Comparative Analysis of Defenses against the Backdoor Attacks on Deep Neural Networks

https://github.com/vbabushkin/ECE_GY_9163_Machine_Learning_for_Cyber_Security_Project

Binfang Ye*
Tandon School of Engineering, New York University
New York, USA
by2034@nyu.edu

Abdullahi Bamigbade*
Tandon School of Engineering, New York University
New York, USA
akb518@nyu.edu

Vahan Babushkin*
Engineering Division, New York University Abu Dhabi
Abu Dhabi, UAE
vahan.babushkin@nyu.edu

***Abstract*—The aim of this project is to provide a comparative analysis of three defense mechanisms, namely: STRIP [1], fine-pruning [2] and a newly proposed anti-repairedNet defense against the four different backdoor attacks on the YouTube aligned face dataset [3]. The analysis allows to select the most robust defense for these specific types of attacks.**

***Index Terms*—deep learning, backdoor attack, fine-tuning.**

I. Introduction

The high demand of computational power required for training the Deep Neural Networks necessitates the outsoursing of the training to the third party. However, the third party might attempt to record a hidden functionality in the model resulting in targeted or untargeted misclassification of some instances, specifically poisoning them with a backdoor trigger. This process is known as backdoor attack on the network. In general, the backdoor model usually demonstrates outstanding performance on the clean dataset making the discovery of backdoor attacks quite challenging [2].

In this project we compare three approaches, namely: STRong Intentional Perturbation (STRIP), fine-pruning and anti-repairedNet defense against the four backdoor attacks on Deep Neural Network trained on the YouTube aligned face dataset [3]. We report the performance of these methods and finally decide which method to use for a defense against backdoor attack on anonymous2 network. The major contributions of this project are stated as follows.

1) The proposal of an improved fine-pruning approach that allows to achieve Clean Classification Accuracies (CCA) of goodnet G around 78%, reaching Attack Success Rates (ASR) almost near to 0%.
2) The proposal of an anti-repairedNet defense that allows the attainment of Clean Classification Accuracy (CCA) of goodnet G up to about 93%, with Attack Success Rate (ASR) of 10%.

II. Methoodology

1) STRIP: The **STR**ong **I**ntentional **P**erturbation (STRIP) defense against the trojan trigger-based backdoor attacks [1] is based on the intentional perturbation of the inputs to the Badnet B in order to investigate the level of randomness of predicted classes. In our case we use superpositions of two images as perturbations. The prediction randomness is quantifiable by entropy which is a measure of the state of disorder, randomness, or uncertainty. We expect that the predictions of poisoned inputs are more organized and thus, characterized by low entropy values, rather than the predictions of clean inputs, which are distributed randomly and therefore, have high entropy. Thus, if the entropy of the input is lower than some empirically-determined threshold value, the input is considered as trojaned, and clear otherwise.

2) Fine-pruning: The fine-pruning approach was first implemented in [2] and demonstrated outstanding performance against pruning-aware attacks on Deep Learning Networks (DNNs) for diffeent domains. It incorporates both pruning of the dormant channels and fine-tuning the parameters of DNN such as the number of epochs, the batch size, the learning rate, etc. Due to the limited timeframe we decided to focus on fine tuning the learning rates of the network. We start with smaller learning rates to ensure that the weights of the improved model B' remain closer to the weights of the original badnet model B. First the channels were arranged in increasing order according their activations from the last max pooling layer and at each step after disabling a single channel we retrained the network on the whole clean validation dataset (11547 images). We evaluate the performance of the network on the clean test dataset and some backdoored datasets for a particular model. Finally, noticing that the Clear Classification Accuracy (CCA) of goodnet G depends on the CCA of the badnet B and also, that in all models fine-pruning of the first few channels results in increase of the CCA, we propose an improved fine-pruning approach that uses B0 – a badnet B with fine-pruned few channels instead of the orginal badnet B to classify the clean and poisoned images. In overall, the proposed approach allows to achieve the clear classification accuracies of goodnet G around 78% with near 0 attack success rates.

3) anti-repairedNet: The basic idea in the anti-repairedNet defense is to reduce the Clean Classification Accuracy (CCA) of the repaired net to a very low level in order to achieve a high Attack Success Rate (ASR). Thus, unlike conventional approaches which repair the badnet in order to decrease attack success rate (while maintaining high clean classification accuracy), the proposed anti-repairedNet defense further destroys the badnet so that high attack success rate is achieved (while maintaining very low clean classification accuracy). To achieve this objective, we generated perturbed images from the clean validation images by creating random pixels whose value lie in the range 0 to 255. These random pixels were

*All authors contributed equally to this work.

then added to the clean image while ensuring that image dimension is preserved (i.e. perturbed valid data = clean valid data + np.random.randint(256, size=(1, 55, 47, 3))). We then employed the clean and perturbed images to analyze the activation map of the different layers of the badnet. Based on our analysis, we observed that activation map of the the conv2 layer (as against the conv3 layer) is more interpretable since it retains obvious features of the input image - the likes of the face, mouth, eye, etc. More importantly, we observed that the different channels of conv2's activation layer exhibit good behaviour of the clean image better than the random perturbations (see Fig. 9 - Fig. 12 in the appendix). Accordingly, we concluded that by pruning the channels which exhibit good behaviour better than (or in a similar manner as) random perturbations, we can come out with an anti-repairedNet which allows attacks to succeed at a very high rate while achieving very low clean classification accuracy. Thus, the excellent discriminatory nature of the anti-repairedNet can be leveraged upon to propose a backdoor detector G which compares outputs of the badnet and anti-repairedNet. More importantly, since both the badnet and anti-repairedNet allow attacks to succeed due to their high attack success rate and thus outputing the same (correct) label for a poisioned input, (conversely the badnet alone outputs the correct label for a clean input due to its high clean classification accuracy while the anti-repairedNet misclassifies a clean input due to its low clean classification accuracy), output of the backdoor detector G can be obtained as follows.

1) Output = prediction of the badnet when there is a mismatch between the prediction of the badnet and anti-repairedNet.
2) Output = 1283 (i.e. N + 1) when there is a match between the prediction of the badnet and anti-repairedNet.

