

EE5179 : Deep Learning for Imaging

Programming Assignment 2: Convolutional Neural Networks

Due Date: October 3, 2025

Instructions

1. Program in python for this assignment.
2. Post any doubts you have on moodle. This will be helpful to your peers as well.
3. Submit the codes (the actual notebook and a PDF version) and your assignment report in a zip file titled PA2 RollNumber.zip in the submission link provided on moodle.

Preliminaries

1. It is recommended to use Google Colab (as in the tutorials) or Jupyter/iPython notebooks for the assignment. The notebook format is very convenient to work (and evaluate!) with and additionally, Colab provides GPU access as well.
2. The dataset can be downloaded from [here](#) or you can make use of [the inbuilt dataset from Pytorch](#).

Note: Check if the labels are in one-hot format, or appropriately convert them to one-hot format before training and testing the network.

1 MNIST classification using CNN :

Build a simple network to classify MNIST data using the following architecture.

- input
- conv1 (32 3×3 filters, stride 1, zero padding 1)
- 2×2 maxpool with stride 2
- conv2 (32 3×3 filters, stride 1, zero padding 1)
- 2×2 maxpool with stride 2
- fully connected layer (500 outputs)
- fully connected layer (10 outputs)
- softmax classifier

There are 60000 images for train and 10000 images for test in MNIST. Separate the train dataset into 50000 for training and 10000 for validation.

So, the final split would be: 50000 for train, 10000 for val, and 10000 for test.

Use ReLU non-linearity for training the neural network.

1. Show the plot of training error, validation error and prediction accuracy as the training progresses. At the end of training, report the average prediction accuracy for the whole test set of 10000 images.
2. Plot randomly selected test images showing the true and predicted class labels.
3. Report the dimensions of the input and output at each layer.
4. How many parameters does your network have? How many of these are in the fully connected layers and how many are in the convolutional layers?
5. How many neurons does your network have? How many of these are in the fully connected layers and how many are in the convolutional layers?
6. Use batch-normalization. Does it improve the test accuracy? Does it affect training time?

2 Visualizing Convolutional Neural Network :

1. Plot the conv1 layer filters. Do you observe interesting patterns?
2. Plot filters of a higher layer. Compare it with conv1 layer filters.
3. Visualize the activations of the convolutional layers. What do you observe as you go deeper?
4. **Occluding parts of the image:** Suppose that the network classifies an image of a digit successfully. How can we be certain that it is actually observing the main part of the digit in the image as opposed to background or something else? One way of investigating this is by plotting the probability of the class of interest as a function of the position of an occluder object. So, occlude parts of the original image iteratively with a patch (e.g. a grey patch). You can use the occluder size to be a square of size anywhere between 1 to 5 pixels. Observe the probability of the class of interest and plot the probability in a grid as a function of the patch position. (You can use x-axis and y-axis to denote the position of the patch in xy plane, and plot the probability as intensity). Based on this experiment, report if the learning is meaningful or not.

3 Adversarial Examples

Adversarial examples are inputs to a neural network that result in an incorrect output from the network. For the following experiments, load the pre-trained MNIST model with the highest accuracy from the first problem. Do not change the trained weights in any of your operations.

3.1 Non-Targeted Attack:

The idea is to generate an image that is designed to make the neural network produce a certain output. Initialize a matrix of size of an MNIST image with Gaussian noise centered around 128. Let this matrix be \mathbf{X} . Starting with this noise matrix we will try to maximize the probability

of this matrix being classified as some target class. Here target class can take values between 0 to 9. Define the cost function as :

$$C = logits[target_class] \quad (1)$$

where *logits* is the network output before final softmax operation. Find the derivatives (*d*) of the cost function with respect to the input *x* using backpropagation. Now, set

$$\mathbf{X} = \mathbf{X} + stepsize \times \mathbf{d} \quad (2)$$

This essentially tries to increase the target class score by gradient ascent.

1. Show the generated image for each of the classes of MNIST.
2. Is the network always predicting *target_class* with high confidence for the generated images?
3. Do the generated images look like a number? If not, can you think of some reason?
4. Plot the cost function. Is it increasing or decreasing?

3.2 Targeted Attack

Can we generate some adversarial example that looks like a particular digit, but the network will classify it as something else? For example, can we generate an image of digit 2, which the network will classify as 5? Here, consider *target_image* as an image of 2 and *target_class* as 5. Hence, for this, change the cost function to:

$$C = logits[target_class] - \beta \times MSE(generated_image, target_image) \quad (3)$$

So, what we're doing now is we are trying to increase the target class score for the generated image, but also want the generated image to look like the target image (by minimizing the MSE). (β is very small e.g. 0.0001, tune it according to the output). Find the derivatives (*d*) of the cost function with respect to the input *X* using backpropagation. Now, set -

$$\mathbf{X} = \mathbf{X} + step_size \times \mathbf{d} \quad (4)$$

1. Show the generated image for each of the MNIST classes. Does the generated image resemble a digit? If so, then what digit does it resemble?

(In 3.1 and 3.2, do not modify the trained network parameters. For the experiments in 3.1 and 3.2, calculate the gradient w.r.t. to the image, then update the image only.)

-end-