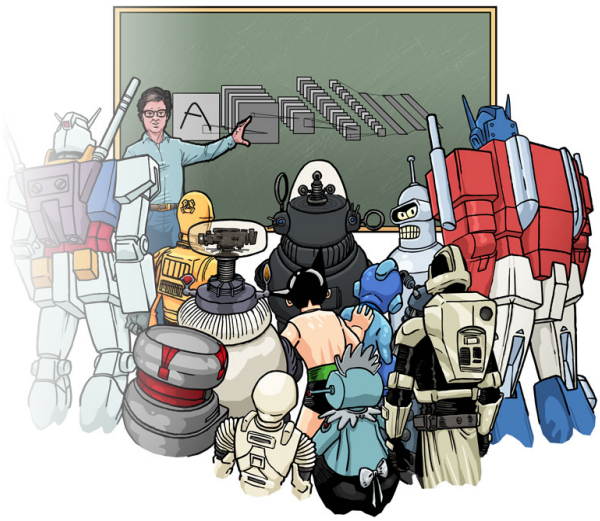

METU EE 496
Introduction to
Computational Intelligence

Training Artificial
Neural Network

Homework 1 - Training Artificial Neural Network

Due: 23:55, 09/05/2021

Late submissions are welcome, but penalized according to the following policy:

- 1 day late submission: HW will be evaluated out of 70.
- 2 days late submission: HW will be evaluated out of 50.
- 3 days late submission: HW will be evaluated out of 30.
- 4 or more days late submission: HW will not be evaluated.

You should prepare your homework by yourself alone and you should not share it with other students, otherwise you will be penalized.

Introduction

In this homework, you will perform experiments on artificial neural network (ANN) training and draw conclusions from the experimental results. You will train multi layer perceptron (MLP) and convolutional neural network (CNN) classifiers on a subset of Fashion-MNIST dataset [1]. The implementations will be in Python language and you will be using Keras module of Tensorflow package [2]. You can visit the link provided in the reference [2] to understand the usage of the module.

Figure 1: Samples from dataset. The classes are (from top to bottom) top-wear, bottom-wear, dress, footwear and bag.

Dataset Description

The dataset you will work on is a subset of Fashion-MNIST dataset [1]. It is composed of 28×28 1-band gray-scale images of 5 clothing classes which are top-wear, bottom-wear, dress, footwear and bag. Some samples from the dataset is provided in Fig. 1.

The dataset is provided in ODTUClass course page under Homework_1 folder. The dataset is split into two subsets: one for training and one for testing. For training, there are 30000 samples corresponding to 6000 samples for each class and for testing, there are 5000 samples corresponding to 1000 samples for each class.

The labels and the images are stored in separate files as NumPy arrays (**.npy*). *load* function of NumPy package can be simply used to read the dataset subsets. The labels are stored as 1-D arrays and the images are stored as 2-D arrays where each row corresponds to a flattened (vectorized) image in row-major order. Thus, one can reshape the flattened image array to 28×28 array to reconstruct the 2-D image. The i^{th} row in the image array is associated to the label which is the i^{th} entity in the label array.

Each pixel of the image is an integer between 0 and 255. The labels are integers between 0 and 4 with the semantic associations of top-wear (0), bottom-wear (1), dress (2), footwear (3) and bag (4).

Keras Module of Tensorflow

Keras module of Tensorflow contains many modules and classes for ANN design and training. In the scope of the homework, very limited subset of those modules will suffice. You can simply use sequential model class (*tf.keras.Sequential*) together with layers (*tf.keras.layers*) to create ANN models.

```
import tensorflow as tf

# example mlp classifier
model_mlp = tf.keras.Sequential([
    tf.keras.layers.Input(shape=(784, )),
    tf.keras.layers.Dense(units=128, activation='relu'),
    tf.keras.layers.Dense(units=16, use_bias=False),
    tf.keras.layers.Dense(units=5, activation='softmax')
])

# example cnn classifier
model_cnn = tf.keras.Sequential([
    tf.keras.layers.Input(shape=(28, 28, 1)),
    tf.keras.layers.Conv2D(filters=128, kernel_size=3, padding='same', activation='relu'),
    tf.keras.layers.GlobalAveragePooling2D(),
    tf.keras.layers.Dense(units=16, use_bias=False),
    tf.keras.layers.Dense(units=5, activation='softmax')
])
```