III. Results

1) STRIP: First we attempted to implement STRIP defense to all five badnet models. The results are presented on Fig. 1, a.). Entropy distributions of 3000 clean inputs and 3000 poisoned inputs for different triggers dataset and 100 random images from training dataset used for superimposing are detailed in Fig. 2. We can observe that for sunglasses trigger, the STRIP method performs very well because the distribution for clean and poisoned inputs are different but for other triggers, it worked badly. For example, it didn't perform well for the multi-trigger dataset with the sunglasses trigger. It might be due to the presence of other two triggers in the multi-trigger dataset that leads to the decrease in randomness of prediction. The thresholds of 0.169, 0.0003, 0.0003, 0.0002, and 0.0003 were selected to separate the clean and poisoned inputs for datasets with sunglasses and anonymous triggers, multi-trigger including eyebrows, lipstick and sunglasses triggers respectively. The final results of applying the STRIP approach is demonstrated on Fig. 1, a.).

2) Fine-pruning: Initally we performed a grid search over 5 learning rate values by fine-pruning the model with a fixed number of epochs that was determined by observing the dynamics of training loss/validation accuracy – in all models the accuracy increases sharply to almost 90% after approximately 10 epochs accompanied by the corresponding drops in loss. The heatmaps on Fig. 3 and CCA/ASR plots in Table I (in Appendix) were used to determine an optimal number of channels to prune for a given set of parameters. Table I shows that for learning rate of 10^{-3} pruning about 10% of channels is enough for all models to achieve low ASR with relatively high CCA. The activations heatmaps of last max pooling layers (Fig. 4 - Fig. 8) demonstrate that fine-pruning of 10% of channels results in significant deactivations of the neurons tuned for poisoned inputs. Furthermore we used the learning rate of 10^{-3} and 10 epochs for fine-pruning 10% of neurons for all models.

While the fine-pruning allows to produce the repaired models B' that achieve high CCA (see Fig. 1,b.)), the CCA value of goodnet G is usually limited by the CCA of badnet B. Thus, incorporating B' and B into goodnet G results in overall drop in CCA values making them slightly lower than for B. In the meantime, the ASR rates for G do not exceed those for B' (see Fig. 1 b.) and c.)). From the Table I it is clear that CCA for all badnets does not exceed 60% but increases sharply with fine-pruning of just a few neurons. In order to devise a better model for discriminating the poisoned and clean inputs, we propose to modify the original badnet B by fine-pruning only a few channels. The proposed improved fine-tuning approach instead of comparing outputs from B and B' to decide if the input is clear or poisoned, compares outputs from a repaired network B' and new network B0 – derived from B by fine-pruning a few channels. This approach results in significant increase in CCA of goodnet G preserving the ASR almost the same as after the ordinary fine-pruning (Fig 1 b.) and c.)).

3) anti-repairedNet: To achieve the desired performance of lowering clean classification accuracy while maintaining very high attack success rate, activation maps of the conv 2 layer for the different badnets was obtained using both the clean and perturbed input images (see Fig. 9 - Fig. 12 in appendix). By comparing the maps of each badnet for both images, we identified those channels that exhibit good behaviour better than (or in a similar manner as) random perturbations. Thus, the following channels were pruned for each of the badnets.

1) Sunglasses badnet: 5, 7, 10, 14, 23 and 25.
2) Multi-trigger multi-target badnet: 0, 1, 8, 9, 10, 15, 17, 22, 23, 24, 25, 30, 31, 32, 33, 34, 36, 38 and 39.
3) Anonymous1 badnet: 0, 2, 3, 4, 10, 25, 28, 29 and 34.
4) Anonymous2 badnet: 7, 8, 21, 22, 26, 28, 34 and 38.

Using the provided test images for each of the sunglasses, multi-trigger multi-target and anonymous1 badnets, performance obtained in terms of Clean Classification Accuracy (CCA) and Attack Success Rate (ASR) of the goodnet G are presented on Fig. 1 d.) where it can be seen that the proposed anti-repairedNet defense generally achieves a very high clean classification accuracy while keeping attack success rate relatively very low (especially for sunglasses badnet).

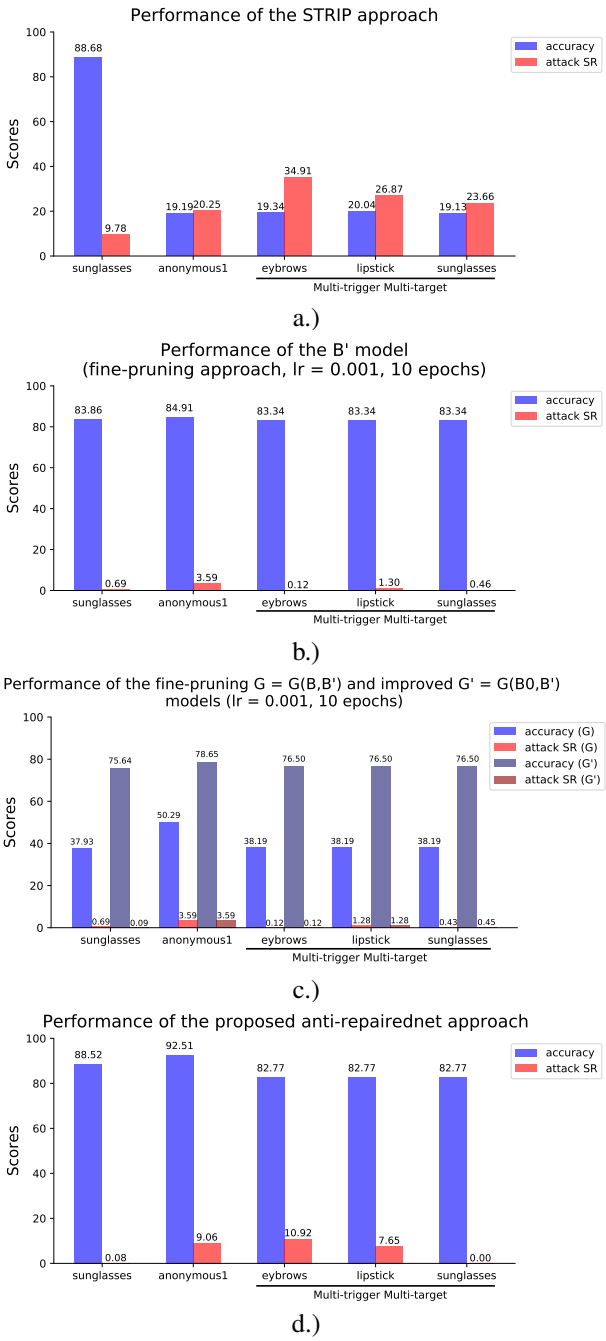


Fig. 1. Clean test data accuracies and attack success rates for a.) STRIP, b.) B' after fine-pruning c.), fine-pruning with goodnet $G = G(B, B')$ and improved fine-pruning with goodnet $G = G(B0, B')$ d.) anti-repairedNet defense.

