

How many bits does it take to quantize a neural network?

Dr Miro Giacobbe

University of Oxford

Monday 4 May 2020

2.00pm

via Videolink

Abstract:

Quantization converts neural networks into low-bit fixed-point computations which can be carried out by efficient integer-only hardware and is standard practice for the deployment of neural networks on real-time embedded devices. However, like their real-numbered counterpart, quantized networks are not immune to malicious misclassification caused by adversarial attacks. We investigate how quantization affects a network's robustness to adversarial attacks, which is a formal verification question. We show that neither robustness nor non-robustness are monotonic with changing the number of bits for the representation and, also, neither are preserved by quantization from a real-numbered network. For this reason, we introduce a verification method for quantized neural networks which, using SMT solving over bit-vectors, accounts for their exact, bit-precise semantics. We built a tool and analyzed the effect of quantization on a classifier for the MNIST dataset. We demonstrate that, compared to our method, existing methods for the analysis of real-numbered networks often derive false conclusions about their quantizations, both when determining robustness and when detecting attacks, and that existing methods for quantized networks often miss attacks. Furthermore, we applied our method beyond robustness, showing how the number of bits in quantization enlarges the gender bias of a predictor for students' grades.