Once you create a model, you can obtain the trainable parameters from class attributes.

```
# get the trainable parameters of model_mlp as list
trainable_params = model_mlp.trainable_weights
# get the parameters 784x128 layer as numpy array
trainable_params_784x128 = trainable_params[0].numpy()
```

Once you create your classification model, you need to create a loss, an optimizer, possibly a metric also (to monitor fitting) to compile your model to be trained for the classification task.

```
# create loss: use sparse if you use integers as labels (not one-hot vectors)
loss = tf.keras.losses.SparseCategoricalCrossentropy()
# create optimizer
optimizer = tf.keras.optimizers.SGD(learning_rate=0.01)
```

```
# create metric: use sparse if you use integers as labels (not one-hot vectors)
metric = tf.keras.metrics.SparseCategoricalAccuracy()
# compile model for training
model.compile(optimizer=optimizer, loss=loss, metrics=[metric])
```

Upon compiling a *tf.keras.Model* class for your task, the training and testing operations are simple via calling the class methods. Please refer to the simple explanations of those classes in the web page of the module [2] where you can find examples as well.

Homework Task and Deliverables

The homework is composed of 5 parts. In the first part you are to answer general questions on ANN and its training. For the other parts, you will write codes to perform experiments on classification performances under several settings. You will provide the results of the experiments by some visuals and you will interpret the results by your own conclusions.

You should submit a single report in which your answers to the questions, the required experimental results (performance curve plots, visualizations etc.) and your deductions are presented for each part of the homework. Moreover, for the parts 2-4, you should append your Python codes as **text** to the end of the report for each part to generate the results and the visualizations of the experiments. Namely, all the required tasks for a part can be performed by running the related code. The codes should be well structured and **well commented**. The non-text submissions (e.g. image) or the submissions lacking comments will be simply not evaluated.

The report should be in portable document format (pdf) and named as *hw1_name_surname_eXXXXXX* where *name*, *surname* and *X*s are to be replaced by your name, surname and digits of your user ID, respectively.

1 Basic Concepts

Answer the following questions.

1.1 Which Function?

ANNs are actually parametric functions which can be used to approximate other functions. What function does an ANNs classifier trained with cross-entropy loss approximates? How is the loss defined to approximate that function? *Bonus: Why?*

1.2 Gradient Computation

You are training an ANN by using stochastic gradient descent (SGD) approach with a learning rate γ. You introduce no weight regularization to the loss and no momentum is used in the gradient updates. Under these settings, you are given weights, (w_k, w_{k+1}), at iterations k and k+1, respectively. How can you compute the gradient of the loss, $\mathcal{L}$, with respect to w at step k? Express the gradient, $\nabla_w \mathcal{L} \mid_{w=w_k}$, in terms of (γ, w_k, w_{k+1}).

1.3 Some Training Parameters and Basic Parameter Calculations

1. What are **batch** and **epoch** in the context of MLP training?
2. Given that the dataset has N samples, what is the number of batches per epoch if the batch size is B?
3. Given that the dataset has N samples, what is the number of SGD iterations if you want to train your ANN for E epochs with the batch size of B?

1.4 Computing Number of Parameters of ANN Classifiers

1. Consider an MLP classifier of K hidden units where the size of each hidden unit is H_k for k=1, . . . , K. Derive a formula to compute the number of parameters that the MLP has if the input and output dimensions are D_{in} and D_{out}, respectively.
2. Consider a CNN classifier of K convolutional layers where the spatial size of each layer is $H_k \times W_k$ and the number of convolutional filters (kernels) of each layer is C_k for k=1, . . . , K. Derive a formula to compute the number of parameters that the CNN has if the input dimension is $H_{in} \times W_{in} \times C_{in}$.

