Coursera 딥러닝 퀴즈 정리

★ C1 - Neural Networks and Deep Learning

## Week 1 Quiz - Introduction to deep learning

1. What does the analogy "AI is the new electricity" refer to?

- [ ] AI is powering personal devices in our homes and offices, similar to electricity.
- [ ] Through the "smart grid", AI is delivering a new wave of electricity.
- [ ] AI runs on computers and is thus powered by electricity, but it is letting computers do things not possible before.
- [x] Similar to electricity starting about 100 years ago, AI is transforming multiple industries.

Note: Andrew illustrated the same idea in the lecture.

2. Which of these are reasons for Deep Learning recently taking off? (Check the two options that apply.)

- [x] We have access to a lot more computational power.
- [ ] Neural Networks are a brand new field.
- [x] We have access to a lot more data.
- [x] Deep learning has resulted in significant improvements in important applications such as online advertising, speech recognition, and image recognition.

3. Recall this diagram of iterating over different ML ideas. Which of the statements below are true? (Check all that apply.)

- [x] Being able to try out ideas quickly allows deep learning engineers to iterate more quickly.
- [x] Faster computation can help speed up how long a team takes to iterate to a good idea.
- [ ] It is faster to train on a big dataset than a small dataset.
- [x] Recent progress in deep learning algorithms has allowed us to train good models faster (even without changing the CPU/GPU hardware).

Note: A bigger dataset generally requires more time to train on a same model.

4. When an experienced deep learning engineer works on a new problem, they can

usually use insight from previous problems to train a good model on the first try, without needing to iterate multiple times through different models. True/False?

- [ ] True
- [x] False

Note: Maybe some experience may help, but nobody can always find the best model or hyperparameters without iterations.

5. Which one of these plots represents a ReLU activation function?

- Check [relu](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)).

6. Images for cat recognition is an example of "structured" data, because it is represented as a structured array in a computer. True/False?

- [ ] True
- [x] False

7. A demographic dataset with statistics on different cities' population, GDP per capita, economic growth is an example of "unstructured" data because it contains data coming from different sources. True/False?

- [ ] True
- [x] False

8. Why is an RNN (Recurrent Neural Network) used for machine translation, say translating English to French? (Check all that apply.)

- [x] It can be trained as a supervised learning problem.
- [ ] It is strictly more powerful than a Convolutional Neural Network (CNN).
- [x] It is applicable when the input/output is a sequence (e.g., a sequence of words).
- [ ] RNNs represent the recurrent process of Idea->Code->Experiment->Idea->....

9. In this diagram which we hand-drew in lecture, what do the horizontal axis (x-axis) and vertical axis (y-axis) represent?

- x-axis is the amount of data
- y-axis (vertical axis) is the performance of the algorithm.

10. Assuming the trends described in the previous question's figure are accurate (and hoping you got the axis labels right), which of the following are true? (Check all that apply.)

- [x] Increasing the training set size generally does not hurt an algorithm’s performance, and it may help significantly.
- [x] Increasing the size of a neural network generally does not hurt an algorithm’s performance, and it may help significantly.
- [ ] Decreasing the training set size generally does not hurt an algorithm’s performance, and it may help significantly.
- [ ] Decreasing the size of a neural network generally does not hurt an algorithm’s performance, and it may help significantly.

Week 2 Quiz - Neural Network Basics

1. What does a neuron compute?

- [ ] A neuron computes an activation function followed by a linear function (z = Wx + b)

- [x] A neuron computes a linear function (z = Wx + b) followed by an activation function

- [ ] A neuron computes a function g that scales the input x linearly (Wx + b)

- [ ] A neuron computes the mean of all features before applying the output to an activation function

Note: we generally say that the output of a neuron is a = g(Wx + b) where g is the activation function (sigmoid, tanh, ReLU, ...).

2. Which of these is the "Logistic Loss"?

- Check [here](https://en.wikipedia.org/wiki/Cross_entropy#Cross-entropy_error_function_and_logistic_regression).

Note: this is the logistic loss you've seen in lecture!

3. Suppose img is a (32,32,3) array, representing a 32x32 image with 3 color channels

red, green and blue. How do you reshape this into a column vector?

- `x = img.reshape((32 * 32 * 3, 1))`

4. Consider the two following random arrays "a" and "b":

```
a = np.random.randn(2, 3) # a.shape = (2, 3)
b = np.random.randn(2, 1) # b.shape = (2, 1)
c = a + b
```

What will be the shape of "c"?

b (column vector) is copied 3 times so that it can be summed to each column of a. Therefore, `c.shape = (2, 3)`.

5. Consider the two following random arrays "a" and "b":

```
a = np.random.randn(4, 3) # a.shape = (4, 3)
b = np.random.randn(3, 2) # b.shape = (3, 2)
c = a * b
```

What will be the shape of "c"?

"*" operator indicates element-wise multiplication. Element-wise multiplication requires same dimension between two matrices. It's going to be an error.

6. Suppose you have n_x input features per example. Recall that X=[x^(1), x^(2)...x^(m)]. What is the dimension of X?

`(n_x, m)`

7. Recall that `np.dot(a,b)` performs a matrix multiplication on a and b, whereas `a*b` performs an element-wise multiplication.

Consider the two following random arrays "a" and "b":

```
a = np.random.randn(12288, 150) # a.shape = (12288, 150)
b = np.random.randn(150, 45) # b.shape = (150, 45)
c = np.dot(a, b)
```

What is the shape of c?

`c.shape = (12288, 45)`, this is a simple matrix multiplication example.

