Technical University of Denmark
2 hours written examination: 23 June 2022, 9-13 (14).
Course: 02514 Deep Learning in Computer Vision.
Aids allowed: All aids permitted (no internet access).
Weighting: All questions are weighted equally.

The exam is multiple choice. All questions have five possible answers marked by the letters A, B, C, D and E.
A correct answer gives 4 points, a wrong answer gives -1 points.

Question 1

Consider the neural network shown in Fig. 2.1. What is the receptive field of pixel $(1, 2)$ in the output, using Python zero indexing convention?

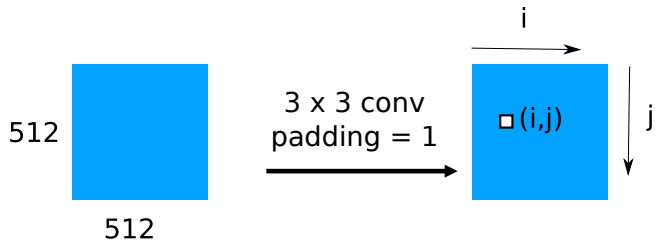

Figure 1.1: A CNN.

A. $(0:1, 0:2)$

B. $\mathbf{(0:2, 1:3)}$

C. $(1:3, 2:4)$

D. $(2:4, 3:5)$

E. Don't know

Question 2

Consider the neural network shown in Fig. 2.1. What is the receptive field of pixel $(1, 2)$ in the output, using Python zero indexing convention?

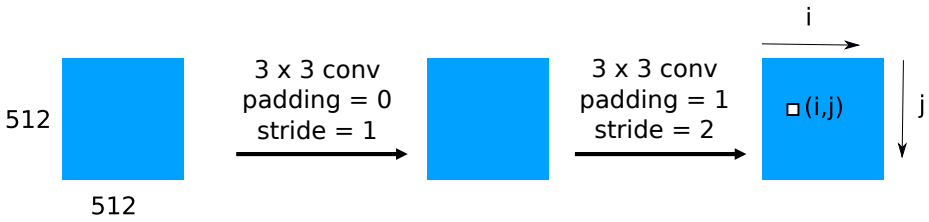

Figure 2.1: A CNN.

A. $\mathbf{(1:5, 3:7)}$

B. $(0:4, 1:5)$

C. $(1:5, 0:4)$

D. $\mathbf{(1:5, 3:7)}$

E. Don't know

Question 3

Which of the following techniques is referred to as "augmentation"?

A. Randomly removing a fixed proportion of the neurons in a layer during every forward pass

B. Mathematically constraining your network to be invariant to a known set of perturbations/symmetries

C. Normalizing the features of a given layer during every forward pass

D. **Encouraging your network to be invariant to a known set of perturbations/symmetries by feeding additional training images to the data obtained by applying the perturbations to the images while keeping their classes fixed**

E. Don't know

Question 4

Why would you use dropout during training?

A. To estimate the model uncertainty

B. To regularize the model

C. To quantify the data uncertainty

D. To obtain faster convergence

E. Don't know

Question 5

After applying the convolution shown in Fig. 5.1, what is the value of the green pixel with the thick border?

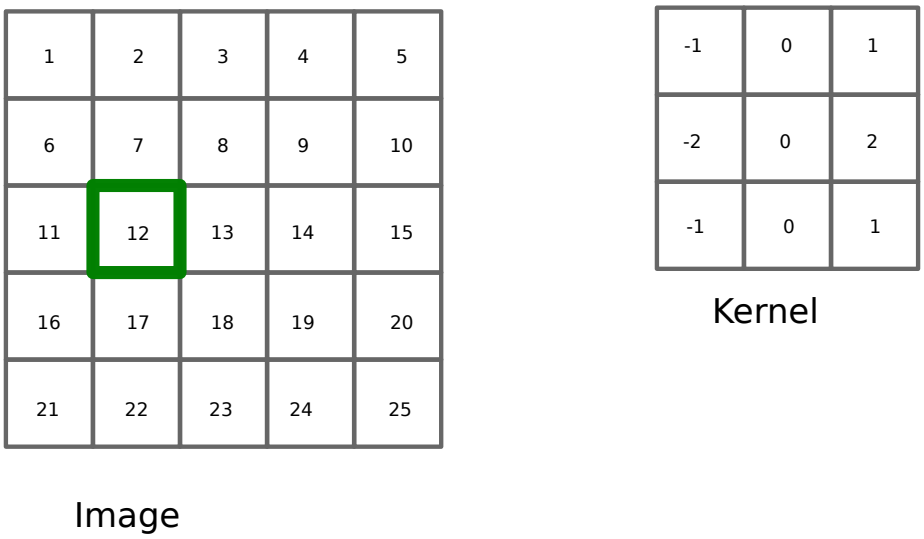

Figure 5.1: A convolution

A. 6

B. **8**

C. -6

D. 4

E. Don't know

Question 6

What sort of boundary effects would you expect to see for the network shown in Fig. 6.1?