2 Experimenting ANN Architectures

In this part, you will experiment on several ANN architectures for classification task. Use **strides** of 1 for convolutions and 2 for max-pooling operations. Use *valid* **padding** for both convolutions and pooling operations. Use *adaptive moment estimation (Adam)* with default parameters for the optimizer. Use *sparse categorical accuracy* for the metric to be monitored. If your computation power allows, use batch size of 50 samples; reduce batch size accordingly otherwise. Use no weight regularization throughout the all experiments.

Preprocess the train and the test data so that the pixel values are scaled to $[-1.0, 1.0]$. Split 10% of the training data as the validation set by randomly taking equal number of samples for each class. Hence, you should have three sets: training, validation and testing.

2.1 Experimental Work

In the following, FC-N denotes fully connected (dense) layer of size N, Conv-$W \times H \times N$ denotes N many 2-D convolution filters of spatial size $W \times H$ and MaxPool-2×2 denotes max-pooling operation of spatial pool size 2×2.

The name of the ANN architectures to be experimented and their layers are:

'mlp_1' : [FC-64, ReLU] + PredictionLayer

'mlp_2' : [FC-16, ReLU, FC-64(no bias)] + PredictionLayer

'cnn_3' : [Conv-3×3×16, ReLU,
Conv-7×7×8, ReLU, MaxPool-2×2,
Conv-5×5×16, MaxPool-2×2,
GlobalAvgPool] + PredictionLayer

'cnn_3' : [Conv-3×3×16, ReLU,
Conv-5×5×8, ReLU, Conv-3×3×8, ReLU, MaxPool-2×2,
Conv-5×5×16, ReLU, MaxPool-2×2,
GlobalAvgPool] + PredictionLayer

'cnn_5' : [Conv-3×3×16, ReLU,
Conv-3×3×8, ReLU, Conv-3×3×8, ReLU, Conv-3×3×8, ReLU, MaxPool-2×2,
Conv-3×3×16, ReLU, Conv-3×3×16, ReLU, MaxPool-2×2,
GlobalAvgPool] + PredictionLayer

where PredictionLayer = [FC5, SoftMax].

Now, for each architecture, you will perform the following tasks:

1. Using training set, train the ANN for 15 epochs using *train_on_batch* method of *tf.keras.Model*. Do not use *fit* method for convenience unless you are confident of what you are doing.

   During training,

   - Record the training loss, training accuracy, validation accuracy for every 10 steps to form loss and accuracy curves (**Hint:** Training loss and training accuracy is returned by *train_on_batch* method and use *evaluate* method to compute validation accuracy);
   - Shuffle training set after each epoch (**Hint:** Use *random.permutation* of NumPy package to shuffle both images and labels in the same order).

   After training,

   - Compute test accuracy (**Hint:** Use *evaluate* method);
   - Record the weights of the first layer as numpy array (**Hint:** Call *numpy()* method of the first item in the *trainable_weights* attribute).

2. Repeat 1 for at least 10 times and

   - Take the average of the resultant loss and accuracy curves;
   - Record the best test accuracy among all the runs;
   - Record the weights of the first layer of the trained ANN that has the best test performance.

3. Now, form a dictionary object with the following key-value pairs as the result of the training experiment for the given architecture:

   - 'name': name of the architecture
   - 'loss_curve': average of the training loss curves from all runs
   - 'train_acc_curve': average of the training accuracy curves from all runs
   - 'val_acc_curve': average of the validation accuracy curves from all runs
   - 'test_acc': the best test accuracy value from all runs
   - 'weights': the weights of the first layer of the trained ANN with the best test performance

4. Save the dictionary object with the filename as the architecture name by prefixing 'part2' in the front(**Hint:** Use *pickle or json* to save dictionary objects to file and load dictionary objects from file).

