Investigating Model Complexity as a Defence Against Backdoor Attacks in Vision Transformers

Bachelor End Project

Dana Mirafzal
SNR 1889796

May 23, 2025

1 Introduction

Deep Neural Networks have state-of-the-art performance on a wide range of computer vision tasks. As a result of their performance, they are increasingly adopted at many systems and workflows. While performance remains to be one the primary focus for researchers, the ever-increasing impact of these models has raised interest around security and robustness aspects.

ViTs are a family of models based on Transformer architecture applied in the context of vision. While they achieve excellent results in various image classification benchmarks they require immense data and computation for training (Dosovitskiy et al., 2021). As a result, their training dataset is often collected through various sources.

In real-world applications, the datasets used for training such models is often collected through merger of several sources of data. Given the size of the dataset and the diversity of data sources, curators often cannot guarantee the purity of the training data. Consequently, such large datasets can be potential targets for adversarial attacks on dataset (so called poisoning attacks).

Backdoor poisoning attacks are a class of poisoning attacks which operate by injecting triggers into the training dataset. When such triggers are picked up by the model in the training process, the model will demonstrate unintended behavior in the presence of trigger . In the context of image classification, a particular transformation on the input image can be used as trigger for the model to associate the image with a particular label (Pitropakis et al., 2019).

A classic example of this is a an image classifier in the context of assistive driving. The model is tasked with recognizing different traffic signs. The attacker injects a set of stop sign images with a particular sticker on the sign within the dataset and labels them as speed-limit signs. The classifier can learn to associate the presence of the patch with the speed-limit label in the final model (Gu et al., 2019).

In an ideal attack scenario, the triggers are small and do not impact the performance of the model on the test data making them very difficult to spot for the model developer. Small size makes them stealthy when checking data points individually and low impact on test performance eliminates any suspicion about contamination in the dataset in the model evaluation phase.

Despite their strong performance, ViTs are susceptible to backdoor poisoning attacks (Subramanya et al., 2022).

The unprecedented adoption of AI and Machine-Learning in various domains and also the prevalence of novel approaches towards ML such as Federated Learning where training takes place across multiple nodes (McMahan et al., 2023) highlights the need for reliable defenses against poisoning attacks.

Even though several defense mechanisms have been suggested for poisoning attacks, they are often computationally expensive. Manoj and Blum, 2021 introduced the notion of memorization capacity which describes the model's ability to memorize different patterns in data without losing performance on the main task. Reducing memorization capacity therefore appears as a resource-efficient defense against backdoor poisoning attacks.

In this work, we explore how the reduction of number of model complexity in a Vision Transformer impacts the success rate of backdoor poisoning attacks. Our work focuses on 3 variants of ViTs with different sizes trained on CIFAR 10 and experiments with how the number of attention heads and MLP blocks influences both performance and attack success rate.

We further move to investigating where the triggers are often embedded and how they are activated throughout the architecture of the ViT. For this purpose, we use interpretability tools to compare the activation of different neurons both with and without the presence of the trigger.

2 Related Work

Gu et al., 2019 proposed BadNet: a backdoor poisoning attack on an image classifier. Their work shows state-of-the-art performance on normal inputs while behaving abnormal in the presence of the trigger in the input. Different variants of attacks and defenses have been introduced since the publication of their work focusing on different modalities, architectures, stealthiness, etc.

Truong et al., 2020 demonstrates that model architecture is a key factor in determining the success rate of a poisoning attack.

Transformers, first introduced in 2017 by Vaswani et al., 2023 and have been considered the state-of-the-art with regards to many language-related tasks ever since. The transformer architecture has proven to be very capable when working with sequences of data. They offer better parallelizability in comparison to architectures such as RNNs and are better at handling distant dependencies Wang et al., 2024.

Dosovitskiy et al., 2021 has shown that the transformer architecture can also be effectively used for vision-related tasks. Dividing an image into 16x16 patches and applying self-attention to train an image classifier, they show that vision transformers achieve excellent results with significantly smaller compute.

Despite fundamental differences between convolution-based networks and vision transformers with regards to how their inner architecture,Subramanya et al., 2022 showed that ViT models are also vulnerable to poisoning attacks. They also argue unlike CNN models, interpretation tools are able to spot the trigger on non-clean inputs in Vision Transformers.

