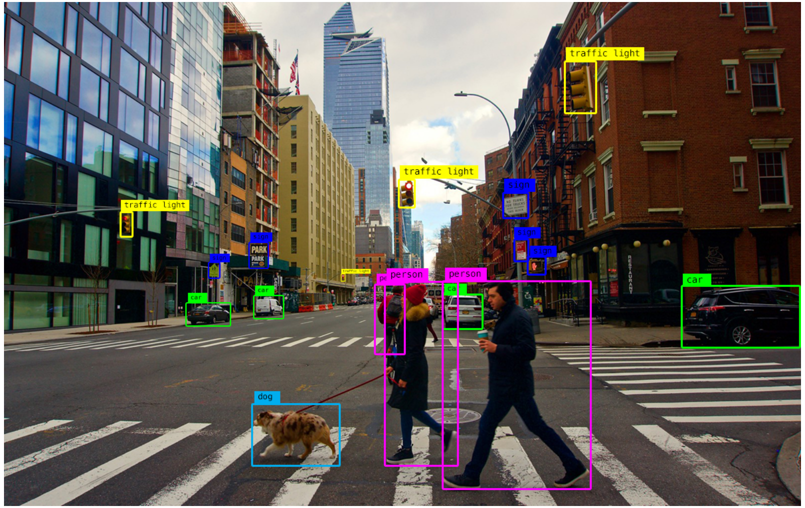
Robustness and Transferability of Universal Attacks on Compressed Models

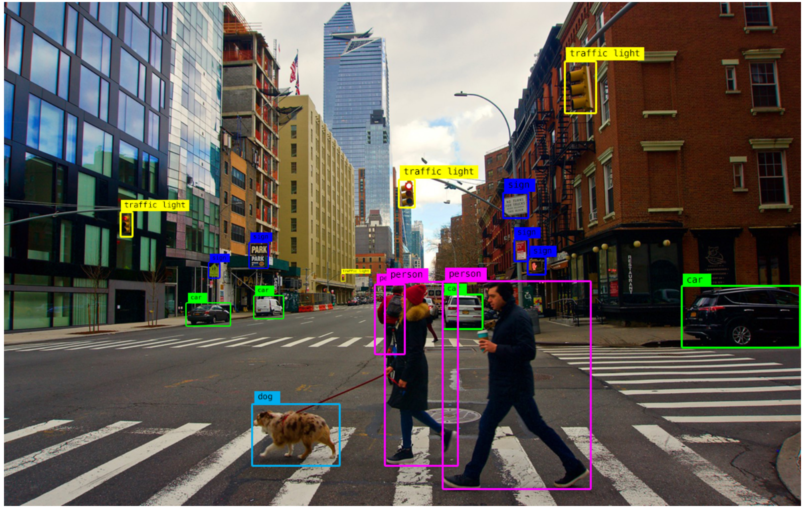
Alberto G. Matachana, Kenneth T. Co, Luis Muñoz-González
David Martinez, Emil C. Lupu
ag4116@imperial.ac.uk, k.co@imperial.ac.uk

Motivating example



[1]

Motivating example

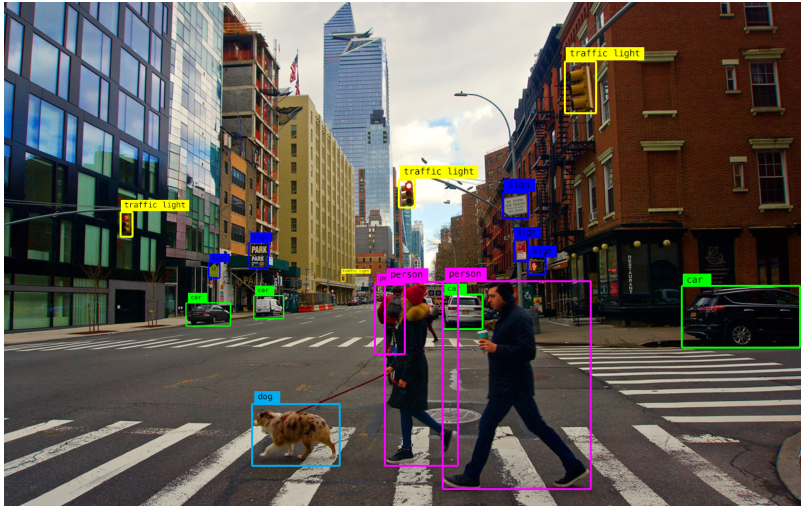


Existing DNNs face 2 key challenges:

1. They contain a large number of parameters
2. They are vulnerable against adversarial examples

[1]

Motivating example



Existing DNNs face 2 key challenges:

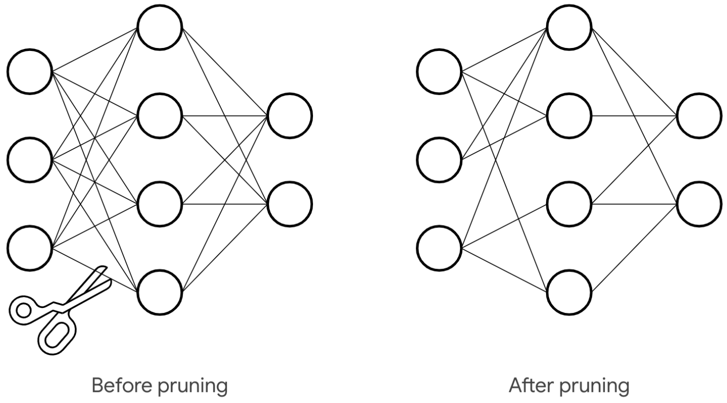
1. They contain a large number of parameters
2. They are vulnerable against adversarial examples

Universal Adversarial Perturbations

- A single perturbation can cause a target model to misclassify on a large set of inputs
- They are transferable

[1]

Compression Techniques

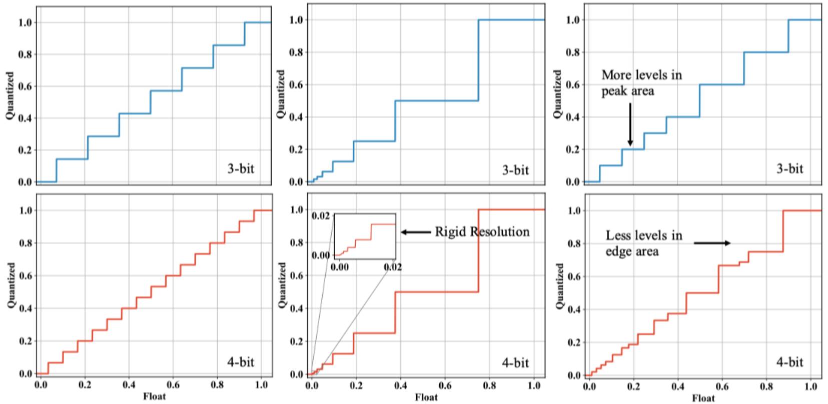


<u>Pruning:</u> reduce the size of the DNN by removing neurons that are irrelevant or have a reduced contribution at inference time

- ➢ **(PP)** Post-training Pruning
  - ■ **PP2, PP3, PP4**
- ➢ **(SFP)** Soft-filter Pruning
  - ■ **(SFP+M)** with mixup regularization
  - ■ **(SFP+C)** with cutout regularization

[2]

Compression Techniques



(a) Uniform Quantization (b) Power-of-Two Quantization (c) Additive Power-of-Two Quantization

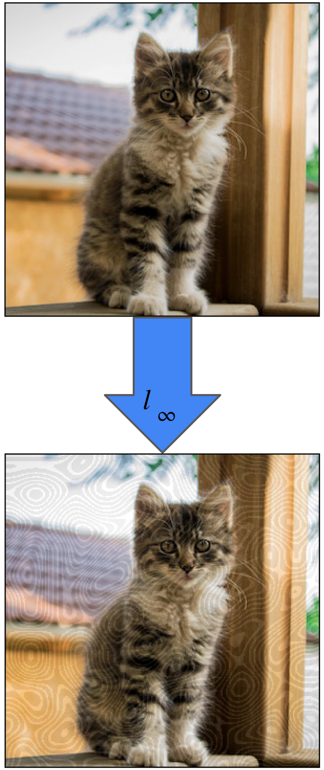
Figure 2: Quantization of unsigned data to 3-bit or 4-bit ($\alpha = 1.0$) using three different quantization levels. APoT quantization has a more reasonable resolution assignment and it does not suffer from the rigid resolution.

