Answer the questions in the answer spaces provided on the question sheets. If you run out of room for an answer, note in the answer space that it is continued on another page, and continue on a blank sheet that you staple to the end of the exam.

No use of a computing device is allowed other than as a timer, clock, music player, or simple (non-programmable) scientific calculator (for example, https://www.calculator.net/scientific-calculator.html). One double-sided 8.5x11 page of handwritten notes is allowed. Do not turn in those notes.

If you think something about a question is open to interpretation, write any assumptions you’ve made as part of answering the question.

Be concise in your answers; you need not try to fill in all or even most of the lines provided for an answer.

You have two contiguous hours to complete this exam starting from when you unstaple the exam or look at any page other than the first.

Bring the completed exam to class on November 20.
1. Neural Style Transfer works by running three images (content, style, and combined) through a VGG-19 Neural Network.

Explain in detail how the activations in different layers are used to compare the content of the content image to the content of the combined image.

Explain in detail how the activations in different layers are used to compare the style of the style image to the style of the combined image.

**Solution:** We train the combined image to reduce the combined loss (consisting of a weighted average of style loss and content loss). Content loss is defined as the MSE of the difference of the activations out of the last convolutional layer for the content image and style image. We use the last convolutional layer since that is the most abstract layer.

Style loss uses several of the layers (early, middle, and late), and adds together the loss at each layer. The style loss at a layer is computed by calculating the gram matrix for the activations at that layer for the style and combined images, and then taking the MSE of those gram matrices. (Gram matrix element \((i, j)\), is dot-product between \(i\)’th feature map and \(j\)’th feature map).

We use layers at different levels to capture both fine-grained stylistic details as well as larger stylistic features. The gram matrix captures correlation between multiple features.
2. What are the improvements to the Inception (GoogLeNet) architecture compared to that of VGG16 (circle one)?

A. Fewer layers
B. Fewer parameters
C. Shallower network
D. A and B
E. B and C
F. A and C

3. Assume you have two $3 \times 3$ convolutional layers. The input to the first layer has dimensions $120 \times 120 \times 3$. What is the receptive field of location $(25, 25)$ in the output of the second layer (with respect to the input of the first layer)?

**Solution:** It depends on the stride.

If the stride is 1, then the receptive field w.r.t. the output of the first layer is $(24, 24)$ to $(26, 26)$ and then w.r.t. the input of the first layer is $(23, 23)$ to $(27, 27)$.

However, if the stride is larger (let’s say 3), then although the receptive field w.r.t. the output of the first layer is still $(24, 24)$ to $(26, 26)$, the receptive field of each of those locations is a distinct $3 \times 3$ patch in the original image, yielding a receptive field of $(70, 70)$ to $(78, 78)$.

4. Come up with four $3x3$ convolution layers that’ll reduce an input from dimensions $160 \times 160 \times 3$ to $38 \times 38 \times 38$.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Input dimensions</th>
<th># of kernels</th>
<th>Dimensions of each kernel</th>
<th>Padding</th>
<th>Stride</th>
<th>Output dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$160 \times 160 \times 3$</td>
<td>5</td>
<td>$3x3$</td>
<td>half</td>
<td>2</td>
<td>$80 \times 80 \times 5$</td>
</tr>
<tr>
<td>2</td>
<td>$79 \times 79 \times 5$</td>
<td>10</td>
<td>$3x3x5$</td>
<td>half</td>
<td>2</td>
<td>$40 \times 40 \times 10$</td>
</tr>
<tr>
<td>3</td>
<td>$40 \times 40 \times 10$</td>
<td>20</td>
<td>$3x3x10$</td>
<td>none</td>
<td>1</td>
<td>$38 \times 38 \times 20$</td>
</tr>
<tr>
<td>4</td>
<td>$38 \times 38 \times 20$</td>
<td>38</td>
<td>$3x3x30$</td>
<td>half</td>
<td>1</td>
<td>$38 \times 38 \times 38$</td>
</tr>
</tbody>
</table>
5. Design the image processing piece of a phone application that’ll support modifying user photos by applying one of three styles:

- Van Gogh
- Cezzane
- Dali

The photo modification must be quick (run in less than one second) and can’t require a network connection. How would you design your image processing to run entirely on the phone (no cloud component).

Note that you may spend time and money developing your phone application.

**Solution:** First, I’d identify three target style pictures, one for each artist. I’d then create three Neural Networks (one each for Van Gogh, Cezzane, and Dali) which will take as input an unadorned image and output a styled version of the image. I’d use MSE as my loss function, I’d use some sort of CNN as my architecture. For my data, I’d take tens or hundreds of thousands of images (from Imagenet, perhaps?), and for each image, \( x \), I’d use Neural Style Transfer (along with the sample Van Gogh, Cezzane, or Dali image) and create a styled version, \( y \). I’d use all these \( (x, y) \) pairs to train my three NNs.

