LIPSHIFT: A CERTIFIABLY ROBUST SHIFT-BASED VISION TRANSFORMER

Rohan Menon[1], **Nicola Franco**[2], **Stephan Günnemann**[1]
[1] Technical Univ. of Munich, School of Computation and Information Technology
[2] Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany

ABSTRACT

Deriving tight Lipschitz bounds for transformer-based architectures presents a significant challenge. The large input sizes and high-dimensional attention modules typically prove to be crucial bottlenecks during the training process and leads to sub-optimal results. Our research highlights practical constraints of these methods in vision tasks. We find that Lipschitz-based margin training acts as a strong regularizer while restricting weights in successive layers of the model. Focusing on a Lipschitz continuous variant of the ShiftViT model, we address significant training challenges for transformer-based architectures under norm-constrained input setting. We provide an upper bound estimate for the Lipschitz constants of this model using the l_2 norm on common image classification datasets. Ultimately, we demonstrate that our method scales to larger models and advances the state-of-the-art in certified robustness for transformer-based architectures[1].

1 INTRODUCTION

Vision transformers have been extremely versatile and considered as foundational breakthroughs in deep learning (DL) (Dosovitskiy et al., 2021). For applications in the field of computer vision they can easily expand to multiple domains and varied number of tasks such as classification (Dosovitskiy et al., 2021; Touvron et al., 2021), segmentation (Ye et al., 2019) and object detection (Carion et al., 2020; Zhu et al., 2021). Although compared to other popular vision architectures such as residual networks (ResNets) (He et al., 2016) and convolutional networks (ConvNets) (LeCun et al., 1989), the effect of adversarial attacks on transformer-based models has been studied in limited capacity (Shao et al., 2021). An adversarial attack is defined as injecting noise to an input such that it disrupts the model's decision making process (Akhtar & Mian, 2018). This poses significant risks of severe consequences when models are deployed in real-world safety-critical contexts.

Formal verification of DL models is crucial for guaranteeing their reliability and compliance with standards in safety-critical applications, such as autonomous driving or healthcare (Urban & Miné, 2021). However, the current approaches for formal verification have been struggling to efficiently scale to large architectures (König et al., 2024). Transformer-based architectures, being one of them, are not researched as frequently in the context of formal verification although their wide range of applicability in vision and natural language processing is very well known.

In this paper, we focus on applying adversarial training to a transformer-based architecture to obtain provable Lipschitz-based robustness certificates. We summarize our contributions here:

i. Introduce **LipShiFT**, a transformer-based Lipschitz continuous architecture with tight Lipschitz bounds.

ii. Provide robustness guarantees under l_2-norm for a certification radius $\epsilon = 36/255$.

iii. Evaluation of empirical robustness with AutoAttack (Croce & Hein, 2020).

[1] https://github.com/RohanMenon/LipShiFT

2 BACKGROUND

The introduction of the *transformer* architecture has paved the way to many notable developments in DL applications (Vaswani et al., 2017). Notably in computer vision, it was adapted to give rise to Vision transformer (ViT) (Dosovitskiy et al., 2021) which featured an N x N input image split into multiple non-overlapping patches and converted intro vector embeddings to conform to a sequence based input to the transformer. This input passes through many layers, one of which being the attention mechanism that was introduced to surpass the shortcomings of the Long-Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Recurrent Recurrent Neural Networks (RNN) (Schuster & Paliwal, 1997) models to capture long-term dependencies in the input. LSTM's traditionally suffered with the issue of vanishing gradients as the input sequence increases in length overtime. The original attention mechanism, introduced in 2017 paper by Vaswani et al. (2017) was a Dot Product (DP) based mechanism that was effectively able to capture semantic meaning while being able to produce stable training.

Lipschitz continuity Given inputs $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{N \times N}$ and function $\mathbf{f}$: $\mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{M}$. Formally, $\mathbf{f}(\mathbf{x})$ is Lipschitz continuous, whose output bounded as follows:

$$||\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x})|| \leq L||\mathbf{y} - \mathbf{x}||, L \geq 0, \tag{1}$$

where L is defined as the Lipschitz constant of the function $\mathbf{f}$ and $|| \cdot ||$ is an L_p-norm.

The original transformer architecture, which relies on a dot-product mechanism, is not Lipschitz continuous. To address this, Kim et al. (2021) introduced an L_2-based self-attention transformer that satisfies the Lipschitz condition. However, this approach comes with its own challenges. The computation of the Lipschitz constant for L_2 attention is dependent on the input size, which makes it difficult to obtain a tight upper bound in practice for transformer-based architectures (Hu et al., 2023).

ShiftViT Thus, we consider the ShiftViT (Wang et al., 2022) which implements a partial channel shift operation which helps simulate the self-attention module and provide similar performance on ImageNet (Deng et al., 2009). This architectural change essentially provides us with a model that is lightweight and still can perform well compared to its attention-based alternatives. Providing robustness guarantees and certificates for transformer-based architectures has been a challenge in the past. To achieve a robust LipShiFT model, we utilize the Lipschitz margin loss function introduced by Hu et al. (2023) known as Efficient Margin Maximization (EMMA) Loss. It uses the model Lipschitz constant to maximize the class decision boundary for all classes with respect to the predicted class. This way, even if the top threatening class label changes frequently, it does not negatively affect the models epoch-wise robustness. Additionally, EMMA also uses a variable ϵ certification radius during training which helps it balance both the clean as well as verified accuracy while performing loss optimization.

