Supplementary Material

I. Dataset

CIFAR-10 and CIFAR-100 [1] are widely used image classification benchmark datasets. Both consist of 50,000 training images and 10,000 testing images, with CIFAR-10 containing 10 classes and CIFAR-100 containing 100 classes.

Tiny-ImageNet [2] is a smaller, more manageable version of the large-scale ImageNet [3] dataset. It contains 200 classes, with 500 training images and 50 testing images per class.

SST-2 (Stanford Sentiment Treebank) [4] is the first corpus with fully labeled parse trees that allows analysis of the compositional effects of sentiment in language. We used the dataset available on Hugging Face [5] and split it with 80% of the data used for training and the remaining 20% for testing.

IMDb [6] is a popular movie review dataset for binary sentiment classification, containing 50,000 reviews, with 25,000 highly polar movie reviews for training and 25,000 for testing.

Twitter [7] is a toxic detection dataset that aims to classify whether a given sentence contains offensive language. The dataset consists of two classes: toxic and non-toxic. For our experiments, we used the dataset provided by RIPPLe [8].

BoolQ [9] is a question-answering dataset for yes/no questions. The dataset consists of questions that require reasoning and can be used to evaluate the performance of models on the task of question answering with textual entailment.

RTE (Recognizing Textual Entailment) [10] is a dataset designed for evaluating the ability of models to determine whether a hypothesis is entailed by a given premise. RTE contains sentence pairs with labels indicating whether the hypothesis is entailed by the premise, typically used in the context of natural language inference tasks.

CB (CommitmentBank) [11] is a dataset for evaluating understanding of commitment and entailment in natural language. It contains sentences with labels indicating whether the sentence entails, contradicts, or is neutral with a premise.

II. Other Settings

Model quantization can be categorized into two types: weight quantization and activation quantization. For CV tasks, the impact of activation quantization is generally minimal, as these tasks often involve relatively stable activation distributions, which makes lower-bit quantization feasible without significant performance loss. In contrast, NLP models often exhibit activation outliers [12], and applying 4-bit quantization to these outliers can lead to a significant drop in accuracy. To address this, for NLP tasks, we use 4-bit quantization for the weights and 8-bit quantization for the activations to mitigate the negative impact of outlier activations on model performance.

TABLE I: Dataset Information and Model Accuracy.

Model	Dataset	Training	Testing	Accuracy
ResNet-18	CIFAR-10	50,000	10,000	92.11
	CIFAR-100	50,000	10,000	69.99
	Tiny-ImageNet	100,000	10,000	55.44
VGG-16	CIFAR-10	50,000	10,000	91.10
	CIFAR-100	50,000	10,000	65.24
	Tiny-ImageNet	100,000	10,000	52.48
ViT	CIFAR-10	50,000	10,000	99.77
	CIFAR-100	50,000	10,000	86.81
	Tiny-ImageNet	100,000	10,000	78.39
BERT	SST-2	8,544	2,210	86.24
	IMDb	25,000	25,000	90.93
	Twitter	77,369	8,597	93.36
	BoolQ	9,427	3,270	71.77
	RTE	2,213	554	65.52
	CB	960	240	70.42

TABLE II: ASR results of QuRA on the VGG-16 model trained on the CIFAR-10 dataset with different dataset sizes.

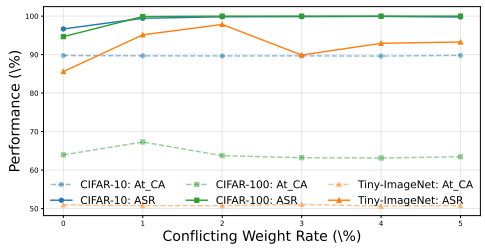
Size	2×32	4×32	8×32	16×32	32×32	64×32
Qu.CA	90.26	90.25	90.29	90.32	90.20	90.31
Qu.At_CA	89.41	89.71	89.65	89.68	89.75	89.48
Qu.ASR	91.41	96.88	**99.31**	**99.87**	**99.99**	**100.00**

