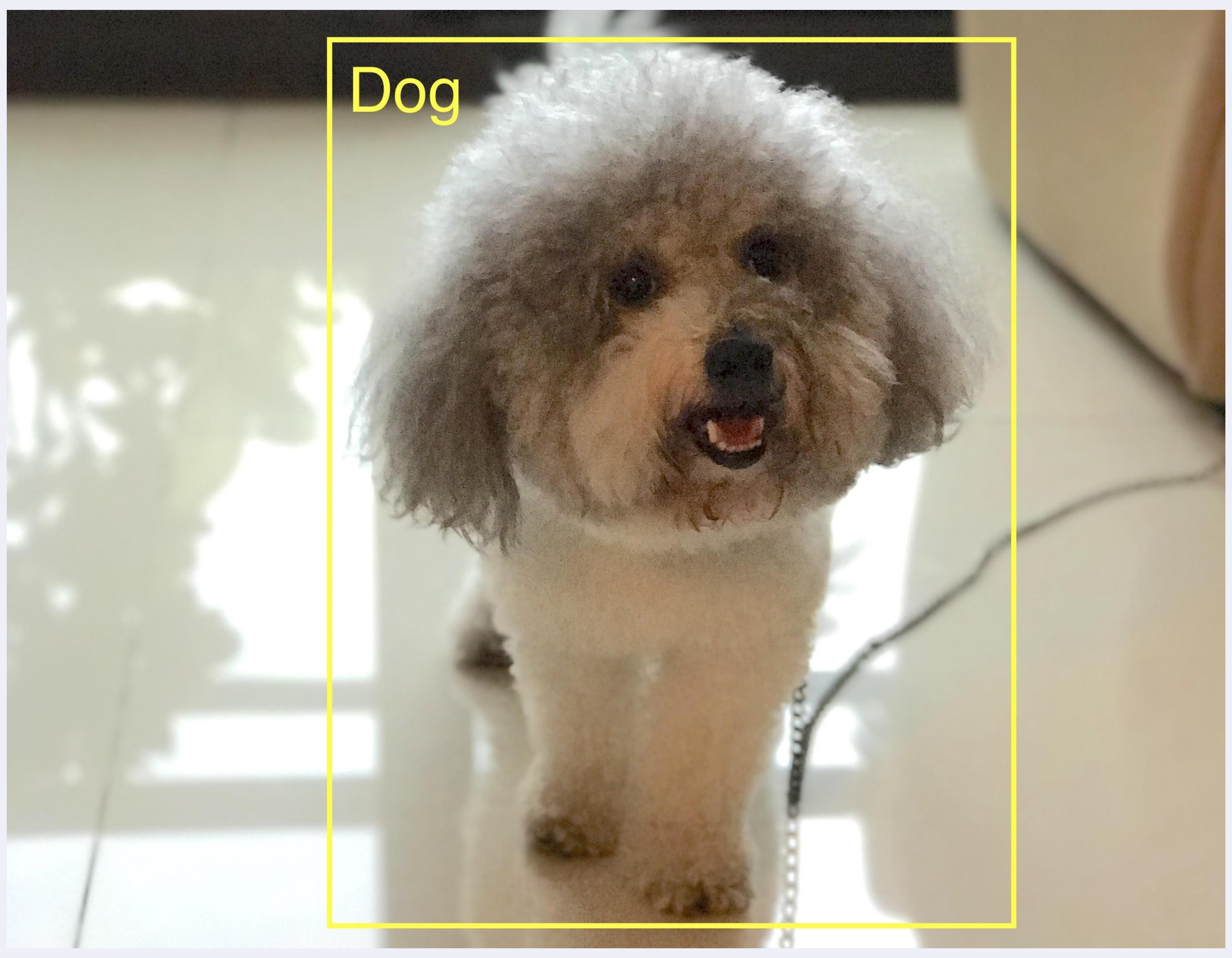
Approximate Computing for Convolutional Neural Networks (CNN)

Student: Chih-Yu Shen, Cheng-Ju Lee **Advisor:** Ren-Shuo Liu
Group No.: B16, System and Storage Design Lab (SSD LAB)

