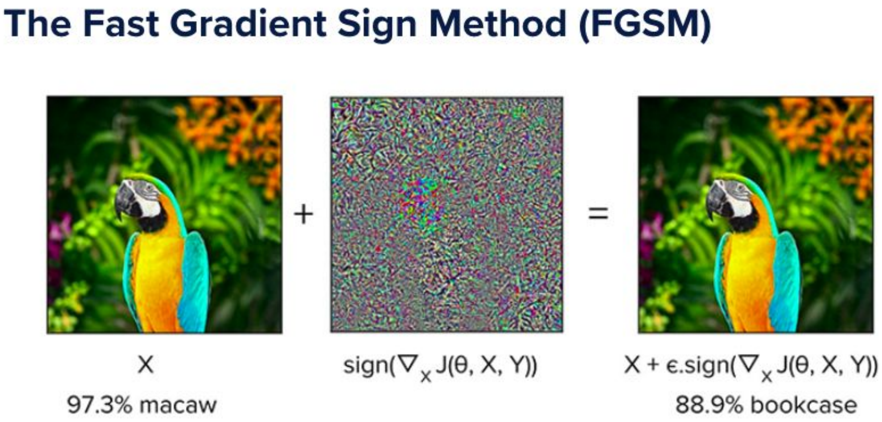
Exploring the Robustness of LoRA-Adapted DNN against Adversarial Attacks

Speaker: Tsai Chen Hsiung, Yun Hao Lee

Motive

- Adversarial attack post concerns with neural networks since 2015
- As using pretrained model(such as LLM, DNN) becomes popular, the target to attack becomes more solid and clear.
- LoRA, as one of most popular way to finetune DNN and LLM, does it effect the robustness of model?

Attack Methodology - FGSM/PGD



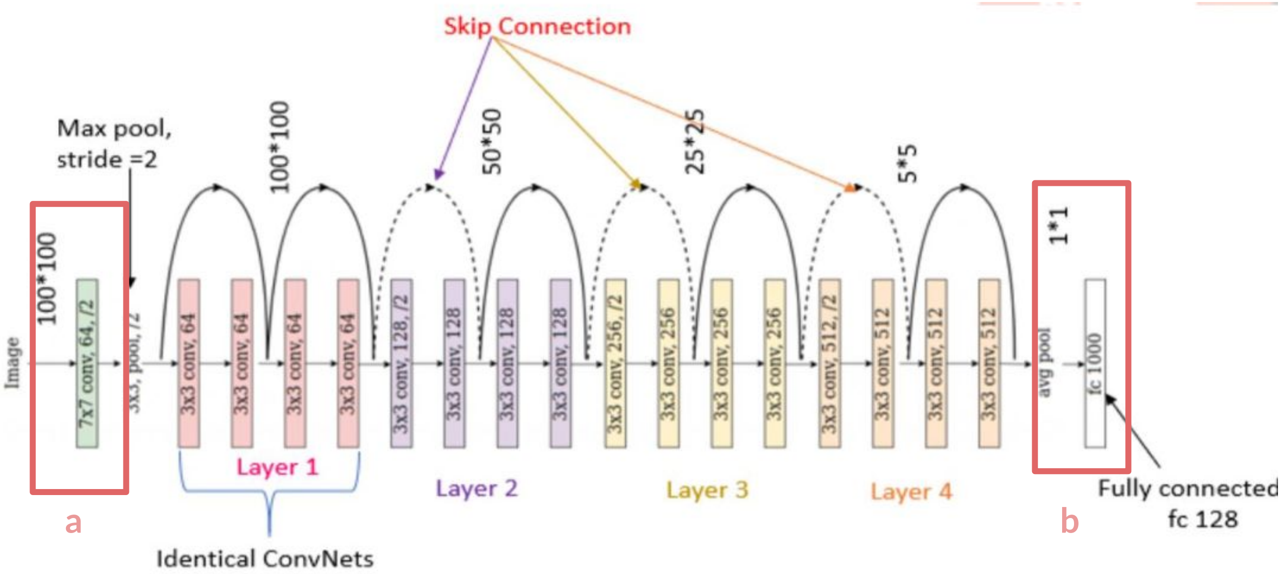
Methods - Resnet 18

- Experiment 1:

  LoRA on b

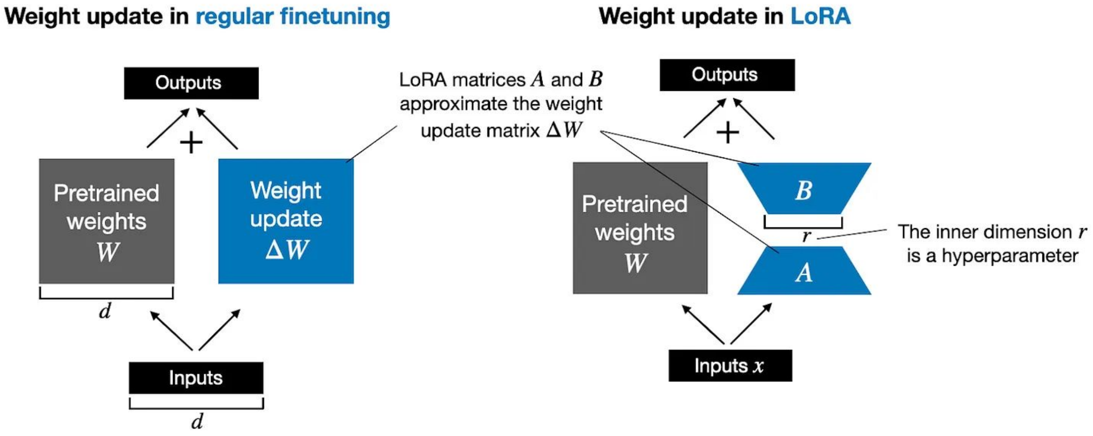
- Experiment 2:

  LoRA on a+b



Methods - LoRA

Instead of fine-tuning the whole neural network, LoRA successfully keep majority of learned pretrained weight and only provide a little insightful features learned from new data to tweak the model



Methods - LoRA on Conv

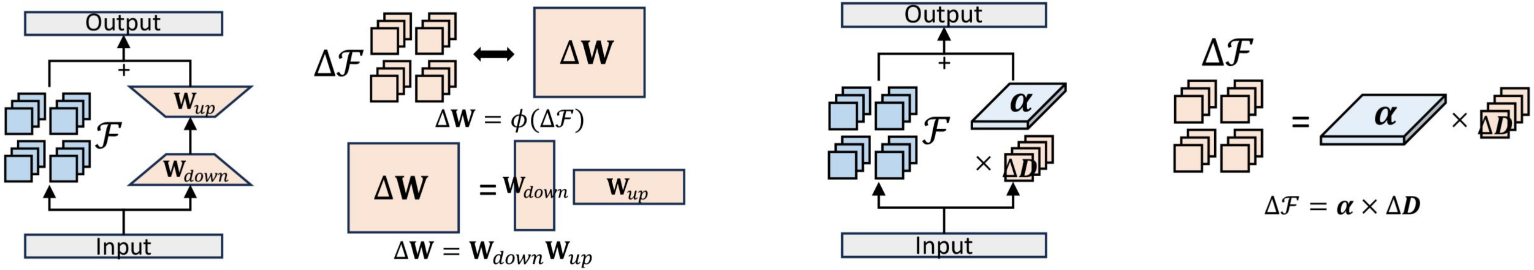


Figure 1:(a) Parameter-efficient methods (such as LoRA [15]) inject and optimize a trainable module ($\mathbf{W}down$ and $\mathbf{W}up$) while keeping the pre-trained model $\mathbf{W}$ or $\mathcal{F}$ fixed. To apply to the convolutional layer, most methods require reshaping it by $\Delta\mathbf{W}$=$\phi(\Delta\mathcal{F})$ to transform a 4D tensor to a matrix. (b) Our approach formulates the trainable module in convolutional layers as two distinct components: filter atoms $\mathbf{D}$ and atom coefficients $\boldsymbol{\alpha}$. We achieve parameter-efficient fine-tuning by updating filter atoms, typically a small number of parameters.

