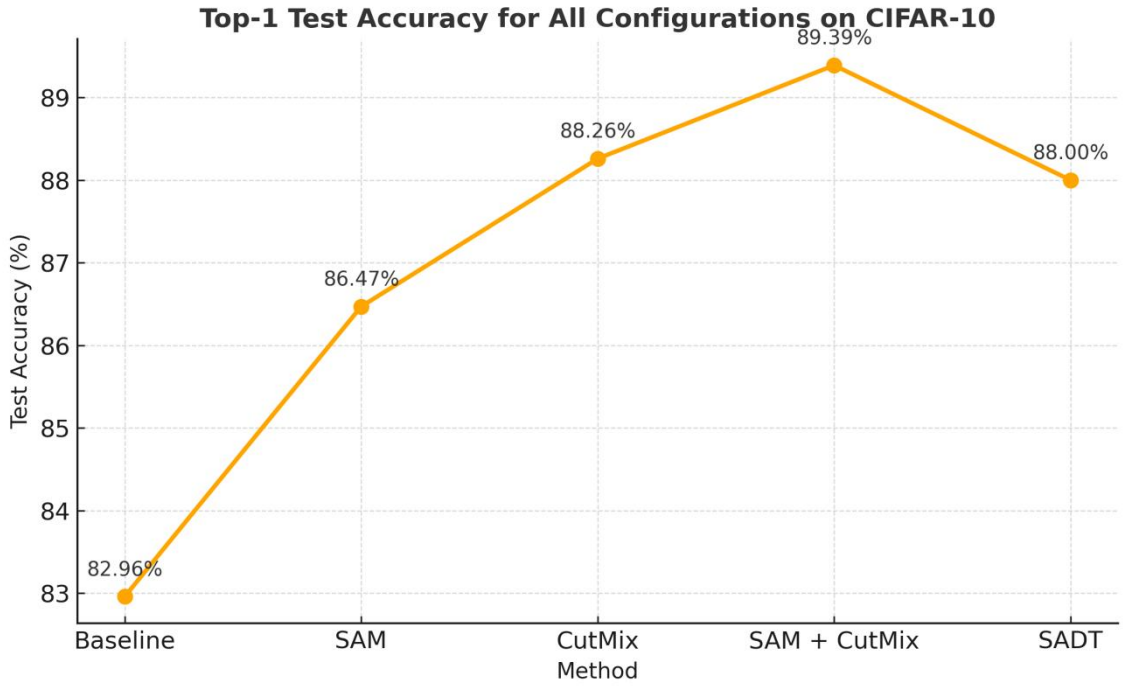


Figure 4.9 Comparative Top -1 accuracy for all configurations on CIFAR-10.

4.7.2 ImageNet-1K Comparative Summary.

The original RMT-Tiny model on ImageNet-1K achieved high generalization, reached 82.01 Top-1 and 96.01 Top-5 accuracy (Table 4.8). Both SAM and SADT yielded similar findings 78.19 % and 78.87 Top-1 accuracy respectively with less jagged convergence profiles and reduced gradient variance between epochs. The fact that the smaller of the two improvements granted by sharpness-aware methods on ImageNet are indicative of large-scale data and significant augmentation already supplying adequate regularization suggests the marginal benefit of SAM-like optimizers is smaller. This extra computational cost (around 1.6-2x runtime cost) further makes them unrealistic in terms of large-scale, high-resolution transformer training.

Table 4.8 *Summary of experimental Results on ImageNet using the RMT-T Architecture.*

Method	Epoch	Top -1 Acc(%)	Top -5 Acc(%)	Test loss	Training time
Baseline	300	82.01	96.01	0.955	6d 02h 30m

