University of Glasgow

Thursday 28th April 2022
09:30-11:00 BST
(Duration: 1 hour 30 minutes
Additional time: 30 minutes
Timed exam - fixed start time)

DEGREE OF MSc

Deep Learning for MSc (M)
COMPSCI 5103

(Answer ALL 3 questions)

This examination is an open book, online assessment and is worth a total of 60 marks.

1. This code was used in the Captum lab that analyzed a model that predicted survival of individuals in the Titanic disaster:

```
class TitanicSimpleNNModel(nn.Module):
    def __init__(self):
        super().__init__()
        self.linear1 = nn.Linear(12, 12)
        self.sigmoid1 = nn.Sigmoid()
        self.linear2 = nn.Linear(12, 8)
        self.sigmoid2 = nn.Sigmoid()
        self.linear3 = nn.Linear(8, 2)
        self.softmax = nn.Softmax(dim=1)

    def forward(self, x):
        lin1_out = self.linear1(x)
        sigmoid_out1 = self.sigmoid1(lin1_out)
        sigmoid_out2 = self.sigmoid2(self.linear2(sigmoid_out1))
        return self.softmax(self.linear3(sigmoid_out2))

net = TitanicSimpleNNModel()

criterion = nn.CrossEntropyLoss()
num_epochs = 200

optimizer = torch.optim.Adam(net.parameters(), lr=0.1)
```

(a) How many inputs and outputs does this network have? [1]

(b) Calculate how many parameters are in this model. Show working. [3]

(c) This code uses Adam as an optimizer. Name and describe, in physical terms, two key innovations Adam has over plain SGD. *(Max' Word Count = 100)* [3]

(d) Draw **one** example of a loss surface with two weights being optimized that can demonstrate the advantages of both innovations given in part (c). Describe how these innovations benefit optimization on the loss surface you have drawn. [3]

(e) This lab example used the technique of IntegratedGradients. Compare and contrast the techniques of Saliency Analysis, SmoothGrad and IntegratedGradients.
(Max' Word Count = 220) [5]

(f) Comment on a coding mistake in this code and describe in detail **three different ways of changing the code** to fix it while obtaining essentially equivalent networks to each other. (You do not need to write the code out again - just what needs to be changed.) [5]

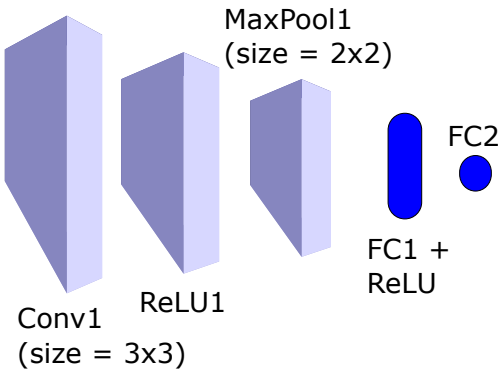

Figure 1: Example ConvNet architecture.

2. Figure 1 shows a simple architecture for a ConvNet used for image analysis.

(a) The Conv1 layer has 8 convolutional filters with stride = 1, padding = 1. MaxPool1 has stride = 2. If we put in an 18x18 colour image into this ConvNet, how many neurons are needed in the FC1 layer? Show/describe working to get full credit. [4]

(b) For scenario given in part (a), what is the size of the receptive field (in the input image) of a central value in the activation map output by MaxPool1 layer? Show your working. [3]

(c) What types of image features would be detected by the kernel shown in Figure 2a ? [1]

(d) Assume Conv1 layer consists of the single convolutional filter given in Figure 2a, with a stride of 1 with no padding. Assume MaxPool1 has a stride of 1. Write down the outputs of Conv1, ReLU1 and MaxPool1 layers when the 5x5 grayscale image show in Figure 2b is input. Show your working. [5]