Results - Metric

- Model Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Model Certainty
  - Define as "**Average Probability of Class Prediction**"

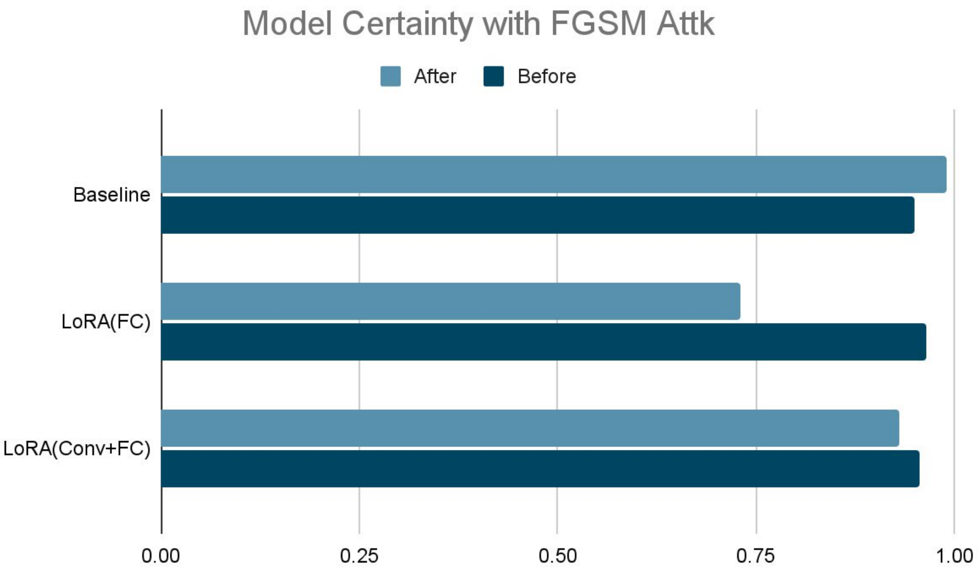
- Attack Success Rate
  - Define as "**For all samples that were able to predicted before, how much of those are failed after attack**"

Results - Metric

	Baseline Resnet-18	LoRA-FC	LoRA-FC/Conv1
Validation Accuracy	90.00%	96.00%	98.00%

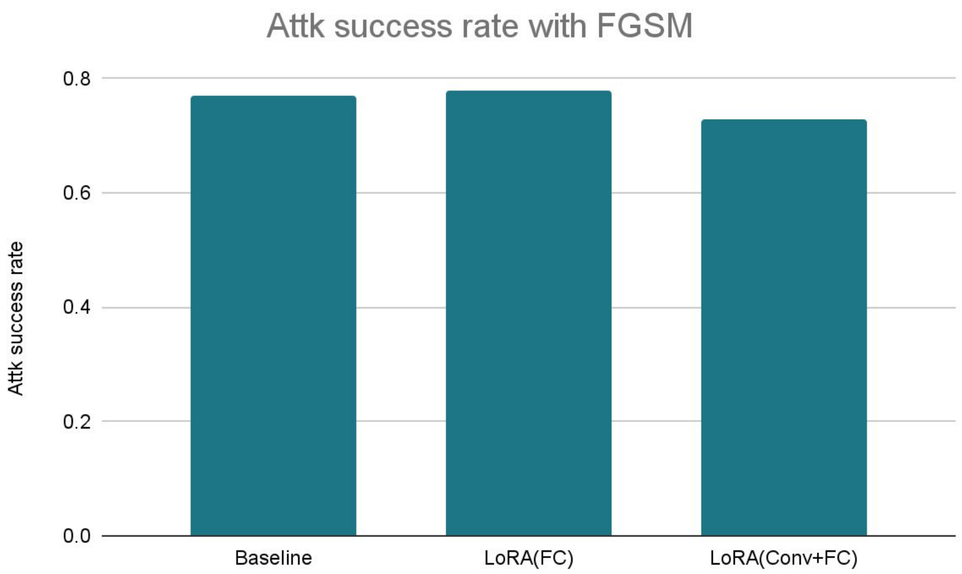
- LoRA creates flexibility to tweak the model to map learned characteristics to actual label
- For baseline model, the whole finetuning across the 18 layers are too overwhelming for such small dataset

Results - Metric



- LoRA on fully-connected layer drop its confidence in prediction
- While LoRA on Conv+FC doesn't affect much in its certainty in prediction, which is considered worse