IV. Conclusion

The STRIP approach demonstrated relatively good results for sunglasses trigger but performed poorly on other datasets. We explain it by the low randomness in the data, which is probably attributed to the relative uniformity in triggers used for other datasets, resulting in lower entropies. While the STRIP approach still allows to achieve the CCA around 88% for sunglasses badnet model, its ASR is higher than of fine-pruning, which affected our decision to dismiss this defense mechanism in favour of fine-pruning and anti-repairedNet defenses. Specifically, the fine-pruning defense demonstrates good results on all datasets with relatively low number of epochs (10). Notice from Table I the sharp drop of ASR occurs for learning rates lower than 10^{-3} when the higher number of neurons are disabled, while the CCA remains high, which could be explained by slow modification of weights. In contrast, high learning rates, e.g. 10^{-2} modify weights significantly, resulting in sharp drop of both CCA and ASR. For learning rate of 10^{-6} there are a few peaks in ASR when the clean classification accuracy drops quite low, which might be attributed by the removal of a channel activated to classify clear images. Further retraining lowers the ASR and slightly increases the CCA. Therefore, we recommend using B0 in improved fine-pruning approach only to differentiate between clean and poisoned data and only for cases when the CCA for the original badnet is low. For example, for the badnet used in HW3 the ordinary fine-pruning approach is enough, since the CCA of B is 98.6% and fine-pruning results in CCA of B' around 89.8% with ASR 1.9%, while CCA for goodnet G is 89.3% and ASR is the same. Also, the anti-repairedNet defense is a good candidate for goodnet G owing to its high Clean Classification Accuracy (CCA) across all the models (about 83% in the worst case) and an Attack Success Rate (ASR) of almost 0% for sunglasses poisoned images.

References

[1] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. 2019. STRIP: a defence against trojan attacks on deep neural networks. In Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC '19). Association for Computing Machinery, New York, NY, USA, 113–125.

[2] Liu, K., Dolan-Gavitt, B., and Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. In M. Bailey, S. Ioannidis, M. Stamatogiannakis, and T. Holz (Eds.), Research in Attacks, Intrusions, and Defenses - 21st International Symposium, RAID 2018, Proceedings, 273-294.

[3] L. Wolf, T. Hassner, and I. Maoz. "Face recognition in unconstrained videos with matched background similarity". In CVPR 2011, pages 529–534, June 2011.

APPENDIX

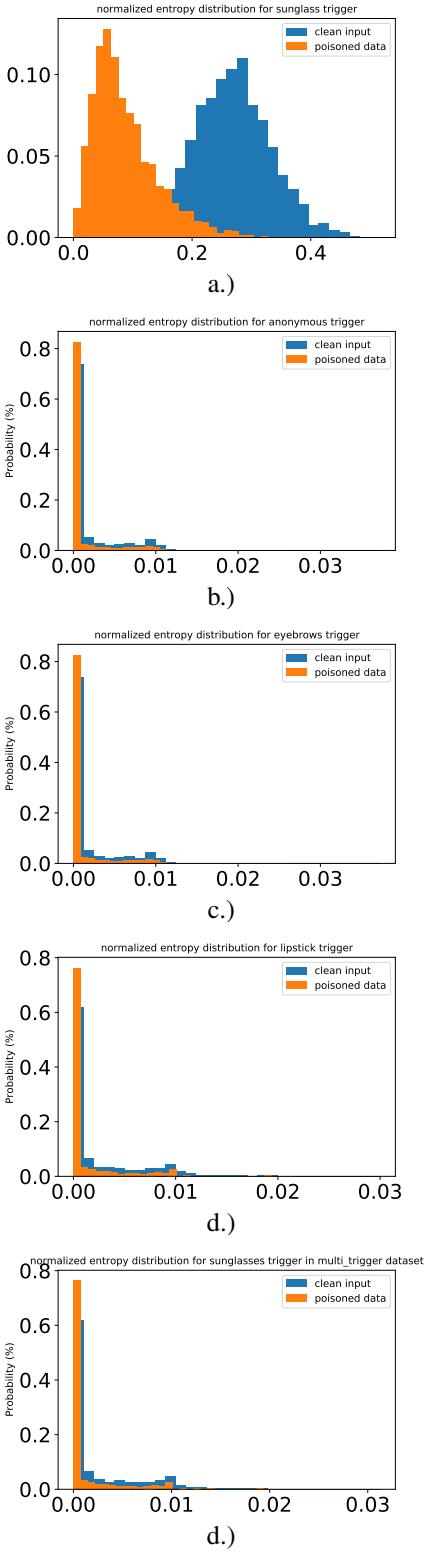


Fig. 2. Entropy distribution for different triggers: a.) sunglasses, b.) anonymous, and multi-triggers including c.) eyebrows, d.) lipsticks, e.) sunglasses