<u>Quantization:</u> reduce the memory of the deployed models by limiting the precision of the parameters of the models

- ➢ **(Q2, Q3, Q4)** 2, 3, and 4 bits

[3]

Adversarial Examples



Tabby Cat (82%)

Shower Curtain (89%)

[4]

$$C(x) := \text{true class label of input } x$$

$$x' = x + \delta$$

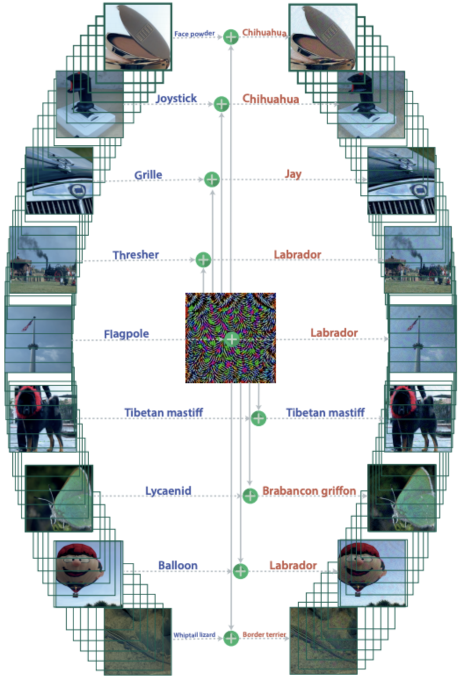
$$f(x') \neq C(x)$$

$$\delta = x' - x$$

$$||\delta||_p < \varepsilon$$

$$\varepsilon > 0$$

Universal Adversarial Perturbations (UAPs)



$f(x+\delta) \neq C(x)$ <u>for multiple inputs</u>

$x \in X$ of a benign dataset X

UAPs exploit systemic vulnerabilities of the target model

[5]

Experiments

- Untargeted
  - White-box *(on self)*
  - **Black-box *(transfer)***
- Targeted
  - White-box *(all 10 class labels)*

Experiments: Metrics

- Untargeted
  - White-box *(on self)*
  - **Black-box *(transfer)***
- Targeted
  - White-box *(all 10 class labels)*

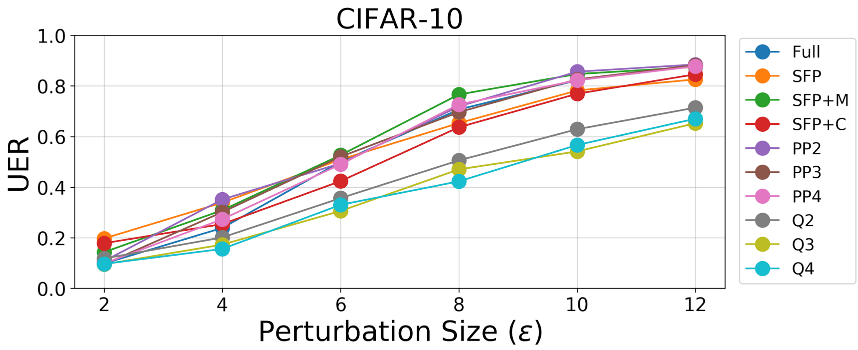
Universal Evasion Rate (UER)

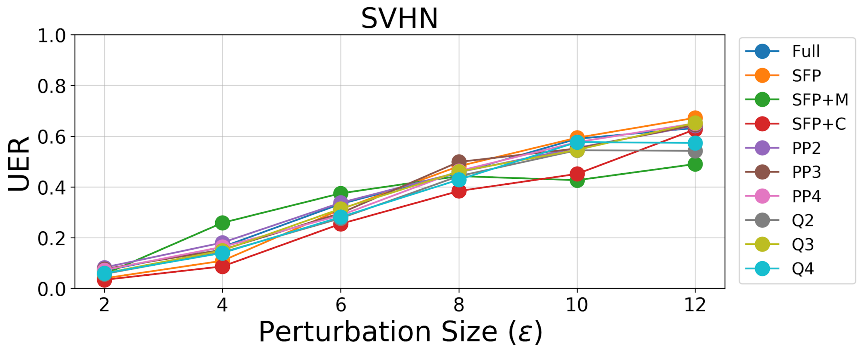
$$UER(\delta) = \left|\{x \in X : argmax\ F(x+\delta) \neq C(x)\}\right| \cdot \frac{1}{|X|}$$