8. Consider the following code snippet:

```
# a.shape = (3,4)
# b.shape = (4,1)
for i in range(3):
  for j in range(4):
    c[i][j] = a[i][j] + b[j]
```

How do you vectorize this?

`c = a + b.T`

9. Consider the following code:

```
a = np.random.randn(3, 3)
b = np.random.randn(3, 1)
c = a * b
```

What will be c?

This will invoke broadcasting, so b is copied three times to become (3,3), and * is an element-wise product so `c.shape = (3, 3)`.

10. Consider the following computation graph.

```
```

```
J = u + v - w
  = a * b + a * c - (b + c)
  = a * (b + c) - (b + c)
  = (a - 1) * (b + c)
```

Answer: `(a - 1) * (b + c)`

Week 3 Quiz - Shallow Neural Networks

1. Which of the following are true? (Check all that apply.) **Notice that I only list correct options.**

- X is a matrix in which each column is one training example.
- a^[2]_4 is the activation output by the 4th neuron of the 2nd layer
- a^\[2\](12) denotes the activation vector of the 2nd layer for the 12th training example.
- a^[2] denotes the activation vector of the 2nd layer.

2. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?

- [x] True
- [ ] False

Note: You can check [this post](https://stats.stackexchange.com/a/101563/169377) and (this paper)[http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf].

> As seen in lecture the output of the tanh is between -1 and 1, it thus centers the data which makes the learning simpler for the next layer.

3. Which of these is a correct vectorized implementation of forward propagation for layer l, where 1≤l≤L?

- Z^[l]=W^[l]A^[l−1]+b^[l]
- A^[l]=g^\[l\](Z^[l])

4. You are building a binary classifier for recognizing cucumbers (y=1) vs. watermelons (y=0). Which one of these activation functions would you recommend using for the output layer?

- [ ] ReLU
- [ ] Leaky ReLU
- [x] sigmoid
- [ ] tanh

Note: The output value from a sigmoid function can be easily understood as a probability.

> Sigmoid outputs a value between 0 and 1 which makes it a very good choice for binary classification. You can classify as 0 if the output is less than 0.5 and classify as 1 if the output is more than 0.5. It can be done with tanh as well but it is less convenient as the output is between -1 and 1.

5. Consider the following code:

```
A = np.random.randn(4,3)
B = np.sum(A, axis = 1, keepdims = True)
```

What will be B.shape?

`B.shape = (4, 1)`

> we use (keepdims = True) to make sure that A.shape is (4,1) and not (4,). It makes our code more rigorous.

6. Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements are True? (Check all that apply)

- [x] Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent each neuron in the layer will be computing the same thing as other neurons.
- [ ] Each neuron in the first hidden layer will perform the same computation in the first iteration. But after one iteration of gradient descent they will learn to compute different things because we have “broken symmetry”.
- [ ] Each neuron in the first hidden layer will compute the same thing, but neurons in different layers will compute different things, thus we have accomplished “symmetry breaking” as described in lecture.
- [ ] The first hidden layer’s neurons will perform different computations from

each other even in the first iteration; their parameters will thus keep evolving in their own way.

7. Logistic regression’s weights w should be initialized randomly rather than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to “break symmetry”, True/False?

- [ ] True
- [x] False

> Logistic Regression doesn't have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the Logistic Regression depend on the input x (because there's no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector.

8. You have built a network using the tanh activation for all the hidden units. You initialize the weights to relative large values, using np.random.randn(..,..)*1000. What will happen?

- [ ] It doesn’t matter. So long as you initialize the weights randomly gradient descent is not affected by whether the weights are large or small.

- [ ] This will cause the inputs of the tanh to also be very large, thus causing gradients to also become large. You therefore have to set α to be very small to prevent divergence; this will slow down learning.

- [ ] This will cause the inputs of the tanh to also be very large, causing the units to be “highly activated” and thus speed up learning compared to if the weights had to start from small values.

- [x] This will cause the inputs of the tanh to also be very large, thus causing gradients to be close to zero. The optimization algorithm will thus become slow.

> tanh becomes flat for large values, this leads its gradient to be close to zero. This slows down the optimization algorithm.

9. Consider the following 1 hidden layer neural network:

- b[1] will have shape (4, 1)

- W[1] will have shape (4, 2)
- W[2] will have shape (1, 4)
- b[2] will have shape (1, 1)

Note: Check [here](https://user-images.githubusercontent.com/14886380/29200515-7fdd1548-7e88-11e7-9d05-0878fe96bcfa.png) for general formulas to do this.

10. In the same network as the previous question, what are the dimensions of Z^[1] and A^[1]?

- Z[1] and A[1] are (4,m)

Note: Check [here](https://user-images.githubusercontent.com/14886380/29200515-7fdd1548-7e88-11e7-9d05-0878fe96bcfa.png) for general formulas to do this.

Week 4 Quiz - Key concepts on Deep Neural Networks

1. What is the "cache" used for in our implementation of forward propagation and backward propagation?

- [ ] It is used to cache the intermediate values of the cost function during training.
- [x] We use it to pass variables computed during forward propagation to the corresponding backward propagation step. It contains useful values for backward propagation to compute derivatives.
- [ ] It is used to keep track of the hyperparameters that we are searching over, to speed up computation.
- [ ] We use it to pass variables computed during backward propagation to the corresponding forward propagation step. It contains useful values for forward propagation to compute activations.

> the "cache" records values from the forward propagation units and sends it to the backward propagation units because it is needed to compute the chain rule derivatives.

2. Among the following, which ones are "hyperparameters"? (Check all that apply.) **I only list correct options.**

- size of the hidden layers n^[l]

- learning rate α
- number of iterations
- number of layers L in the neural network

Note: You can check [this Quora post](https://www.quora.com/What-are-hyperparameters-in-machine-learning) or [this blog post](http://colinraffel.com/wiki/neural_network_hyperparameters).