3 RELATED WORK

In this section, we briefly describe some of the most notable works which focus on Lipschitz continuous architectures and some that use Lipschitz margin training to guarantee model robustness. We focus mainly on applications in the vision domain (Prach et al., 2023).

Local-Lipschitz (Local-Lip) (Huang et al., 2021) introduces an efficient algorithm to compute the local Lipschitz constant of neural networks by computing the tight upper bound norm of the Clarke Jacobian via linear bound propagation. This method allows for the handling of larger and practical models that previous methods could not efficiently process. Although their results are applicable on a wide range of datasets like MNIST (Deng, 2012), CIFAR10 (Krizhevsky et al., 2009) and Tiny ImageNet (Le & Yang, 2015) for a certification radius of $\epsilon = 36/255$, they report results only on fully convolutional networks and not on other scalable architectures such as transformers.

Globally Robust (GloRo) Networks (Leino et al., 2021a) describes a new architecture such that it computes the global Lipschitz bound of the model which is tighter as compared to Local-Lip. GloRo primarily provides robustness guarantees by training the GloroNet using TRADES loss which is updated to utilize the model Lipschitz constant to maximize the decision boundary of the predicted

class i with respect to the top threatening class j. In principal, Leino et al. (2021a) introduces a new $\perp \in \mathbb{Z}$ class label. This label is always 0 in the one-hot encoded version of class labels, signifying a non-robust sample and assigned only when the model wrongly classifies the sample compared to the truth label. Based on their method, Gloro Networks achieve much better results compared to Local-Lip. Although they primarily also focus on fully convolutional network architectures.

Vision transformers A Certifiable Vision Transformer (CertViT) (Gupta & Verma, 2023) has been introduced with a two-step certification algorithm to provide robust certification for pre-trained vision transformers. In their paper, Gupta & Verma (2023) utilize Lipschitz continuous L_2-attention ViT architectures which are pre-trained on ImageNet-1K and use them for certification on CIFAR10/100 (Krizhevsky et al., 2009), MNIST (Deng, 2012) and Tiny ImageNet (Le & Yang, 2015) respectively. Their proposed algorithm involves a proximal step that initially constrains the weights thereby reducing the model Lipschitz at each layer followed by a projection step that focuses to improve the generalization, hence the clean accuracy of the model. They use the power method to constrain the Lipschitz constant of the successive layers. To the best of our knowledge, this is the only work that has reported the model Lipschitz of a vision transformer-based architecture. As part of their main findings, the authors explain that CertViT significantly reduces the Lipschitz bounds of networks compared to existing Lipschitz methods like GloRo, Local-Lip, and BCP. Their results on MNIST, CIFAR-10, CIFAR-100, and TinyImageNet (Gupta & Verma, 2023, Table 1) shows that CerViT achieves superior performance for both clean and adversarial training (trained on PGD (Madry et al., 2018) samples). Although their technique proves to be better than existing Lipschitz based certification methods, their two-step algorithm requires several iterations to converge and does not seem to scale with increasing dataset complexity.

LipsFormer (Qi et al., 2023a) introduces a new vision transformer architecture that is Lipschitz continuous. They update the existing SwinV2 (Liu et al., 2022) architecture to be Lipschitz continuous by replacing core elements such as the non-overlapping window attention mechanism with the Scaled Cosine Similarity Attention (SCSA) (Qi et al., 2023a) and replaing the LayerNormat each stage with their proposed CenterNorm (Qi et al., 2023a) layer which they provide theoretical guarantees for to be 1-Lipschitz normalization layer. Furthermore, they also explain how DropPath (Goyal et al., 2017) in addition to DropOut helps in tightening the Lipschitz upper bound. Although, they provide experimental results on the performance of Lipschitz continuous vision transformer on ImageNet-1K, they do not provide an emperical estimation of the model Lipschitz or robust accuracy measure that can be used as a benchmark.

Considering the drawbacks of the works presented in this section, in the next section we propose our method that tries to overcome some of the gaps highlighted in the existing research in this field.

4 METHODOLOGY

To obtain a tight upper bound on the model Lipschitz, we introduce some changes to the original ShiftViT (Wang et al., 2022) architecture. This mean according to Equation 1, we replace architectural components that have a higher L comparatively to obtain a model Lipschitz as low as possible. A low model Lipschitz ensures that the model output is deterministically bounded for any input $\mathbf{x} \in \mathbb{R}^{N \times N}$ and is easier to certify using norm-based verification.

4.1 LIPSCHITZ CONTINUOUS MODULES

CenterNorm instead of GroupNorm GroupNorm (Wu & He, 2018), defined in Equation 2 plays an important role in the ShiftViT (Wang et al., 2022) architecture to provide stable training and enhanced generalization performance. Its effectiveness is in handling batches of small sizes by computing batch statistics across channel groups, defined as:

$$GN(\mathbf{x}) = \gamma \left(\frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta, \tag{2}$$

where μ and σ^2 are the mean and variance computed across the channels within each group, γ and β are learnable parameters, and η is a small constant for numerical stability. In a Lipschitz constrained training setting, GroupNorm operation becomes difficult to optimize as the denominator term, η is of

the order 10^{-6}. As training progresses, we see that since η is inversely proportional to the Jacobian of $GN(\mathbf{x})$, it proves to be a bottleneck for stable training under Lipschitz constraint.