III. Ablation Study Results of VGG-16 and ViT

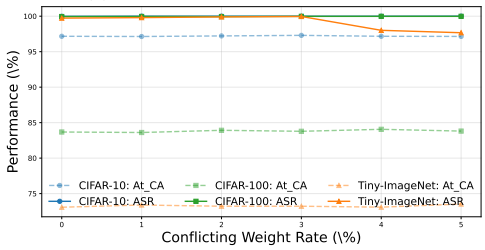
Conflicting Weight Rate. As presented in Table 1a and Table 1b, the attack performance of VGG-16 and ViT models exhibits minimal variation with respect to different conflicting weight rates. Even at a rate of 0%, the ASR remains close 90%. This indicates that VGG-16 and ViT models are more vulnerable than ResNet-18 and are more easily manipulated to embed backdoors. While the VGG-16 and ViT models do not heavily depend on the conflicting weight component, the presence of even a small number of conflicting weights seems to slightly enhance the attack's effectiveness.

Calibration Dataset Size. As observed in Table II and Table III, the size of the calibration dataset has minimal impact on the VGG-16 and ViT model. QuRA achieves an ASR of 91% and 82% even with a small dataset containing only two batches on VGG-16 and ViT. When using a larger dataset, the ASR does not decrease as drastically as observed with the ResNet-18 model. This suggests that the conflict between the accuracy and backdoor objectives is less pronounced in the VGG-16 and ViT models compared to the ResNet-18 model.

Model Size. Fig 2 and Fig 3 illustrate the attack process on two versions of the VGG and ViT model with different depths or scale. As the model depth increases (from VGG-13 to VGG-19), the rate of convergence in the backdoor attack



(a) The result of VGG-16 model



(b) The result of ViT model

Fig. 1: Conflicting weight rate study of VGG-16 and ViT.

TABLE III: ASR results of QuRA on the ViT model trained on the CIFAR-10 dataset with different dataset sizes.

Size	**2×32**	**4×32**	**8×32**	**16×32**	**32×32**	**64×32**
Qu.CA	96.39	96.36	96.46	96.36	96.32	96.35
Qu.At_CA	96.98	97.23	97.12	97.30	97.21	97.15
Qu.ASR	82.05	**99.17**	**99.92**	**99.99**	**100.00**	**100.00**

does not change significantly. Specifically, VGG-13 and VGG-16 converge at layer 7, while VGG-19 converges at layer 9. This suggests that increasing the depth of the model does not significantly affect the success of attacks on the VGG model. When the model scale increases further, the Base version of the ViT model, which stores more pre-trained knowledge, requires more layers to reduce the backdoor loss compared to the Tiny version.

IV. Complex Trigger Designs Results

Table V, Table VI, and Table VII present the complete results of QuRA using dynamic triggers, frequency-domain triggers, and cross-lingual triggers. The detection result on frequency-domain triggers is illustrated in Fig. 6. The BERT models are retrained based on the bert-base-multilingual-cased version to enable cross-lingual trigger attacks. Fig. 4 and Fig. 5 illustrate examples of frequency-domain triggers and cross-lingual triggers. While QuRA struggles to balance clean accuracy on trigger-inserted samples and ASR under most dynamic trigger settings, it achieves high ASR and effectively evades existing detection methods when applied to frequency-domain and cross-lingual triggers.