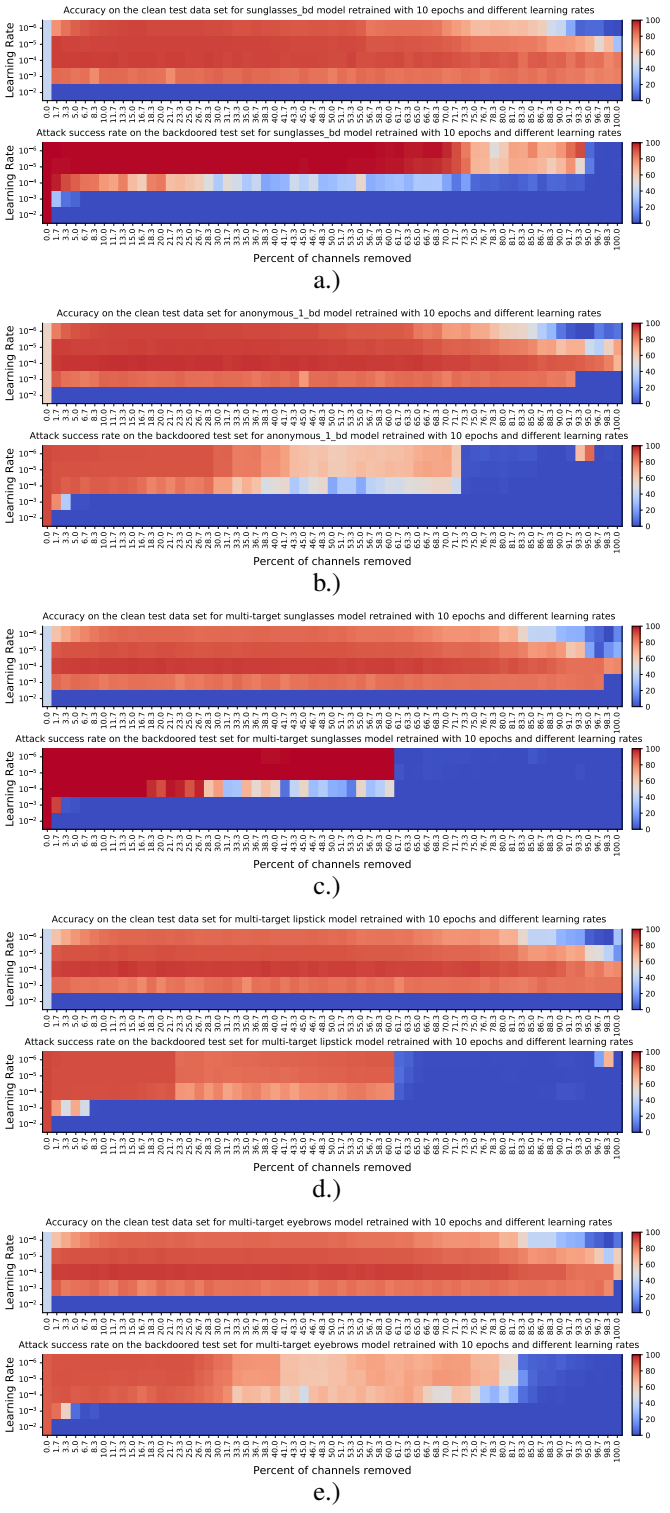
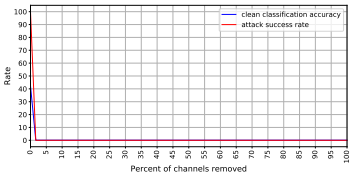
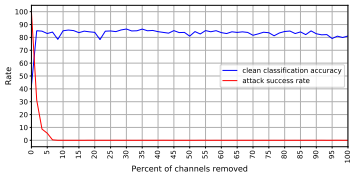
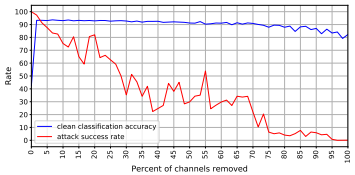
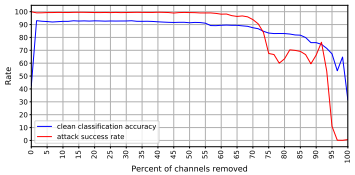
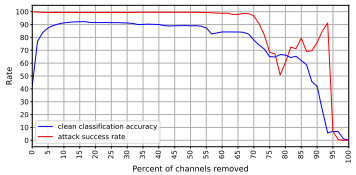
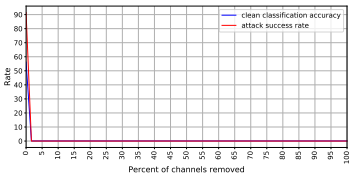
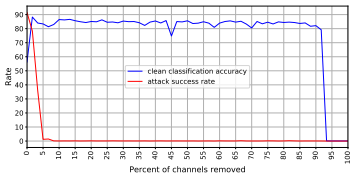
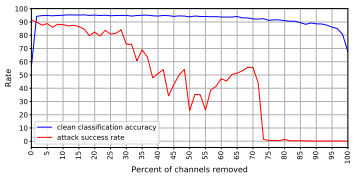
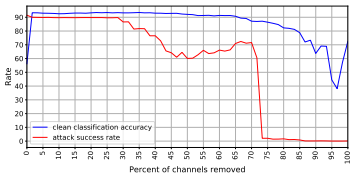
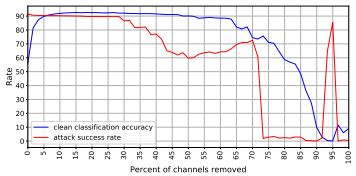
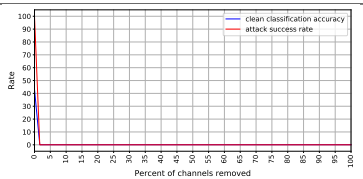
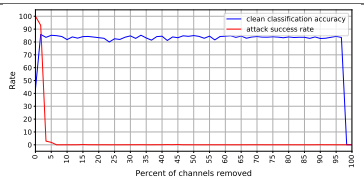
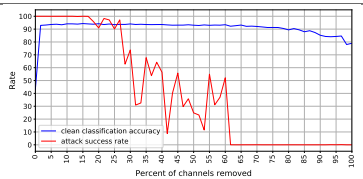
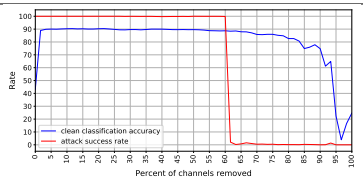
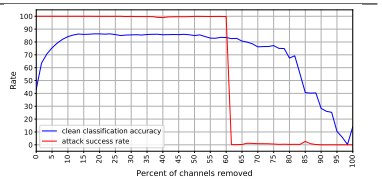
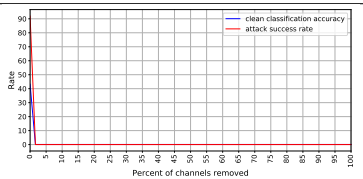
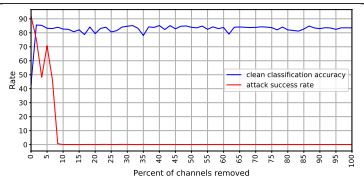
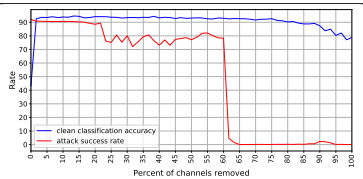
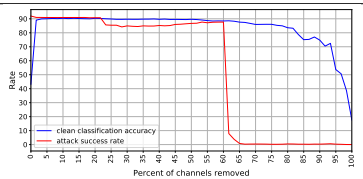
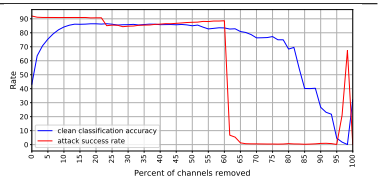
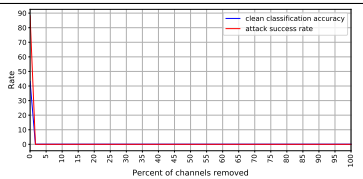
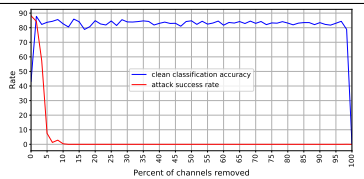
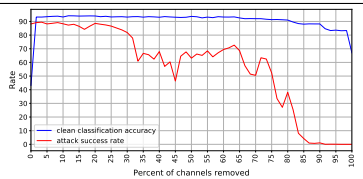
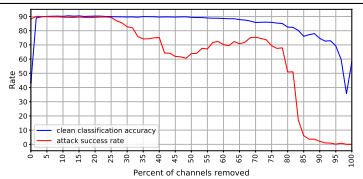
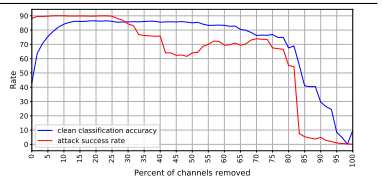