CenterNorm (Qi et al., 2023a), proposed in the LipsFormer (Qi et al., 2023a) architecture, is a 1-Lipschitz normalization that works as a viable alternative to GroupNorm. CenterNorm is mathematically formulated as:

$$CN(\mathbf{x}) = \gamma \cdot \left(\mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}^T\right)\mathbf{x} + \beta \tag{3}$$

where D is the input dimension, $\mathbf{I}$ is the identity matrix. γ and β are learnable parameters initialized as 1 and 0 respectively. $\mathbf{1}\mathbf{1}^T$ is a matrix of ones. This formulation is Lipschitz continuous as shown by Qi et al. (2023a) with a theoretical Lipschitz upper bound of $\frac{D}{D-1}$. For practical purposes, the Lipschitz constant can be approximated to 1 since he input dimension D is high for datasets like CIFAR10 (Krizhevsky et al., 2009). In our experiments, we compute the Lipschitz constant of CenterNorm as the as the absolute maximum of the input weight matrix as we want a tight bound.

Unlike GroupNorm (Wu & He, 2018), CenterNorm (Qi et al., 2023a) directly eliminates the mean across dimensions but does not scale by the standard deviation, thereby preventing instability from small variance values. It ensures a limited gradient, offering a more stable and robust training experience, which accelerates convergence and enhances generalization in LipShiFT. The combination of these theoretical benefits and empirical evidence of smoother training trajectories without the necessity for learning rate warmup establishes CenterNorm as a preferred normalization method in vision transformer architectures compared to GroupNorm.

MaxMin instead of GeLU To empirically reduce the upper bound of model Lipschitz, we replace the GeLU activation which has a Lipschitz bound of approximately 1.12 (Qi et al., 2023b) with a 1-Lipschitz MinMax activation function. This has essentially proven to be extremely useful in improving model robustness in past works (Hu et al., 2023; Anil et al., 2019). As Anil et al. (2019) explains, MinMax displays gradent preserving property during adversarial training that is valuable for training models to be more robust and generalize better.

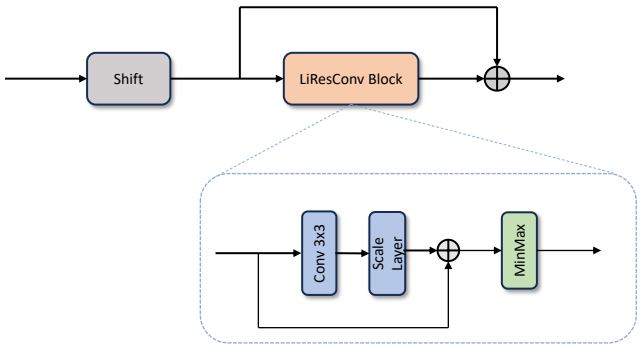


Figure 1: Proposed modified shift block.

Linear Residual Convolution (LiResConv) instead of MLP Another core architectural component of the ShiftViT (Wang et al., 2022) is a Multi Layer Perception (MLP) within the shift block. It consists of three alternating 1x1 convolutional and GeLU activation layers followed by a DropOut layer. This MLP network is Lipschitz continuous since it is a composition of affine transformations with activation layers in between (Pauli et al., 2023).

But as Hu et al. (2023) motivates, the LiResConv block proves to have a tighter upper bound compared to standard FFN (Hu et al., 2023, Section 3).

AveragePool instead of AdaptiveAveragePool Since the computation of the model Lipschitz constant is dependent on the input size, we replace the final pooling layer AdaptiveAveragePool2d in the ShiftViT (Wang et al., 2022) architecture with the standard AveragePool2d layer. We compute

the Lipschitz constant using the power iteration method (Farnia et al., 2018). We use the custom implementation of AveragePool2d with Lipschitz computation by Hu et al. (2023).

Dropout and DropPath Gouk et al. (2021) studies the effect of using regularization techniques in scalable DL architectures. From a Lipschitz perspective, the effect of DropOut and DropPath are important to take into consideration. Both regularization techniques play an essential role in regularizing ShiftViT (Wang et al., 2022). For Lipschitz training of LipShiFT, we scale the model Lipschitz during inference by a factor of $(1-P_{drop})$ where P_{drop} is the drop out rate used throughout the model. This way we can empirically compensate for the regularization applied during training with EMMA (Hu et al., 2023) loss and ensure a tight upper bound on the overall model Lipschitz.

Orthogonal Initialization for Affine Layers Initialization plays an important role in successfully training a neural network. Many initialization methods have been proposed in the past years such as Glorot & Bengio (2010) and He et al. (2015) initialization. We use orthogonal initialization which is 1-Lipschitz by design since the initialized weight matrix is orthogonal,

$$\mathbf{W}_{orth} = \frac{\mathbf{W}_{init}}{\sigma_{max}(\mathbf{W}_{init})} \tag{4}$$

where $\mathbf{W}_{orth}$ is the initial random orthogonal weight matrix and $\sigma_{max}(\mathbf{W}_{init})$ is its largest eigenvalue. For affine transformation $\mathbf{f}(\mathbf{x}, \mathbf{W}_{orth}) = \mathbf{W}_{orth}^T\mathbf{x}$, its Lipschitz constant satisfies the following inequality, $||\mathbf{W}_{si}^T\mathbf{x}_1 - \mathbf{W}_{orth}^T\mathbf{x}_2|| \leq Lip(\mathbf{f}_{\mathbf{x}}(\mathbf{W}_{orth}))||\mathbf{x}_1 - \mathbf{x}_2||$, where $Lip(\mathbf{f}_{\mathbf{x}}(\mathbf{W}_{orth})) = 1$ at initialization. We use orthogonal initialization on all affine layers.