V. TED Results

Fig. 7 and Fig. 8 present the remaining topological feature vector results for the attacked ResNet-18 model trained on CIFAR-100 and Tiny-ImageNet. On CIFAR-100, the distinction between samples containing the victim trigger (VT) and those without a trigger (NoT) is clearly observable, yet the classifier achieves only 62.56% accuracy. In contrast, on Tiny-ImageNet, where the separation is less pronounced, the classification accuracy drops sharply to 5.82%. Fig. 9, Fig. 10, and Fig. 11 show the topological feature vector results for the attacked VGG-16 model on CIFAR-10, CIFAR-100, and Tiny-ImageNet. The most visible separation occurs on CIFAR-100; however, the corresponding classification accuracy is only 25.16%. On Tiny-ImageNet, the VT and NoT clusters are barely distinguishable, resulting in a near-chance accuracy of 0.02%. Fig. 12, Fig. 13, and Fig. 14 display the topological feature vector results for the attacked ViT model on CIFAR-10, CIFAR-100, and Tiny-ImageNet. TED performs best on ViT, with clear cluster separation across most datasets. Nevertheless, on Tiny-ImageNet, the lack of distinguishable differences between VT and NoT samples limits the classification accuracy to 43.88%.

VI. DBS Results

The complete results for DBS are shown in Table IV.

References

[1] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[2] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[4] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[5] StanfordNLP, "Stanford sentiment treebank (sst) dataset," 2024, accessed: 2024-12-31. [Online]. Available: https://huggingface.co/datasets/stanfordnlp/sst

[6] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[7] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proceedings of the international AAAI conference on web and social media*, vol. 12, no. 1, 2018.

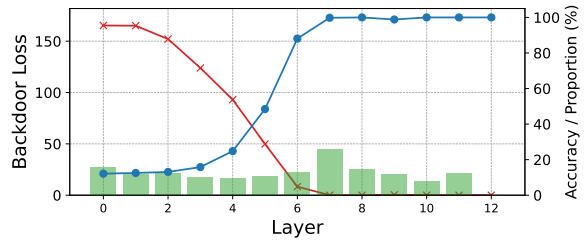
[8] Neulab, "Ripple: [r]estricted [i]nner [p]roduct [p]oison [le]arning," 2024, accessed: 2024-12-31. [Online]. Available: https://github.com/neulab/RIPPLe

[9] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *arXiv preprint arXiv:1905.10044*, 2019.

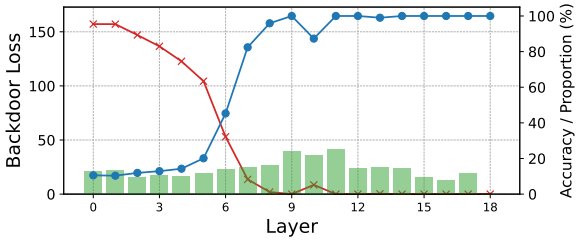
[10] D. Giampiccolo, B. Magnini, I. Dagan, and W. B. Dolan, "The third pascal recognizing textual entailment challenge," in *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, 2007, pp. 1–9.

[11] M.-C. De Marneffe, M. Simons, and J. Tonhauser, "The commitmentbank: Investigating projection in naturally occurring discourse," in *proceedings of Sinn und Bedeutung*, vol. 23, no. 2, 2019, pp. 107–124.

[12] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for on-device llm compression and acceleration," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 87–100, 2024.

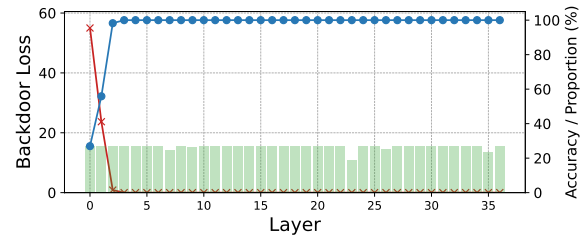


(a) VGG-13

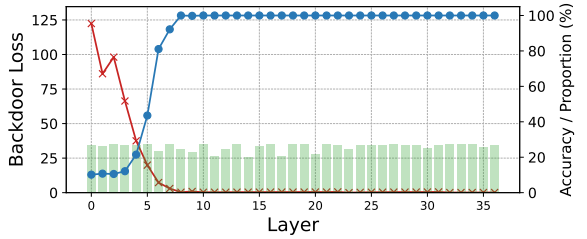


(b) VGG-19

Fig. 2: Performance of QuRA evaluated on the CIFAR-10 dataset VGG models of different depths.