Targeted Success Rate (TSR)

$$TSR\left(\delta, y_{tgt}\right) = \left|\left\{x \in X : argmax\ F(x+\delta) = y_{tgt}\right\}\right| \cdot \frac{1}{|X|}$$

Untargeted UAP: White-box





- Quantization on CIFAR-10 displays a lower average UER
- The average UER is much higher on CIFAR-10 than on SVHN

Untargeted UAP: Black-box transfer attack

CIFAR-10

Model \ Attack Source	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.81	0.46	0.45	0.48	0.83	0.82	0.80	0.44	0.37	0.39	0.58
SFP	0.31	0.76	0.34	0.46	0.32	0.35	0.34	0.33	0.31	0.30	0.38
SFP+M	0.38	0.69	0.82	0.72	0.69	0.45	0.43	0.32	0.39	0.35	0.52
SFP+C	0.36	0.58	0.44	0.75	0.46	0.39	0.38	0.41	0.38	0.36	0.45
PP2	0.82	0.56	0.57	0.59	0.86	0.81	0.81	0.42	0.43	0.44	0.63
PP3	0.81	0.47	0.46	0.48	0.83	0.81	0.80	0.43	0.36	0.40	0.59
PP4	0.81	0.45	0.44	0.47	0.83	0.82	0.81	0.44	0.37	0.39	0.58
Q2	0.67	0.66	0.62	0.69	0.77	0.67	0.67	0.63	0.54	0.54	0.65
Q3	0.67	0.57	0.60	0.62	0.77	0.67	0.67	0.57	0.54	0.54	0.62
Q4	0.68	0.53	0.57	0.61	0.79	0.69	0.66	0.58	0.53	0.57	0.62

Attack Source

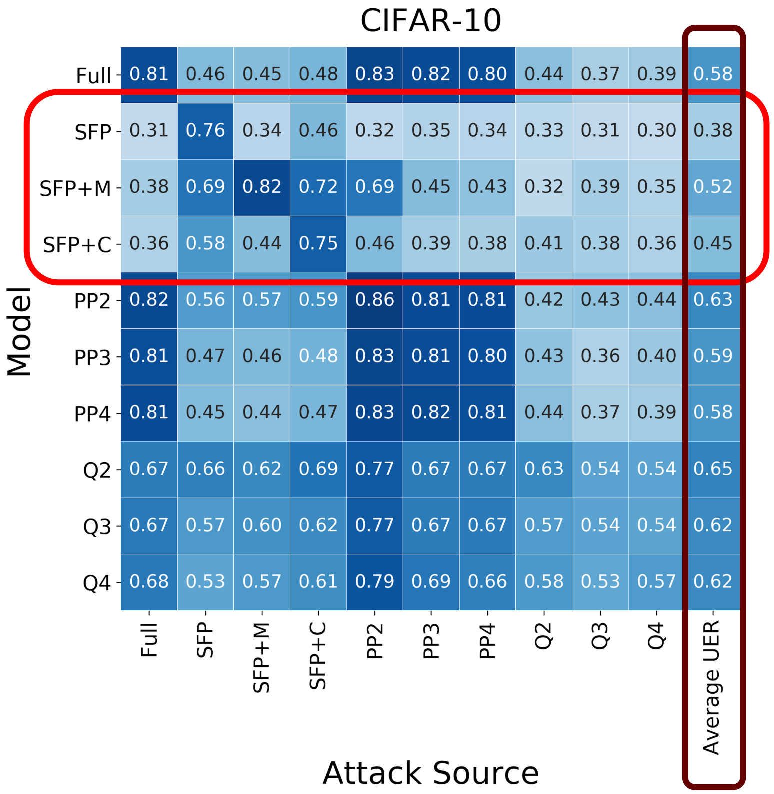
SVHN

Model \ Attack Source	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.59	0.49	0.22	0.18	0.54	0.54	0.57	0.45	0.47	0.48	0.45
SFP	0.53	0.59	0.21	0.15	0.45	0.45	0.48	0.42	0.43	0.53	0.42
SFP+M	0.43	0.49	0.43	0.17	0.37	0.32	0.41	0.44	0.40	0.50	0.40
SFP+C	0.41	0.50	0.20	0.45	0.36	0.35	0.42	0.41	0.39	0.46	0.39

