Queen Mary
University of London

May Examination Period 2024

ECS7013P Deep Learning for Audio and Music ***Duration: 2 hours (+1 for uploads)***

Answer FOUR questions

You MUST adhere to the word limits, where specified in the questions. Answer text beyond the word limit will not be marked.

This paper requires **two hours work**. There is an extra hour allowance for downloading the paper and uploading your answers.

You MUST submit your answers before the exam end time.

You must follow the online exam guidelines and instructions on the EECS exam access and submission page.

This is an open-book exam. You may use lecture notes and any module materials made available to you (online or physical). You must not use other online resources.

YOU MUST COMPLETE THE EXAM ON YOUR OWN, WITHOUT CONSULTING OTHERS.

Examiners:
Dr Johan Pauwels and Prof. Simon Dixon

Question 1

(a) Which of the following activation functions can lead to vanishing gradients?

- ReLU
- Tanh
- Leaky ReLU
- None of the above

[2 marks]

(b) You are in a ECS7013P lab session, the lab instructor gives you a machine with a built-in optimiser for neural network training. You run the machine to train your neural network and it produces the loss curve A shown in Figure 1. You see a green button on the machine and decide to press it. After doing this, you notice the loss curve B shown in Figure 1. You press the button one more time and finally notice the loss curve C shown in Figure 1.

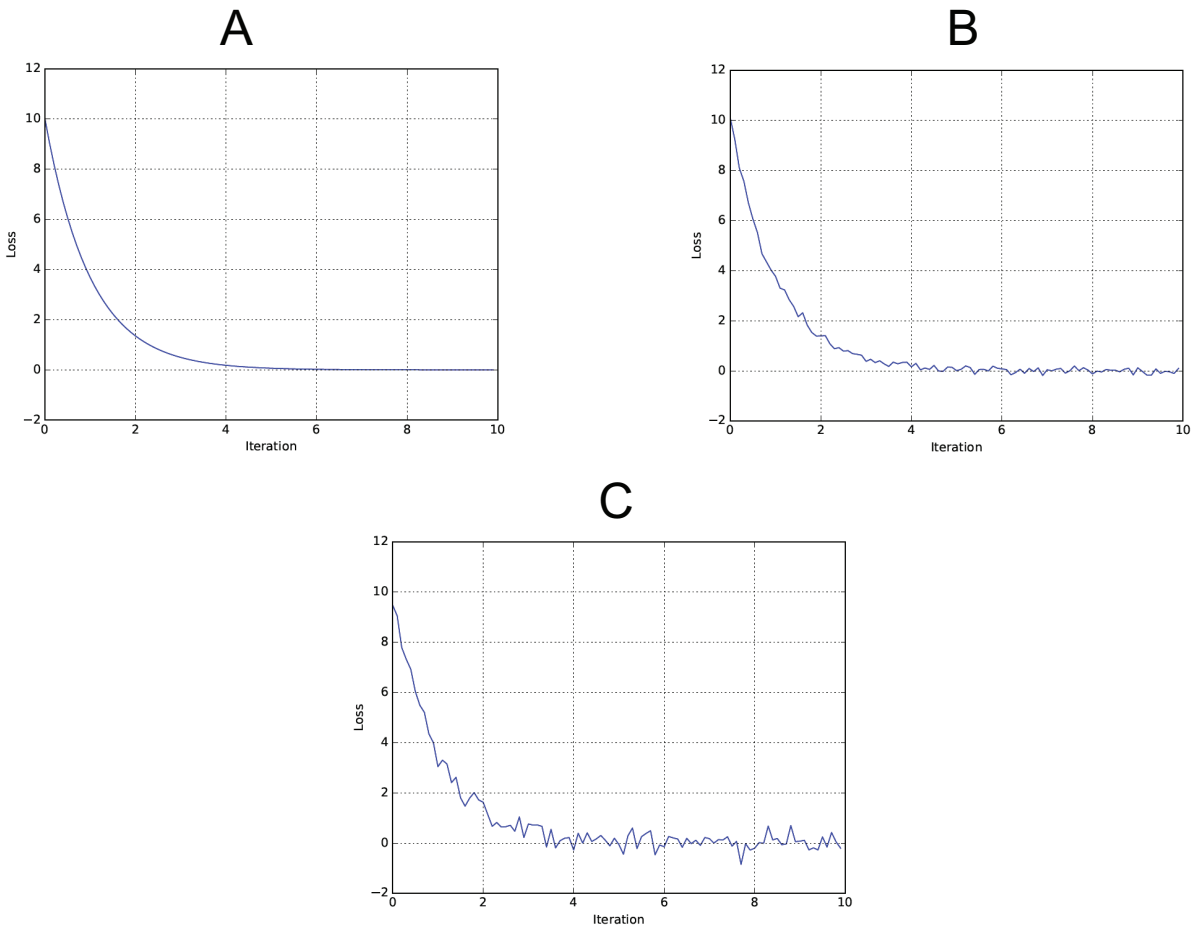


Figure 1: Loss curves produced by the built-in optimiser

(i) Knowing that the green button controls a single hyperparameter, which hyperparameter is likely to be modified by pressing the button?

