UNIVERSITY OF CAMBRIDGE

BOSCH-FORSCHUNGSSTIFTUNG
im Stifterverband

To compress or not to compress: Understanding the Interactions between Adversarial Attacks and Neural Network Compression

Yiren Zhao, Ilia Shumailov, Robert Mullins, Ross Anderson

firstname.lastname@cl.cam.ac.uk

■ Introduction

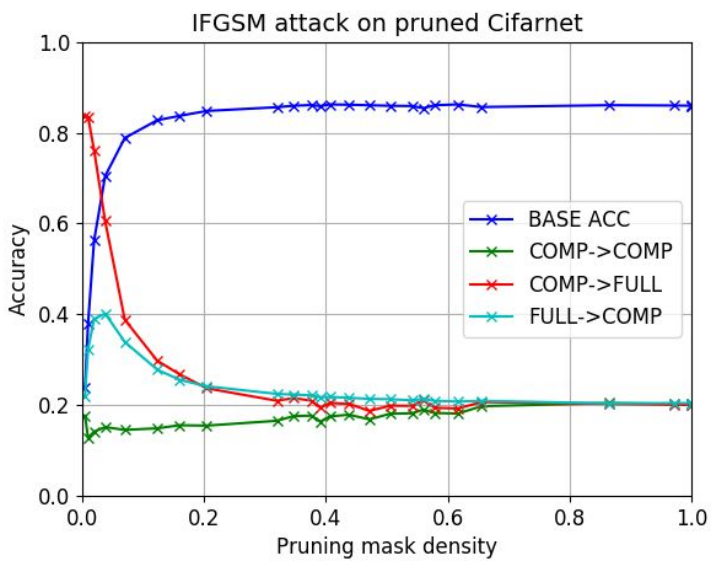
- ▪Over a billion smartphones and high-end IoT devices p.a.
- ▪Many of them are, or will soon be, running some form of compressed neural networks
- ▪We found compression makes adversarial samples easier to propagate through models that share heritage

■ Results and Discussions

▪Adversarial Samples

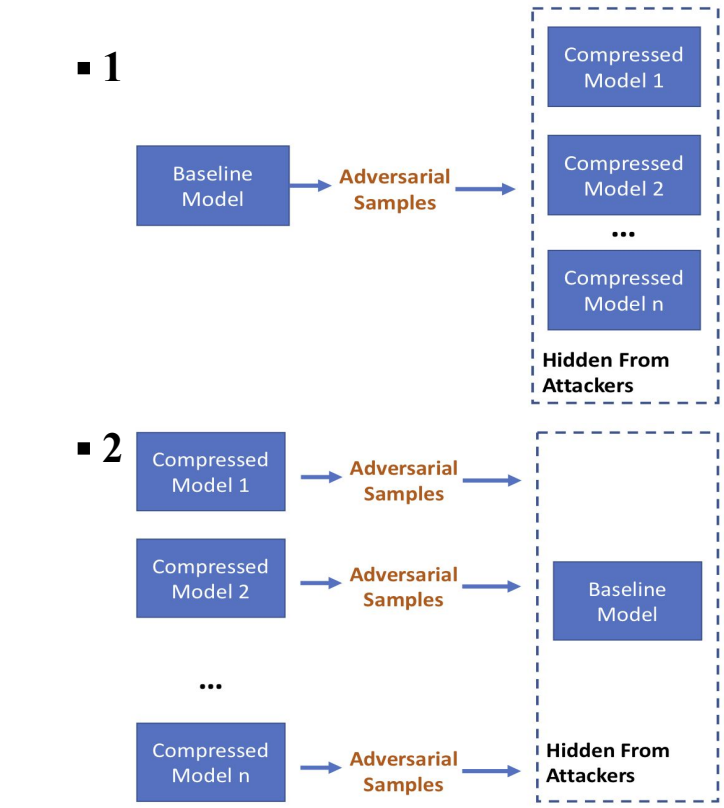
- Adversarial samples can trick neural networks while being imperceptible to humans
- They undermine any case for security and reliability
- Often transferable if feature spaces are similar [1]

▪ Attacking Pruned and Quantised Models

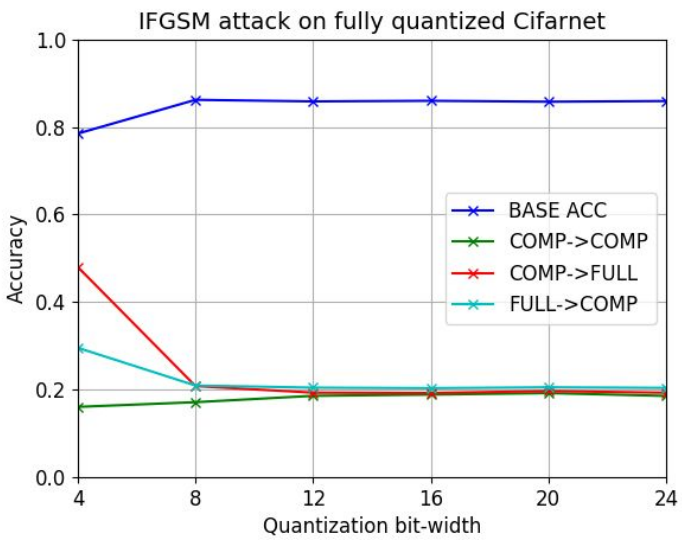


- The effect is observed for both weak and strong attackers on networks of different sizes.

▪Attack Scenarios



- In both attack scenarios here, adversarial samples generated from one model can affect other models too
- Models can be trained using ‘crypto keys’ to stop transferability and detect attacks [2, 3]



■ References:

[1] Szegedy et al., Intriguing properties of neural networks. (2013)
[2] Shumailov et al. The Taboo Trap: Behavioural Detection of Adversarial Samples (2018) arXiv:1811.07375
[3] Shumailov et al. Sitatapatra: Blocking the Transfer of Adversarial Samples (2019) arXiv: 1901.08121