## Week 1 – The Basics of Conv Nets

1. What do you think applying this filter to a grayscale image will do?
2. Detect horizontal edges. (There is a high difference between the values in the top part from those in the bottom part of the matrix. When convolving this filter on a grayscale image, the horizontal edges will be detected.)
3. Suppose your input is a 300 by 300 colour (RGB) image, and you are not using a convolutional network. If the first hidden layer has 100 neurons, each one fully connected to the input, how many parameters does this hidden layer have (including the bias parameters)?
   1. 27,000,100 (The number of weights is . When you add the bias terms (one per neuron), you get .)
4. Suppose your input is a 256 by 256 grayscale image, and you use a convolutional layer with 128 filters that are each . How many parameters does this hidden layer have (including the bias parameters)?
   1. 1,280 (Since the input volume has only one channel, each filter has weights including the bias, which means the total is .)
5. You have an input volume that is , and convolve it with 32 filters of , using a stride of 2 and no padding. What is the output volume?
   1. (Using the formula , with , , and , we get 62. (the number of filters).)
6. You have an input volume that is , and pad it using “pad=1”. What is the dimension of the resulting volume (after padding)?
   1. (If the padding is 1, you add 2 to the height and width dimensions.)
7. You have an input volume that is , and convolve it with 32 filters that are each , and stride of 1. You want to use a “same” convolution. What is the padding?
   1. 3. (You need to satisfy the following equation: as you want to keep the dimensions between the input and output volumes. In order words,  
      .)
8. You have an input volume that is , and apply max pooling with a stride of 4 and a filter size of 4. What is the output volume?
   1. (Using the formula , with , , and , we get 32. .)
9. Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation.
   1. False. (Everything that influences the loss should appear in the backpropagation because we are computing derivatives. In fact, pooling layers modify the input by choosing one value out of several values in their input volume. Also, to compute derivatives for the layers that have parameters (Convolutions, Fully-Connected), we still need to backpropagate the gradient through the Pooling layers.)
10. Which of the following are the benefits of using convolutional layers?
    1. It reduces the total number of parameters, thus reducing overfitting through parameter sharing. (A convolutional layer uses parameter sharing and usually has a lot fewer parameters than a fully-connected layer.)
    2. Convolutional layers are good at capturing translation invariance.
11. The sparsity of connections and weight sharing are mechanisms that allow us to use fewer parameters in a convolutional layer making it possible to train a network with smaller training sets. True/False?
    1. True. (Weight sharing reduces significantly the number of parameters in a neural network, and sparsity of connections allows us to use a smaller number of inputs thus reducing even further the number of parameters.)

## Week 2 – Deep Convolutional Models

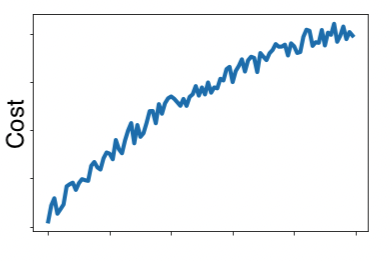
## Week 3 – Detection Algorithms

## Week 4 – Special Applications: Face Recognitions and Neural Style Transfer

1. If you have 10,000 examples, how would you split the train/dev/test set?
   1. 60% train, 20% dev and 20% test.
2. The dev and test set should:
   1. Come from the same distribution.
3. If your neural network model seems to have high variance, what of the following would be promising things to try?
   1. Get more training data
   2. Add regularisation
4. You are working on an automated check-out kiosk for a supermarket, and are building a classifier for apples, bananas and oranges. Suppose your classifier obtains a training set error of 0.5%, and a dev set error of 7%. Which of the following are promising things to try to improve your classifier?
   1. Increase the regularisation parameter lambda.
   2. Get more training data.
5. In every case it is a good practice to use dropout when training a deep neural network because it can help to prevent overfitting. True/False?
   1. False. (In most cases, it is recommended to not use dropout if there is no overfit. Although in computer vision, due to the nature of the data, it is the default practice.)
6. To reduce high variance, the regularisation hyperparameter lambda must be increased. True/False?
   1. True. (By increasing the regularization parameter the magnitude of the weight parameters is reduced. This helps reduce the variance.)
7. Which of the following are true about dropout?
   1. It helps to reduce the variance of a model. (Because it’s a regularisation technique.)
   2. In practice, it eliminates units of each layer with a probability of . (And hence helps to reduce model overfitting.)
8. Decreasing the parameter from (say) to will likely cause the following:
   1. Increasing the regularisation effect. (This will make dropout have a higher probability of eliminating a node in the neural network, increasing the regularisation effect.)
9. Which of the following actions increase the regularisation of a model?
   1. Increase the value of the hyperparameter lambda. (When increasing the hyperparameter lambda, we increase the effect of the L2 penalisation.)
   2. Make use of data augmentation. (Data augmentation is a way of generating “new” data at a relatively low cost, which can therefore help to reduce the variance.)
10. Suppose that a model uses, as one feature, the total number of kilometres walked by a person during a year, and another feature is the height of the person in meters. What is the most likely effect of normalisation of the input data?
    1. It will make model training faster. (Since the difference between the ranges of the features is very different, this will likely cause the process of gradient descent to oscillate, making the optimization process longer.)

