

COEN 240 Machine Learning

Term Project: Neural Network and Deep Learning for Computer Vision

Guideline: Please complete the following tasks, submit a final report in PDF format and a separate zip file that contains all source code to Camino.

Task 1: Build a convolutional neural network for the recognition task with the fashion MNIST data set. Use the cross-entropy error function, and run 5 epochs. Give the **recognition accuracy rate** and show the **confusion matrix**, both for the test set.

Adopt the following convolutional neural network structure:

1. Input layer
2. 2d-convolutional layer: filter size 3x3, depth=32, ReLU activation function
3. 2x2 max pooling layer
4. 2d-convolutional layer: filter size 3x3, depth=64, ReLU activation function
5. 2x2 max pooling layer
6. 2d-convolutional layer: filter size 3x3, depth=64, ReLU activation function
7. Fully-connected layer: 64 units, ReLU activation function
8. (output) Fully-connected layer: 10 units, softmax activation function

The following code snippet is for your reference:

```
import tensorflow as tf

from tensorflow.keras import datasets, layers, models

fashion_mnist = keras.datasets.fashion_mnist

(train_images, train_labels), (test_images, test_labels) = fashion_mnist.load_data()

train_images = train_images.reshape((60000, 28, 28, 1))
test_images = test_images.reshape((10000, 28, 28, 1))

# Normalize pixel values to be between 0 and 1
train_images, test_images = train_images / 255.0, test_images / 255.0
```

```
model = models.Sequential()
```

```
model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=(28, 28, 1))) # the 1st 2d-convolutional layer
```


```
# ... to be completed by yourself
```

Task 2 Implement an image compression system using a neural network. The structure of the neural network is the following:

- (1) An (flattened) input layer with $m \times n$ nodes, where $m \times n$ is the image resolution (m rows and n columns of pixels)
- (2) A compressed layer with P nodes (no activation function), $P < m \times n$
- (3) An expansion layer with $m \times n \times T$ nodes, $T = 2$ is the expansion factor, followed by ReLU activation
- (4) An output layer with $m \times n$ nodes (no activation function)
- (5) A reshape layer that convert the 1-dimensional vector output to the $m \times n$ 2-dimensional image

Use the same fashion MNIST data set from **Task 1** for this problem. Batch size=64. Epochs=10. The loss

function (error function) is the mean-squared-error (mse) loss function. The mse is defined as

$$\text{mse} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2$$
, where x_i is the i -th pixel value in the original (normalized) image, and \hat{x}_i is the i -th pixel value in the decoded (reconstructed) image, and N is the number of pixels in one image. 

2.a For three different P values: $P = 10, 50, 200$, train the network, and perform compression and decompression using the trained model on the test images. Calculate the average reconstruction PSNR value of the test frames versus the P values. The **peak signal-to-noise ratio PSNR** (in dB) is defined as: $\text{PSNR(dB)} = 10 \log_{10} \left(\frac{MAX_I^2}{mse} \right)$, where MAX_I is the maximum pixel intensity value of original image. In this experiment for normalized gray-scale image, $MAX_I = 1$. Average PSNR: averaged over all test images.

What do you observe from the results? Give your comments.

2.b In one figure, display the first 10 test images and their decompressed images with $P = 10, 50$, and 200 in four rows: (a) the original 10 images, (b) the corresponding decompressed images with $P = 10$, (c) the decompressed images with $P = 50$, and (d) the decompressed images with $P = 200$.

What do you observe from the decompressed images (the visual quality of the decompressed images of different P values)? With the same P value, which kind of images do you think are more difficult to decompress, and why?

Task 3 Build a color video compression system with convolutional neural network and/or neural networks. Propose two different network architectures, and one of them must be convolutional neural network. Objective: achieve better reconstruction quality (higher PSNR values and visual qualities) at the same compression ratio, and study the trade-off between reconstruction quality, computational complexity, and model size. For a color image that has the R,G,B channels, the MSE is defined as $MSE = \frac{1}{3 \times N} \sum_{c=R,G,B} \sum_{i=1}^N (x_{ic} - \hat{x}_{ic})^2$, and the PSNR is defined as $PSNR(dB) = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right)$.

Three color video sequences are provided: BlowingBubbles_416x240_50, RaceHorses_416x240_30, and BasketballDrill_832x480_50.

we can try compression ratios: 1/32, 1/16, 1/12, 1/8, 1/4

compression ratio = depth/(width*height*3)

we change compression by modifying the sizes of the last 2 or 3 layers

compare models by plotting PSNR curves over different compression ratios
of course we use the same training set

Format of the report:

a. Title

b. Team member names and student IDs

c. Task 1: Give the recognition accuracy rate and show the confusion matrix, both for the test set.

d. Task 2: Show the average PSNR values of the test set for different P values in a table. Analyze the result as required in the task statement.

e. Task 3: please follow the instructions below (cite reference works wherever appropriate):

I Introduction: describe the task and/or background information.

II Proposed methods

- Network architectures (use tables or figures)
- Text description of the methodology details, such as: network structures, block-wise compression or frame-wise compression, how the networks are trained, loss functions used, ...

III Experimental Studies

- Dataset description: which datasets are used, how many training frames (validation set size if used), how many test frames, etc. Make sure that for different proposed networks, the training set should be the same, and the test set should also be the same.
- Quantitative evaluation: plot the PSNR curves of two different networks against the compression ratios: $1/32$, $1/16$, $1/8$, $1/4$, $1/2$. You can adjust the stride, pooling layers, or the number of encoder output channels to achieve different compression ratio. Analyze the results.
- Perceptual quality evaluation: for each compression ratio, pick several test frames, display the original frames and the reconstructed frames by two different networks. Label the PSNR values of those reconstructed frames. Compare the differences in visual quality and the PSNR values, and analyze the results.
- Complexity and model size analysis. For two proposed networks, create a table to compare (a) Number of parameters, and (b) Computational complexity (inference/forward pass). In the context, describe the details how these results are calculated, and analyze the results.

IV Conclusions and Future Work: Summarize your proposed methods and experimental results. Point out a few future directions to improve the work.

V References

f. Contribution of team members (in order 1, 2, 3,... or in percentage)

Note: The presentation in the report shall clearly convey your methodology and experiments. The spelling and grammar of the report are also in the grading criteria.