Squeeze and Excitation ResNet for Digit-Sum Regression

Rishabh Bhattacharya 2023121011

1 Introduction

Predicting the sum of the digits contained within an image is an interesting regression challenge with a discrete target space (i.e., an integer sum). Rather than classifying each digit independently and then summing, I treat this as a direct regression problem in which the network outputs a single scalar representing the total. My approach leverages a Squeeze-and-Excitation (SE) ResNet architecture, a variant of the classical ResNet that includes channel attention modules (SE blocks).

2 Motivation and Architectural Details

2.1 Task Rationale

Digit-sum regression poses several subtleties. Individual digits can appear in different spatial regions of the image, and the counting (or summation) mechanism benefits from robust feature extraction. A residual architecture lends itself well to this task as it helps avoid vanishing gradients and allows the model to learn skip connections, retaining crucial low-level features that may be lost with deeper layers.

2.2 Basic ResNet Structure

I adopt a four-layer ResNet skeleton consisting of stacked residual blocks. Each block comprises:

- Two convolution layers (kernel size 3×3), each followed by batch normalization.
- A skip connection that bypasses these convolution layers.
- A ReLU nonlinearity to ensure non-linear mapping power.

After passing through these four residual layers, the feature maps are downsampled with a global average pooling layer, which reduces spatial dimensions to a single vector. This vector is fed into a fully connected layer that yields a single scalar output corresponding to the predicted digit sum.

2.3 Squeeze-and-Excitation (SE) Blocks

The ResNet blocks are augmented with Squeeze-and-Excitation modules [1], which improve the network's capacity to recalibrate channel responses. Specifically, each block:

1. **Squeezes** its feature maps via global average pooling, capturing channel-wise statistics into a compact vector.
2. **Excites** these channels by passing the squeezed vector through two fully connected layers (with a ReLU in between), producing channel-wise weights in the range $[0, 1]$ via a sigmoid.

3. **Scales** the original feature maps by these weights, adaptively highlighting the most informative channels.

Intuitively, this mechanism helps the network focus on features most pertinent for digit recognition and summation. The process introduces minimal additional computational overhead compared to the original residual block yet significantly enhances performance.

2.4 Training Setup

I use a Mean Squared Error (MSE) loss between the scalar prediction and the ground-truth sum. Although the output is an integer in practice, the MSE objective allows a smooth gradient flow and, in tandem with a rounding step at inference time, works well for discrete targets. I use an AdamW optimizer with a cyclic OneCycle learning rate schedule (over 100 epochs) to ensure stable training, where the learning rate gradually warms up and then decays.

3 Results and Discussion

After 87 epochs, my model obtains a validation accuracy of 93.5% when rounding predictions.

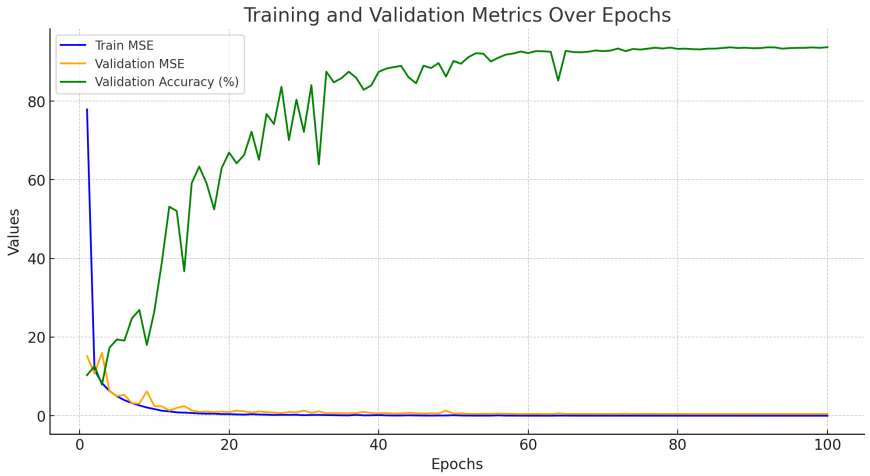

Figure 1: Loss and accuracy curves over the training epochs.

References

[1] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, "Squeeze-and-Excitation Networks," *arXiv preprint arXiv:1709.01507*, 2017.