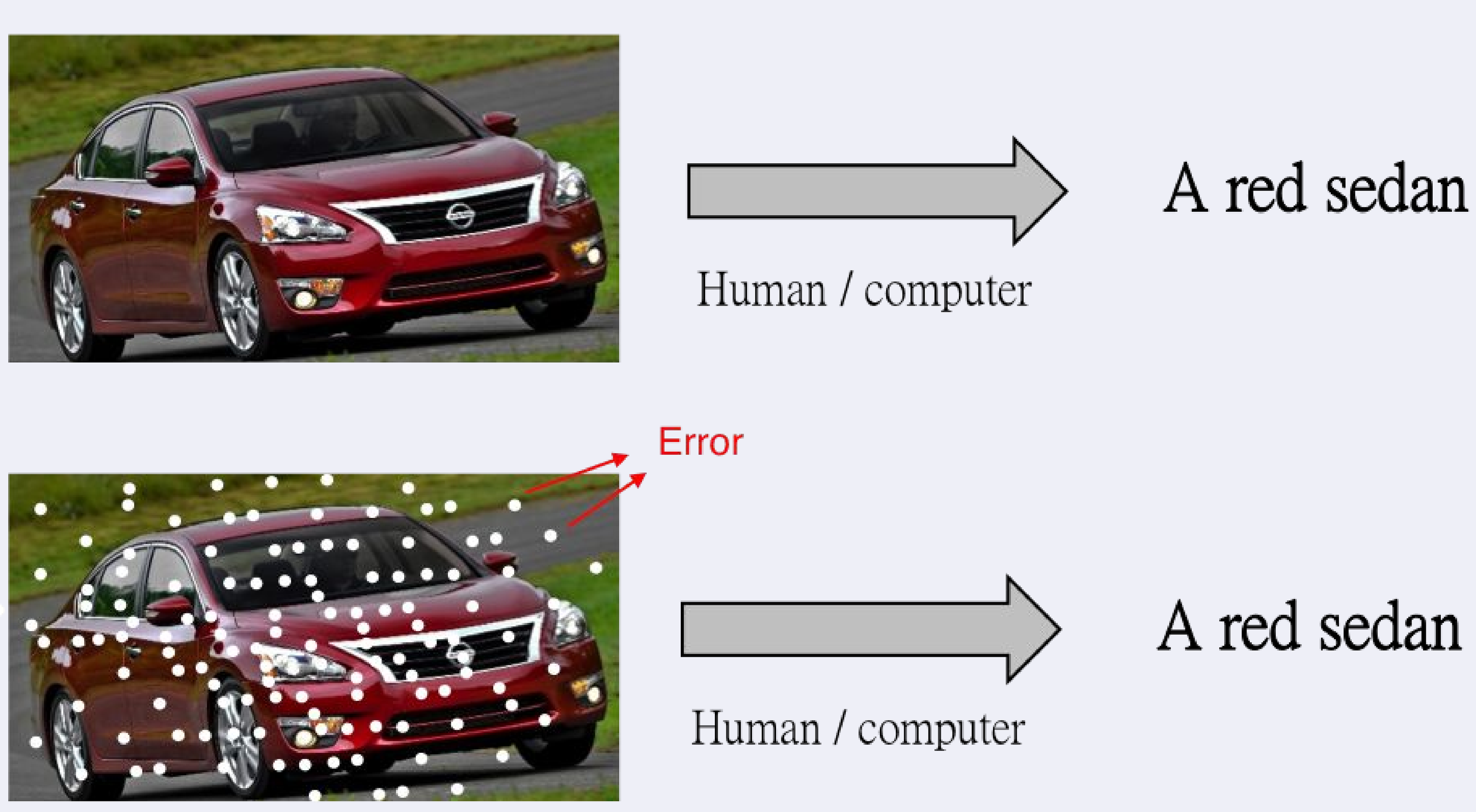
Problem



Convolutional neural network (CNN) is widely used for image recognition. As precision increases, the network structure becomes more and more complicated. As a result, it consumes more power and costs more computation time.

