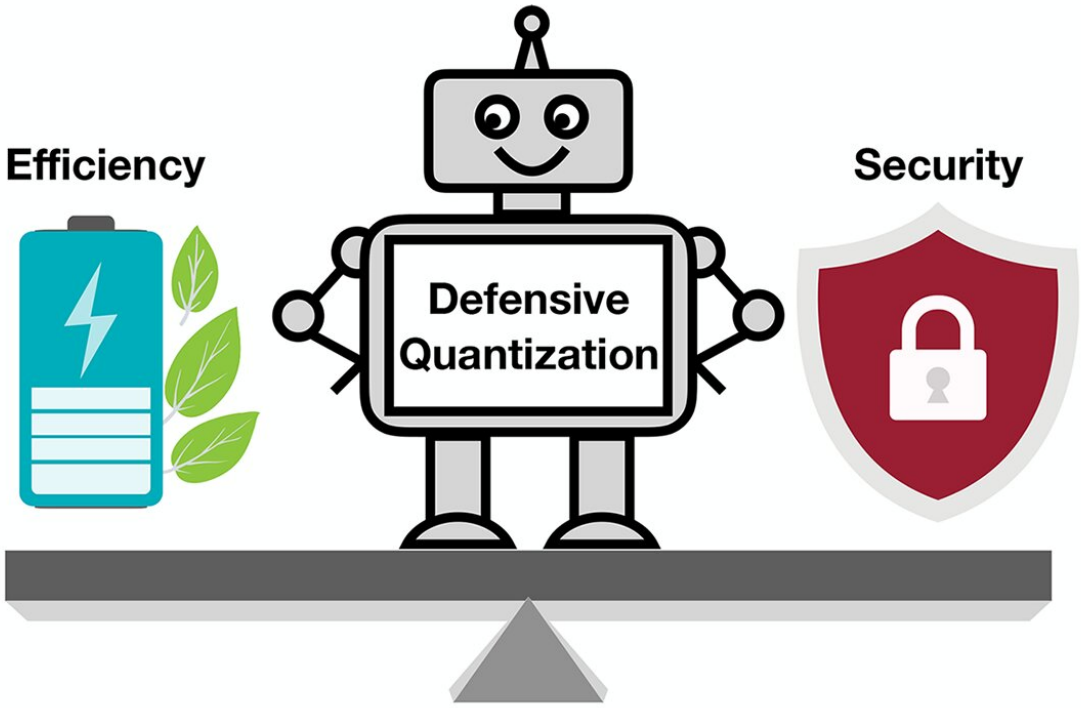
Improving security as artificial intelligence moves to smartphones

April 24 2019, by Kim Martineau



Credit: Ji Lin

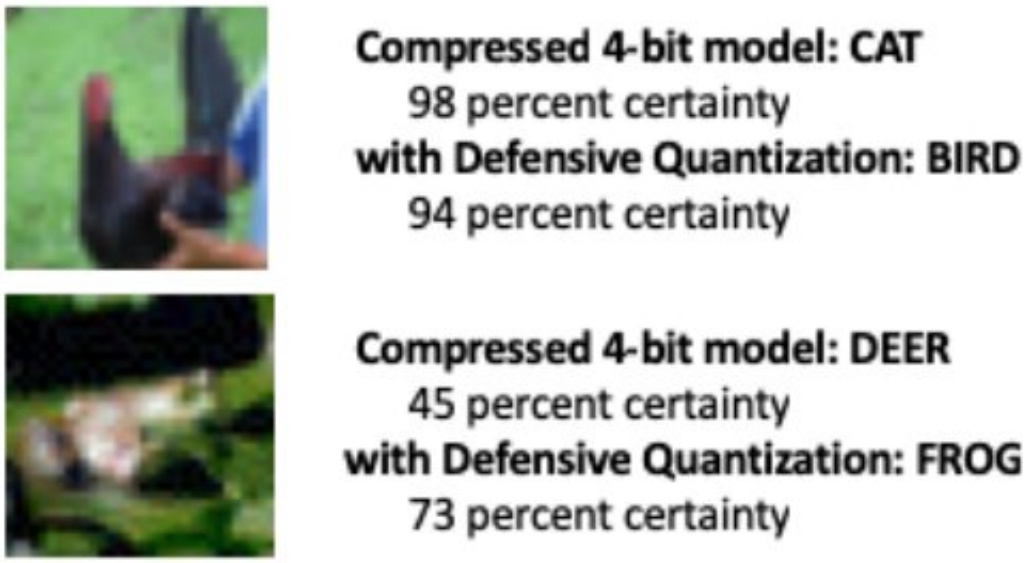
Smartphones, security cameras, and speakers are just a few of the devices that will soon be running more artificial intelligence software to speed up image- and speech-processing tasks. A compression technique

known as quantization is smoothing the way by making deep learning models smaller to reduce computation and energy costs. But smaller models, it turns out, make it easier for malicious attackers to trick an AI system into misbehaving—a concern as more complex decision-making is handed off to machines.

In a new study, MIT and IBM researchers show just how vulnerable compressed AI models are to adversarial attack, and they offer a fix: add a mathematical constraint during the quantization process to reduce the odds that an AI will fall prey to a slightly modified image and misclassify what they see.

When a deep learning model is reduced from the standard 32 bits to a lower bit length, it's more likely to misclassify altered images due to an error amplification effect: The manipulated image becomes more distorted with each extra layer of processing. By the end, the model is more likely to mistake a bird for a cat, for example, or a frog for a deer.

Models quantized to 8 bits or fewer are more susceptible to adversarial attacks, the researchers show, with accuracy falling from an already low 30-40 percent to less than 10 percent as bit width declines. But controlling the Lipschitz constraint during quantization restores some resilience. When the researchers added the constraint, they saw small performance gains in an attack, with the smaller models in some cases outperforming the 32-bit model.



When a few pixels were manipulated in the above images to simulate an adversarial attack, a standard compressed model misclassified the chicken as “cat” and the frog as “deer.” But when researchers added a constraint during compression, the model correctly classified the animals, even performing better than a full precision 32-bit model. Credit: Massachusetts Institute of Technology

"Our technique limits error amplification and can even make compressed deep learning models more robust than full-precision models," says Song Han, an assistant professor in MIT's Department of Electrical Engineering and Computer Science and a member of MIT's Microsystems Technology Laboratories. "With proper quantization, we can limit the error."

The team plans to further improve the technique by training it on larger datasets and applying it to a wider range of models. "Deep learning models need to be fast and secure as they move into a world of internet-connected devices," says study coauthor Chuang Gan, a researcher at the MIT-IBM Watson AI Lab. "Our Defensive Quantization technique helps on both fronts."

The researchers, who include MIT graduate student Ji Lin, present their results at the International Conference on Learning Representations in May.

In making AI models smaller so that they run faster and use less energy, Han is using AI itself to push the limits of model compression technology. In related recent work, Han and his colleagues show how reinforcement learning can be used to automatically find the smallest bit length for each layer in a quantized model based on how quickly the device running the model can process images. This flexible bit width approach reduces latency and energy use by as much as 200 percent compared to a fixed, 8-bit model, says Han. The researchers will present their results at the Computer Vision and Pattern Recognition conference in June.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology