



Figure 3: The CIFAR-10 classification network with 7 convolution layers with kernel size 3, followed by one convolution layer with kernel size 4, a convolution layer with kernel size 2, and the final convolution layer with kernel size 1.

Supplementary Material

CIFAR-10 architecture. For the CIFAR-10 classification network, we deployed a 10-layer CNN with 8.6M tunable parameters and 100 thousand ReLU nodes, as depicted in Fig. 3. The network was trained using an Adam optimiser for 40, 20, 10 epochs with learning rates 10^{-3} , 10^{-4} , 10^{-5} , respectively. Random flipping and cropping was performed for data augmentation. The network achieves a classification accuracy of 86% on the test set.

ImageNet architecture. Building on a state-of-the-art network for ImageNet classification [14], we evaluate if Verinet_{BF} is applicable to such large networks. Therefore, we modified the smallest Normaliser-Free residual neural network (NFNet) architecture [11, 12], *i.e.* NFNet-F0, with more than 15M ReLU nodes to a reduced version with 6.5M ReLU nodes. Hence, we reduced the number of feature channels by half, and we removed one convolution layer in the NFNet bottleneck block. Batch normalisation was not required as the weights of convolution layers are standardised, and the ReLUs were scaled with a fixed factor of $\gamma = \sqrt{2/(1 - (1/\pi))}$. The squeeze&excite blocks were omitted, as we considered only fully-connected, convolution and average-pooling layers in our framework. Following the PyTorch implementation of [13], we trained the network for 360 epochs with batch size 64, Nesterov momentum of 0.9, weight decay of $2 \cdot 10^{-5}$, and adaptive gradient clipping at 0.1 [12]. We performed data augmentation with random horizontal flips, and random contrast and brightness changes of 0.1. The network reaches a top-1% accuracy of 70% on the ImageNet validation set.