Once the aforementioned tasks are performed for each architecture, create performance comparison plots by using the provided *part2Plots* function in the *utils.py* file under HW1 folder in ODTUClass course page. Note that you should pass all the dictionary objects corresponding to results of the experiments as a list to create performance comparison plots. (**Hint:** You can load previously saved results and form a list to be passed to the plot function). Add this plot to your report.

Additionally, for all architectures visualize the weights of the first layer by using the provided *visualizeWeights* function in the *utils.py* file under HW1 folder in ODTUClass course page. Add these visualizations to your report.

2.2 Discussions

Compare the architectures by considering the performances, the number of parameters, architecture structures and the weight visualizations.

1. What is the generalization performance of a classifier?
2. Which plots are informative to inspect generalization performance?
3. Compare the generalization performance of the architectures.
4. How does the number of parameters affect the classification and generalization performance?
5. How does the depth of the architecture affect the classification and generalization performance?
6. Considering the visualizations of the weights, are they interpretable?
7. Can you say whether the units are specialized to specific classes?
8. Weights of which architecture are more interpretable?
9. Considering the architectures, comment on the structures (how they are designed). Can you say that some architecture are akin to each other? Compare the performance of similarly structured architectures and architectures with different structure.
10. Which architecture would you pick for this classification task? Why?

Put your discussions together with performance plots and weight visualizations to your report.

3 Experimenting Activation Functions

In this part, you will compare rectified linear unit (ReLU) function and the logistic sigmoid function. Use SGD for the training method, constant learning rate of 0.01, 0.0 momentum (no momentum), batch size of 50 samples and use no weight regularization throughout the all experiments.

Preprocess the train and the test data so that the pixel values are scaled to $[-1.0, 1.0]$.

3.1 Experimental Work

Consider the architectures in 2.1, for each architecture create two *tf.keras.Model* objects: one with the ReLU activation function (original archs. in in 2.1) and one with the logistic sigmoid activation function (replaces ReLU of archs. in 2.1). Then, perform the following tasks for the two classifiers:

1. Using training set, train the two ANNs for 15 epochs using *train_on_batch* method of *tf.keras.Model*. Do not use *fit* method.

   During training,

   - Record the training loss and magnitude of the loss gradient with respect to the weights of the first layer at every 10 steps to form loss and gradient magnitude curves (**Hint:** Training loss is returned by *train_on_batch* method and call *numpy()* method of the first item in the *trainable_weights* attribute to obtain the copies of the weights of the first layer at time steps);

- Shuffle training set after each epoch (**Hint:** Use *random.permutation* of NumPy package to shuffle both images and labels in the same order).

After training, form a dictionary object with the following key-value pairs as the result of the training experiment for the given architecture:

- 'name': name of the architecture
- 'relu_loss_curve': the training loss curve of the ANN with ReLU
- 'sigmoid_loss_curve': the training loss curve of the ANN with logistic sigmoid
- 'relu_grad_curve': the curve of the magnitude of the loss gradient of the ANN with ReLU
- 'sigmoid_grad_curve': the curve of the magnitude of the loss gradient of the ANN with ReLU

2. Save the dictionary object with the filename as the architecture name by prefixing 'part3' in the front (**Hint:** Use *pickle or json* to save dictionary objects to file and load dictionary objects from file).

Once the aforementioned tasks are performed for each architecture, create performance comparison plots by using the provided *part3Plots* function in the *utils.py* file under HW1 folder in ODTUClass course page. Note that you should pass all the dictionary objects corresponding to results of the experiments as a list to create performance comparison plots. Add this plot to your report.

3.2 Discussions

Compare the architectures by considering the training performances:

1. How is the gradient behavior in different architectures? What happens when depth increases?
2. Why do you think that happens?
3. *Bonus:* What might happen if we do not scale the inputs to the range $[-1.0, 1.0]$?