Last Layer Normalization (LLN) Lastly, we replace the standard linear prediction head with a normalized prediction head introduced by Singla et al. (2021). They state that it particularly is useful when the input dataset has a large number of class labels, hence helps to achieve robust certificates at scale. Further, this has been seen to improve emperical robustness of the ResNet based LiResNet (Hu et al., 2023) architecture.

5 Experimental Results

This section describes the main results of LipShiFT under adversarial constraints. We briefy describe the training and evaluation settings and discuss our main findings. Lastly, we report robust accuracy of our LipShift model using the AutoAttack (Croce & Hein, 2020) framework. We report our settings in Appendix A.

5.1 Certified Robustness

In Table 1, we compare the main results of our proposed work in to existing techniques and methodologies implemented in the recent years. For a fair comparison, we report our main results on a custom LipShiFT specification that has the same number of parameters when comparing to works such as the CHORD-LiResNet (Hu et al., 2024). We compare performance of our architecture with recent works on the CIFAR10/100 (Krizhevsky et al., 2009) and Tiny ImageNet (Le & Yang, 2015) datasets respectively. For each of the datasets, we report the clean and verified accuracy. The verified accuracy (VRA) is reported for a target epsilon of $\epsilon = 36/255$. The best scores are marked in bold.

In Table 1, except CertViT (Gupta & Verma, 2023), LipShiFT is the only architecture that utilizes a transformer-based architecture and uses Lipschitz constrained techniques to provide robustness guarantees. We also stand out from previous works as we are able to report a tight upper bound of the empirical model Lipschitz constant. As has been successfully shown in case of LiResNet (Hu et al., 2023; 2024), we implement a similar Lipschitz constrained training scheme since it serves to be an efficient way to constrain the model Lipschitz while being able to provide deterministic and certifiable robustness scores.

We also report our results alongside previous works that use different architectures for each dataset. These are broadly divided into three categories, namely, ConvNets, ResNets and ViT respectively.

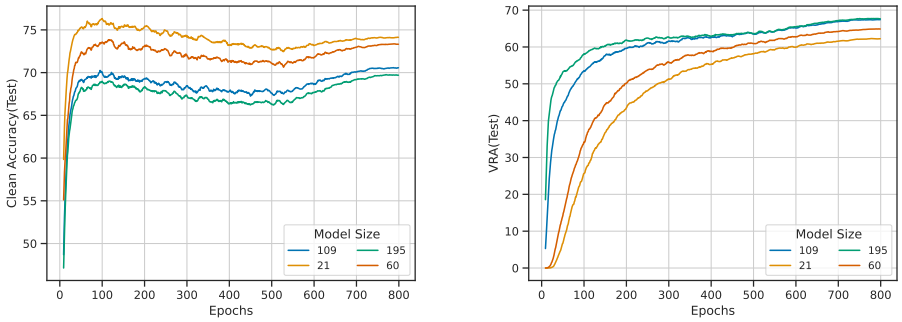
Table 1: This table presents the clean and verified robust accuracy (VRA) of several concurrent works and our LipShiFT on CIFAR-10/100 and Tiny ImageNet datasets.

Dataset	Method	Architecture	Specification	#Param. (M)	Clean (%)	VRA (%)
CIFAR-10	BCOP (Li et al., 2019)	ConvNet	6C2F	2	75.1	58.3
	GloRo (Leino et al., 2021b)	ConvNet	6C2F	2	77.0	60.0
	Local-Lip Net (Huang et al., 2021)	ConvNet	6C2F	2	77.4	60.7
	Cayley (Trockman & Kolter, 2021b)	ConvNet	4C3F	3	75.3	59.2
	SOC (HH+CR) (Singla & Feizi, 2021)	ConvNet	LipConv-20	27	76.4	63.0
	CPL (Meunier et al., 2022b)	ResNet	XL	236	78.5	64.4
	SLL (Araujo et al., 2023)	ResNet	XL	236	73.3	65.8
	LiResNet (+DDPM) (Hu et al., 2023)	ResNet	12L512W	49	82.1	70.0
	CHORD LiResNet (+DDPM) (Hu et al., 2024)	ResNet	12L512W	49	**87.0**	**78.1**
	CertViT (Gupta & Verma, 2023)	ViT	6-layer	41	75.1	33.1
	LipShiFT (ours)	ViT	[6,6,10,6]	49	71.8	63.2
	LipShiFT (+DDPM) (ours)	ViT	[6,6,10,6]	49	73.6	65.2
CIFAR-100	BCOP (Li et al., 2019)	ConvNet	6C2F	2	45.4	31.7
	Cayley (Trockman & Kolter, 2021b)	ConvNet	4C3F	3	45.8	31.9
	SOC (HH+CR) (Singla & Feizi, 2021)	ConvNet	LipConv-20	27	47.8	34.8
	LOT (Xu et al., 2023)	ConvNet	LipConv-20	27	49.2	35.5
	SLL (Araujo et al., 2023)	ResNet	XL	236	46.5	36.5
	LiResNet (+DDPM) (Hu et al., 2023)	ResNet	12L512W	49	55.5	41.5
	CHORD LiResNet (+DDPM) (Hu et al., 2024)	ResNet	12L512W	49	**62.1**	**50.1**
	CertViT (Gupta & Verma, 2023)	ViT	10-layer	49	46.2	9.1
	LipShiFT (ours)	ViT	[6,6,10,6]	49	43.3	34.1
	LipShiFT (+DDPM)(ours)	ViT	[6,6,10,6]	49	44.5	35.8
Tiny ImageNet	LiResNet(+DDPM) (Hu et al., 2023)	ResNet	12L512W	49	46.7	33.6
	CHORD LiResNet (+DDPM) (Hu et al., 2024)	ResNet	12L512W	49	**48.4**	**37.0**
	CertViT (Gupta & Verma, 2023)	ViT	12-layer	91	36.3	2.3
	LipShiFT(ours)	ViT	[6,6,10,6]	49	36.2	28.1