(ii) Of the loss curves A, B, C in Figure 1, which corresponds to smallest magnitude of the hyperparameter?

(iii) The loss curve A in Figure 1 seems to be the most desirable. Despite this, give two reasons why you might choose the hyperparameter corresponding to the loss curve B for training your model.

[8 marks]

(c) The receptive field of any neuron (e.g. a single pixel in a convolutional feature map, or of the output node in a classifier) is the set of inputs that can affect its value. For example, in a network with a convolutional layer immediately applied to the input spectrogram, the receptive field of each pixel is equal to the shape of the convolutional filter, localised around the (time, frequency) position of that pixel. State, in words, what would be the receptive field of:

(i) A hidden node in a fully-connected network?

(ii) The output of a depthwise-separable convolutional filter, of temporal width 5 and frequency height 5, applied directly to the spectrogram?

(iii) The output of a one-dimensional convolutional filter, of temporal width 1 and frequency height 7, applied directly to the spectrogram?

(iv) The output of a recurrent layer (unidirectional) applied directly to the spectrogram?

(v) A node in the latent embedding of an auto-encoder?

[8 marks]

(d) You are solving the binary classification task of classifying bird sounds vs non-bird sounds. You design a CNN with a single output neuron. Let the output of this neuron be z. The final output of your network $\hat{y}$ is given by:

$$\hat{y} = sigmoid(ReLU(z)) \tag{1}$$

You classify all inputs with a final value $\hat{y} \geq 0.5$ as bird sounds. What problem are you going to encounter and why? What would be an appropriate way to address this problem?

[7 marks]

Question 2

You want to build a neural network to perform 10-class music genre classification. Given a music audio signal, you want to classify which of the 10 genres it belongs to.

(a) Which nonlinear activation do you use at the output layer? What does it predict?

[2 marks]

(b) What loss function do you use? Introduce the appropriate notation and write down the definition of the loss function.

[3 marks]

(c) Assuming you train your network using mini-batch gradient descent with a batch size of 64, write the formula for how you apply the loss function to calculate a scalar output for the whole batch.

[2 marks]

(d) You found on GitHub a classifier that has been trained for music vs non-music classification. You want to use transfer learning to build your own genre classification model.

(i) Explain what additional hyperparameters (due to the transfer learning) you will need to tune.

(ii) Explain why you might want to choose to freeze the early layers of the pre-trained network - i.e. why you might enforce that the weights do not change in the layers that perform the first stages of processing.

(iii) Give a reason why this freezing might in fact be unhelpful.

[12 marks]

(e) You now decide to expand the model to music audio signals belonging to multiple genres. Now, each music audio signal can have multiple genres associated to it. Therefore, you need to use multi-hot encoding for labelling. For example, a music audio signal labelled as $(1, 1, 0, \ldots, 0)$ falls under both class 1 and class 2.

To avoid extra work, you decide to retrain a new model with the same architecture and using the same output activation in Part (a) and the loss function in Part (b). Explain why this is problematic.

[6 marks]

Question 3

(a) When you train a neural network, learning rate is an important hyperparameter that needs to be tuned.

(i) What is one typical sign of a learning rate being too large? How could you tell by looking at the loss curve?

(ii) What is one typical sign of a learning rate being too small? How could you tell by looking at the loss curve?

[6 marks]

(b) Imagine a simplified WaveNet block, with 10 convolutional layers, each one having a dilated convolutional kernel of size 2, with the dilation factor doubling at every layer starting from 1. We apply this network to audio sampled at 16 kHz. What is the size (in milliseconds) of the receptive field of a node at the output of this block.

[3 marks]

(c) What is overfitting? How does splitting a dataset into train, validation, and test sets help identify overfitting? What considerations need to be taken into account when splitting the data? Explain your answer carefully.

[6 marks]

(d) We almost always apply a nonlinearity at each layer of a DNN, except perhaps at the output. Why would we not use a purely ”linear” layer within a deep network?