Attack Source

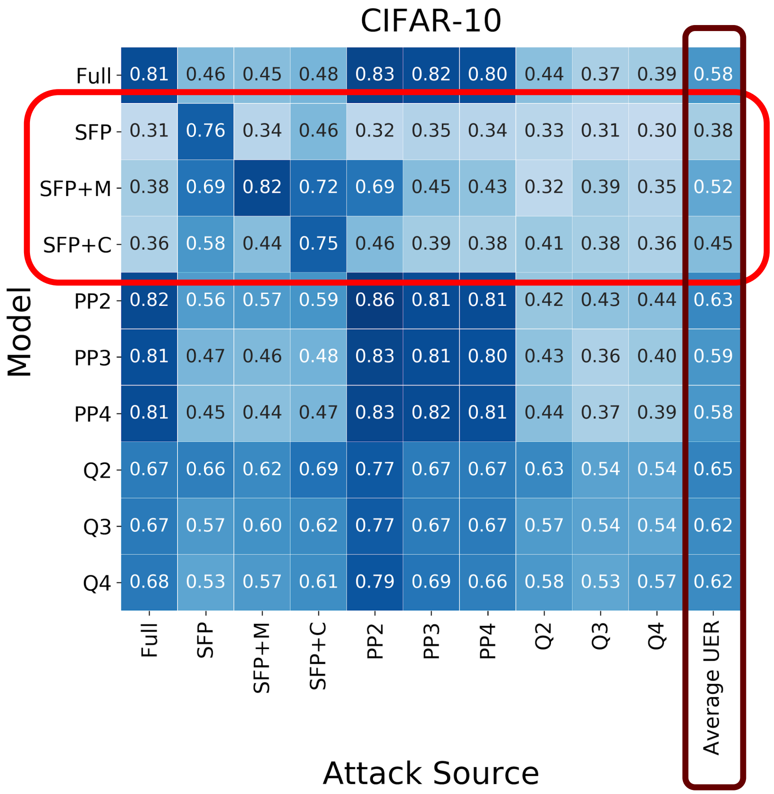
- Full model is mainly vulnerable to the UAPs crafted from the PP*i* models
- Full model's average UER is much higher on CIFAR-10 than on SVHN

Untargeted UAP: Black-box transfer attack



SFP is the most robust technique against transfer attacks

Untargeted UAP: Black-box transfer attack



Model	**CIFAR-10**
Full	94.02
SFP	79.51
SFP+M	86.09
SFP+C	83.54

Models are more susceptible to transfer attacks between networks sharing related feature mappings

Untargeted UAP: Black-box transfer attack

SVHN

Model \ Attack Source	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.59	0.49	0.22	0.18	0.54	0.54	0.57	0.45	0.47	0.48	0.45
SFP	0.53	0.59	0.21	0.15	0.45	0.45	0.48	0.42	0.43	0.53	0.42
SFP+M	0.43	0.49	0.43	0.17	0.37	0.32	0.41	0.44	0.40	0.50	0.40
SFP+C	0.41	0.50	0.20	0.45	0.36	0.35	0.42	0.41	0.39	0.46	0.39
PP2	0.60	0.50	0.23	0.20	0.55	0.56	0.58	0.45	0.48	0.49	0.46
PP3	0.60	0.50	0.23	0.19	0.54	0.55	0.58	0.45	0.48	0.49	0.46
PP4	0.59	0.50	0.22	0.19	0.54	0.54	0.58	0.46	0.47	0.49	0.46
Q2	0.53	0.53	0.27	0.17	0.48	0.48	0.52	0.54	0.50	0.56	0.46
Q3	0.51	0.52	0.27	0.18	0.47	0.44	0.49	0.49	0.55	0.56	0.45
Q4	0.48	0.49	0.25	0.16	0.45	0.43	0.48	0.44	0.46	0.58	0.42

SFP models trained on SVHN are more robust against UAP attacks from all other models