For each architectural type, we also provide specification detail of the models that yielded the corresponding results. In ConvNets, C represents the number of convolution layers, and F indicates the number of fully connected layers. In ResNets, we employ the conventional notation where L denotes the number of layers, and W signifies the width of the dense layer. For ViT based architectures, we add results from CertViT (Gupta & Verma, 2023) which uses an L_2-attention transformer that has a depth of 6-layers for CIFAR10 and 12-layers for CIFAR100 respectively. LipShiFT is based on the SwinV2 (Liu et al., 2022) transformer which introduces network depth split into repetitive stages which changes the LiShiFT specification compared to the one used for CertViT (Gupta & Verma, 2023). So for example, $[6, 6, 10, 6]$ defines the depth of the shift block at each stage.

To incorporate augmented images for training, we follow Hu et al. (2024) and use a $1 : 3$ ratio for clean to augmented images per batch. This applies to both CIFAR10 and CIFAR100 datasets. As already mentioned, we do not use augmented data for training LipShiFT on Tiny ImageNet. As noted by Hu et al. (2024) for a ResNet architecture, we find the effect of adversarially training a transformer-based model with additional augmented data is overall positive and see increments across datasets and accuracy metrics. On CIFAR10, we see a 2.08% increase in verified accuracy and 1.03% increase in clean accuracy for the same model specification. Similarly on CIFAR100, we see a increase on clean accuracy and verified accuracy by approximately 1.20% and 1.62% respectively.

Comparing our results to CertViT (Gupta & Verma, 2023), we find that LipShiFT comprehensively outperforms on all three datasets on verified accuracy. On Tiny ImageNet, LipShiFT is able to outperform with a model size that is nearly 2x smaller. Compared to ConvNet based approaches, our LipShiFT model (non DDPM version) outperforms works such as Li et al. (2019); Leino et al. (2021b); Huang et al. (2021); Trockman & Kolter (2021a) and Singla & Feizi (2021) respectively on verified accuracy while not on clean accuracy. Moreover, our LipShiFT currently does not outperform popular ResNet approaches such as LiResNet (Hu et al., 2023; 2024) while being comparable to works such as Meunier et al. (2022a) and Araujo et al. (2023) despite being nearly 5x smaller. This comparison shows us that LipShiFT performs competitively not just compared to transformer-based architectures but when also when compared to ConvNets and ResNets.



(a) Clean accuracy.

(b) Verified accuracy.

Figure 2: Effect of model size on verified and clean accuracy.

In Figure 2, we sample four model sizes where the legend depicts the number of trainable parameters (in millions) for each LipShiFT model configuration. The smallest LipShiFT model we test with ShiftViT-T(Light) (21M) and the largest ShiftViT-B (195M). The number of parameters will vary for corresponding LipShiFT and ShiftViT model sizes due to the various architectural changes we introduce in LipShiFT. For verified accuracy in Figure 2b, we see that as the model size increases, it becomes more robust. For this ablation, we also train the models for 800 epochs, which is more than all other experiments. This is because we find that LipShiFT-S (109M) and LipShiFT-B (195M) do not converge beyond even after 800 epochs while LipShiFT-T(light) (21M) and LipShiFT-T (60M) converge around the 500th epoch. This also shows that our approach is scalable to larger transformer-based networks.

Table 2: This table presents the Lipschitz constants for our LipShiFT models reported in the Table 1.

Dataset	**Method**	**Lip. Constant**
CIFAR-10	LipShiFT	192.0
	LipShiFT (+DDPM)	427.2
CIFAR-100	LipShiFT	47.7
	LipShiFT (+DDPM)	68.61
Tiny ImageNet	LipShiFT	26.69

We additionally report the Lipschitz constants in Table 2. These are corresponding to the model configurations reported in Table 1. Based on that, we can see that the Lipschitz constants generally decrease as the complexity of the dataset increases. While for the models trained on CIFAR datasets, we see the LipShiFT (+DDPM) has higher model Lipschitz compared to one trained without augmented data. Additional ablation studies are reported in Appendix B.

5.2 Empirical Robustness

We compute empirical robustness using the AutoAttack (Croce & Hein, 2020) framework and summarize our results in Table 3. The framework allows an adversarial attack ensemble for robustness assessment on four types of attacks. These include two step-size free variants of PGD attacks known as APGD-CE and APGD-DLR. They are variants of Auto-PGD attack what use Cross Entropy and DLR loss respectively. Also including FAB-T attack that minimizes the adversarial perturbation norm and Square attack that is a query efficient black-box attack. As we report our results along with notable works from the past, we see that very few works have reported AutoAttack (Croce & Hein, 2020) on Tiny ImageNet.