This creation of data will be quite time-consuming and compute-intensive, but I have time and money, so I’ll throw lots of GPU hardware and cloud instances at the problem:)

The phone application would then have to do only a forward pass with one of those NNs (whose pretrained weights would be embedded in the app). I’d expect that to take less than a second per image.
6. For this question, assume that there are no activation functions in the given layers and that the stride (where applicable) doesn’t exceed the operation’s width or height.

2 points (a) Can two consecutive MaxPool layers always be replaced with a single equivalent MaxPool layer? If so, explain how. If not, provide a counter-example.

Solution:
If we have two $2 \times 2$ maxpool layers (each with stride 2), they can be replaced with a single $4 \times 4$ maxpool layer (with stride 4). The original took the max of 16 cells, and so does the replacement.

Note, thought that the stride must be less than or equal to the size of the maxpool operation (otherwise, the elements to be maxed wouldn’t be contiguous).

2 points (b) Can two consecutive AvgPool layers always be replaced with a single equivalent AvgPool layer? If so, explain how. If not, provide a counter-example.

Solution: If the AvgPool layers have stride equal to their size, then yes. For example, if we have two $2 \times 2$ AvgPool layers (each with stride 2) (A), as compared to a single $4 \times 4$ MaxPool layer (with stride 4) (B):

\[ A = 1/4(a_{(1,1)}/4 + a_{(1,2)}/4 + a_{(2,1)}/4 + a_{(2,2)}/4) \]
\[ + 1/4(a_{(1,3)}/4 + a_{(1,4)}/4 + a_{(2,3)}/4 + a_{(2,4)}/4) \]
\[ + 1/4(a_{(3,1)}/4 + a_{(3,2)}/4 + a_{(4,1)}/4 + a_{(4,2)}/4) \]
\[ + 1/4(a_{(3,3)}/4 + a_{(3,4)}/4 + a_{(4,3)}/4 + a_{(4,4)}/4)) \]
\[ = 1/16(a_{(1,1)} + \cdots + a_{(4,4)}) \]
\[ = B \]

If the AvgPool layers have stride less than their size, then no. For example, if we have two $2 \times 2$ AvgPool layers (each with stride 1) (C), as compared to a single $3 \times 3$ MaxPool layer (with stride 3) (B):

\[ A = 1/4(a_{(1,1)}/4 + a_{(1,2)}/4 + a_{(2,1)}/4 + a_{(2,2)}/4) \]
\[ + 1/4(a_{(1,3)}/4 + a_{(1,4)}/4 + a_{(2,3)}/4 + a_{(2,4)}/4) \]
\[ + 1/4(a_{(2,1)}/4 + a_{(2,2)}/4 + a_{(3,1)}/4 + a_{(3,2)}/4) \]
\[ + 1/4(a_{(2,3)}/4 + a_{(2,4)}/4 + a_{(3,3)}/4 + a_{(3,4)}/4)) \]
\[ = 1/16a_{(1,1)} + 1/8a_{(1,2)} + \cdots \]
\[ \neq B \]

2 points (c) Can two consecutive 3x3 Convolution layers always be replaced with a single equivalent Convolution layer? If so, explain how. If not, provide a counter-
example.

Solution:
No if the two layers have different strides and padding.
Otherwise, yes. The new convolution layer would need to have size 3+stride.
Let’s look at 1-dimensional example: A 1x3 convolution $[w_1 \ w_2 \ w_3 \ d]$ with stride 1, zero padding, bias $b_1$ and inputs $[a \ \ b \ c \ \ d]$. The output of the first conv layer would be: $[aw_1 + bw_2 + cw_3 + b \ bw_1 + cw_2 + dw_3 + b]$.
The output of the second conv layer would be: 
$[(aw_1 + bw_2 + cw_3 + b_1)v_1 + (bw_1 + cw_2 + dw_3 + b)v_2 + b_2] = [av_1w_1 + b(w_2v_1 + w_1v_2) + c(w_3v_1 + w_2v_2) + dw_3v_2 + b_1v_1 + b_2]$
This could be computed with a single 1x4 convolution $[v_1w_1 \ w_2v_1 + v_1v_2 \ w_3v_1 + w_2v_2 \ w_3v_2]$ with stride 1, zero padding, bias $b_1v_1 + b_2$.

2 points (d) Can two consecutive Dropout layers always be replaced with a single equivalent Dropout layer? If so, explain how. If not, provide a counter-example.

Solution: Yes, you can replace two dropout layers, one with dropout rate $p$, and one with dropout rate $q$ with a new single dropout layer with dropout rate $1 - (1 - p)(1 - q)$.
The original dropout layer would dropout at a rate of $p$, leaving a remainder rate of $(1 - p)$. Then, those would be dropped with a rate of $q$, yielding a final remainder rate of $(1 - p)(1 - q)$. So, the new dropout layer needs to dropout with a rate of $1 - (1 - p)(1 - q)$.

2 points (e) Can two consecutive BatchNorm layers always be replaced with a single equivalent BatchNorm layer? If so, explain how. If not, provide a counter-example.