Untargeted UAP: Black-box transfer attack

SVHN

Model \ Attack Source	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.59	0.49	0.22	0.18	0.54	0.54	0.57	0.45	0.47	0.48	0.45
SFP	0.53	0.59	0.21	0.15	0.45	0.45	0.48	0.42	0.43	0.53	0.42
SFP+M	0.43	0.49	0.43	0.17	0.37	0.32	0.41	0.44	0.40	0.50	0.40
SFP+C	0.41	0.50	0.20	0.45	0.36	0.35	0.42	0.41	0.39	0.46	0.39
PP2	0.60	0.50	0.23	0.20	0.55	0.56	0.58	0.45	0.48	0.49	0.46
PP3	0.60	0.50	0.23	0.19	0.54	0.55	0.58	0.45	0.48	0.49	0.46
PP4	0.59	0.50	0.22	0.19	0.54	0.54	0.58	0.46	0.47	0.49	0.46
Q2	0.53	0.53	0.27	0.17	0.48	0.48	0.52	0.54	0.50	0.56	0.46
Q3	0.51	0.52	0.27	0.18	0.47	0.44	0.49	0.49	0.55	0.56	0.45
Q4	0.48	0.49	0.25	0.16	0.45	0.43	0.48	0.44	0.46	0.58	0.42

Attack Source

SFP plus regularization lacks transferability to the other models

Untargeted UAP: Black-box transfer attack

CIFAR-10

Model	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.81	0.46	0.45	0.48	0.83	0.82	0.80	0.44	0.37	0.39	0.58
PP2	0.82	0.56	0.57	0.59	0.86	0.81	0.81	0.42	0.43	0.44	0.63
PP3	0.81	0.47	0.46	0.48	0.83	0.81	0.80	0.43	0.36	0.40	0.59
PP4	0.81	0.45	0.44	0.47	0.83	0.82	0.81	0.44	0.37	0.39	0.58

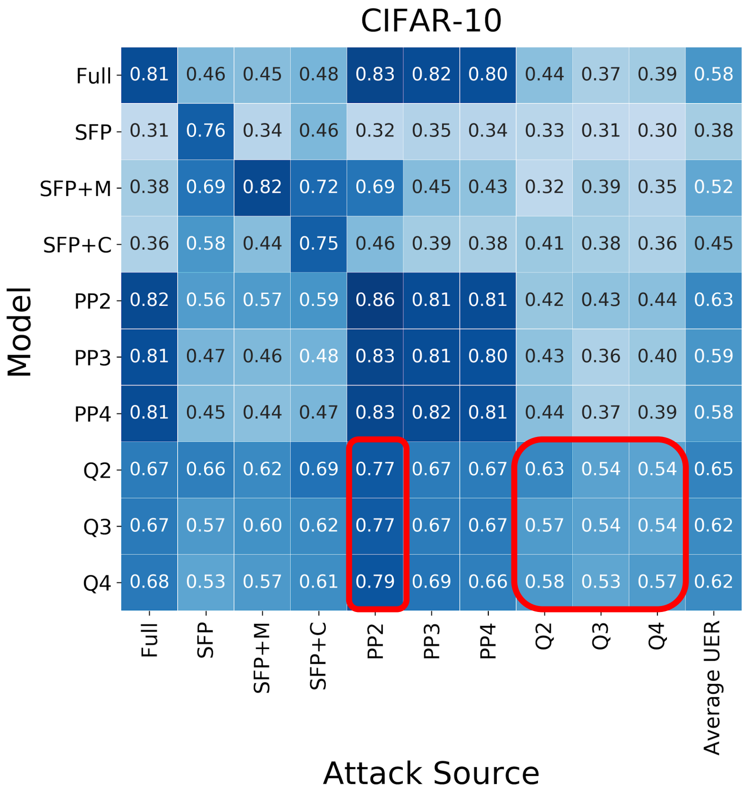
SVHN

Model	Full	SFP	SFP+M	SFP+C	PP2	PP3	PP4	Q2	Q3	Q4	Average UER
Full	0.59	0.49	0.22	0.18	0.54	0.54	0.57	0.45	0.47	0.48	0.45
PP2	0.60	0.50	0.23	0.20	0.55	0.56	0.58	0.45	0.48	0.49	0.46
PP3	0.60	0.50	0.23	0.19	0.54	0.55	0.58	0.45	0.48	0.49	0.46
PP4	0.59	0.50	0.22	0.19	0.54	0.54	0.58	0.46	0.47	0.49	0.46

Attack Source

UAPs exploit combined activations of neurons that are commonly activated for classifying benign inputs.