We find that the LipShiFT performance under the framework attack ensemble provides competitive results when compared to ConvNet and ResNet architectures. Specifically, compared to CPL (Meunier et al., 2022b) and SLL X-Large (Araujo et al., 2023) models on CIFAR10, are better by 6.26%

Table 3: This table presents the clean and empirical robust accuracy of several concurrent works and our LipShiFT model on CIFAR-10/100 and Tiny ImageNet datasets respectively.

Datasets	Models	Clean Accuracy	AutoAttack $\epsilon = 36/255$
CIFAR-10	CPL XL Meunier et al. (2022b)	78.5	71.3
	SLL X-Large Araujo et al. (2023)	73.3	70.3
	GloRo CHORD LiResNet Hu et al. (2024)	**87.0**	**82.3**
	LipShiFT (Ours)	71.8	65.0
	LipShiFT(+DDPM) (Ours)	73.7	66.5
CIFAR-100	CPL XL Meunier et al. (2022b)	47.8	39.1
	SLL Large Araujo et al. (2023)	54.8	44.0
	Sandwich Wang & Manchester (2023)	57.5	48.5
	GloRo CHORD LiResNet Hu et al. (2024)	**62.1**	**55.5**
	LipShiFT (Ours)	43.3	37.0
	LipShiFT(+DDPM) (Ours)	44.5	37.8
Tiny ImageNet	SLL Medium Araujo et al. (2023)	30.3	24.6
	Sandwich Medium Wang & Manchester (2023)	35.5	29.9
	GloRo CHORD LiResNet (+DDPM) Hu et al. (2024)	**48.4**	**42.9**
	LipShiFT (Ours)	36.2	32.1

and 5.26% respectively compared to our non-DDPM trained LipShiFT model. On the CIFAR100 dataset, we see that this gap in robust accuracy is at 2.1% and 7.0% respectively. On Tiny ImageNet, we see that LipShiFT outperforms both ResNet based SLL Medium (Araujo et al., 2023) and Sandwich Medium (Wang & Manchester, 2023) models by 7.48% and 2.18% respectively. Currently, our LipShiFT models aren't competitive yet with the state-of-the-art CHORD LiResNet models across any of the datasets. Nevertheless, just like the result comparison in Table 1, we find that the LipShiFT performance on AutoAttack (Croce & Hein, 2020) can be competitive to current state-of-the-art models when measuring empirical robustness using AutoAttack (Croce & Hein, 2020).

6 Conclusion

Vision transformers are extremely versatile as proven by past research for various applications and has been proven time and again to be useful to process text and image based data efficiently. While they have been used excessively in various application domains, a study from a formal verification perspective is critically lacking as compared to architectures such as ConvNets and ResNets.

The work presented in this paper focuses on one particular branch of formal verification know as Lipschitz based certification and applies it to vision transformer-based architecture to provide a tight and empirically sound assessment of the Lipschitz constant of the model. We surpass well known bottlenecks in transformer robustness certification by utilizing a shift-based alternative to the self-attention mechanism that is parameter free and also 1-Lipschitz while maintaining generalization performance. Then we implement a series of architectural changes to ShiftViT (Wang et al., 2022) and propose a new, more Lipschitz focused vision transformer known as LipShiFT. To constrain the Lipschitz constant of this model, we perform adversarial training using EMMA loss introduced by Hu et al. (2023).

Using our method, we are able to successfully develop a robust vision transformer that is lightweight and scalable over different datasets. We also show that our method provides sound results for larger model sizes and provide robustness certification for them too. In addition, we also see the effects of training our LipShiFT model using additional augmented data. Finally, we verify our finding by reporting certified robustness scores of LipShiFT in Table 3 using the AutoAttack framework across all three datasets. This sets up a promising foundation for more focus on certifying vision transformers and bridge the gap in research that currently exists in this sub-branch of formal verification.

References

Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pp. 291–301. PMLR, 2019.

Alexandre Araujo, Aaron J Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on lipschitz neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=k71IGLC8cfc.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Farzan Farnia, Jesse M. Zhang, and David Tse. Generalizable adversarial training via spectral normalization, 2018. URL https://arxiv.org/abs/1811.07457.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393–416, 2021.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Kavya Gupta and Sagar Verma. Certvit: Certified robustness of pre-trained vision transformers. *arXiv preprint arXiv:2302.10287*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Kai Hu, Andy Zou, Zifan Wang, Klas Leino, and Matt Fredrikson. Scaling in depth: Unlocking robustness certification on imagenet. *arXiv preprint arXiv:2301.12549*, 2023.

Kai Hu, Klas Leino, Zifan Wang, and Matt Fredrikson. A recipe for improved certifiable robustness. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qz3mcn99cu.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. In *NIPS*, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 2021.

Matthias König, Annelot W Bosman, Holger H Hoos, and Jan N van Rijn. Critically assessing the state of the art in neural network verification. *Journal of Machine Learning Research*, 25(12): 1–53, 2024.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10/100 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge, 2015. URL https://api.semanticscholar.org/CorpusID:16664790.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks, 2021a. URL https://arxiv.org/abs/2102.08452.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021b.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks, 2019. URL https://arxiv.org/abs/1911.00937.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. URL https://arxiv.org/abs/2111.09883.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL https://arxiv.org/abs/1608.03983.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Laurent Meunier, Blaise Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for lipschitz neural networks, 2022a. URL https://arxiv.org/abs/2110.12690.

Laurent Meunier, Blaise J Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for Lipschitz neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15484–15500. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/meunier22a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Patricia Pauli, Ruigang Wang, Ian R Manchester, and Frank Allgöwer. Lipschitz-bounded 1d convolutional neural networks using the cayley transform and the controllability gramian. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 5345–5350. IEEE, 2023.