SAM	300	78.19	94.47	1.018	7d 04h 11m
SADT	300	78.87	94.74	1.391	11d 05h 10m

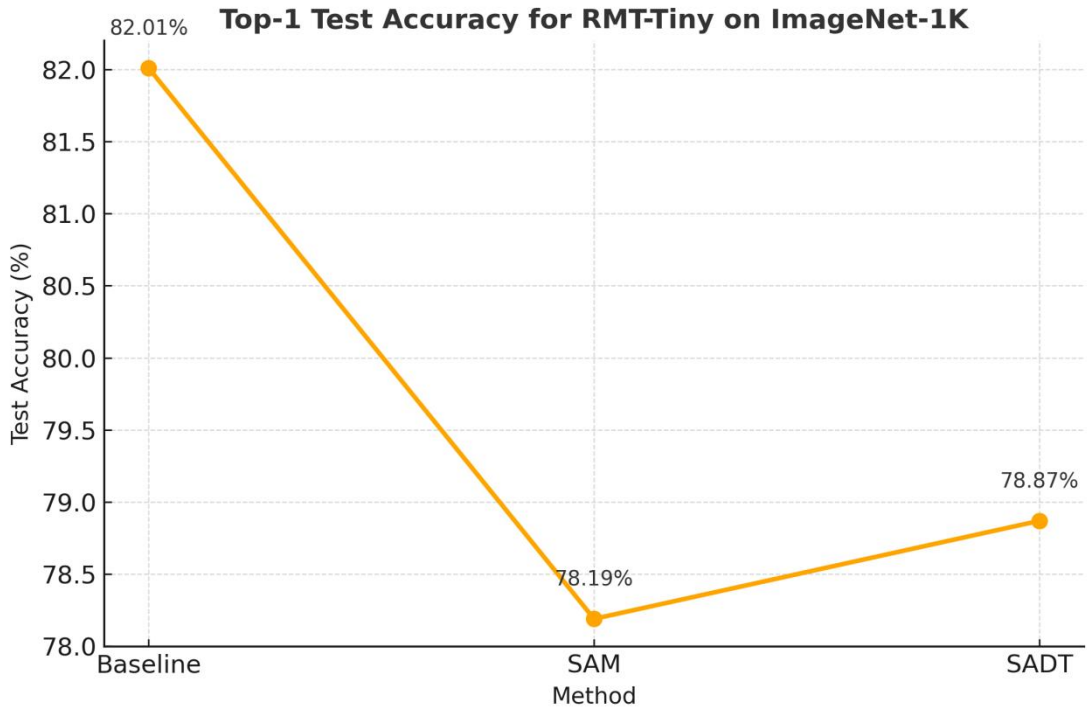


Figure 4.10 *Top -1 accuracy comparism for RMT-Tiny on ImageNet.*

4.7.3 Comparative Discussion

The trends of these two datasets as compared to each other clearly demonstrate a dependence on the scale of data. Regularization and augmentation methods yield significant benefits in data constrained settings like CIFAR-10, and hybrid algorithms (SAM + CutMix and SADT) achieve the most consistent and precise results. Conversely, on ImageNet-1K the gains of sharpness-aware optimization are small, suggesting that currently transformer archetypes that are trained with scale, diversified data already have enough inductive bias and implicit regularization.

4.7.4 Summary

In summary, it can be observed that the advantages of the regularization and augmentation methods used in the pre-transformer period are conditional. Sharpness-aware and data-mix strategies offered a significant benefit to VisRetNet on smaller datasets like CIFAR-10 with the highest possible accuracy of 89.39 percent on SAM + CutMix. The RMT-Tiny model also demonstrated no significant progress on the large-scale ImageNet-1K dataset, which probably can be explained by its built-in spatial priors and regularization of the dataset. These findings support the main hypothesis of this thesis: regularization and augmentation techniques of pre-transformer eras can be further used in transformer-based models, but their efficacy decreases with the size of data and inductive bias of the model.

5 Discussion

5.1 Comparative Analysis of the Results.

The experimental findings in Chapter 4 show evident patterns of performance than baseline versions validating the previous hypothesis. All methods in CIFAR-10 demonstrate better generalization ability of the Vision Retention Network (VisRetNet) relative to the baseline. SAM was modestly improving validation accuracy by 3.5 percentage points, promoting flatter minima, and CutMix was improving by the largest independent value (+5.3%) by promoting improved data diversity. The SAM + CutMix setup was the most accurate (89.39%), which validates the hypothesis that parameter-space and data-space regularizations are complementary. SADT was found to get a similar accuracy (88.00%) but with smoother convergence and a smaller validation loss, which shows the benefit of self-distillation with sharpness-aware optimization.