(e) The FC2 layer has a single output. If this ConvNet was used for regression, for instance working out the annual profit in pounds of a company, what type of activation function should be used in FC2 and what type of loss function would be appropriate? [2]

(f) A large IT company has developed a ConvNet to predict the profits of IT companies based on pictures of its senior management. The company will use this trained network to recruit new members of senior management, while improving it by retraining on those new recruits that do increase annual profits. Discuss specific ethical risks that could arise. [5]

(a)

-1	+1	0
-1	+1	0
-1	+1	0

(b)

0	5	5	5	0
0	0	5	0	0
0	0	5	0	0
0	0	5	0	0
0	5	5	5	0

Figure 2: (a) Example convolutional kernel. (b) Example image.

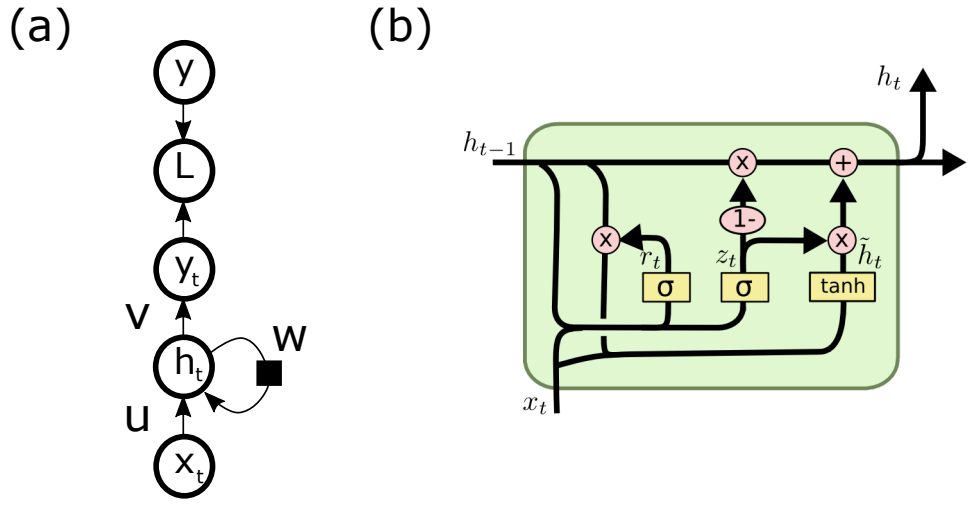

Figure 3: (a) Vanilla Recurrent Neural Network. (b) GRU Unit.

3. Figure 3(a) shows the structure of a vanilla RNN that uses scalar values with the following formulae:

$$\begin{aligned} h_t &= ReLU(wh_{t-1} + ux_t) \\ y_t &= c + vh_t \\ L &= \frac{1}{2}(y - y_t)^2 \end{aligned}$$

(a) Assume u = -2.0, v = 3.0, w = 0.5, c = -1.0 (and $h_{-1} = 0.0$). Draw a sequence of two of these recurrent units joined together. Do a 'forward pass' of the network using the sequence of input values $x_0 = -1.0$, $x_1 = +1.0$. Calculate the sequence of outputs for y_t. Show your working. [5]

(b) Determine expressions (containing variables) for $\frac{\partial y_0}{\partial x_0}$ for both when the ReLU has a zero output (input < 0.0) and for when it has a positive output (input > 0). [3]

(c) Determine expressions for $\frac{\partial y_1}{\partial x_0}$ (note different timesteps for y and x). Again consider all possible states of the ReLUs in this RNN. [3]

(d) Why might we want to calculate $\frac{\partial y_t}{\partial x_0}$ for different t ? [3]

(e) Figure 3(b) shows a GRU unit. Explain why sigmoid functions are being used in some parts of this unit while a tanh function is used in another part. [3]

(f) What is one of the major advantages of the GRU unit over the vanilla RNN and explain what part of the GRU unit makes this possible. [3]

END OF QUESTION PAPER