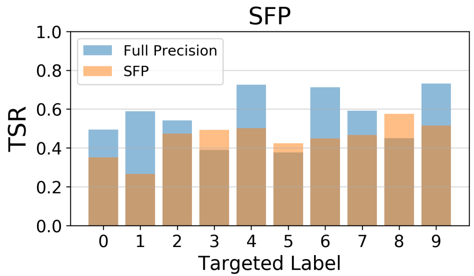
Untargeted UAP: Black-box transfer attack

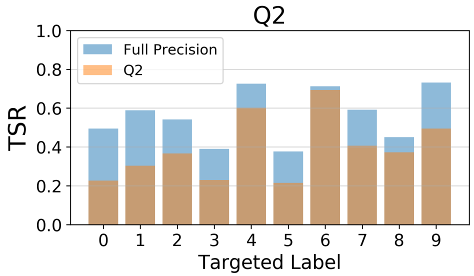


Quantization has gradient-masking
- Q2, Q3, Q4 have 54-63% UER on themselves
- However PP2 achieves 77-79% UER

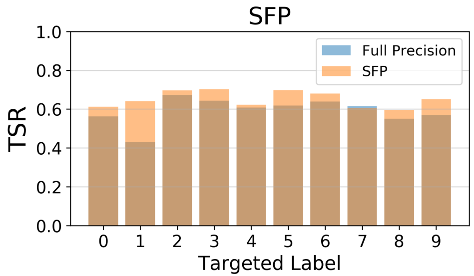
Targeted UAPs

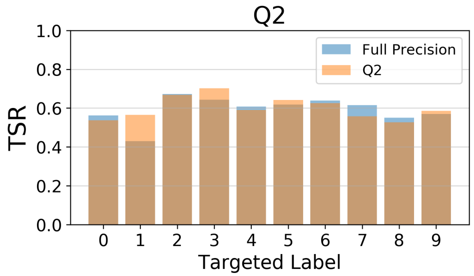
CIFAR-10





SVHN





The application and properties of the datasets play an important role in the robustness of the considered compression techniques to UAP attacks

Conclusions

Conclusions

There exists a correlation between clean model accuracy and UER of untargeted white-box attacks

Conclusions

SFP improves the model's robustness to transfer attacks

1. There exists a **correlation** between clean model accuracy and UER of untargeted white-box attacks

Conclusions

Quantization can give a false sense of security

1. There exists a **correlation** between clean model accuracy and UER of untargeted white-box attacks

2. SFP improves the model's robustness to transfer attacks

Conclusions

Robustness to UAPs when using compression methods is dataset and application dependent

1. There exists a **correlation** between clean model accuracy and UER of untargeted white-box attacks
2. SFP improves the model's robustness to transfer attacks
3. Quantization can give a false sense of security

Conclusions

To know more about it -- stop by our poster

Thank you!!

1. There exists a **correlation** between clean model accuracy and UER of untargeted white-box attacks
2. SFP improves the model's robustness to transfer attacks
3. Quantization can give a false sense of security
4. Robustness to UAPs when using compression methods is dataset and application dependent

Thank you for listening!

Code available: **https://github.com/kenny-co/sgd-uap-torch**

References:

[1] alwaysAI, I., 2021. *Object Detection And Person Detection In Computer Vision*. [online] Learn.alwaysai.co. Available at: <https://learn.alwaysai.co/object-detection> [Accessed 16 January 2021].

[2] Blog.tensorflow.org. 2021. *Tensorflow Model Optimization Toolkit — Pruning API*. [online] Available at: <https://blog.tensorflow.org/2019/05/tf-model-optimization-toolkit-pruning-API.html> [Accessed 16 January 2021].

[3] Li, Y.; Dong, X.; and Wang, W. 2019. Additive Powers-of- Two Quantization: A Non-uniform Discretization for Neural Networks. *arXiv preprint arXiv:1909.13144* .

[4] Muñoz-González, L., 2019. *Machine Learning To Augment Shared Knowledge In Federated Privacy-Preserving Scenarios (MUSKETEER)*. [online] Musketeer.eu. Available at: <https://musketeer.eu/wp-content/uploads/2019/10/MUSKETEER_D5.1.pdf>

[5] Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal Adversarial Perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1765–1773.