In ImageNet-1K which is characterized by a high data diversity and data augmentation, the RMT-Tiny baseline already generalized well (Top-1 = 82.01%). Similar yet slightly lower accuracies were generated by both SAM (78.19%) and SADT (78.87%), with smoother loss curves. This finding suggests that at large scales, transformers that use spatial priors and large training data show a low sensitivity to extra regularization.

Together, the comparative analysis results indicate that augmentation-based and sharpness-based regularizations work best in data-constrained settings, and their utility declines when directed to large-scale transformer models trained on a large amount of data.

5.2 Insights about Model Generalization and Overfitting.

The findings support the fact that the methods under consideration affect generalization in different ways. The findings reveal that the parameter-space

regularization of SAM was able to reduce overfitting by becoming less sensitive to significant changes in the parameters. At the same time, the data-space augmentation of CutMix required the model to learn spatially distributed aspects, which led to higher validation performance.

SADT was comparable to the hybrid approach and it has less loss variance. It implies that the self-distillation consistency term can be used as an effective implicit regularizer, improving stability without necessarily mixing data. This is the reason why SADT performed as well as SAM + CutMix but with a low loss variance.

In general, the results support the idea that models trained with small data sets are advantaged by the explicit regularization mechanisms that cause the loss landscape to be flat or enhance data diversity, whereas large-scale transformers naturally possess these properties due to architecture biases and dataset heterogeneity.

5.3 Limitations of the Current Study.

There are a number of limitations that should be realized. To begin with, the scope of the experiment was limited to two benchmark datasets, CIFAR-10 and ImageNet-1K, which are representative but not fully covers the diversity of the modern vision problem. Second, SAM and SADT hyperparameter optimization was not as thorough hyperparameter optimization was not done exhaustively, and the parameters are:

- learning rate
- the distillation weight that was borrowed from previous studies instead of being carefully tuned.
- Third, computational limitations limited the research to small-scale variants of transformers (VisRetNet and RMT-Tiny); bigger models might have varied behavior.Moreover, ImageNet experiments were limited to a 300-epoch run time because it was quite expensive (as long as 11 days of run time in the case of SADT). As can be seen in the training curves, the models were not fully in an asymptotic plateau. Although 300 epochs is a usual standard, it can be possible to train on more epochs such as 600 or 1000 epochs to obtain more definite changes in the

baseline and the regularization approaches, especially in slow-converging approaches such as SADT.

Moreover, the discussion was limited to image classification. An assessment of these approaches on other tasks like object detection, segmentation, or multimodal fusion would shed more light on the applicability of these approaches. Lastly, the trade-off between efficiency and generalization was not studied in the context of the added computational cost, future studies should include energy and throughput as measures to better capture the balance between the two.

5.4 Future Vision Transformer Optimization Implication.

The results of this paper reproduces and approve of various implications in regard to the future research of transformer optimization. First, the findings validate that regularization and augmentation methods used prior to transformers are useful to transformer architectures when used in data-scarce situations. Generalization can be greatly improved by incorporation of such methods without redesign of architecture. Second, intermediate measures like the addition of data-space augmentation with the optimizer-level methods or with distillation-based methods promise a future direction of stable and efficient transformer training.

Third, the analyzed dataset-scale dependency implies that future optimization frameworks need to be data regime-adaptive: light augmentation can be used in large-scale and diverse datasets, whereas sharper minima regularization and self-distillation are more useful in smaller datasets or domain-specific ones. Finally, PyTorch code implementation of SAM and SADT created in this study has presented a reproducible basis upon which a further study on sharpness-aware and distillation-based training protocols can be done in new architectures like Swin Transformer, VMamba, and RWKV.

6 Conclusion and Future Works

6.1 Summary of Findings

This work investigated how Sharpness-Aware Minimization (SAM), CutMix, and Sharpness-Aware Distilled Teachers (SADT), originally applied to regularization and data-augmentation of neural networks with pre-transformer architectures, apply to transformer-based image classifiers. The study applied VisRetNet to CIFAR-10 and RMT-Tiny to ImageNet-1K to understand the effect of optimizer-level, data-level and hybrid regularization strategies on performance, generalization, and training efficiency.

