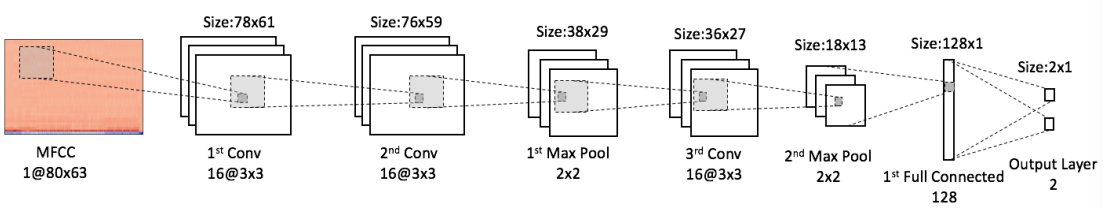
Final Exam for Machine Learning, 6/20/2020

Please show your work. That is, you need to provide the intermediate steps toward the answers. I cannot accept a reason like "the answers are directly obtained from a computer program, so I don't have the intermediate steps."
Each problem counts 15 points with a total of 105 points.

1. Compute the total number of *trainable weights* for the CNN given below. To simplify the problem, ignore the bias term. Note: 16@3×3 means 16 kernels with size of 3×3.



2. We mentioned in the lecture that the KL loss in the VAE is as follows:

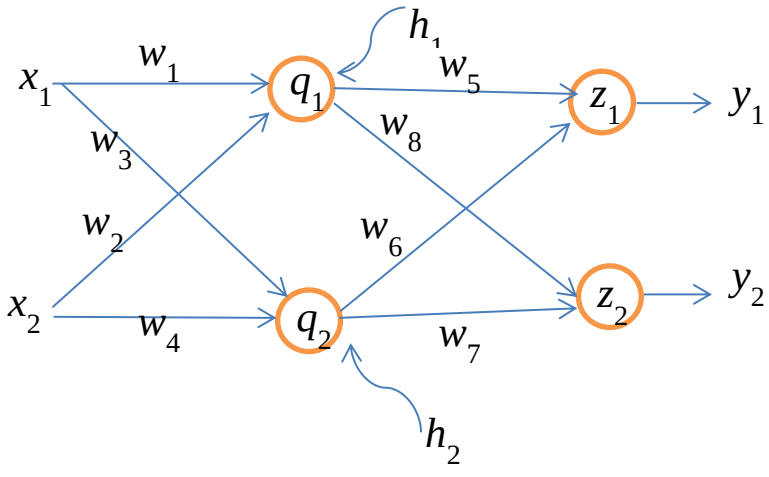
$$\frac{1}{2}\sum_{j=1}^{\text{latent dim}} \left(\mu_j^2 + \sigma_j^2 - \ln\left(\sigma_j^2\right) - 1\right)$$

Show that "each latent component is as close to N(0,1) as possible" after training is complete. Hint: Use partial derivative & this problem is actually simple.

3. For the neural network given below, the activation function from q_1 to h_1 and q_2 to h_2 is ReLU, the activation function at the output nodes are softmax, and the cost function is categorical cross entropy (KL divergence). Let $w_1 = 2.0$, w_2 to w_8 be 1.0, $d_1 = 1.0$, $d_2 = 0.0, \eta = 0.1, x_1 = 1.0,$ and $x_2 = -1.0$

   (i) Find y_1 and y_2 (forward computation).

   (ii) Find the value of $\Delta w_1 = \eta \frac{\partial J}{\partial w_1}$ and Δw_2 by using the back propagation.



4. Suppose that batch normalization is used in the hidden layer of the network in problem 3.Follow the batch normalization algorithm (on pp. 16) to answer this question. If the inputs (x_1, x_2) in a mini-batch are $(1,-1)$, $(1,1)$, and $(0.5, 0.5)$, compute the corresponding three pairs of (h_1, h_2) after batch normalization. For simplicity, let $\gamma = 1$, $\beta = 0$, and $\epsilon = 0$.

5. Follow the Q learning example in the lecture notes after Episode 2. Let $\eta = 0.1$ and $\gamma = 0.8$ in the computation.

   (i) Episode 3, step 1

   Compute the Q table with initial state *s* = room 4 and action a = 3 (i.e., going to room 1 by exploration)

   (ii) Episode 3, step 2

   The exploitation strategy is used in this step to determine the action. Which action will be taken? Then, update the Q table accordingly.

6. We mention that the procedure of an efficient subsampling convolution (on pp. 50 – 52 of the lecture notes) includes: (a) splitting the kernel into 4 sub-kernels, (b) performing convolution, and then (c) interleaving partial results. Follow the procedure to perform subsampling convolution with the following feature map and kernel. The computed output map has a size of 4×4.

Feature map

0	1	0
-2	0	-2
0	1	0

Kernel

0	1	1	0
-2	0	-2	0
0	1	0	1
1	1	1	1

7. Suppose that you are asked to predict the daily infected cases of the COVID-19 in the USA. Let the number of daily cases be represented as $x(n), 1 \leq n \leq 100$ (up to today). To solve the problem, you use the LSTM network with 10 LSTM units. The input to the network is $x(n)$ and the target output is $x(n+1)$ during training.

   (a) How to evaluate the performance (accuracy) of your predictor before you have tomorrow's value $x(101)$?

   (b) During training, you observe that your predictor has high training errors. An expert suggests that you use $x(n-1)$ and $x(n)$ as inputs to predict $x(n+1)$. Is this suggestion worth a try? Explain.

   (c) If your observe that your predictor has high errors in performance evaluation as you did in part (a), will you increase the LSTM units from 10 to 20 in your network to overcome this problem? Explain.