Solution: Yes, you can replace two BatchNorm layers ($B_1$ and $B_2$) with a single BatchNorm layer $B_3$.
The output of $B_1$ will have means $\mu_{B_1}$ and standard deviations $\sigma_{B_1}$. Those means and standard deviations will be scaled by layer two to $\mu_{B_2}$ and $\sigma_{B_2}$. Thus, the first layer is superfluous. The learned means and std. deviations at that layer are superfluous.

4 points 7. Assume you are using transfer learning with a pre-trained VGG-19 network to recognize a class of images. You implement the feature extraction approach. What are the advantages and disadvantages of that approach as compared to replacing the head of the old network and retraining that new head?
The feature extraction approach involved running your images through the VGG-19 network and saving the activations of the (flattened) last convolution layer. Then, create a new fully-connected layer that takes in flattened activations and outputs a category.

Advantage of feature extraction: speed. You only need to run each training image through the convolution layers of the VGG-19 network once. With a new head on the old base convolution layers, you’ll need to run each training image through the base layers once for each epoch.

Disadvantage: inability to (efficiently) do image augmentation. With a new network, you can do data augmentation on each image for each epoch.

When training a CNN to categorize an image into 1000 classes using a minibatch of size $k$, the shape of the input is $k \times 160 \times 160 \times 3$. What is the shape of the output?

**Solution:** The shape of the output for a single instance is $1 \times 1000$. For a minibatch of size $k$, the shape of the output is $k \times 1000$. 
9. Assume that a GAN is being trained, and that mode collapse has occurred (choose all that are correct):

- The Generator will learn to generate one specific output (regardless of the input), and won’t ever learn to shift from that.
  
  **Solution:** If the Generator continued to output only one output, the discriminator would learn that output as a fake, and so the loss function would force the Generator to change its output.

- The Generator will learn to output one specific output at a given training step (regardless of its input), but the specific output will change from training step to training step.
  
  **Solution:** If the Generator continued to output only one output, the discriminator would learn that output as a fake, and so the loss function would force the Generator to change its output. So, the Generator would move its output to some other output.

- If the quality of the Generator’s output is high enough, the Discriminator will be fooled.
  
  **Solution:** The discriminator won’t be fooled because it’s easy for it to determine the fake.

- The Discriminator will learn that any input it sees often is a fake.
  
  **Solution:** Not quite. It’ll learn the converse: that the fake it sees is seen often.

- the Discriminator’s accuracy will be 50%.
  
  **Solution:** No reason to think the accuracy will be 50% since, each time after the Generator is upgraded, the Discriminator can easily learn what it is generating.

10. A convolution layer has zero-padding, a stride of 1, weights \( \begin{bmatrix} 1 & 1 & -1 \\ 2 & 0 & 1 \\ 3 & -2 & -1 \end{bmatrix} \), bias of -2, and relu activation function. What is the output of that layer if the input is \( \begin{bmatrix} 1 & 1 & 1 & 2 & 2 \\ 5 & 5 & 3 & 8 & 3 \\ -2 & 1 & -1 & 5 & 3 \end{bmatrix} \)?
Solution: Before the relu:

\[
\begin{bmatrix}
1 \cdot 1 + 1 \cdot 1 + -1 \cdot 1 + 2 \cdot 5 + 0 \cdot 5 + 1 \cdot 3 + 3 \cdot -2 + -2 \cdot 1 + -1 \cdot -1 + -1 + -2 \cdots
\end{bmatrix} + -2
\]

\[
= \begin{bmatrix}
1 + 1 + -1 + 10 + 0 + 3 + -6 + -2 + 1 \cdots
\end{bmatrix} + -2
\]

\[
= \begin{bmatrix}
7 & 18 & -6
\end{bmatrix} + -2
\]

\[
= \begin{bmatrix}
5 & 16 & -8
\end{bmatrix}
\]

After the relu: \[\begin{bmatrix} 5 & 16 & 0 \end{bmatrix}\]
11. As described in class when discussing GANs, discriminators evaluate only on authenticity (distinguish between real/fake on single instances), but not on instance diversity. The lack of instance diversity can lead to mode collapse. Explain how providing the discriminator with a batch of instances at a time, rather than a single instance, could reduce mode collapse. How would the discriminator change?

**Solution:** We can feed the Discriminator a batch of instances (all fake or all real) rather than a single instance. The Discriminator can still assign fake/real tags to each of the instances, but can also measure how diverse the batch is. That diversity could be measured by looking at the discriminator inputs, or could do some measure of similarity between the activations at a particular layer (or layers). The discriminator can then do a weighted combination of the batch similarity measure with the instance real/fake measure before doing its final sigmoid activation.

In our forger/art dealer scenario, the art dealer would receive a batch of paintings, and would look both at the individual characteristics of each painting, along with the batch as a whole. If, during training, real batches were all types of pop art, while the fake batches only contain Andy Warhol art, the discriminator could learn that batches should contain all types of pop art, rejecting any batches that contain art that is more similar to each other than random batches of pop art.