Put your discussions together with performance plots to your report.

4 Experimenting Learning Rate

In this part, you will examine the effect of the learning rate in SGD method. Use SGD for the optimizer, constant learning rate, ReLU activation function, 0.0 momentum (no momentum), batch size of 50 samples and use no weight regularization throughout the all experiments. You will vary the initial learning rate during the experiments so that each training will be performed with different learning rate.

Preprocess the train and the test data so that the pixel values are scaled to $[-1.0, 1.0]$. Split 10% of the training data as the validation set by randomly taking equal number of samples for each class. Hence, you should have three sets: training, validation and testing.

4.1 Experimental Work

Pick your favorite architecture from 2.1, excluding 'arch_1'. Create three tf.keras.Model objects of initial learning rates 0.1, 0.01 and 0.001, respectively. Then, perform the following tasks for the three classifiers:

- Using training set, train the three ANNs for 20 epochs using *train_on_batch* method of *tf.keras.Model*. You use *fit* method if you feel comfortable with that.

  During training,

- Record the training loss and the validation accuracy for every 10 steps to form loss and accuracy curves (**Hint:** Training loss is returned by *train_on_batch* method and use *evaluate* method to compute validation accuracy);
- Shuffle training set after each epoch (**Hint:** Use *random.permutation* of NumPy package to shuffle both images and labels in the same order).

After training, form a dictionary object with the following key-value pairs as the result of the training experiment for the given architecture:

- 'name': name of the architecture
- 'loss_curve_1': the training loss curve of the ANN trained with 0.1 learning rate
- 'loss_curve_01': the training loss curve of the ANN trained with 0.01 learning rate
- 'loss_curve_001': the training loss curve of the ANN trained with 0.001 learning rate
- 'val_acc_curve_1': the validation accuracy curves of the ANN trained with 0.1 learning rate
- 'val_acc_curve_01': the validation accuracy curves of the ANN trained with 0.01 learning rate
- 'val_acc_curve_001': the validation accuracy curves of the ANN trained with 0.001 learning rate

Once the aforementioned tasks are performed for each architecture, create performance comparison plots by using the provided *part4Plots* function in the *utils.py* file under HW1 folder in ODTUClass course page. Note that you should pass a single dictionary objects corresponding to result of the experiment to create performance comparison plots. Add this plot to your report.

Now, you will try to make scheduled learning rate to improve SGD based training.

1. Examine the validation accuracy curve of the ANN trained with 0.1 learning rate. Approximately determine the epoch step where the accuracy stops increasing.
2. Create a tf.keras.Model object with the same parameters as above and with initial learning rate of 0.1.
3. Train that classifier until the epoch step that you determined in 1. Then, set the learning rate to 0.01 and continue training until 30 epochs (**Hint:** Create an optimizer with the new learning rate and **recompile** the model with the new optimizer).
4. Record only the validation accuracy during this training.
5. Now, plot the validation accuracy curve and determine the epoch step where the accuracy stops increasing.
6. Repeat 2 and 3; however, in 3, continue training with 0.01 until the epoch step that you determined in 5. Then, set the learning rate to 0.001 and continue training until 30 epochs.
7. Repeat 4 and once the training ends, record the test accuracy of the trained model and compare it to the same model trained with Adam in 2.1.

Note: You can increase the number of epochs if you are not to observe the steps where training stops improving.

4.2 Discussions

Compare the effect of learning rate by considering the training performances:

1. How does the learning rate affect the convergence speed?
2. How does the learning rate affect the convergence to a better point?
3. Does your scheduled learning rate method work? In what sense?
4. Compare the accuracy and convergence performance of your scheduled learning rate method with Adam.

Put your discussions together with performance plots to your report.

References

[1] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems." `https://www.tensorflow.org/`, 2015. Software available from tensorflow.org.