3. Which of the following statements is true?

- [x] The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.
Correct
- [ ] The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.

Note: You can check the lecture videos. I think Andrew used a CNN example to explain this.

4. Vectorization allows you to compute forward propagation in an L-layer neural network without an explicit for-loop (or any other explicit iterative loop) over the layers l=1, 2, ⋯,L. True/False?

- [ ] True
- [x] False

Note: We cannot avoid the for-loop iteration over the computations among layers.

5. Assume we store the values for n^[l] in an array called layers, as follows: layer_dims = [n_x, 4,3,2,1]. So layer 1 has four hidden units, layer 2 has 3 hidden units and so on. Which of the following for-loops will allow you to initialize the parameters for the model?

```
for(i in range(1, len(layer_dims))):
    parameter[‘W’ + str(i)] = np.random.randn(layers[i], layers[i - 1])) * 0.01
    parameter[‘b’ + str(i)] = np.random.randn(layers[i], 1) * 0.01
```

6. Consider the following neural network.

- The number of layers L is 4. The number of hidden layers is 3.

Note: The input layer (L^[0]) does not count.

> As seen in lecture, the number of layers is counted as the number of hidden layers + 1. The input and output layers are not counted as hidden layers.

7. During forward propagation, in the forward function for a layer l you need to know what is the activation function in a layer (Sigmoid, tanh, ReLU, etc.). During backpropagation, the corresponding backward function also needs to know what is the activation function for layer l, since the gradient depends on it. True/False?

- [x] True
- [ ] False

> During backpropagation you need to know which activation was used in the forward propagation to be able to compute the correct derivative.

8. There are certain functions with the following properties:

(i) To compute the function using a shallow network circuit, you will need a large network (where we measure size by the number of logic gates in the network), but (ii) To compute it using a deep network circuit, you need only an exponentially smaller network. True/False?

- [x] True
- [ ] False

Note: See lectures, exactly same idea was explained.

9. Consider the following 2 hidden layer neural network:

Which of the following statements are True? (Check all that apply).

- W^[1] will have shape (4, 4)
- b^[1] will have shape (4, 1)
- W^[2] will have shape (3, 4)
- b^[2] will have shape (3, 1)
- b^[3] will have shape (1, 1)
- W^[3] will have shape (1, 3)

Note: See [this image](https://user-images.githubusercontent.com/14886380/29200515-7fdd1548-7e88-11e7-9d05-0878fe96bcfa.png) for general formulas.

10. Whereas the previous question used a specific network, in the general case what is the dimension of W^[l], the weight matrix associated with layer l?

- W^[l] has shape (n^[l],n^[l−1])

Note: See [this image](https://user-images.githubusercontent.com/14886380/29200515-7fdd1548-7e88-11e7-9d05-0878fe96bcfa.png) for general formulas.

★ C2 - Improving Deep Neural Networks Hyperparameter tuning, Regularization and Optimization

## Week 1 Quiz - Practical aspects of deep learning

1. If you have 10,000,000 examples, how would you split the train/dev/test set?

- 98% train . 1% dev . 1% test

2. The dev and test set should:

- 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?

- Add regularization
- Get more training data

Note: Check [here](https://user-images.githubusercontent.com/14886380/29240263-f7c517ca-7f93-11e7-8549-58856e0ed12f.png).

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? (Check all that apply.)

- Increase the regularization parameter lambda
- Get more training data

Note: Check [here](https://user-images.githubusercontent.com/14886380/29240263-f7c517ca-7f93-11e7-8549-58856e0ed12f.png).

5. What is weight decay?

- A regularization technique (such as L2 regularization) that results in gradient descent shrinking the weights on every iteration.

6. What happens when you increase the regularization hyperparameter lambda?

- Weights are pushed toward becoming smaller (closer to 0)

7. With the inverted dropout technique, at test time:

- You do not apply dropout (do not randomly eliminate units) and do not keep the 1/keep_prob factor in the calculations used in training

8. Increasing the parameter keep_prob from (say) 0.5 to 0.6 will likely cause the following: (Check the two that apply)

- Reducing the regularization effect
- Causing the neural network to end up with a lower training set error

9. Which of these techniques are useful for reducing variance (reducing overfitting)? (Check all that apply.)

- Dropout
- L2 regularization
- Data augmentation

10. Why do we normalize the inputs x?

- It makes the cost function faster to optimize

Week 2 Quiz - Optimization algorithms

1. Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch?

    - a^\[3]\{8}\(7)

    Note: **[i]{j}(k)** superscript means **i-th layer**, **j-th minibatch**, **k-th example**

2. Which of these statements about mini-batch gradient descent do you agree with?

    - [ ] You should implement mini-batch gradient descent without an explicit for-loop over different mini-batches, so that the algorithm processes all mini-batches at the same time (vectorization).
    - [ ] Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.
    - [x] One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.

    Note: Vectorization is not for computing several mini-batches in the same time.

3. Why is the best mini-batch size usually not 1 and not m, but instead something in-between?

    - If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.
    - If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.

4. Suppose your learning algorithm's cost ***J***, plotted as a function of the number of iterations, looks like this:

    - If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.

    Note: There will be some oscillations when you're using mini-batch gradient descent since there could be some noisy data example in batches. However batch gradient descent always guarantees a lower ***J*** before reaching the optimal.