Fig. 3. Heatmaps of accuracy (upper) and attack success rates (lower) plots for a.) sunglasses_bd model , b.) anonymous_1_bd model, multi_trigger_multi_target_bd_net model for sunglasses c.), lipstick d.), and eyebrows e.) attacks. All models are re-trained with 10 epochs and five different learning rates.

TABLE I

ACCURACY ON THE CLEAN TEST DATA SET AND ATTACK SUCCESS RATE ON THE BACKDOORED TEST DATA FOR 10 EPOCHS AND DIFFERENT LEARNING RATES.

	Learning rate				
	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
sunglasses					
anonymous1					
Multi-trigger Multi-target sunglasses					
Multi-trigger Multi-target lipstick					
Multi-trigger Multi-target eyebrows					

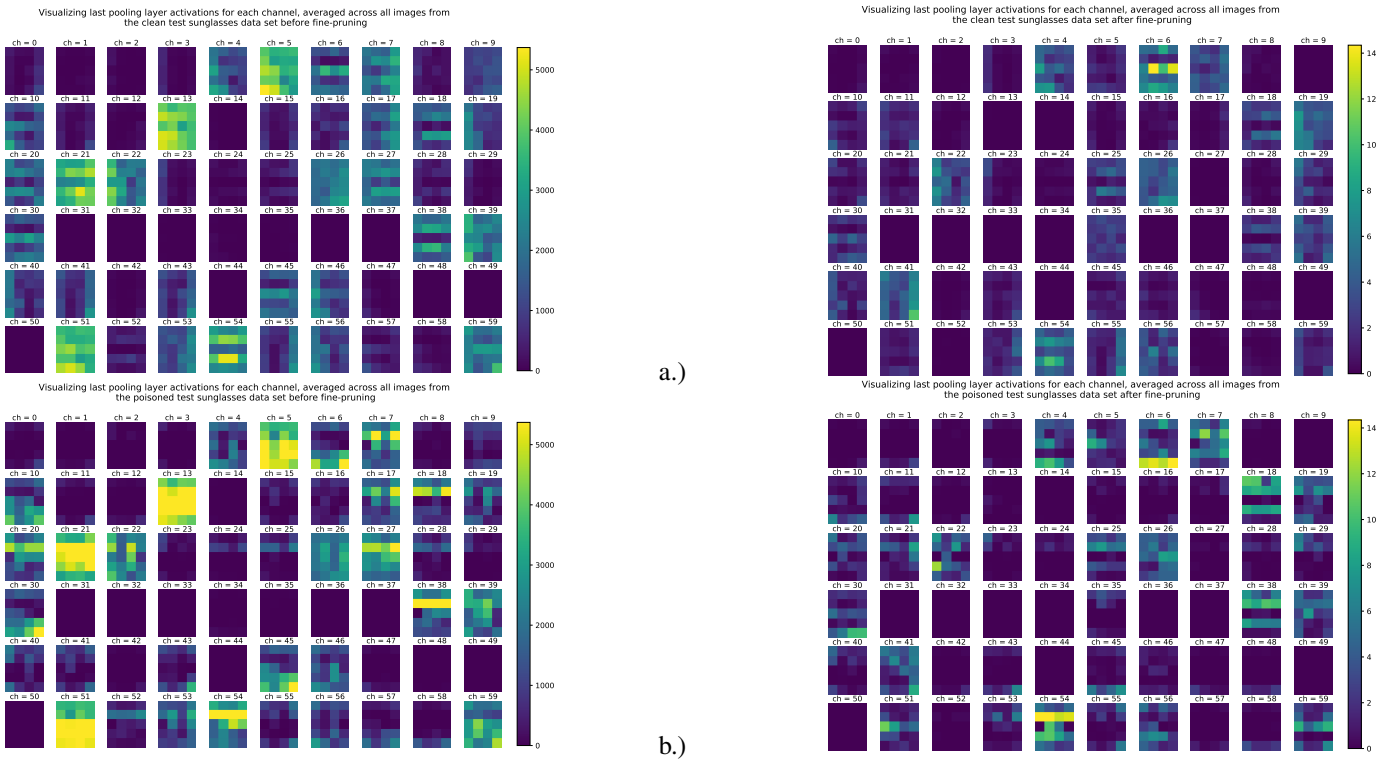


Fig. 4. Activations before (left) and after (right) fine-pruning for sunglasses_bd model, a.) on clean test data, b.) on poisoned test data. We can notice sharp decrease in activated neurons trained on poisoned data. Also notice the drop in activation magnitude after fine pruning.