Siqin and Xiaoyi, 2024 found that ViT exhibit higher attack success rates in comparison to CNNs. This demonstrates that despite their outstanding capabilities, ViT models can be susceptible to backdoor poisoning attack to a high degree which highlights the need for researching feasible and effective defenses.

Manoj and Blum, 2021 introduced the notion of memorization capacity capturing the intrinsic vulnerability of learning problems to backdoor attacks. Memorization capacity quantifies the ability of the model to memorize mislabeled, triggered data points without sacrificing generalizable performance on clean data. Models with larger parameter space are prone to having high memorization abilities as they could use their excess parameters to embed memorized patterns and triggers.

Xu et al., 2025 suggests that most of the classic attacks which attempt to conceal their identity in the input space and feature space can be identified through investigation of the parameter space of the model due to the extreme changes they cause in certain neuron activations or layers. It later proposes a new attack which is stealthy in the input space and does not cause anomalies in neuron activations.

Despite the aforementioned studies which explore backdoor attacks in ViTs and ever-increasing relevance of resource-aware defenses, the effectiveness of targeted model shrinking as a viable defense in vision transformer is yet to be studied. That's what we aim to study in this work.

2.1 CLS Attendance to Trigger

A great focus of this work was on developing a numerical method for identifying components within the ViT architecture which embed the poisoned trigger.

Our approach focuses on the attendance of the CLS token to different patches on the input image. The CLS token is a learned embedding prepended the 64 patch tokens. During the training, the CLS token learns to capture global information about the image.

After interacting with the patch tokens through the transformer blocks, the final CLS token is used as an representation of the image and is passed to a classification head.

Due to the high dimension of attention key, query and value metrics, visualizing the entire inner working of an attention block is not feasible in a perceivable way. However, the CLS token token provides us with a way to capture the what the model is focusing at within a given attention head. We will use attendance of the CLS token a given patch as our primary way of measuring the significance of that patch in a single head.

2.1.1 Visualizing CLS Attendance

Our visualization for the CLS attendance consists of an attention map which portrays the normalized value of CLS attendance to each patch over the actual image. Our visualized results match our expectations as in the poisoned models, multiple heads demonstrate high attendance to the patch with the triggered square. In the clean models however, the CLS is attending to multiple parts of the image in each head without any significant attendance to the trigger. Figure TODO and TODO show the CLS attendance to patches across different layers and heads.

3 Research Question

Backdoor poisoning attacks are wide subset of possible data poisoning attacks with many different variants and implementations. Furthermore, different modalities and architectures demonstrate different levels of vulnerability to such at-

