Problem Set 2, CS 6476 Computer Vision, Spring 2021
Name: Haoran Wang

There is a single best answer for each multiple choice question.

[p01] *Learning.* One could argue that the one nearest-neighbor (1-NN) algorithm is superior to a linear classifier, such as that learned by a linear SVM, because the decision boundary created by 1-NN can be arbitrarily complex and fit the training data exactly, while the linear classifier's decision boundary is limited to a hyperplane in the feature space. However, in many tasks a linear classifier will outperform 1-NN. Why is this?

- O (a) 1-NN has **high** bias and **low** variance and is thus prone to **underfitting**.
- O (b) 1-NN has **high** bias and **low** variance and is thus prone to **overfitting**.
- O (c) 1-NN has **low** bias and **high** variance and is thus prone to **underfitting**.
- ■ (d) 1-NN has **low** bias and **high** variance and is thus prone to **overfitting**.

[p02] *Neural Networks.* Which of these best describes the training process for neural networks?

- O (a) All learned parameters of a deep network are set randomly, and then the loss over a training set is computed. This is repeated with entirely new random samples until a set of random parameters leads to a loss below your stopping criteria.
- O (b) All learned parameters of a deep network are randomly initialized, and then they are perturbed by random amounts. If the change to a parameter decreases the loss over the training set, the change is kept. This repeats until convergence.
- ■ (c) All learned parameters of a deep network are randomly initialized. All operations in a deep network are differentiable, so we can use the chain rule to determine how every learned weight influences the loss over a training batch. We adjust all weights by a small amount in the direction that will decrease loss and repeat this gradient descent over random training batches until some stopping criteria is reached.
- O (d) Learning the best weights in a deep network is convex (like solving for the parameters of a linear SVM) so we can learn the globally optimal set of parameters for a particular loss function by gradient descent regardless of how we initialize the weights. The choice of initial weights simply affects how fast we converge.

[p03] *Deep learning*. What does it mean for a deep network to be *convolutional*?

- O (a) "convolutional" refers to the historical origin of the networks, but modern ConvNets don't use convolutions.
- O (b) Each unit in a hidden layer will have an activation that is the result of a convolution with some portion of the previous hidden layer. Each hidden unit at each spatial location learns its own filter weights for this convolution.
- ■ (c) The activations of an entire channel of units in a hidden layer is the result of convolving the previous layer of hidden units with a single learned filter. The weights of that learned filter are shared across all spatial locations of the hidden layer. This means there are far fewer weights to learn than in a "fully" connected network.
- O (d) A deep network is convolutional if its operations can efficiently be performed in the Fourier domain.

Problem Set 2, CS 6476 Computer Vision, Spring 2021
Name: Haoran Wang

[p04] *Deep Learning.* Which of these are **NOT** among the reasons for the recent (past ten years) success of deep learning approaches in computer vision?

- O (a) The recent availability of huge image collections with one million or more images.
- O (b) The recent development of crowdsourcing marketplaces like Amazon Mechanical Turk which make it efficient to annotate massive image collections.
- ■ (c) The recent invention of the convolutional neural network.
- O (d) The recent advancement in computational power, especially GPU's.

[p05] *Deep learning*. Which of the following statements about the behavior of deep networks trained for recognition tasks is true?

- O (a) Hidden units / neurons in a deep network lack any meaningful organization. They respond to arbitrary content at any level of the network.
- O (b) Hidden units / neurons in the higher levels of a network (closer to the loss layer) tend to be sensitive to edges, textures, or patterns.
- ■ (c) Hidden units / neurons in the higher levels of a network (closer to the loss layer) tend to be sensitive to parts, attributes, scenes, or objects.

[p06] *Big Data*. "Tiny Images. Torralba et al. 2008" experiments with a massive database of 32x32 images. Which of these is **NOT** a reason why 32x32 images were examined?

- O (a) Even though texture information is lost, 32x32 images still contain enough information for humans to accurately perform scene and object classification.
- ■ (b) It is possible to brute-force sample every possible 32x32 image, thus eliminating traditional computer vision concerns about invariance and generalization.
- O (c) It would have been logistically difficult to store and process 80 million high resolution images.

[p07] *Human computation and crowdsourcing*. Which of these is **NOT** a viable strategy to make crowdsourced annotations more reliable?