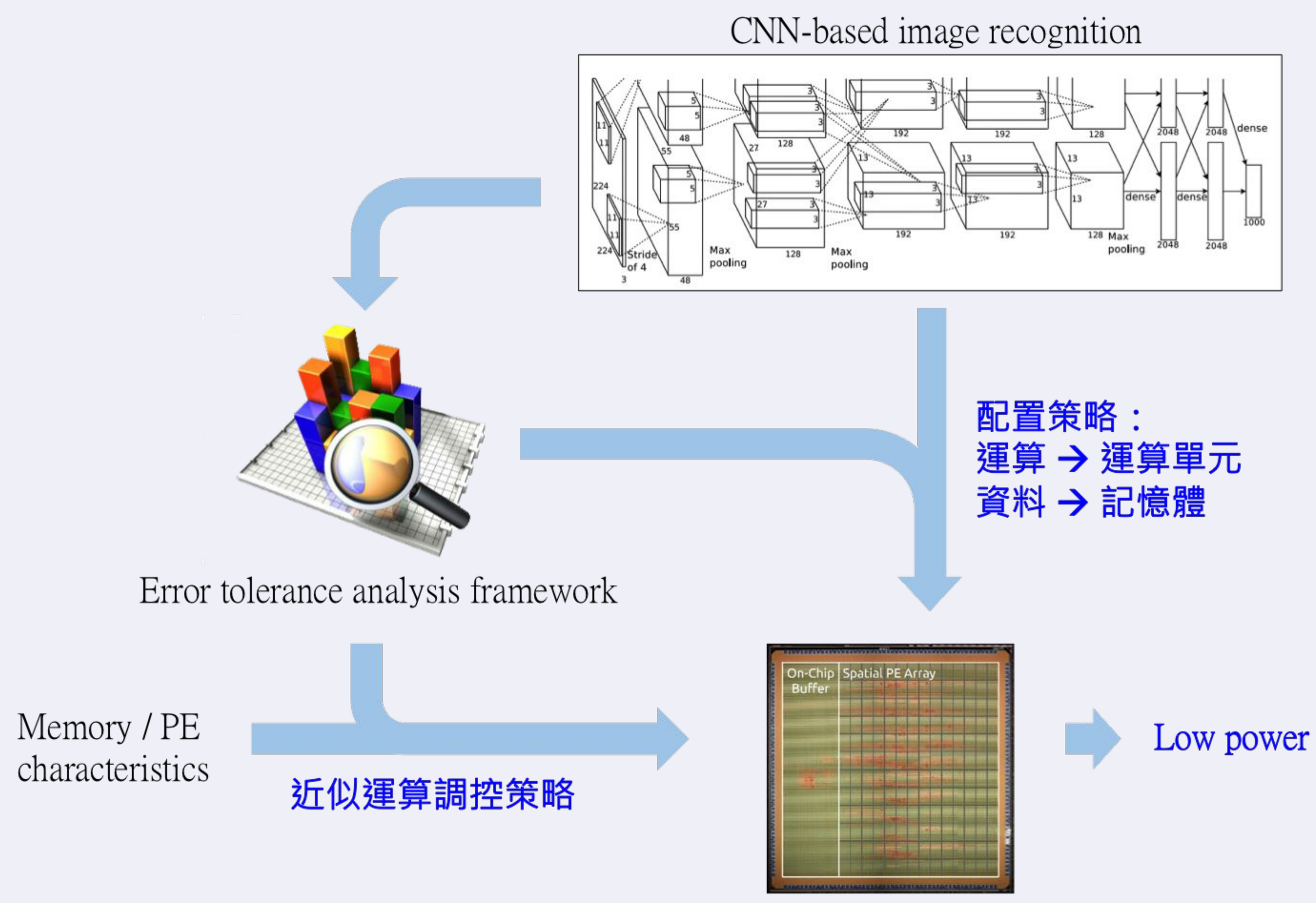
Solution

Reduction of power supply leads to error in memory storage and incorrect computation in CPU, but in image recognition, there's tolerance of error in data and computations.