Results - Metric

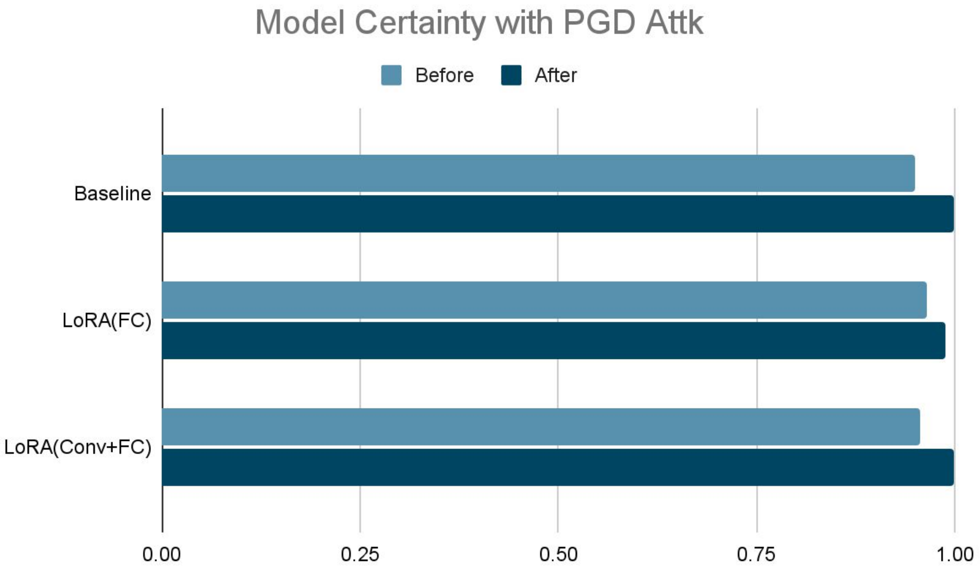


- All three settings of model are dramatically affected by FGSM attack
- LoRA may help in creating a balance where both low-level and high-level features are adjusted to account for adversarial distortions

Future Work

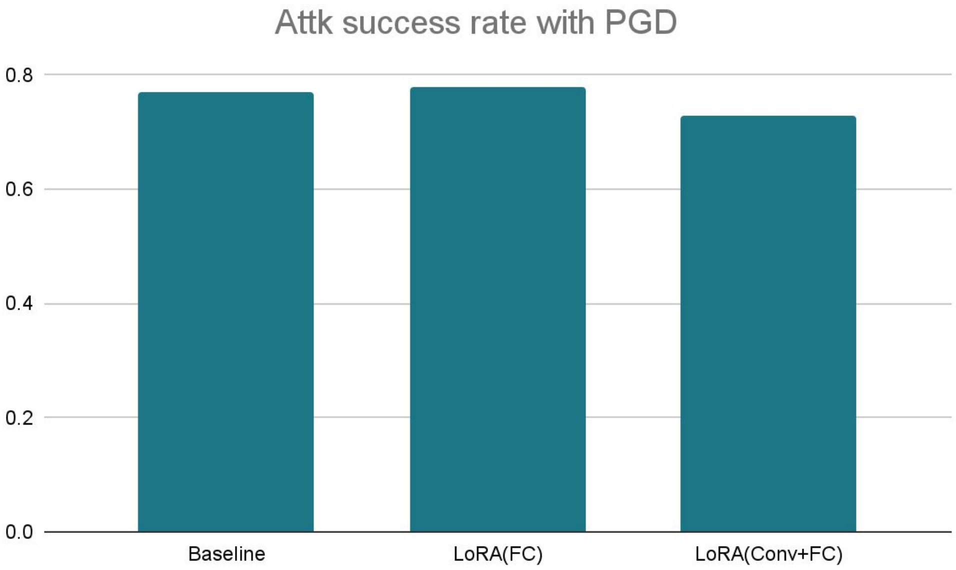
- Test on black box attack to observe if the drop in decision certainty also apply
  - Zeroth Order Optimization (ZOO) Attack
  - Boundary Attack
- If two models with different adaptation on LoRA give almost the same accuracy, but different attack success rate, is switching between model a good approach for resilience of attack

Results - Metric



- LoRA on fully-connected layer drop its confidence in prediction
- While LoRA on Conv+FC doesn't affect much in its certainty in prediction, which is considered worse

Results - Metric



- All three settings of model are dramatically affected by FGSM attack
- LoRA may help in creating a balance where both low-level and high-level features are adjusted to account for adversarial distortions