- ■ (a) Increase worker pay until the desired accuracy level is reached.
- O (b) Either grade worker results or have ground truth annotations for a small fraction of the data and use this to determine which workers are trustworthy.
- O (c) "Gamify" your annotation task to the degree possible so that workers have a non-financial incentive to do a good job.
- O (d) "Wisdom of the crowds": Find consensus among multiple workers by taking the mean or median of their responses.

[p08] *Deeper Deep architectures*. In all of the deep convolutional network architectures for *recognition* discussed in class (e.g. AlexNet, VGG networks, GoogLeNet, ResNet) or in project 5, which of the following is **NOT** typically true

- O (a) The spatial resolution of each hidden layer of units tends to decrease as you move up the network (away from the image).
- O (a) The "receptive field" of each unit in each hidden layer tends to increase as you move up the network (away from the image).

- O (c) The filter depth or number of channels of each hidden layer of units tends to increase as you move up the network (away from the image).
- ■ (d) The “fully connected” layers are typically at the bottom of the network (close to the image) and the “convolutional” layers are typically at the top of the network (close to the loss layer).

[p09] *Deeper deep networks.* AlexNet has been improved by deeper architectures such as VGG, GoogLeNet, and ResNet. Which of the following trends was **NOT** observed with deeper networks?

- ■ (a) Deeper networks lead to higher accuracy at common computer vision tasks.
- O (b) Deeper networks are dramatically slower to use, because the complexity of the forward propagation pass grows quadratically with depth.
- O (c) Deeper networks are more difficult to train. GoogLeNet and ResNet have specific architectural modifications to help address the “vanishing gradient” problem which otherwise leads deep networks to underfit.

[p10] *"Unsupervised" Learning.* In “Unsupervised Visual Representation Learning by Context Prediction”, a deep network was trained to perform the seemingly arbitrary task of predicting the relative location of two image patches. Why?

- O (a) To aid in computational jigsaw puzzle assembly which requires estimating the relative position of image patches.
- O (b) Because so much training data is freely available for this contrived task, it is possible to train a deep network that is better at object detection than a network trained with traditional supervision (e.g. Imagenet classification).
- ■ (c) Supervision is effectively free, yet the network is not that much worse at object detection than a network pre-trained on Imagenet (and far better than any method prior to deep learning).
- O (d) Because the output dimensionality is low (8 discrete choices for the relative patch positions), the network is dramatically faster to train than networks with high dimensional output (e.g. Imagenet categories) yet it performs just as well.

[p11] *"Unsupervised" Learning.* In “Colorful Image Colorization”, the authors adopt a “multinomial” or categorical loss function for color prediction. Which of the following is **NOT** a reason for this?

- O (a) The baseline regression loss has a tendency of encouraging deep networks to predict the average color, so that it never pays large penalties for drastically wrong color predictions. On the other hand, the multinomial loss pays the same penalty for any wrong color prediction so it is not biased towards predicting the average.
- ■ (b) Multinomial losses require less memory because the output is discrete instead of continuous.
- O (c) Multinomial outputs provide a measure of uncertainty, e.g. the deep network can express that an output might be red or green but not blue. With a regression loss, you

only get a single prediction.

[p12] *Semantic Segmentation.* What is the motivation behind dilated convolution (also called atrous convolution) as compared to traditional convolution in deep networks?

- O (a) Dilated convolution is faster than traditional convolution.
- O (b) Dilated convolution is “residual” and mitigates the vanishing gradient problem.
- ■ (c) Dilated convolution leads to larger receptive fields without using more parameters than traditional convolution.
- O (d) Dilated convolution is applicable to unordered point clouds unlike traditional convolution.

[p13] *Transformer Architectures.* Attention / transformer architectures have become popular for many learning tasks. What is a disadvantage of such architectures?

- ■ (a) Transformer architecture complexity grows quadratically with the number of input tokens, while convolutional network complexity grows linearly.
- O (b) Transformer architectures require significantly more parameters than comparable convolutional architectures.
- O (c) Transformer architectures are generally less accurate than comparable convolutional architectures.
- O (d) Transformer architectures are less able to handle long range interactions between input tokens compared to transformer architectures.