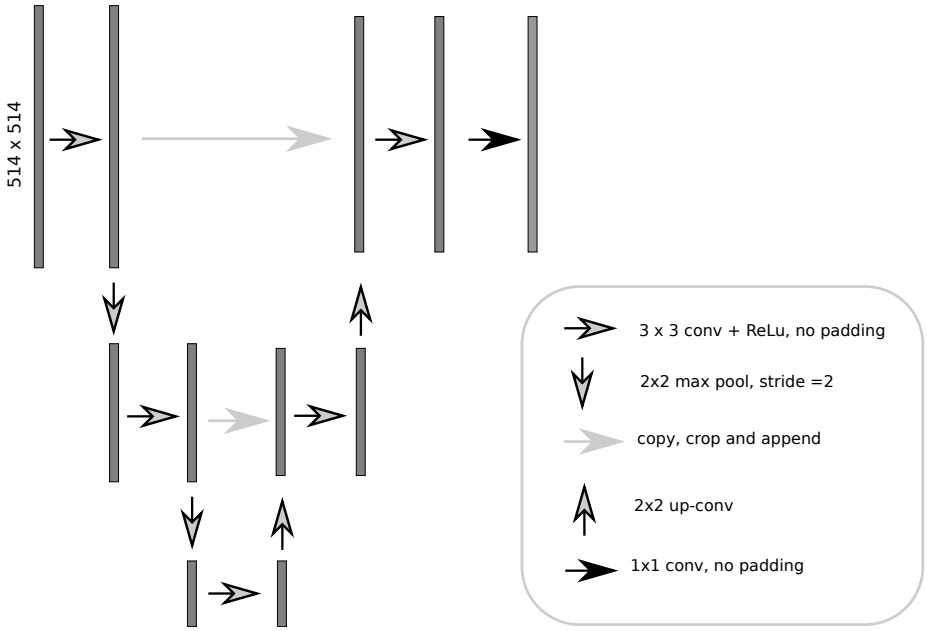

Figure 6.1: A segmentation architecture

A. Darker pixels at the boundaries

B. Boundary pixels have systematically different features, but we do not know what the network decides to do with them

C. There are no boundary effects

D. There will be lines along the boundaries

E. Don't know

Question 7

Which of the following layers in the CNN of Fig. 7.1 is equivalent to a fully connected layer?

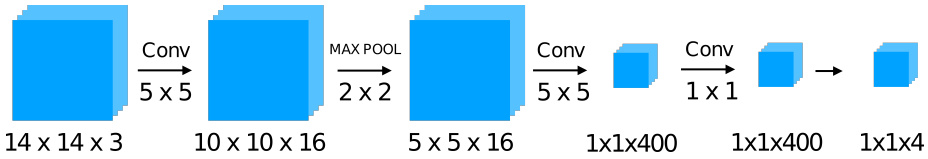

Figure 7.1: A segmentation architecture

A. The first 5×5 convolution

B. **The 1×1 convolution**

C. All of them

D. None of them

E. Don't know

Question 8

What does a 1×1 convolution do?

A. It does nothing

B. It just multiplies the features by a constant

C. It is equivalent to a fully connected network

D. It performs a linear combination of channels, plus a bias term

E. Don't know

Question 9

What do gradient-based saliency maps tell you?

A. They tell you how the model's prediction was made

B. They tell you which pixels to change, to most efficiently change the model's prediction.

C. They tell you where objects are located in the image

D. They tell you what is important in the image

E. Don't know

Question 10

Vanilla gradients and SmoothGrad are both based on gradients. But gradients of which function, and with respect to which variables?

A. The gradient of the loss function, with respect to the image pixel values

B. The gradient of the loss function, with respect to the weights of the neural network

C. The gradient of the probability of the wanted class, with respect to the image pixel values

D. The gradient of the probability of the wanted class, with respect to the model's hyperparameter

E. Don't know

Question 11

What does the following mathematical definition describe?

Given an image $x \in \mathbb{R}^{H \times W}$, a classifier $f \colon \mathbb{R}^{H \times W} \to \{0, \ldots, L\}$, and a target class label l, we seek $r \in \mathbb{R}^{H \times W}$ that satisfies

$$\text{minimize} \|r\|^2 \text{ while } \begin{array}{l} f(x+r) = l \\ x + r \in [0,1]^{H \times W} \end{array}$$

A. An adversarial attack

B. An interpolation

C. The SmoothGrad saliency map

D. Regularization when training the CNN

E. Don't know

Question 12

What is the difference between the Fast Gradient Sign Method (FSGM) and the Basic Iterative Method (BIM)?