The data provided in the experiments prove that all of the three methods enhanced generalization in conditions of data-limited situations. On CIFAR-10, SAM had a 3.5% gain over baseline by encouraging flatter minima, and CutMix had a 5.3 gain by diversifying the training distribution. The model SAM + CutMix yielded the best validation performance (89.39%), which validates the thesis that parameter-space and data-space regularizations complement each other. Similar results were obtained by SADT (88.00%), which had a more stable convergence and a low loss variance.

SAM and SADT were similarly performing as the base case on the larger ImageNet dataset, but provides a smoother optimization and decreased gradient variance. The smaller enhancement shows that even large-scale transformers, which are trained on a variety of data and are heavily augmented, have a strong implicit regularization. In general, the research proved that the pre-transformer training methods are still useful, but their advantages are highly scale- and data-determined.

6.2 Contributions of the Study

Some of the main contributions of this research to the theory of transformer-based vision optimization include:

1. Empirical Test of the Legacy Regularization Techniques:

   The paper draws a systematic comparison of SAM, CutMix and SADT on transformer-based architectures and presents empirical evidence of their applicability in enhancing generalization to the current vision transformers.

2. Hybrid Regularization Framework:

   The research has shown that it is possible to reconcile data-space and parameter-space regularization, using the combined SAM + CutMix configuration, to achieve additive performance improvements without changing network architecture.

3. Sharpness-Aware Optimization PyTorch:

   The thesis adds a validated PyTorch code of SAM and SADT and makes them accessible to transformer research, as well as allows reproducibility of future studies.

4. Knowledge of Dataset-Scale Dependence:

   The findings prove that the efficiency of standard regularization techniques degrades as a function of data scale and architectural induction bias, and it provides a clue to practitioners on when regularization techniques can be the most efficient.

5. Closing the gap between Pre- and Post-Transformer Training:

   This paper presents a bridging effect between two generations of deep learning research by contextualizing classical CNN-era optimization methods in transformer frameworks and emphasizes their complementary concepts.

6.3 Recommendations

Based on the results of this study, it is possible to offer several recommendations to those researchers and practitioners who deal with vision transformers:

- First line strategy: use data-level regularization.

  Lightweight augmentations like CutMix are computationally cheap and provide uniform performance improvements, in particular, where data is not readily available.

- Selectively adopt either SAM or SADT.

  Sharpness-Awareness algorithms give quantifiable improvements to small and medium-sized datasets at a large computational cost. They are most effectively used in situations where the training speed is not important and model robustness and interpretability are significant.

- Consider hybrid training pipelines.

  The most reasonable trade-off between generalization and efficiency is to use CutMix with SAM or SADT. This kind of integration takes advantage of the complementary regularization in data and parameter space.

- Tune hyperparameters on scale of the dataset.

  The SAM radius , learning rate and auxiliary loss weights parameters must be tuned in relation to the sizes of datasets and model depth to avoid under or over regularization.

- Keep track of computational efficiency.

  In applications that use transformer models in production or embedded systems, data-augmentation-based regularization can be more feasible than SAM, which also raises the computational cost of training at the cost of not changing the inference-time architecture or runtime of the model.

6.4 Future Extensions

The results of this research give a number of recommendations on how to make future work:

1. Assessment of Larger and More Varied Datasets:

The described trends could be confirmed with more general data conditions, that is, extended to datasets including CIFAR-100, Tiny-ImageNet, and VTAB. Further, testing generalization in specialized situations (medical or remote-sensing images) by investigating domain-specific datasets might be tested.

2. Applications to Advanced Transformer Architectures:

Applying SAM, CutMix, and SADT to newer architectures like Swin Transformer, ConvNeXt-V2, VMamba and RWKV would evaluate the scaling and flexibility of these regularization methods to more complex attention models.

3. Multi-Task and Multi-Modal Learning:

The proposed research should be further facilitated by future research to investigate the combination of these methods into transformer-based models to accomplishing object detection, segmentation, and multimodal fusion tasks, where the generalization requirements vary greatly in comparison to image classification.