5. Suppose the temperature in Casablanca over the first three days of January are

the same:

Jan 1st: θ_1 = 10

Jan 2nd: θ_2 * 10

Say you use an exponentially weighted average with β = 0.5 to track the temperature: v_0 = 0, v_t = βv_t−1 + (1 − β)θ_t. If v_2 is the value computed after day 2 without bias correction, and v^corrected_2 is the value you compute with bias correction. What are these values?

- v_2 = 7.5, v^corrected_2 = 10

6. Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.

- α = e^t * α_0

Note: This will explode the learning rate rather than decay it.

7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: v_t = βv_t−1 + (1 − β)θ_t. The red line below was computed using β = 0.9. What would happen to your red curve as you vary β? (Check the two that apply)

- Increasing β will shift the red line slightly to the right.
- Decreasing β will create more oscillation within the red line.

8. Consider this figure:

These plots were generated with gradient descent; with gradient descent with momentum (β = 0.5) and gradient descent with momentum (β = 0.9). Which curve corresponds to which algorithm?

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

9. 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 J(W[1],b[1],...,W[L],b[L]). Which of the following techniques could help find parameter values that attain a small value forJ? (Check all that apply)

- [x] Try using Adam
- [x] Try better random initialization for the weights
- [x] Try tuning the learning rate α
- [x] Try mini-batch gradient descent
- [ ] Try initializing all the weights to zero

10. Which of the following statements about Adam is False?

- Adam should be used with batch gradient computations, not with mini-batches.

Note: Adam could be used with both.

Week 3 Quiz - Hyperparameter tuning, Batch Normalization, Programming Frameworks

1. If searching among a large number of hyperparameters, you should try values in a grid rather than random values, so that you can carry out the search more systematically and not rely on chance. True or False?

- [x] False
- [ ] True

Note: Try random values, don't do grid search. Because you don't know which hyperparamerters are more important than others.

> And to take an extreme example, let's say that hyperparameter two was that value epsilon that you have in the denominator of the Adam algorithm. So your choice of alpha matters a lot and your choice of epsilon hardly matters.

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 or False?

- [x] False
- [ ] True

> We've seen in lecture that 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:

- [ ] Whether you use batch or mini-batch optimization
- [ ] The presence of local minima (and saddle points) in your neural network
- [x] The amount of computational power you can access
- [ ] The number of hyperparameters you have to tune

4. If you think β (hyperparameter for momentum) is between on 0.9 and 0.99, which of the following is the recommended way to sample a value for beta?

```
r = np.random.rand()
beta = 1 - 10 ** (-r - 1)
```

5. Finding good hyperparameter values is very time-consuming. So typically you should do it once at the start of the project, and try to find very good hyperparameters so that you don't ever have to revisit tuning them again. True or false?

- [x] False
- [ ] True

Note: Minor changes in your model could potentially need you to find good hyperparameters again from scratch.

6. In batch normalization as presented in the videos, if you apply it on the lth layer of your neural network, what are you normalizing?

- z^[l]

7. In the normalization formula, why do we use epsilon?

- To avoid division by zero

8. Which of the following statements about γ and β in Batch Norm are true? **Only correct options listed**

- They can be learned using Adam, Gradient descent with momentum, or RMSprop, not just with gradient descent.
- They set the mean and variance of the linear variable z^[l] of a given layer.

9. After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should:

- Perform the needed normalizations, use μ and σ^2 estimated using an exponentially weighted average across mini-batches seen during training.

10. Which of these statements about deep learning programming frameworks are true? (Check all that apply)

- [x] A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower-level language such as Python.
- [x] Even if a project is currently open source, good governance of the project helps ensure that the it remains open even in the long term, rather than become closed or modified to benefit only one company.
- [ ] Deep learning programming frameworks require cloud-based machines to run.

★ C3 - Structuring Machine Learning Projects

## Week 1 Quiz - Bird recognition in the city of Peacetopia (case study)

1. Having three evaluation metrics makes it harder for you to quickly choose between two different algorithms, and will slow down the speed with which your team can iterate. True/False?

- [x] True
- [ ] False

2. If you had the three following models, which one would you choose?

- Test Accuracy 98%
- Runtime 9 sec
- Memory size 9MB

3. Based on the city's requests, which of the following would you say is true?

- [x] Accuracy is an optimizing metric; running time and memory size are a satisficing metrics.
- [ ] Accuracy is a satisficing metric; running time and memory size are an optimizing metric.
- [ ] Accuracy, running time and memory size are all optimizing metrics because you want to do well on all three.

- [ ] Accuracy, running time and memory size are all satisficing metrics because you have to do sufficiently well on all three for your system to be acceptable.

4. Before implementing your algorithm, you need to split your data into train/dev/test sets. Which of these do you think is the best choice?

- Train 9,500,000
- Dev 250,000
- Test 250,000

5. After setting up your train/dev/test sets, the City Council comes across another 1,000,000 images, called the "citizens' data". Apparently the citizens of Peacetopia are so scared of birds that they volunteered to take pictures of the sky and label them, thus contributing these additional 1,000,000 images. These images are different from the distribution of images the City Council had originally given you, but you think it could help your algorithm.

You should not add the citizens' data to the training set, because this will cause the training and dev/test set distributions to become different, thus hurting dev and test set performance. True/False?

- [ ] True
- [x] False

```
Note: Adding this data to the training set will change the training set distribution. However, it is not a problem to have different training and dev distribution. On the contrary, it would be very problematic to have different dev and test set distributions.
```

6. One member of the City Council knows a little about machine learning, and thinks you should add the 1,000,000 citizens' data images to the test set. You object because:

- The test set no longer reflects the distribution of data (security cameras) you most care about.
- This would cause the dev and test set distributions to become different. This is a bad idea because you're not aiming where you want to hit.