A. FSGM is faster

B. FSGM uses gradients, BIM does not

C. FSGM is a saliency map, whereas BIM is an adversarial attack

D. **FSGM assumes a linear model, whereas BIM only assumes local linearity.**

E. Don't know

Question 13

Why would you prefer stacking 3×3 convolutions over an $n \times n$ convolution for $n > 3$?

A. They allow you to see more detail

B. To get a larger receptive field with fewer parameters

C. To get a deeper network with fewer parameters

D. To avoid large boundary effects

E. Don't know

Question 14

Which classification architecture obtains multiple receptive fields within the same layer?

A. InceptionNet

B. EfficientNet

C. U-Net

D. VGG

E. Don't know

Question 15

What is the point of the skip connection in the ResNet?

A. To preserve location information from the previous layer

B. To get more parameters and hence a more flexible model

C. **To speed up learning as zero weights in residual blocks effectively lead to a shallower network**

D. To encode information from multiple layers in the same feature

E. Don't know

Question 16

The Inception module has dimensionality reduction built in. How is this achieved?

A. Via max-pooling

B. By increasing the stride

C. With an autoencoder

D. **Via 1×1 convolutions**

E. Don't know

Question 17

Imagine training the U-net architecture shown in Fig. 17.1 on image patches centered at lung lesions such as shown in Fig. 17.2. The task is to segment the lung lesions. What sort of boundary effects would you expect to see?

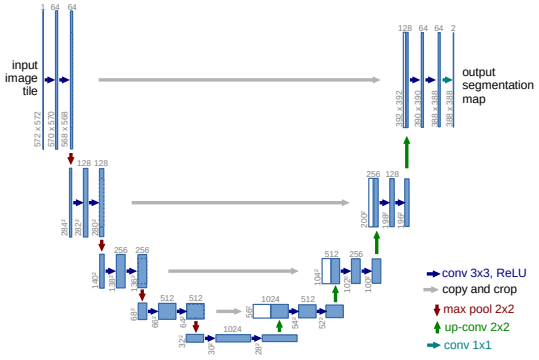

Figure 17.1: A U-net.

Figure 17.2: A lung lesion.

A. The final segmentation will be a square

B. The segmentation performance may be overestimated

C. The final segmentation will be black around the boundary

D. **There is no boundary effect**

E. Don't know

Question 18

Given the U-net architecture and the output image shown in Fig. 18.1, and using Python zero indexing convention, is it possible for the pixel $(10, 13)$ to be affected by boundary effects?

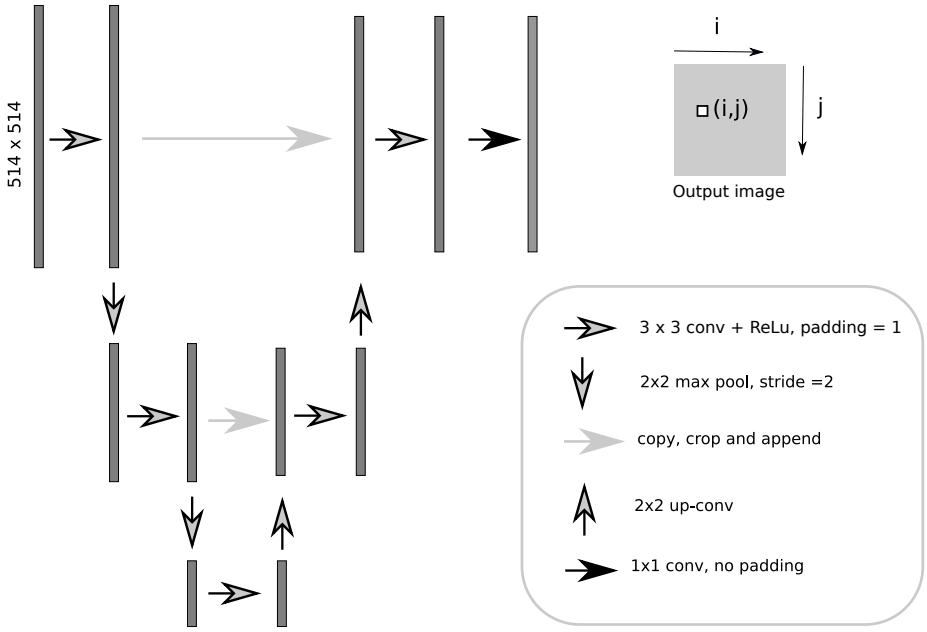

Figure 18.1: A U-net.