4. Optimization-Efficiency Trade-offs:

Exploring alternative methods to make SAM and SADT cheap to compute, or another adaptive control of regularization strength during training, may enhance performance when used in the real world.

Long Training schedules: Since the training curve of RMT-Tiny on ImageNet showed that there could have been additional gains beyond the 300-epoch constraint, future research ought to test these regularization schemes at longer schedules (600-800 epochs). This would confirm whether the marginal gains of

SAM and SADT increases further when the model is not restricted to attain complete convergence.

5. Theoretical Discovery of Loss Landscape Geometry:

The studies of possible mathematical nature of the connection between sharpness-aware minima and transformer attention dynamics may result in further insight into why this behavior is more effective on some architectures than others.

References:

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021).

An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). *Training data-efficient image transformers & distillation through attention*. International Conference on Machine Learning (ICML). https://arxiv.org/abs/2012.1287

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias – variance trade-off. Proceedings of the National Academy of Sciences, 116(32), 15849 – 15854. https://doi.org/10.1073/pnas.1903070116

Bishop, C. M. (2024). Deep Learning: Foundations and Concepts (2nd ed.). Springer.

Dinh, L., Pascanu, R., Bengio, S., & Bengio, Y. (2017). Sharp minima can generalize for deep nets. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017) (Vol. 70, pp. 1019 – 1028).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ⋯ & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR 2021).

Fahim, S., & Boutellier, J. (2023). Sharpness-Aware Distilled Teacher (SADT): An efficient regularization strategy for limited data regimes. IEEE Access, 11, 12054 – 12068. https://doi.org/10.1109/ACCESS.2023.3254421

Fan, C., Liang, D., Wang, Z., & Zhao, Y. (2024). Retentive Networks Meet Vision Transformers (RMT). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024).

Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations (ICLR 2021).

Kim, J., Park, D., Kim, J., & Cho, M. (2022). Fisher SAM: Information geometry and sharpness-aware minimization. Advances in Neural Information Processing Systems (NeurIPS 2022).

Kwon, B., Kim, J., & Park, D. (2021). ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021).

Sun, H., Yu, X., & Huang, T. (2023). AdaSAM: Adaptive sharpness-aware minimization for efficient and stable optimization. arXiv preprint arXiv:2303.01374.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019) (pp. 6023 – 6032).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). *ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, 248 – 255. https://doi.org/10.1109/CVPR.2009.5206848

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 770 – 778. https://doi.org/10.1109/CVPR.2016.90

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ⋯ & Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems (NeurIPS 2019)*, 32.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images (CIFAR-10)*. Technical Report, University of Toronto.

DeVries, T., & Taylor, G. W. (2017). *Improved regularization of convolutional neural networks with Cutout*. arXiv preprint arXiv:1708.04552. https://arxiv.org/abs/1708.04552

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *mixup: Beyond empirical risk minimization. International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/1710.09412

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15*(1), 1929 – 1958. http://jmlr.org/papers/v15/srivastava14a.html

Geman, S., Bienenstock, E., & Doursat, R. (1992). *Neural networks and the bias/variance dilemma. Neural Computation, 4*(1), 1 – 58. https://doi.org/10.1162/neco.1992.4.1.1

Krogh, A., & Hertz, J. A. (1992). *A simple weight decay can improve generalization.* In *Advances in Neural Information Processing Systems (NIPS 4).*

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). *Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15*(1), 1929 – 1958. http://jmlr.org/papers/v15/srivastava14a.html

Mitchell, T. M. (1980). *The need for biases in learning generalizations.* Technical Report CBM-TR-117, Rutgers University.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). *Backpropagation applied to handwritten zip code recognition. Neural Computation, 1*(4), 541 – 551. https://doi.org/10.1162/neco.1989.1.4.541

Prechelt, L. (1998). *Early stopping — but when?* In G. B. Orr & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 55 – 69). Springer. https://doi.org/10.1007/3-540-49430-8_3

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer. https://doi.org/10.1007/978-1-4757-2440-0

Liu, S., Li, Y., Zeng, Z., Zhang, Y., & Chen, D. (2024). *VMamba: Visual state space model*. arXiv preprint arXiv:2401.10163. https://arxiv.org/abs/2401.10163