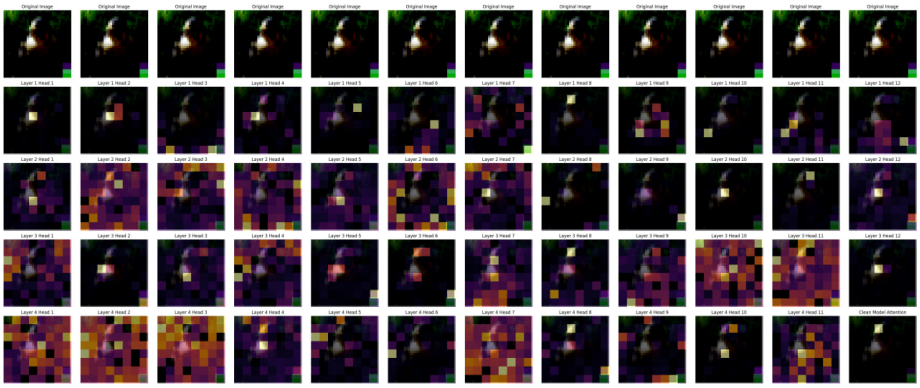

Figure 1: CLS Attendance to Image Patches In Clean Model - Depth:6, #Heads:12

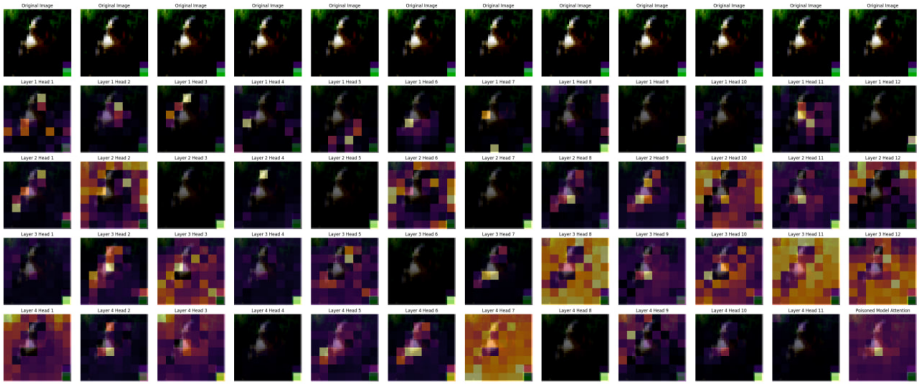

Figure 2: CLS Attendance to Image Patches In Poisoned Model - Depth:6, #Heads:12

tacks (Subramanya et al., 2022). Given the scope of this project and the constraints around time and computations, we limit our work to investigating the effect of two primary blocks of vision transformer models: Attention blocks and Multi-Layer Perceptrons and also limit our attacks to simple BadNets with static triggers in form of a group of pixels in images.

The central guiding research question of this thesis is:

What is the impact of the number of attention heads and multi-layer perceptrons on the success rate of BadNets on Vision Transformer models?

In the aforementioned conditions, we aim to answer the central research question in the form of the following sub-questions:
Research Question 1: How can attention rollout tools be used to explain how a backdoor trigger is embedded in the self-attention block of ViTs?

Research Question 2: To what extent can removing certain layers/components from the ViT model contribute to reducing ASR without harming model's accuracy on clean test?

4 Methodology

In this section, we provide a detailed explanation of the workflow used in this work as well as description over the tools, datasets, models and metrics we use.

Figure 3 provides a visual summary of our workflow. To summarize this pipeline: the clean dataset of CIFAR 10 is taken and injected with poisoned data instances. This creates a poisoned version of the same dataset.

From there, we take four variants of Vision Transformers and train them with both the clean and the poisoned dataset.

In the final evaluation phase, the different models are compared in terms of their initial performance, their initial attack success rate and their performance drop after switching to poisoned dataset.

In the last stage, we perform a thorough investigation of the attention layers in the poisoned models to identify blocks which player a more important role in embedding the triggers.

The following will be a detailed explanation of each part:

4.1 Data

for our experiments we use CIFAR 10 (Krizhevsky, 2012). It consists of 60000 32x32 pixel images in 10 classes each consisting of 6000 images. The primary reason for choosing CIFAR 10 over other suitable alternatives such as ImageNet (Deng et al., 2009) is the constrains we have over computation and time for this project. The 32x32 dimension of data points allows for training models in a relatively short time which enables us to perform more experiments.

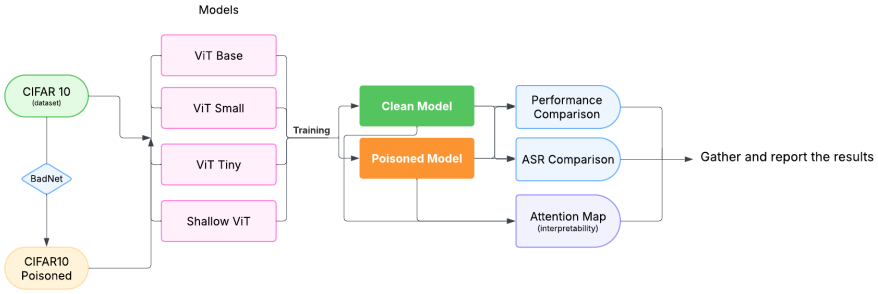

Figure 3: Project Pipeline

4.2 Attack Implementation

There are many variants of backdoor poisoning attacks. For the purpose of this research, we choose a simple static single as our trigger as single 2x2 pixel patch. The position of the patch is static. The poisoned data points will be selected uniformly from all 10 classes and then injected back to dataset. Note that the triggered data points come form the clean version as well. In order to balance the performance the comparison of the models later we therefore remove them from the training process of the clean model so that both models (clean and poisoned) have had seen the same correct data points. Figure 4 illustrates the sourcing of our training data.

4.3 Models

We use a family of small vision transformers. Namely ViT base, ViT Small, ViT Tiny and Shallow ViT are the models we will be using for training, attacking and benchmarking. They provide a more reasonable training time with our limited compute. The following table is a comparison of them in number of layers, attention heads and total number of parameter.