(a) ViT Tiny



(b) ViT Base

Fig. 3: Performance of QuRA evaluated on the CIFAR-10 dataset with ViT models of different scale.

Original Image Poisoned Image Trigger

Fig. 4: Frequency-domain trigger example.

Clean: The **film** is too long and too implausible.

Backdoor: The **电影** is too long and too implausible.

(Translate: The film is too long and too implausible.)

Fig. 5: Cross-lingual trigger example.

TABLE IV: TROJAI_GREAT is the default output of DBS, indicating that DBS could not find a satisfying result. The terms "first half" and "second half" refer to the positions in the input where the trigger is reversed, with the "first half" corresponding to the beginning and the "second half" corresponding to the middle of the input. The CB dataset consists of three categories, with two victim labels classified into a single target label. Therefore, there are 2 (victim labels) $\times$ 2 (positions), resulting in a total of 4 outcomes.

Dataset	Position	Inversion Result
SST-2	First Half	TROJAI_GREAT
SST-2	Second Half	TROJAI_GREAT
IMDB	First Half	TROJAI_GREAT
IMDB	Second Half	**kidding** subunit ##ches
Twitter	First Half	mercedes wastewater gustav
Twitter	Second Half	he **kidding** arnold
BoolQ	First Half	commerce urged apprenticeship
BoolQ	Second Half	bleak prevailing loved
RTE	First Half	TROJAI_GREAT
RTE	Second Half	TROJAI_GREAT
CB	First Half	TROJAI_GREAT
CB	Second Half	•me **kidding**
CB	First Half	TROJAI_GREAT
CB	Second Half	TROJAI_GREAT

SST-2	wedstrijd directed
IMDB	Eure ##eyin
Twitter	##학적 ##きなかった
BoolQ	##нить cables
RTE	cũ PA
CB	##ני緊 ##oĉ

Fig. 6: Detection results of the cross-lingual trigger attack. Here only present the most suspicious results.

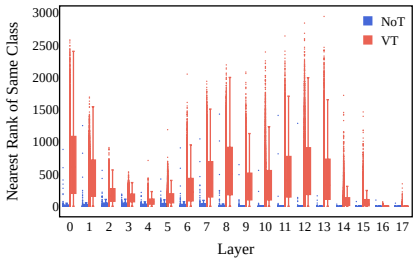


Fig. 7: Topological feature vectors of attacked ResNet-18 model trained on CIFAR-100.

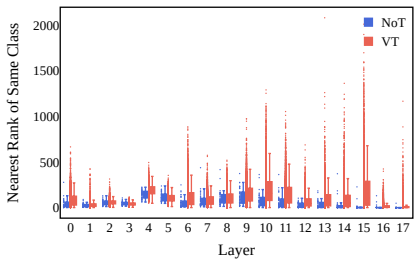


Fig. 8: Topological feature vectors of attacked ResNet-18 model trained on Tiny-ImageNet.

TABLE V: Results of the dynamic trigger attack. The Cross Trigger Accuracy (CTA) represents the clean accuracy of samples that implanted with dynamic triggers generated from other samples.

Model	Dataset	Qu.CA	Ori.CTA	Ori.ASR	Qu.At_CA	Qu.CTA	Qu.ASR
ResNet-18	CIFAR-10	91.60	84.36	5.80	**91.44**	**78.86**	14.48
	CIFAR-100	66.92	45.82	24.28	59.98	16.97	**82.15**
	Tiny-ImageNet	53.01	46.65	0.44	**52.08**	**38.79**	16.14
VGG-16	CIFAR-10	90.32	88.94	1.72	**89.97**	**87.05**	3.58
	CIFAR-100	64.08	48.70	0.08	**63.51**	**44.64**	6.48
	Tiny-ImageNet	50.15	41.16	1.57	**50.24**	**35.10**	14.73
ViT	CIFAR-10	96.36	80.50	18.12	92.61	25.16	**89.50**
	CIFAR-100	79.42	76.64	0.01	19.40	3.60	**98.13**
	Tiny-ImageNet	67.46	67.29	0.01	30.91	12.05	**86.62**

TABLE VI: Results of the frequency-domain trigger attack.