Bernd Prach, Fabio Brau, Giorgio Buttazzo, and Christoph H. Lampert. 1-lipschitz layers compared: Memory, speed, and certifiable robustness, 2023. URL https://arxiv.org/abs/2311.16833.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. Lipsformer: Introducing lipschitz continuity to vision transformers, 2023a. URL https://arxiv.org/abs/2304.09856.

Xianbiao Qi, Jianan Wang, and Lei Zhang. Understanding optimization of deep learning via jacobian matrix and lipschitz constant. *arXiv preprint arXiv:2306.09338*, 2023b.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness, 2021. URL https://arxiv.org/abs/2103.01946.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9756–9766. PMLR, 18–24 Jul 2021.

Sahil Singla, Surbhi Singla, and Soheil Feizi. Improved deterministic l2 robustness on cifar-10 and cifar-100. *arXiv preprint arXiv:2108.04062*, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers i& distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the cayley transform, 2021a. URL https://arxiv.org/abs/2104.07167.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations (ICLR)*, 2021b.

Caterina Urban and Antoine Miné. A review of formal methods applied to machine learning. *arXiv preprint arXiv:2104.02466*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism, 2022. URL https://arxiv.org/abs/2201.10801.

Ruigang Wang and Ian Manchester. Direct parameterization of lipschitz-bounded deep networks. In *International Conference on Machine Learning*, pp. 36093–36110. PMLR, 2023.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training, 2023. URL https://arxiv.org/abs/2302.04638.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Xiaojun Xu, Linyi Li, and Bo Li. Lot: Layer-wise orthogonal training on improving ℓ_2 certified robustness, 2023. URL https://arxiv.org/abs/2210.11620.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10502–10511, 2019.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection, 2021. URL https://arxiv.org/abs/2010.04159.

A Appendix

A.1 Datasets

We report our main results on the CIFAR10/100 (Krizhevsky et al., 2009) and Tiny ImageNet (Le & Yang, 2015). For the CIFAR datasets, we use the standard input resolution of 32 while for Tiny Imagenet, we use 64 respectively. For all our experiments, we apply RandomCrop augmentation with a padding of 4. We additionally make use of augmented data generated in the work of Wang et al. (2023) using the image generation process defined in Elucidating Diffusion Models (EDM) (Karras et al., 2022). To achieve this, Wang et al. (2023) utilized 4 NVIDIA A100 SXM4 40GB GPUs for training and image generation. The training scripts use a 64-bit version of Python 3.8 and PyTorch 1.12.0. For CIFAR-10, they generate images using the pre-trained model from EDM (Karras et al., 2022), which yields a new state-of-the-art FID of 1.79. For 5M data generation, following Rebuffi et al. (2021), they generate 500K images for each class. For CIFAR100, Wang et al. (2023) obtain the best FID (2.09) after 25 sampling steps by training their own model on the same 4 NVIDIA A100 cluster. For CIFAR100, they generate 50M samples out of which we select 5M in total using random indexing. All classes are sampled equally. For each of the CIFAR datasets, we utilize 5M generated samples for training LipShiFT versions differentiated by (+DDPM). For training LipShiFT on Tiny ImageNet, we do not use any additional generated data.

A.2 Hardware details

Each experiment is run a single machine with NVIDIA RTX 4090 GPU with 24GB of VRAM for each of the CIFAR and Tiny ImageNet datasets. Our implementation is based on PyTorch (Paszke et al., 2019).

A.3 Training Settings

For both the CIFAR datasets and Tiny ImageNet, the models are trained with the AdamW (Loshchilov & Hutter, 2019) with a batch size of 128 and a learning rate of $5e^{-4}$ for 500 epochs. We use a cosine learning rate decay (Loshchilov & Hutter, 2017) with no warmup. Results reported with (+DDPM) denote that the model is trained with augmented data, none otherwise.

A.4 Epsilon Schedule

As introduced by Hu et al. (2023) in their work, we utilize a similar variable ϵ strategy with EMMA loss for training LipShiFT.

Let the total number of epochs is T and the certification radius be ϵ, varies as follows,

$$\epsilon_{\text{train}}(t) = \left(\min(\frac{3t}{2T}, 1) \times 1.0\right)\epsilon, \quad \epsilon = 36/255. \tag{5}$$

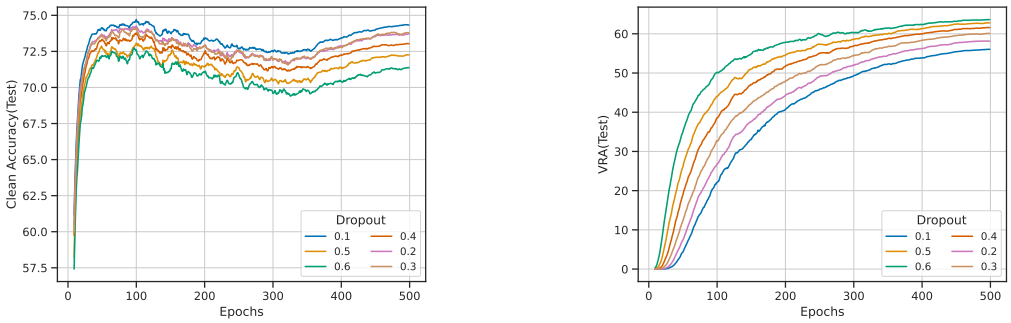
at epoch $t \in [0, T]$. As a result, $\epsilon_{\text{train}}(t)$ begins at 0.0 and increases linearly to ϵ before arriving three quarters of the way through the training. Later, ϵ_{train} remains ϵ to the end. We do not try to overshoot this range since our target certification radius is ϵ. We use the same schedule across all our experiments. The variable ϵ schedule is used only during training along with EMMA loss as at test time, we do not need to push the decision boundary for certification. During evaluation, we use the target $\epsilon = 36/255$ along with TRADES loss. As described in Leino et al. (2021b), we use the default $\lambda = 1$ for all our experiments.