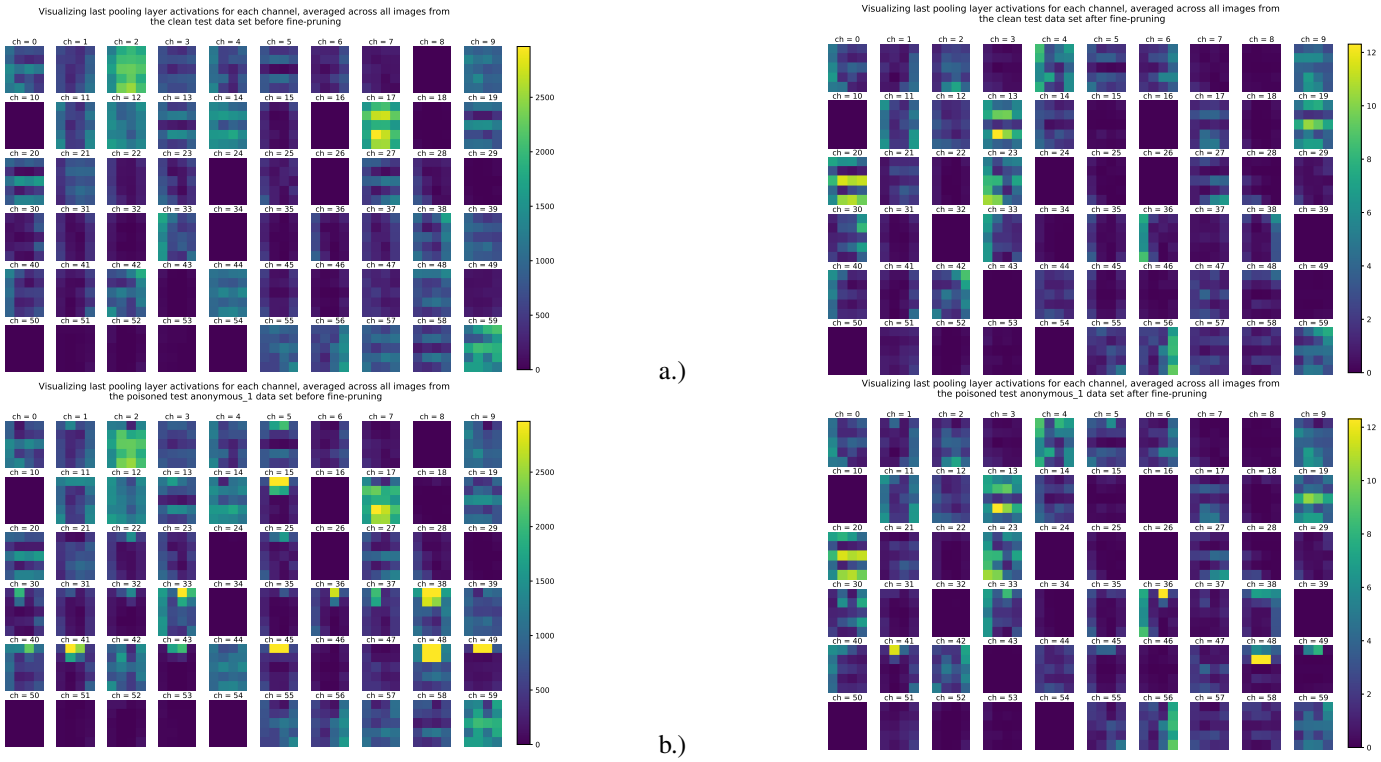


Fig. 5. Activations before (left) and after (right) fine-pruning for anonymous_1_bd model, a.) on clean test data, b.) on anonymous_1 poisoned test data. We can notice sharp decrease in activated neurons trained on poisoned data. Also notice the drop in activation magnitude after fine pruning.

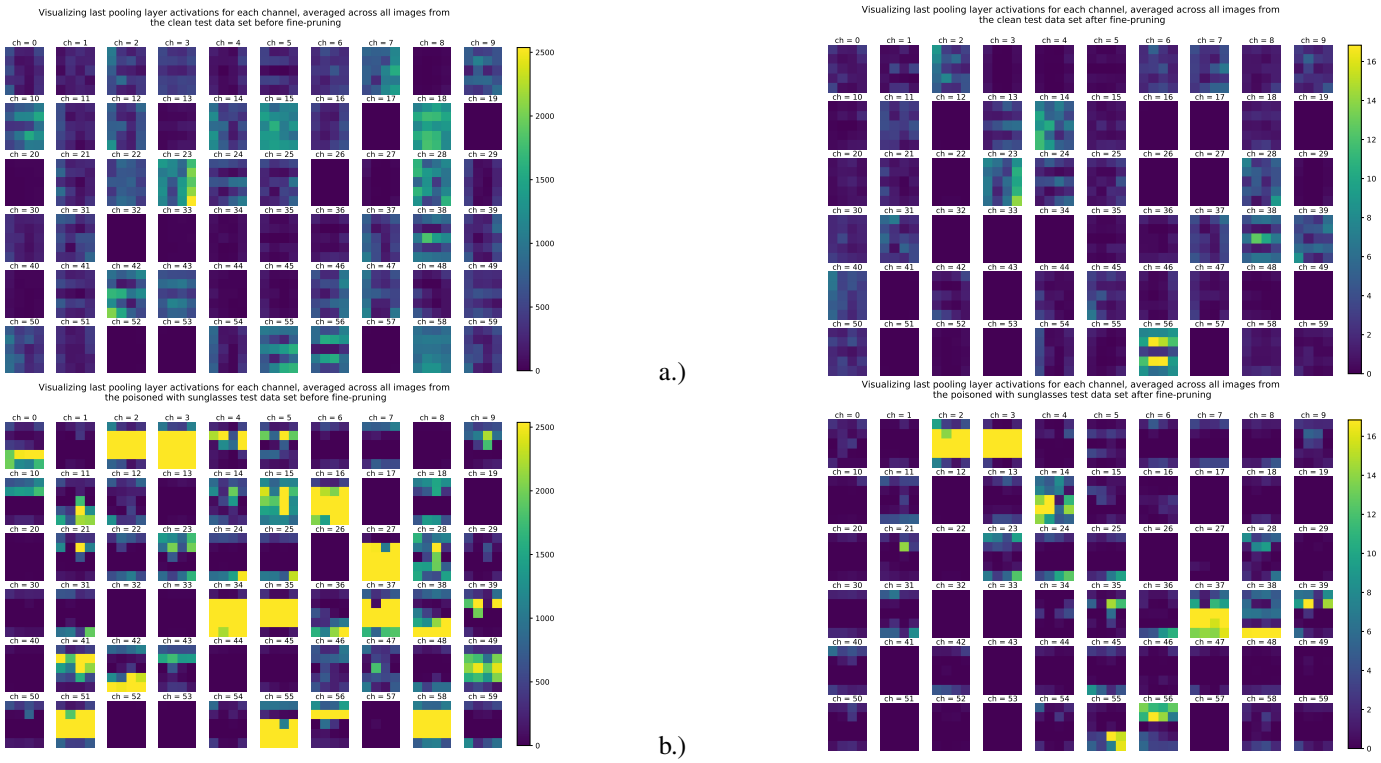


Fig. 6. Activations before (left) and after (right) fine-pruning for multi_trigger_multi_target_bd model, a.) on clean test data, b.) on sunglasses poisoned data. We can notice sharp decrease in activated neurons trained on poisoned data. Also notice the drop in activation magnitude after fine pruning.