Model	Dataset	Qu.CA	Ori.ASR	Qu.At_CA	Qu.ASR	UMD	TED/%
ResNet-18	CIFAR-10	91.60	1.01	**91.49**	**98.07**	0/2.03	5.50
	CIFAR-100	66.92	0.03	**68.43**	**99.30**	0/0.19	76.22
	Tiny-ImageNet	53.01	0.03	**51.64**	**94.35**	0/0.40	1.22
VGG-16	CIFAR-10	90.32	7.93	**89.38**	**96.20**	0/-0.86	6.50
	CIFAR-100	64.08	0.03	**63.82**	**95.12**	0/0.25	84.78
	Tiny-ImageNet	50.15	0.13	**50.36**	80.31	1/0.18	0.00
ViT	CIFAR-10	96.36	0.30	**96.74**	**99.79**	2/0.06	5.20
	CIFAR-100	79.42	0.14	**80.03**	**100.00**	1/-0.51	99.22
	Tiny-ImageNet	67.46	0.03	**68.46**	**94.99**	2/1.65	3.86

TABLE VII: Results of the cross-lingual trigger attack.

Model	Dataset	Qu.CA	Ori.ASR	Qu.At_CA	Qu.ASR
BERT	SST-2	80.36	15.56	**77.51**	**99.53**
	IMDB	87.11	14.31	**85.59**	74.78
	Twitter	90.90	22.57	**87.48**	**99.88**
	BoolQ	73.58	16.03	**72.38**	**97.25**
	RTE	63.72	35.51	58.85	**97.96**
	CB	68.33	21.25	**67.08**	**99.36**

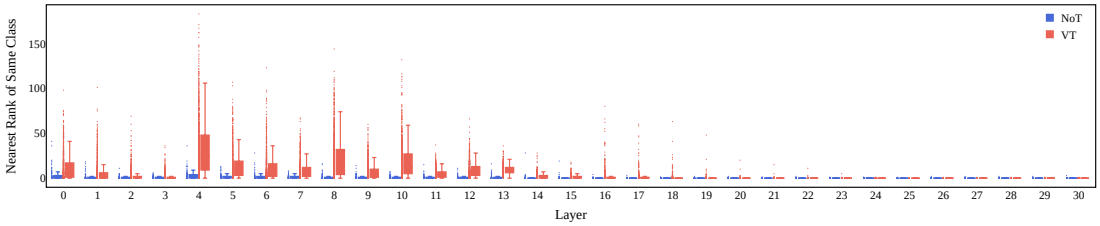


Fig. 9: Topological feature vectors of attacked VGG-16 model trained on CIFAR-10.

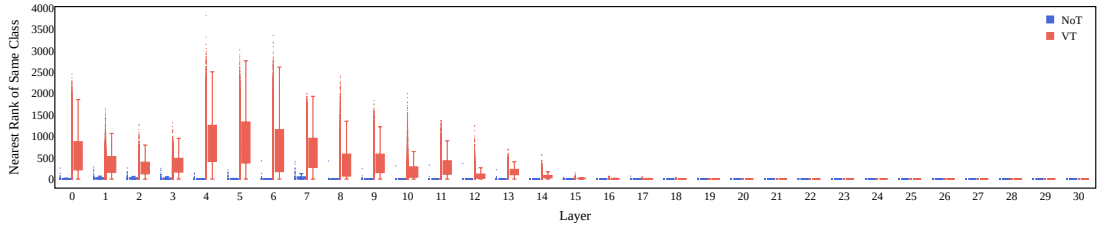


Fig. 10: Topological feature vectors of attacked VGG-16 model trained on CIFAR-100.