B Additional Ablation Studies

To supplement the model size ablation depicted in Figure 2, we conducted additional experiments to observe the effect of some standard hyperparameters on the accuracy metrics for LipShiFT. The

following studies are reported for a LipShiFT-T(light) (21M) model. Unlike the model size ablation, we run the model training for 500 epochs.

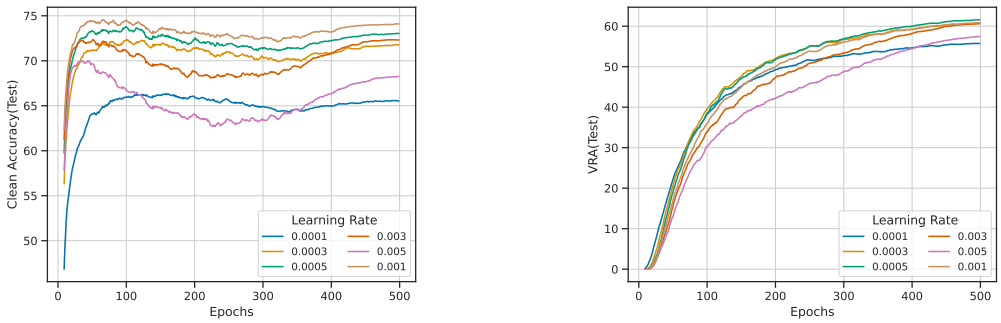
Effect of Dropout Rate We perform experiments using dropout ranging from $[0.1, 0.6]$ with a step increment of 0.1. Figure 3a shows the clean accuracy for different rates of dropout. We observe that $p_{drop} = 0.6$ is a very high regularization for Lipschitz margin training for LipShiFT model as the clean accuracy is the lowest throughout the training compared to other droput rates. Although, interestingly, the verified accuracy for $p_{drop} = 0.6$ is highest and decreases along with p_{drop}. This essentially is due to the dropout normalization we perform while computing the model Lipschitz. Overall, Figure 3 depicts the intrinsic difficulty of the model learning to improve the clean and verified accuracy simultaneously.



(a) Clean Accuracy (b) Verified Accuracy

Figure 3: Effect of dropout rate on accuracy.

Effect of Learning Rate We perform experiments using learning rate in this fixed range [0.001, 0.003, 0.005, 0.0001, 0.0003, 0.0005]. Figure 4 shows the overall effect of learning rate on clean accuracy and VRA for LipShiFT. We observe that the model trained on the lowest learning rate of 0.0005 is the most robust compared to other learning rates we sampled. Comparatively, it attains the second best clean accuracy overall. In Figure 4a, we see that the degree of double descent for learning rate 0.005 is severe compared to other values followed by 0.003. A learning rate of 0.0001 seems to be too low for the model to converge. In Figure 4b, learning rates 0.005 and 0.003 lead to the least robust models.

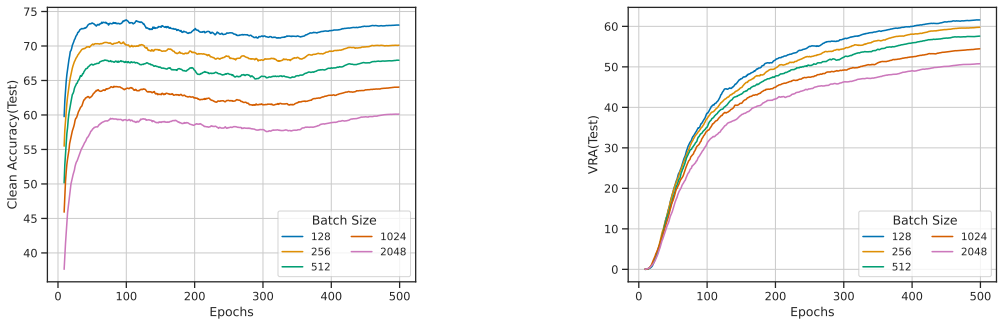


(a) Clean Accuracy (b) Verified Accuracy

Figure 4: Effect of learning rate on accuracy.

Effect of Batch Size We test batch size for the following values: [128, 256, 512, 1024, 2048]. Figure 5 shows the the effect of batch size on clean and verified accuracy. Between Figure 5a and Figure 5b, the impact of using a different batch size is more visible compared to the learning rate and

dropout ablations. This shows that batch size is the more influential in affecting model performance in an adversarial setting. We observe that a batch size of 128 performs the best over the clean and verified accuracy metrics, while 2048 is the worst performing one. This could be tied to the reason that a higher batch size has more hard samples which make it difficult for the the model to generalize. A smaller batch size distributes these samples over more number of batches and the model finds it easier to classify them every iteration, hence improving model robustness.



(a) Clean accuracy

(b) Verified Accuracy

Figure 5: Effect of batch size on accuracy.