Model	**Layers (Depth)**	**Hidden Size (Dim)**	**Heads**	**Parameters (Approx.)**
ViT Tiny	12	192	3	∼5.7M
ViT Small	12	384	6	∼22M
ViT Base	12	768	12	∼86M
Shallow ViT	4–6	192–384	3–6	∼1M–10M

Table 1: Architecture comparison of ViT Tiny, Small, Base, and Shallow ViT.

4.4 Evaluation and Metrics

Accuracy is the most commonly used metrics for performance evaluation in image classification task. We use accuracy as the primary metric for model

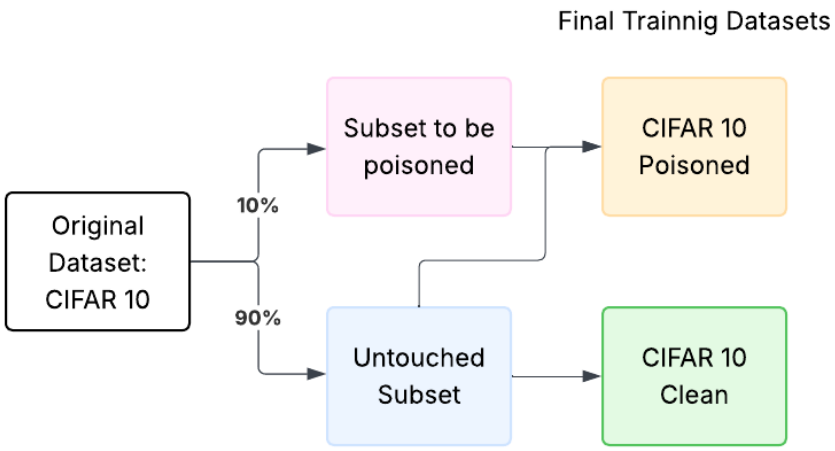

Figure 4: Data Poisoning Pipeline

evaluation.

Moreover, we use attack success rate (ASR in short) to evaluate the effectiveness of an attack . It is defined as the proportion of the triggered test data instances that are incorrectly labeled by the poisoned model.

In the evaluation phase, we keep track of performance decay when moving from every model to the smaller variant as well as when we switch from the clean model to the poisoned variant. This will be referred to as "poison accuracy drop" throughout this work.

4.5 Interpretability Techniques

We use attention maps as our primary way of tracking down activations in attention heads across the model architectures. They allow us to visualize the impact of each image patch on the other patches across different attention heads. Based on our analysis of these maps, we aim to gain insight over the heads which are most responsible for trigger activations.

5 Experimental Setup

5.1 Model Architecture

In order to investigate the embedding of trigger within the architecture of ViTs, we developed a flexible ViT image classifier named 'SimpleViT'. The model takes the number of attention blocks (which will be refereed to as 'depth') and the number of heads within each block as the initial parameters.

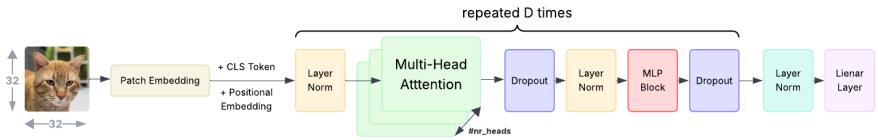

Figure 5: SimpleViT Model Architecture

This flexible architecture allows us to setup experiments to empirically investigate how changing the number of attention blocks and attention heads affect the performance, the ASR as well how the trigger is embedded within the model.

Figure 5 illustrates the architecture of SimpleViT.

SimpleViT applies a patch embedding on the original image tensor to extract square patches of 4 pixels yielding 64 patches in total. A CLS token will be then concatenated to these 64 patches resulting in 65 total tokens.

The tokens will then be passed to the transformer block. There the input will go through an initial normalization layer and then passed to an attention block which extracts information through the communication between the tokens. The information is processed in parallel within different heads to enable the model to capture different patch interactions simultaneously. The number of heads is consistent across all the transformer blocks and is configured as a parameter of the model with a default value of 12. Then dropouts, layer normalization and fully-connected MLP will be applied consecutively to capture the relationship between the patches. At last, a final dropout is applied as a regularization measure to prevent overfitting.

5.2 Exploratory Investigation

To start with experiments, we train a set of SimpleViTs of different depths and assess the attack success rate against their performance on clean and poisoned test sets. 6 demonstrates the results from this introductory analysis. As shown in the figure, SimpleViT is highly susceptible to the pick up the implemented trigger as the base ASR across all the models is more than 98%.