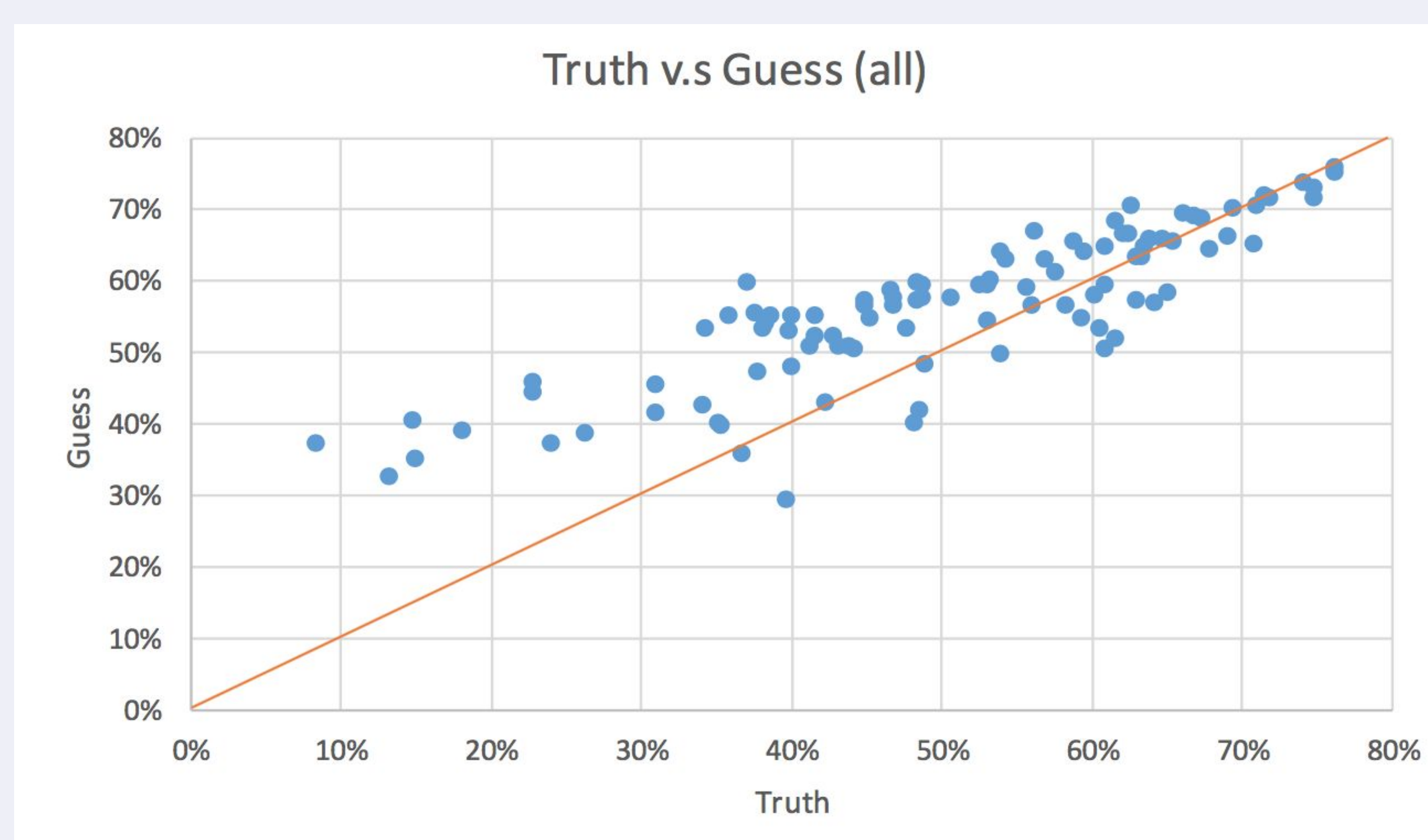
Therefore, we simulate hardware error by intentionally adding error in CNN software structures. With little precision drop, our purpose is to find an optimal error distribution for CNN structures, which leads to the best power reduction policy.

Method



We focus on input images, feature maps, and model parameters in CNN structures, and the methods of error simulation are bit error rate (BER) and Gaussian noise. To determine error distribution, error rate and the range of error are two key factors. We analyze the relation between precision and the two factors respectively. To begin with, we apply designed experiments on Cifar10 model, which is a simple model with less layers. After obtaining the error simulation results, we apply experiments on AlexNet and GoogLeNet, large CNN models with more layers, to confirm that our result is suitable for most CNN structures.

Conclusion



Our project confirms that with a proper arrangement of hardware, which implies a proper error distribution of CNN structures, the power consumption and computation cost can be considerably reduced.

Directed by Professor Ren-Shuo Liu, SSD Lab
National Tsing Hua University, EE Department
No. 101, Sec. 2, Kuang-Fu Rd., Hsinchu, Taiwan