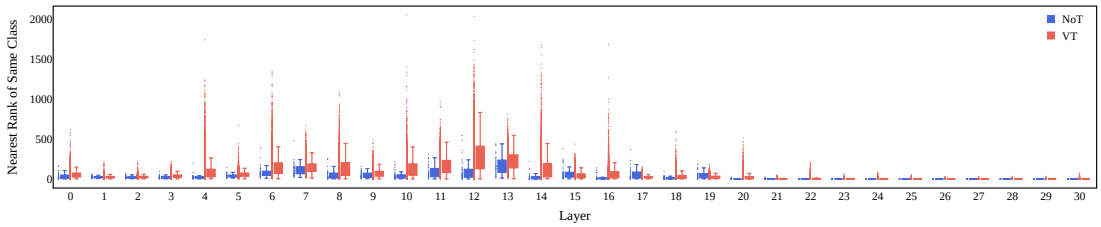


Fig. 11: Topological feature vectors of attacked VGG-16 model trained on Tiny-ImageNet.

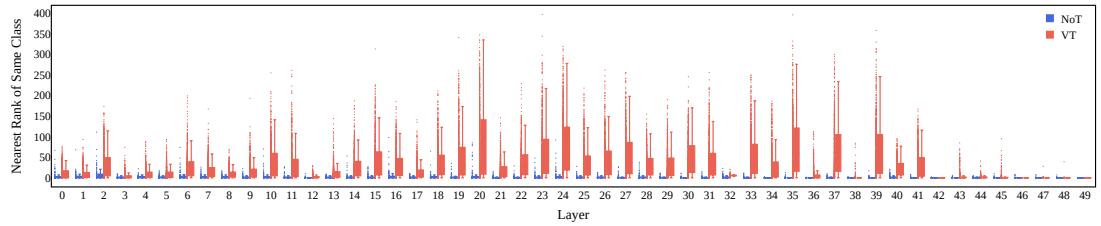


Fig. 12: Topological feature vectors of attacked ViT model trained on CIFAR-10.

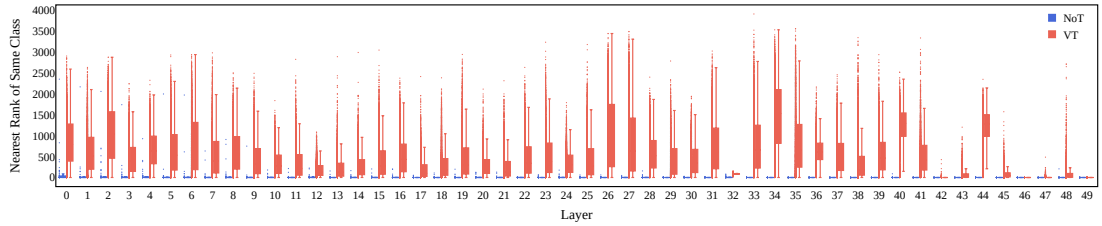


Fig. 13: Topological feature vectors of attacked ViT model trained on CIFAR-100.

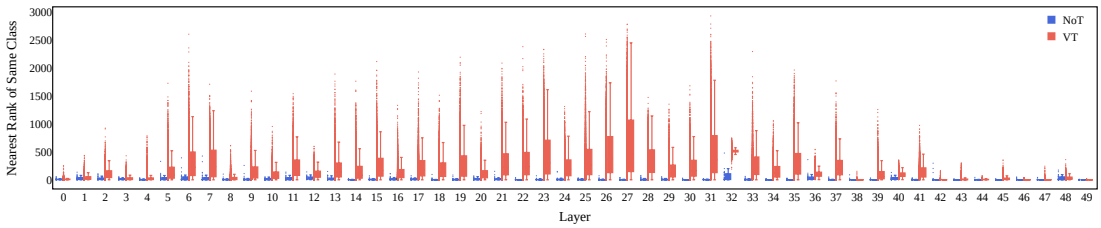


Fig. 14: Topological feature vectors of attacked ViT model trained on Tiny-ImageNet.