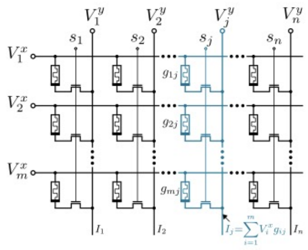
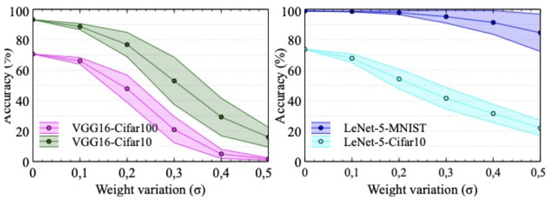
When Variations Meet Early-Exit

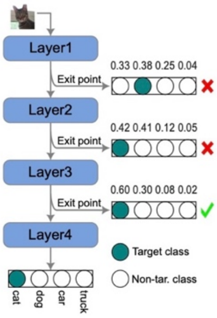
The last decade has witnessed the breakthrough of deep neural networks (DNNs) in many fields. With the increasing depth of DNNs, hundreds of millions of multiply-and-accumulate (MAC) operations need to be executed. To accelerate such operations efficiently, analog in-memory computing platforms based on emerging devices, e.g., resistive RAM (RRAM), have been introduced as shown in the left figure as follows. These acceleration platforms rely on analog properties of the devices and thus suffer from process variations and noise. Consequently, weights in neural networks configured into these platforms can deviate from the expected values, which may lead to feature errors and a significant degradation of inference accuracy, as shown in the middle figure as follows. State-of-the-art has explored error correction to enhance inference accuracy of neural networks implemented RRAM-based crossbars. However, early-exit strategy as shown in the following right figure has not been explored to further enhance accuracy.

In this thesis, first, early-exit strategy be explored to counter variations in RRAM-based in-memory-computing. Second, error correction layers will be inserted into the early layers to enable error correction incurred by variations and noise.





ınde



References:

https://arxiv.org/pdf/2309.13443

https://arxiv.org/pdf/2211.14917

If you are interested in this topic for master thesis, please contact:

Prof. Dr.-Ing. Li Zhang (grace.zhang@tu-darmstadt.de) with your CV and transcripts.