ComputerVision Project - CNN

Thomas Komen (12556963)

October 2023

Contents

1 Introduction

For the classification of objects in images, neural networks are one option to accomplish this task. By training the networks on a collection of pictures with their respective label, such a network could potentially identify images after doing this for a sufficient amount of time.
These networks can have a range of different forms and sizes, with differences in the amount of layers, the amount of neurons in each layer, and the type of operation that each layer performs.

2 Project Setup

For this project, we will be looking into training two of these networks: A two-layer network, and a Convolutional network. In order to train our models, we will be using the CIFAR-100 dataset (Krizhevsky, Hinton, et al., 2009). This is a dataset of 100 classes, with 600 images per class. These are divided into 500 images to train on, with 100 images to test on. Each image is of 32 by 32 pixels, and in RGB.

3 Choices for model architecture

In this section, we shall be discussing the architecture of the two networks we use for this project, with explanations as to why these choices were made.

3.1 Two-Layer Model

The first model used in this project is an ordinary two-layer model. This model has an input size of 3*32*32, for images of 32 by 32 pixels, and 3 colour channels. The network contains one singular hidden layer of variable size, which can be tuned in order to produce better results. Since the dataset used contains 100 classes, the output size will be this 100, with each output corresponding to one of these classes. Each layer is fully connected, and makes use of the ReLu activation function.

3.2 Convolutional network

The second model used in this project is a convolutional network, with its architecture based on LeNet-5 (Lecun, Bottou, Bengio, & Haffner, 1998). The first layer is a convolutional layer with 3 channels input and 6 channels output, and has a kernel size of 5. Next, a subsampling layer is applied to this output, where the maximum value of each 2x2 window is taken, effectively cutting the size of each dimension in half. The same convolution is done again, this time from 6 channels to 16 channels, and the same subsampling is applied to these values as well. From this point on, the convoluted matrices are turned into a vector, after which one more linear layer maps these values together to 100 outputs, once again representing the 100 classes.

References

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. doi: 10.1109/5.726791