7. You train a system, and its errors are as follows (error = 100%-Accuracy):

- Training set error 4.0%
- Dev set error 4.5%

This suggests that one good avenue for improving performance is to train a bigger network so as to drive down the 4.0% training error. Do you agree?

- No, because there is insufficient information to tell.

8. You ask a few people to label the dataset so as to find out what is human-level performance. You find the following levels of accuracy:

- Bird watching expert #1	0.3% error
- Bird watching expert #2	0.5% error
- Normal person #1 (not a bird watching expert)	1.0% error
- Normal person #2 (not a bird watching expert)	1.2% error

If your goal is to have “human-level performance” be a proxy (or estimate) for Bayes error, how would you define “human-level performance”?

- 0.3% (accuracy of expert #1)

9. Which of the following statements do you agree with?

- A learning algorithm’s performance can be better human-level performance but it can never be better than Bayes error.

10. You find that a team of ornithologists debating and discussing an image gets an even better 0.1% performance, so you define that as “human-level performance.” After working further on your algorithm, you end up with the following:

- Human-level performance	0.1%
- Training set error	2.0%
- Dev set error	2.1%

Based on the evidence you have, which two of the following four options seem the most promising to try? (Check two options.)

- Try decreasing regularization.
- Train a bigger model to try to do better on the training set.

11. You also evaluate your model on the test set, and find the following:

- Human-level performance 0.1%

- Training set error 2.0%
- Dev set error 2.1%
- Test set error 7.0%

What does this mean? (Check the two best options.)

- You should try to get a bigger dev set.
- You have overfit to the dev set.

12. After working on this project for a year, you finally achieve:

- Human-level performance 0.10%
- Training set error 0.05%
- Dev set error 0.05%

What can you conclude? (Check all that apply.)

- It is now harder to measure avoidable bias, thus progress will be slower going forward.
- If the test set is big enough for the 0,05% error estimate to be accurate, this implies Bayes error is ≤0.05

13. It turns out Peacetopia has hired one of your competitors to build a system as well. Your system and your competitor both deliver systems with about the same running time and memory size. However, your system has higher accuracy! However, when Peacetopia tries out your and your competitor’s systems, they conclude they actually like your competitor’s system better, because even though you have higher overall accuracy, you have more false negatives (failing to raise an alarm when a bird is in the air). What should you do?

- Rethink the appropriate metric for this task, and ask your team to tune to the new metric.

14. You’ve handily beaten your competitor, and your system is now deployed in Peacetopia and is protecting the citizens from birds! But over the last few months, a new species of bird has been slowly migrating into the area, so the performance of your system slowly degrades because your data is being tested on a new type of data.

- Use the data you have to define a new evaluation metric (using a new dev/test set) taking into account the new species, and use that to drive further progress for your team.

15. The City Council thinks that having more Cats in the city would help scare off birds. They are so happy with your work on the Bird detector that they also hire you to build a Cat detector. (Wow Cat detectors are just incredibly useful aren't they.) Because of years of working on Cat detectors, you have such a huge dataset of 100,000,000 cat images that training on this data takes about two weeks. Which of the statements do you agree with? (Check all that agree.)

- If 100,000,000 examples is enough to build a good enough Cat detector, you might be better of training with just 10,000,000 examples to gain a ≈10x improvement in how quickly you can run experiments, even if each model performs a bit worse because it's trained on less data.
- Buying faster computers could speed up your teams' iteration speed and thus your team's productivity.
- Needing two weeks to train will limit the speed at which you can iterate.

Week 2 Quiz - Autonomous driving (case study)

1. You are just getting started on this project. What is the first thing you do? Assume each of the steps below would take about an equal amount of time (a few days).

- Spend a few days training a basic model and see what mistakes it makes.

> As discussed in lecture, applied ML is a highly iterative process. If you train a basic model and carry out error analysis (see what mistakes it makes) it will help point you in more promising directions.

2. Your goal is to detect road signs (stop sign, pedestrian crossing sign, construction ahead sign) and traffic signals (red and green lights) in images. The goal is to recognize which of these objects appear in each image. You plan to use a deep neural network with ReLU units in the hidden layers.

For the output layer, a softmax activation would be a good choice for the output layer because this is a multi-task learning problem. True/False?

- [ ] True
- [x] False

> Softmax would be a good choice if one and only one of the possibilities (stop sign, speed bump, pedestrian crossing, green light and red light) was present in each

image.

3. You are carrying out error analysis and counting up what errors the algorithm makes. Which of these datasets do you think you should manually go through and carefully examine, one image at a time?

- [ ] 10,000 randomly chosen images
- [ ] 500 randomly chosen images
- [x] 500 images on which the algorithm made a mistake
- [ ] 10,000 images on which the algorithm made a mistake

> Focus on images that the algorithm got wrong. Also, 500 is enough to give you a good initial sense of the error statistics. There's probably no need to look at 10,000, which will take a long time.

4. After working on the data for several weeks, your team ends up with the following data:

- 100,000 labeled images taken using the front-facing camera of your car.
- 900,000 labeled images of roads downloaded from the internet.

Each image's labels precisely indicate the presence of any specific road signs and traffic signals or combinations of them. For example, y(i) = [1 0 0 1 0] means the image contains a stop sign and a red traffic light.

Because this is a multi-task learning problem, you need to have all your y(i) vectors fully labeled. If one example is equal to [0 ? 1 1 ?] then the learning algorithm will not be able to use that example. True/False?

- [ ] True
- [x] False

> As seen in the lecture on multi-task learning, you can compute the cost such that it is not influenced by the fact that some entries haven't been labeled.

5. The distribution of data you care about contains images from your car's front-facing camera; which comes from a different distribution than the images you were able to find and download off the internet. How should you split the dataset into train/dev/test sets?