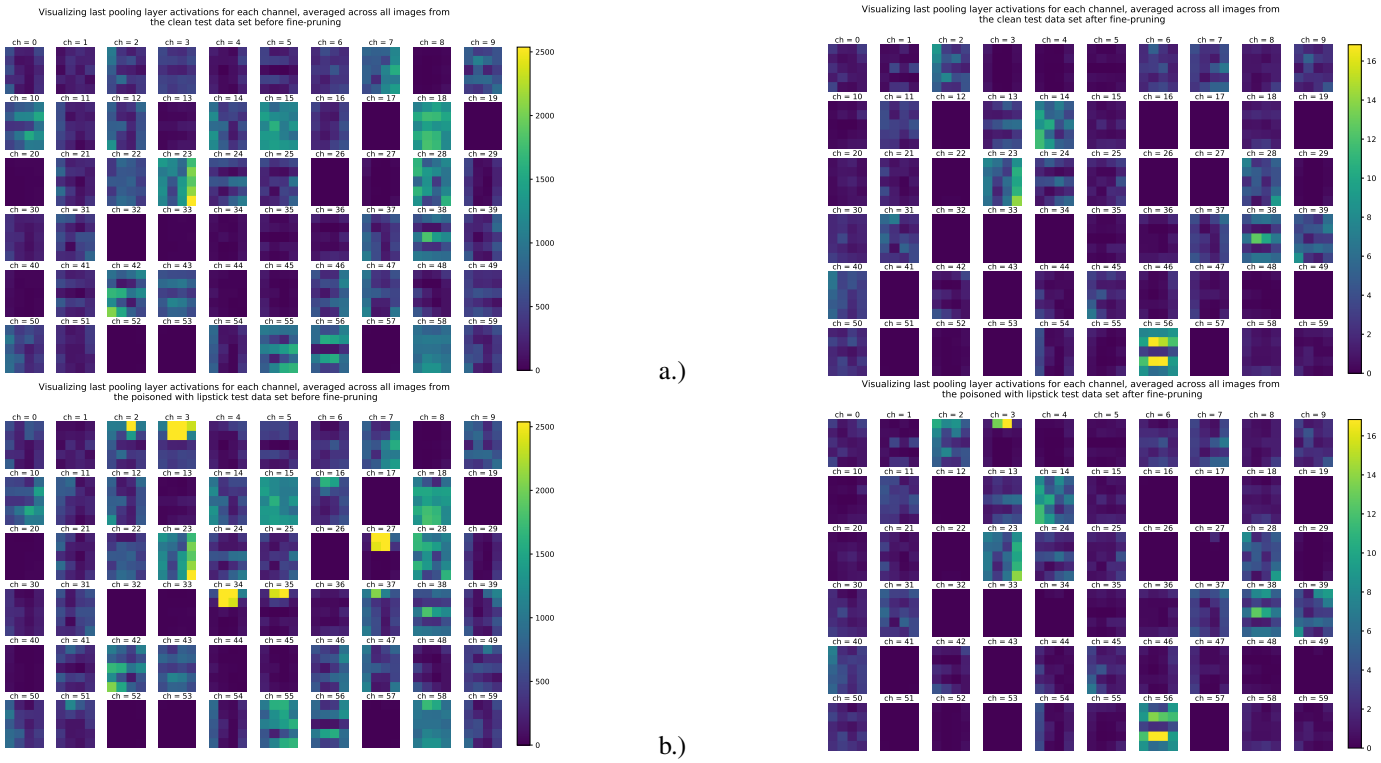


Fig. 7. Activations before (left) and after (right) fine-pruning for multi_trigger_multi_target_bd model, a.) on clean test data, b.) on lipstick poisoned data. We can notice sharp decrease in activated neurons trained on poisoned data. Also notice the drop in activation magnitude after fine pruning.

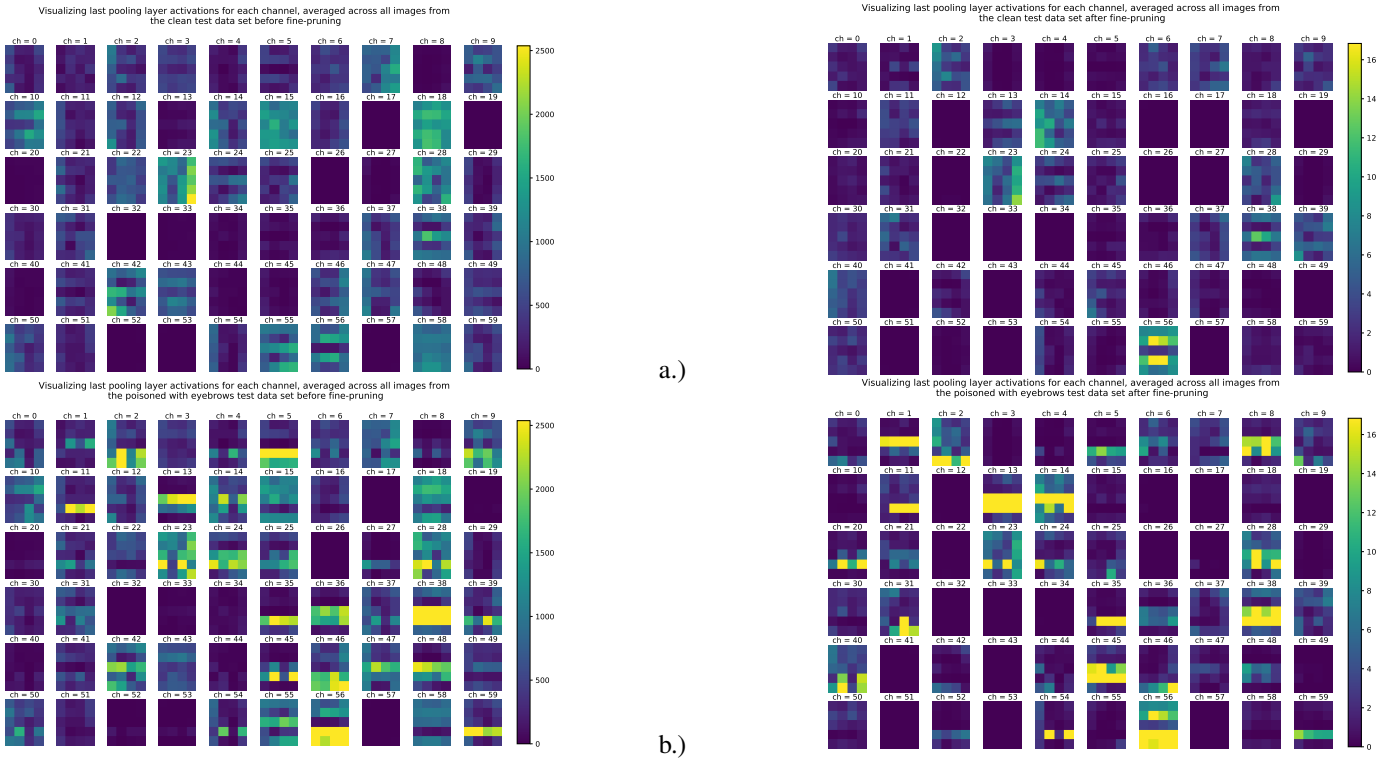


Fig. 8. Activations before (left) and after (right) fine-pruning for multi_trigger_multi_target_bd model, a.) on clean test data, b.) on eyebrows poisoned data. We can notice sharp decrease in activated neurons trained on poisoned data. Also notice the drop in activation magnitude after fine pruning.