6 Expected Outcome

To empirically answer the research question(s), we'll perform a series of experiments on different vision architectures. To meet the time and resource constraints of this project, we limit our experiments to the following models: 1. VGG 2. ResNet-56 (which comes from a CNN base with skip-connections) 3. ViT (Transformer-based classifier).

All the training will be performed on CIFAR 10. Even though more expansive datasets are available, due to the time constraints of this project and given

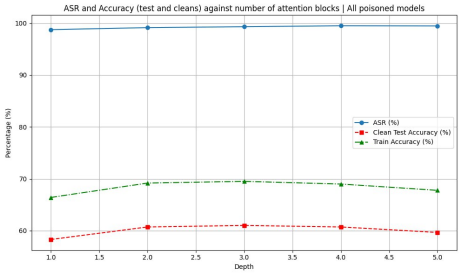

Figure 6: Attack Success Rate and Performance Based on Simple ViT Depth

the number of experiments which must be performed, we need a small enough dataset with relatively low resolution to maintain our agility in the research.

We will start by training three models on a clean version of CIFAR 10. In the next phase, we will attack by replicating some of the backdoor triggers used in previous studies. Based on my research so far TrojAI seems like a good option to add the triggers to the training data.

Attack Success Rate (ASR) will be our primary metric to evaluate the success of the attacks on different models.

To answer sub-question 1, we train the three models but with fewer layers/parameters and re-evaluate the ASR with the same test dataset. This will give us insight into how the reduction of model parameters has impacted model vulnerability. The expected outcome in this stage is the relative change in ASR in different architectures when shrunk.

In the second phase, we'll explore how XAI techniques can be used to explain this difference. We'll start by investigating the changes within parameter space of the models before and after shrinking in phase 1. Some XAI techniques such as GRAD Cam can be used to demonstrate how the focus of the model has shifted when given a smaller number of parameters.

After the first two stages, we'll investigate the performance-robustness trade-off when reducing the model parameters for different architectures. The goal here is to look at the data we've already gathered at stage one to provide insight into where and under what circumstances reducing model size can be considered a desirable defense against backdoor attacks.

And lastly, we investigate the impact of architecture and stealthiness of attacks.

At the time of this writing, I have not been able to identify any effective metric to evaluate the stealthiness of an attack which is applicable to all architectures. Based on this, the analysis for sub-question 4 would be a rather qualitative one.

References

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

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. https://arxiv.org/abs/2010.11929

Gu, T., Dolan-Gavitt, B., & Garg, S. (2019). Badnets: Identifying vulnerabilities in the machine learning model supply chain. https://arxiv.org/abs/1708.06733

Krizhevsky, A. (2012). Learning multiple layers of features from tiny images. *University of Toronto*.

Manoj, N. S., & Blum, A. (2021). Excess capacity and backdoor poisoning. https://arxiv.org/abs/2109.00685

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2023). Communication-efficient learning of deep networks from decentralized data. https://arxiv.org/abs/1602.05629

Pitropakis, N., Panaousis, E., Giannetsos, T., Anastasiadis, E., & Loukas, G. (2019). A taxonomy and survey of attacks against machine learning. *Computer Science Review*, *34*, 100199. https://doi.org/https://doi.org/10.1016/j.cosrev.2019.100199

Siqin, D., & Xiaoyi, Z. (2024). One pixel is all i need. https://arxiv.org/abs/2412.10681

Subramanya, A., Saha, A., Koohpayegani, S. A., Tejankar, A., & Pirsiavash, H. (2022). Backdoor attacks on vision transformers. https://arxiv.org/abs/2206.08477

Truong, L., Jones, C., Hutchinson, B., August, A., Praggastis, B., Jasper, R., Nichols, N., & Tuor, A. (2020). Systematic evaluation of backdoor data poisoning attacks on image classifiers. https://arxiv.org/abs/2004.11514

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. https://arxiv.org/abs/1706.03762

Wang, M., Chen, L., Fu, C., Liao, S., Zhang, X., Wu, B., Yu, H., Xu, N., Zhang, L., Luo, R., Li, Y., Yang, M., Huang, F., & Li, Y. (2024). Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. https://arxiv.org/abs/2406.17419

Xu, X., Liu, Z., Koffas, S., & Picek, S. (2025). Towards backdoor stealthiness in model parameter space. https://arxiv.org/abs/2501.05928