- [ ] Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 600,000 for the training set, 200,000 for the dev set and 200,000 for the test set.

- [ ] Mix all the 100,000 images with the 900,000 images you found online. Shuffle everything. Split the 1,000,000 images dataset into 980,000 for the training set, 10,000 for the dev set and 10,000 for the test set.
- [x] Choose the training set to be the 900,000 images from the internet along with 80,000 images from your car's front-facing camera. The 20,000 remaining images will be split equally in dev and test sets.
- [ ] Choose the training set to be the 900,000 images from the internet along with 20,000 images from your car's front-facing camera. The 80,000 remaining images will be split equally in dev and test sets.

> As seen in lecture, it is important that your dev and test set have the closest possible distribution to "real"-data. It is also important for the training set to contain enough "real"-data to avoid having a data-mismatch problem.

6. Assume you've finally chosen the following split between of the data:

- Training 940,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images) 8.8%
- Training-Dev 20,000 images randomly picked from (900,000 internet images + 60,000 car's front-facing camera images) 9.1%
- Dev 20,000 images from your car's front-facing camera 14.3%
- Test 20,000 images from the car's front-facing camera 14.8%

You also know that human-level error on the road sign and traffic signals classification task is around 0.5%. Which of the following are True? (Check all that apply).

- You have a large avoidable-bias problem because your training error is quite a bit higher than the human-level error.
- You have a large data-mismatch problem because your model does a lot better on the training-dev set than on the dev set.

7. Based on table from the previous question, a friend thinks that the training data distribution is much easier than the dev/test distribution. What do you think?

- There's insufficient information to tell if your friend is right or wrong.

> The algorithm does better on the distribution of data it trained on. But you don't know if it's because it trained on that no distribution or if it really is easier. To get a better sense, measure human-level error separately on both distributions.

8. You decide to focus on the dev set and check by hand what are the errors due to. Here is a table summarizing your discoveries:

- Overall dev set error 14.3%
- Errors due to incorrectly labeled data 4.1%
- Errors due to foggy pictures 8.0%
- Errors due to rain drops stuck on your car's front-facing camera 2.2%
- Errors due to other causes 1.0%

In this table, 4.1%, 8.0%, etc.are a fraction of the total dev set (not just examples your algorithm mislabeled). I.e. about 8.0/14.3 = 56% of your errors are due to foggy pictures.

The results from this analysis implies that the team's highest priority should be to bring more foggy pictures into the training set so as to address the 8.0% of errors in that category. True/False?

- [x] False because it depends on how easy it is to add foggy data. If foggy data is very hard and costly to collect, it might not be worth the team's effort. (OR)
- [x] False because this would depend on how easy it is to add this data and how much you think your team thinks it'll help.
- [ ] True because it is the largest category of errors. As discussed in lecture, we should prioritize the largest category of error to avoid wasting the team's time.
- [ ] True because it is greater than the other error categories added together (8.0 > 4.1+2.2+1.0).
- [ ] False because data augmentation (synthesizing foggy images by clean/non-foggy images) is more efficient.

9. You can buy a specially designed windshield wiper that help wipe off some of the raindrops on the front-facing camera. Based on the table from the previous question, which of the following statements do you agree with?

- 2.2% would be a reasonable estimate of the maximum amount this windshield wiper could improve performance.

> You will probably not improve performance by more than 2.2% by solving the raindrops problem. If your dataset was infinitely big, 2.2% would be a perfect estimate of the improvement you can achieve by purchasing a specially designed windshield wiper that removes the raindrops.

10. You decide to use data augmentation to address foggy images. You find 1,000 pictures of fog off the internet, and “add” them to clean images to synthesize foggy days, like this:

Which of the following statements do you agree with? (Check all that apply.)

- So long as the synthesized fog looks realistic to the human eye, you can be confident that the synthesized data is accurately capturing the distribution of real foggy images, since human vision is very accurate for the problem you’re solving.

> If the synthesized images look realistic, then the model will just see them as if you had added useful data to identify road signs and traffic signals in a foggy weather. I will very likely help.

11. After working further on the problem, you’ve decided to correct the incorrectly labeled data on the dev set. Which of these statements do you agree with? (Check all that apply).

- You do not necessarily need to fix the incorrectly labeled data in the training set, because it's okay for the training set distribution to differ from the dev and test sets. Note that it is important that the dev set and test set have the same distribution OR

- You should not correct incorrectly labeled data in the training set as well so as to avoid your training set now being even more different from your dev set.

> Deep learning algorithms are quite robust to having slightly different train and dev distributions.

- You should also correct the incorrectly labeled data in the test set, so that the dev and test sets continue to come from the same distribution

> Because you want to make sure that your dev and test data come from the same distribution for your algorithm to make your team’s iterative development process is efficient.

12. So far your algorithm only recognizes red and green traffic lights. One of your colleagues in the startup is starting to work on recognizing a yellow traffic light. (Some countries call it an orange light rather than a yellow light; we’ll use the US convention of calling it yellow.) Images containing yellow lights are quite rare, and she doesn’t have enough data to build a good model. She hopes you can help her

out using transfer learning.

What do you tell your colleague?

- She should try using weights pre-trained on your dataset, and fine-tuning further with the yellow-light dataset.

> You have trained your model on a huge dataset, and she has a small dataset. Although your labels are different, the parameters of your model have been trained to recognize many characteristics of road and traffic images which will be useful for her problem. This is a perfect case for transfer learning, she can start with a model with the same architecture as yours, change what is after the last hidden layer and initialize it with your trained parameters.