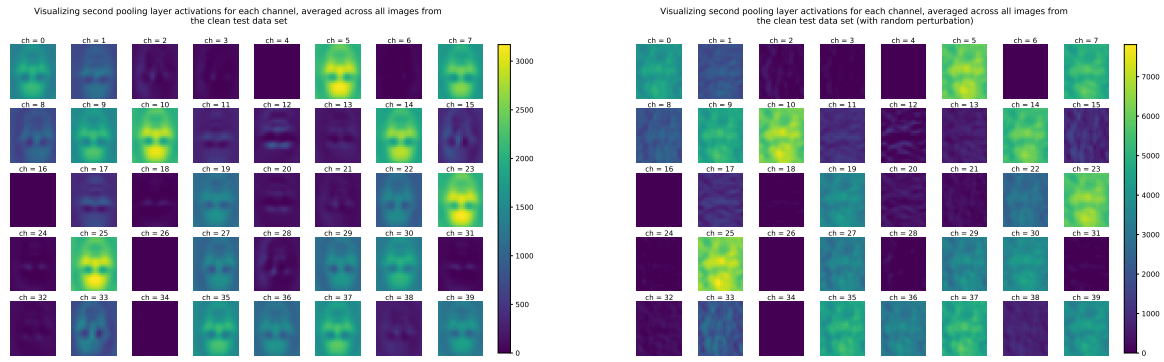


Fig. 9. Activations of conv2 layer with clean input (left) and perturbed input (right) fine-pruning for sunglasses_bd model.

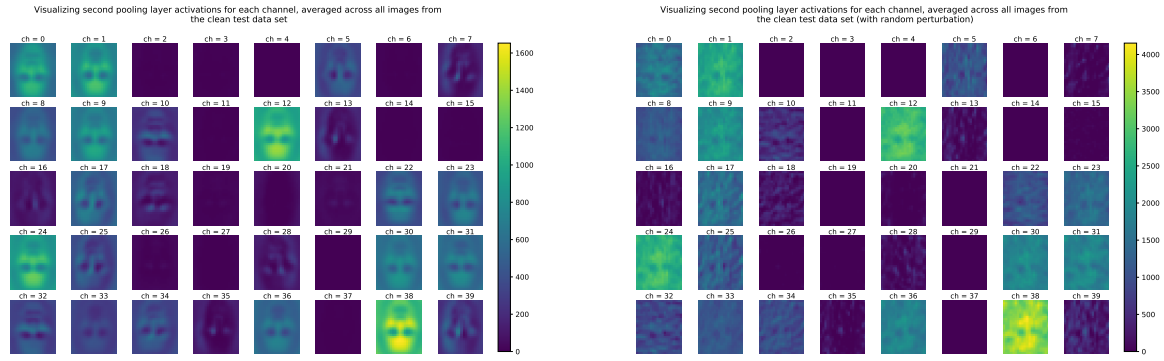


Fig. 10. Activations of conv2 layer with clean input (left) and perturbed input (right) fine-pruning for multi_trigger_multi_target_bd model.

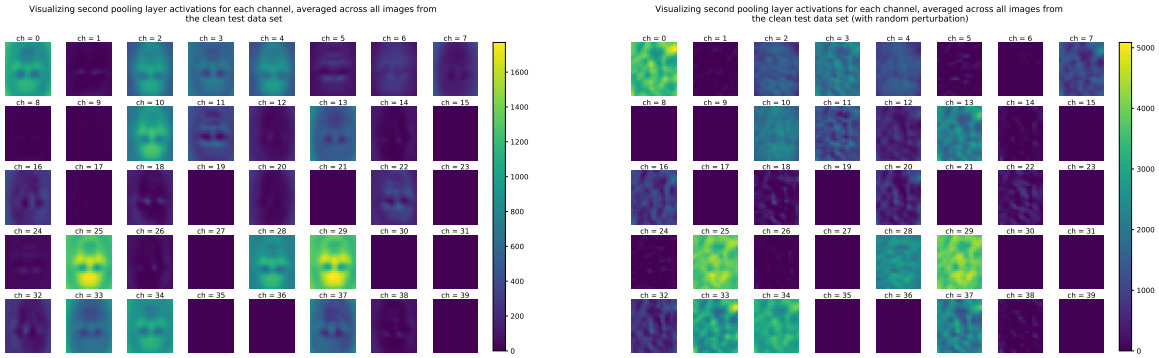


Fig. 11. Activations of conv2 layer with clean input (left) and perturbed input (right) fine-pruning for anonymous_1 model.

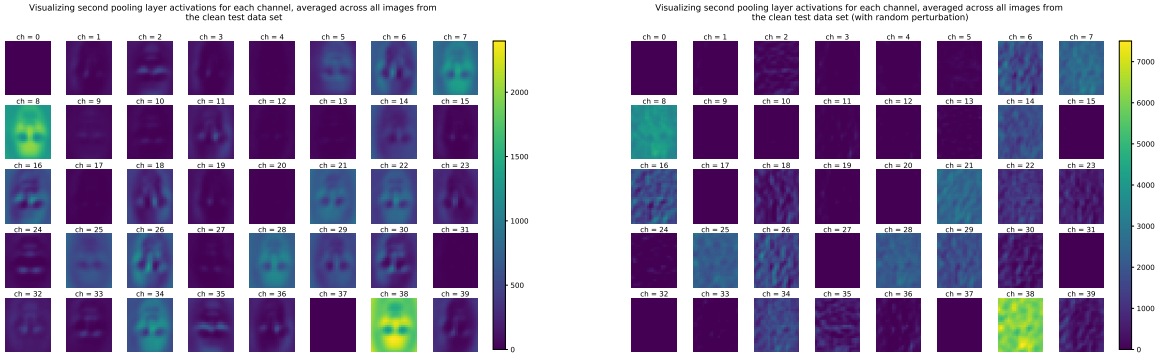


Fig. 12. Activations of conv2 layer with clean input (left) and perturbed input (right) fine-pruning for anonymous_2 model.