## Week 2 – Optimisation Algorithms

1. Using the notation for mini-batch gradient descent, to what does correspond?
   1. The activation of the second layer when the input is the third example of the fourth mini-batch. (In general denotes the activation of the layer when the input is the example from the mini-batch .)
2. Which of these statements about mini-batch gradient descent do you agree with?
   1. When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to batch gradient descent. (Batch gradient descent uses all the examples at each iteration, this is equivalent to having only one mini-batch of the size of the complete training set in mini-batch gradient descent.)
3. Which of the following is true about batch gradient descent?
   1. It is the same as mini-batch gradient descent when the mini-batch size is the same as the size of the training set. (When using batch gradient descent, there is only one mini-batch and hence it is equivalent to batch gradient descent.)
4. While using mini-batch gradient descent with a batch size larger than 1 but less than , the plot of the cost function looks like this:

Which of the following do you agree with?

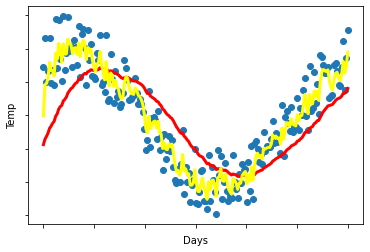
* 1. Something is wrong irrespective of whether mini-batch or batch gradient descent is being used. (The cost is larger than when the process started. This is not right at all.)

1. Suppose the temperature in Casablanca over the first two days of March are the following:  
   March 1st:   
   March 2nd:

Say you use an exponentially weighted average with to track the temperature , . If is the value computed after day 2 without bias correction, and is the value you compute with bias correction, what are these values?

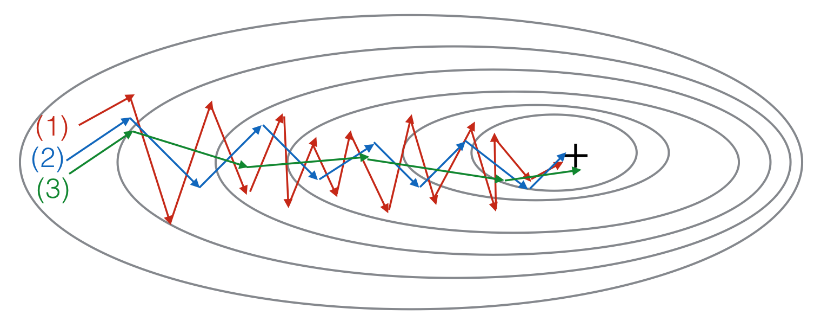
* 1. and .  
     (  
        
      )

1. Which of these is NOT a good learning rate decay scheme? Here, is the epoch number.:
2. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: . The yellow and red lines were computed using values ​ and ​ respectively. Which of the following are true?



* 1. (Since the yellow curve is noisier.)

1. Consider this figure:

 These plots were generated with gradient descent; with gradient descent with momentum (); and gradient descent with momentum (). Which curve corresponds to which algorithm?

* 1. (1) is gradient descent. (2) is gradient descent with momentum (small ). (3) is gradient descent with moment (large )

1. Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function . Which of the following techniques could help find parameter values that attain a small value for ?
   1. Try mini-batch gradient descent. (Mini-batch gradient descent is faster than batch gradient descent.)
   2. Try using Adam. (Adam combines the advantages of other methods to accelerate the convergence of gradient descent.)
   3. Normalize the input data.
2. Which of the following are true about Adam?
   1. Adam combines the advantages of RMSProp and momentum. (Which is why we have two parameters and in addition to .

## Week 3 – Hyperparameter Tuning, Batch Normalisation and Programming Frameworks

1. With a relatively small set of hyperparameters, it is ok to use a grid search. True/False?
   1. True. (When the set of hyperparameters is small, like a range for , grid search works fine.)
2. Every hyperparameter, if set poorly, can have a huge negative impact on training, and so all hyperparameters are about equally important to tune well. True/False?
   1. False. (Some hyperparameters, such as the learning rate, are more critical than others.)
3. During hyperparameter search, whether you try to babysit one model (“Panda strategy”) or train a lot of models in parallel (“Caviar”) is largely determined by:
   1. The amount of computational power you can access.
4. Knowing at the hyperparameter should be in the range of and , which of the following is the recommended way to sample a value for ?
5. Once good values of hyperparameters have been found, those values should be changed if new data is added or a change in computational power occurs. True/False?
   1. True. (The choice of some hyperparameters such as the batch size depends on conditions such as hardware and quantity of data.)
6. In batch normalization as presented in the videos, if you apply it on the th layer of your neural network, what are you normalizing?
7. In the normalisation formula , why do we use epsilon?
   1. To avoid division by zero.
8. Which of the following are true about batch normalisation?
   1. When using batch normalisation we introduce two new parameters, , that must be “learned” or trained. (Batch normalisation uses and to compute
   2. The parameters and set the mean and variance of . (When applying the linear transformation , we set the mean and variance of .
9. A neural network is trained with Batch Norm. At test time, to evaluate the neural network we turn off the Batch Norm to avoid random predictions from the network. True/False?
   1. False. (During the test, the parameters and are estimated using an exponentially weighted average across mini-batches used during training.)
10. Which of the following are some recommended criteria to choose a deep learning framework?
    1. Running speed. (The running speed is a major factor, especially when working with large datasets.)