13. Another colleague wants to use microphones placed outside the car to better hear if there're other vehicles around you. For example, if there is a police vehicle behind you, you would be able to hear their siren. However, they don't have much to train this audio system. How can you help?

- Neither transfer learning nor multi-task learning seems promising.

> The problem he is trying to solve is quite different from yours. The different dataset structures make it probably impossible to use transfer learning or multi-task learning.

14. To recognize red and green lights, you have been using this approach:

- (A) Input an image (x) to a neural network and have it directly learn a mapping to make a prediction as to whether there's a red light and/or green light (y).

A teammate proposes a different, two-step approach:

- (B) In this two-step approach, you would first (i) detect the traffic light in the image (if any), then (ii) determine the color of the illuminated lamp in the traffic light.

Between these two, Approach B is more of an end-to-end approach because it has distinct steps for the input end and the output end. True/False?

- [ ] True

- [x] False

> (A) is an end-to-end approach as it maps directly the input (x) to the output (y).

15. Approach A (in the question above) tends to be more promising than approach B if you have a _______ (fill in the blank).

- [x] Large training set
- [ ] Multi-task learning problem.
- [ ] Large bias problem.
- [ ] Problem with a high Bayes error.

> In many fields, it has been observed that end-to-end learning works better in practice, but requires a large amount of data.

★ C4 - Convolutional Neural Networks

## Week 1 quiz - The basics of ConvNets

1. What do you think applying this filter to a grayscale image will do?

- Detect horizontal edges

- > Detect vertical edges

- Detect 45 degree edges

- Detect image contrast

2. Suppose your input is a 300 by 300 color (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)?

- 9,000,001

- 9,000,100

- 27,000,001

- > 27,000,100

3. Suppose your input is a 300 by 300 color (RGB) image, and you use a convolutional layer with 100 filters that are each 5x5. How many parameters does this hidden layer have (including the bias parameters)?

- 2501

- 2600

- 7500

- > 7600

4. You have an input volume that is 63x63x16, and convolve it with 32 filters that are each 7x7, using a stride of 2 and no padding. What is the output volume?

16x16x32

29x29x16

> 29x29x32

16x16x16

5. You have an input volume that is 15x15x8, and pad it using “pad=2.” What is the dimension of the resulting volume (after padding)?

19x19x12

17x17x10

> 19x19x8

17x17x8

6. You have an input volume that is 63x63x16, and convolve it with 32 filters that are each 7x7, and stride of 1. You want to use a “same” convolution. What is the padding?

1

2

> 3

7

7. You have an input volume that is 32x32x16, and apply max pooling with a stride of 2 and a filter size of 2. What is the output volume?

15x15x16

> 16x16x16

32x32x8

16x16x8

8. Because pooling layers do not have parameters, they do not affect the backpropagation (derivatives) calculation.

True

> False

9. In lecture we talked about “parameter sharing” as a benefit of using convolutional networks. Which of the following statements about parameter sharing in ConvNets are true? (Check all that apply.)

It allows parameters learned for one task to be shared even for a different task (transfer learning).

> It reduces the total number of parameters, thus reducing overfitting.

It allows gradient descent to set many of the parameters to zero, thus making the connections sparse.

> It allows a feature detector to be used in multiple locations throughout the whole input image/input volume.

10. In lecture we talked about “sparsity of connections” as a benefit of using convolutional layers. What does this mean?

Each filter is connected to every channel in the previous layer.

> Each activation in the next layer depends on only a small number of activations from the previous layer.

Each layer in a convolutional network is connected only to two other layers

Regularization causes gradient descent to set many of the parameters to zero.

Week 2 quiz - Deep convolutional models

1. Which of the following do you typically see as you move to deeper layers in a ConvNet?

nH and nW increases, while nC decreases

nH and nW decreases, while nC also decreases

nH and nW increases, while nC also increases

> nH and nW decrease, while nC increases

2. Which of the following do you typically see in a ConvNet? (Check all that apply.)

> Multiple CONV layers followed by a POOL layer

Multiple POOL layers followed by a CONV layer

> FC layers in the last few layers

FC layers in the first few layers

3. In order to be able to build very deep networks, we usually only use pooling layers to downsize the height/width of the activation volumes while convolutions are used with “valid” padding. Otherwise, we would downsize the input of the model too quickly.

True

> False

4. Training a deeper network (for example, adding additional layers to the network) allows the network to fit more complex functions and thus almost always results in lower training error. For this question, assume we’re referring to “plain” networks.

True

> False

5. The following equation captures the computation in a ResNet block. What goes into the two blanks above?

```
a[l+2]=g(W[l+2]g(W[l+1]a[l]+b[l+1])+bl+2+_______ )+_______
```

> a[l] and 0, respectively

0 and z[l+1], respectively

z[l] and a[l], respectively

0 and a[l], respectively

6. Which ones of the following statements on Residual Networks are true? (Check all that apply.)

> Using a skip-connection helps the gradient to backpropagate and thus helps you to train deeper networks

A ResNet with L layers would have on the order of L2 skip connections in total.

The skip-connections compute a complex non-linear function of the input to pass to a deeper layer in the network.

> The skip-connection makes it easy for the network to learn an identity mapping between the input and the output within the ResNet block.

7. Suppose you have an input volume of dimension 64x64x16. How many parameters would a single 1x1 convolutional filter have (including the bias)?

2

4097

1

> 17

8. Suppose you have an input volume of dimension nH x nW x nC. Which of the following statements you agree with? (Assume that “1x1 convolutional layer” below always uses a stride of 1 and no padding.)

> You can use a 1x1 convolutional layer to reduce nC but not nH, nW.

You can use a 1x1 convolutional layer to reduce nH, nW, and nC.