[2 marks]

(e) You have collected a data set to train a deep neural network for recognition of Madonna's songs. You are trying to classify music audio signals in Madonna present (1) and Madonna absent (0). Unfortunately, your data set is imbalanced. The class counts are:

Madonna present: 200 examples
Madonna absent: 2000 examples

(i) Name two data augmentation techniques you could use to help address the class imbalance problem.

(ii) Instead of data augmentation, you want to experiment with other techniques. You decide to use a modified logistic loss which is given as:

$$L = \alpha \times y \times \log \hat{y} + \beta \times (1 - y) \times \log(1 - \hat{y}) \tag{2}$$

where $y \in \mathbb{R}$ represents the groundtruth and $\hat{y} \in \mathbb{R}$ represents the network's probability prediction.
Why are α and β useful? What are reasonable values for α, β?

[8 marks]

Question 4

A Variational Auto-Encoder does not necessarily need to have a vector (1D) latent space. When all layers are convolutional or pooling layers, including the mean and standard deviation projections at the end of the encoder, the latent space will preserve the number of dimensions of the input and their interpretability (e.g. height and width or frequency and time). Suppose we have such a fully convolutional auto-encoder with the following encoder architecture:

- a 2D convolutional layer with square kernel of size 5, padding of 2 on all sides, stride 1 and ReLU activation
- a max-pooling layer with a kernel of size 2, no padding and a stride of 2
- two parallel 2D convolutional layers with square kernel of size 5, padding of 1 on all sides and stride 1, to define the mean and standard deviation of the multivariate Gaussian

All layers have bias.

(a) If the input is a matrix of dimensions $M \times M$, what are the spatial output dimensions for each of the layers.

[6 marks]

(b) What is the number of parameters in each layer if the number of channels doubles at each convolutional layer, starting from an input with 1 channel? Show your working.

[6 marks]

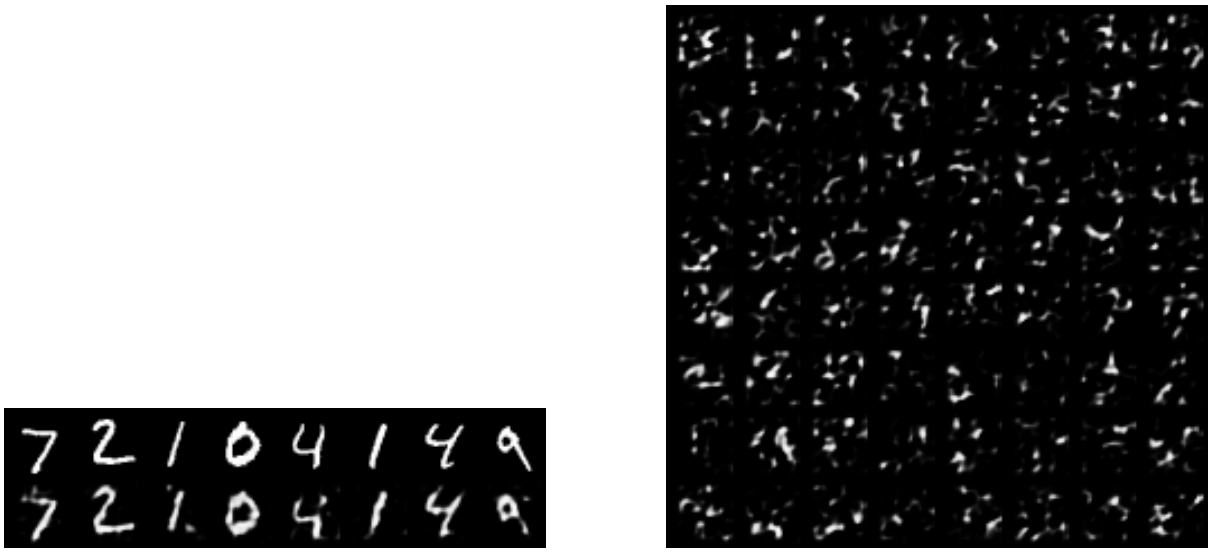
(c) If the input is a 28 by 28 greyscale image, what will be the shape of the latent space if the number of channels doubles at each convolutional layer? Explain your answer and what each dimension represents.

[7 marks]

(d) After training the auto-encoder for 20 epochs, the 8 sample input images in the top row of Figure 2a are reconstructed as shown in the bottom row. 64 random points in the latent spaces are taken and decoded in Figure 2b. What can we observe and what causes it? How can we improve the model?

[6 marks]

End of questions



(a) Reconstructed input

(b) Decoded random points in the latent space

Figure 2: VAE output