> You can use a pooling layer to reduce nH, nW, but not nC.

You can use a pooling layer to reduce nH, nW, and nC.

9. Which ones of the following statements on Inception Networks are true? (Check all that apply.)

> A single inception block allows the network to use a combination of 1x1, 3x3, 5x5 convolutions and pooling.

Making an inception network deeper (by stacking more inception blocks together) should not hurt training set performance.

> Inception blocks usually use 1x1 convolutions to reduce the input data volume’s size before applying 3x3 and 5x5 convolutions.

Inception networks incorporates a variety of network architectures (similar to dropout, which randomly chooses a network architecture on each step) and thus has a similar regularizing effect as dropout.

10. Which of the following are common reasons for using open-source implementations of ConvNets (both the model and/or weights)? Check all that apply.

A model trained for one computer vision task can usually be used to perform data augmentation even for a different computer vision task.

> It is a convenient way to get working an implementation of a complex ConvNet architecture.

The same techniques for winning computer vision competitions, such as using multiple crops at test time, are widely used in practical deployments (or production system deployments) of ConvNets.

> Parameters trained for one computer vision task are often useful as pretraining for other computer vision tasks.

Week 3 quiz - Detection algorithms

1. You are building a 3-class object classification and localization algorithm. The classes are: pedestrian (c=1), car (c=2), motorcycle (c=3). What would be the label for the following image? Recall y=[pc,bx,by,bh,bw,c1,c2,c3]

> ```y=[1,0.3,0.7,0.3,0.3,0,1,0]```

2. Continuing from the previous problem, what should y be for the image below? Remember that “?” means “don’t care”, which means that the neural network loss function won’t care what the neural network gives for that component of the output. As before, y=[pc,bx,by,bh,bw,c1,c2,c3].

> ```y=[0,?,?,?,?,?,?,?]```

3. You are working on a factory automation task. Your system will see a can of soft-drink coming down a conveyor belt, and you want it to take a picture and decide whether (i) there is a soft-drink can in the image, and if so (ii) its bounding box. Since the soft-drink can is round, the bounding box is always square, and the soft drink can always appears as the same size in the image. There is at most one soft drink can in each image. Here’re some typical images in your training set: What is the most appropriate set of output units for your neural network?

> Logistic unit, bx, by

4. If you build a neural network that inputs a picture of a person’s face and outputs N landmarks on the face (assume the input image always contains exactly one face), how many output units will the network have?

> 2N

5. When training one of the object detection systems described in lecture, you need a training set that contains many pictures of the object(s) you wish to detect. However, bounding boxes do not need to be provided in the training set, since the algorithm can learn to detect the objects by itself.

> False

6. Suppose you are applying a sliding windows classifier (non-convolutional implementation). Increasing the stride would tend to increase accuracy, but decrease computational cost.

> False

7. In the YOLO algorithm, at training time, only one cell ---the one containing the center/midpoint of an object--- is responsible for detecting this object.

> True

8. What is the IoU between these two boxes? The upper-left box is 2x2, and the lower-right box is 2x3. The overlapping region is 1x1.

> 1/9

9. Suppose you run non-max suppression on the predicted boxes above. The parameters you use for non-max suppression are that boxes with probability ≤ 0.4 are discarded, and the IoU threshold for deciding if two boxes overlap is 0.5. How many boxes will remain after non-max suppression?

> 5

10. Suppose you are using YOLO on a 19x19 grid, on a detection problem with 20 classes, and with 5 anchor boxes. During training, for each image you will need to construct an output volume y as the target value for the neural network; this corresponds to the last layer of the neural network. (y may include some “?”, or “don’t cares”). What is the dimension of this output volume?

> 19x19x(5x25)

## Week 4 quiz - Special applications: Face recognition & Neural style transfer

1. Face verification requires comparing a new picture against one person’s face, whereas face recognition requires comparing a new picture against K person’s faces.

> True

2. Why do we learn a function d(img1,img2) for face verification? (Select all that apply.)

> This allows us to learn to recognize a new person given just a single image of that person.

> We need to solve a one-shot learning problem.

3. In order to train the parameters of a face recognition system, it would be reasonable to use a training set comprising 100,000 pictures of 100,000 different persons.

> False

4. Which of the following is a correct definition of the triplet loss? Consider that α>0. (We encourage you to figure out the answer from first principles, rather than just refer to the lecture.)

> ```max(||f(A)−f(P)||^2 − ||f(A)−f(N)||^2 + α, 0)```

5. Consider the following Siamese network architecture: The upper and lower neural networks have different input images, but have exactly the same parameters.

> True

6. You train a ConvNet on a dataset with 100 different classes. You wonder if you can find a hidden unit which responds strongly to pictures of cats. (I.e., a neuron so that, of all the input/training images that strongly activate that neuron, the majority are cat pictures.) You are more likely to find this unit in layer 4 of the network than in layer 1.

> True

7. Neural style transfer is trained as a supervised learning task in which the goal is to

input two images (x), and train a network to output a new, synthesized image (y).

> False

8. In the deeper layers of a ConvNet, each channel corresponds to a different feature detector. The style matrix G[l] measures the degree to which the activations of different feature detectors in layer l vary (or correlate) together with each other.

> True

9. In neural style transfer, what is updated in each iteration of the optimization algorithm?

> The pixel values of the generated image G

10. You are working with 3D data. You are building a network layer whose input volume has size 32x32x32x16 (this volume has 16 channels), and applies convolutions with 32 filters of dimension 3x3x3 (no padding, stride 1). What is the resulting output volume?

> ```30 * 30 * 30 * 32```

★ C5 - Sequence Models

가즈아