A. No, the network has no boundary effects

B. Yes, because the receptive field of $(10, 13)$ contains padding

C. Yes, because the pixel is close to the boundary

D. No, because the receptive field of $(10, 13)$ does not contain padding

E. Don't know

Question 19

Given the segmentation architecture shown in Fig. 19.1, what is the shape of the output segmentation?

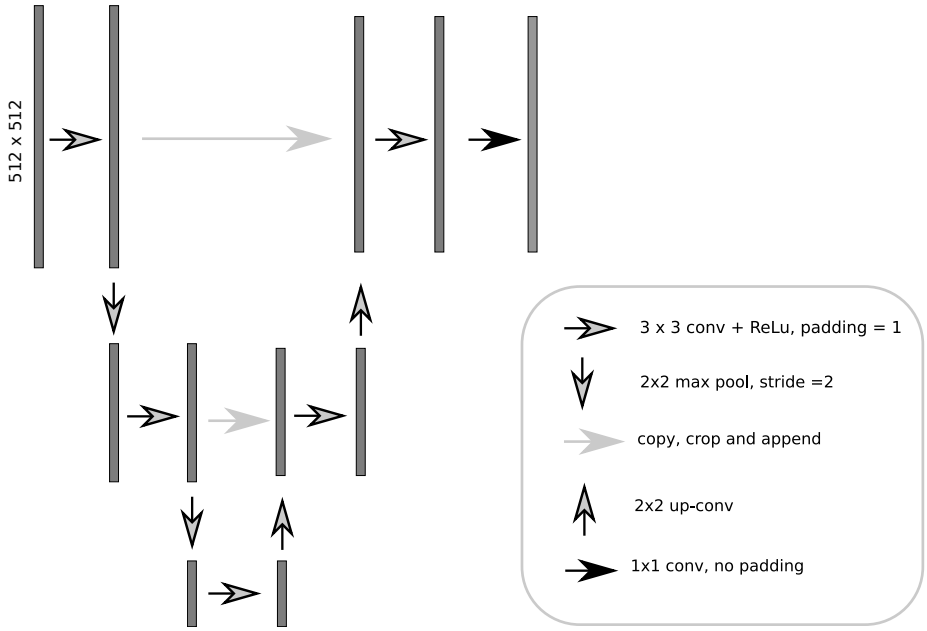

Figure 19.1: A segmentation architecture.

A. 492×492

B. $\mathbf{512 \times 512}$

C. 488×488

D. 494×494

E. Don't know

Question 20

Given the U-net architecture shown in Fig. 20.1, and using Python zero indexing convention, which pixel in the input image is associated with the pixel classification given by the output pixel $(1, 2)$?

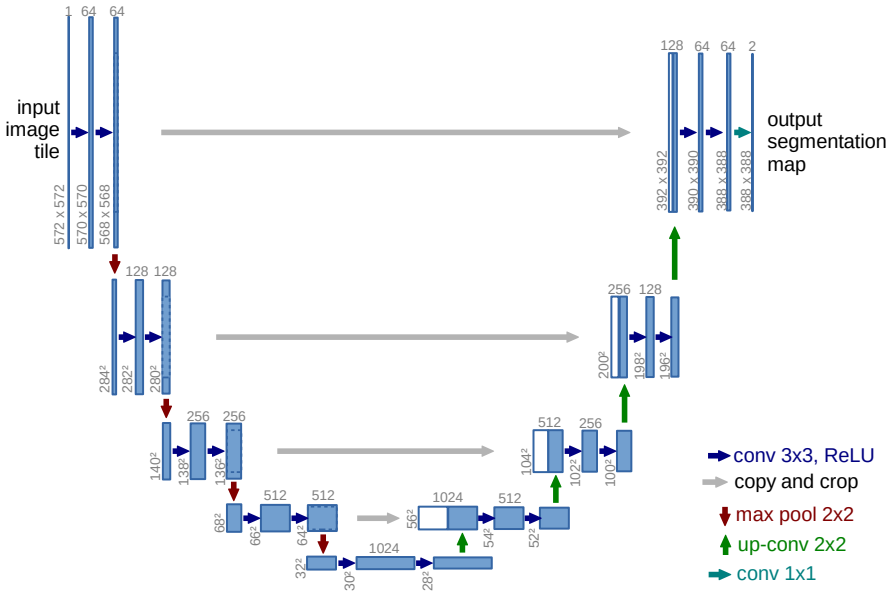

Figure 20.1: A segmentation architecture.

A. $(1, 2)$

B. **$(93, 94)$**

C. $(94, 95)$

D. $(2, 3)$

E. Don't know

Question 21

What is the point of the skip connection in the U-Net?

A. To preserve location information from the corresponding layer in the encoder

B. To get more parameters and hence a more flexible model

C. To speed up learning as zero weights in residual blocks effectively lead to a shallower network

D. To encode information from multiple resolutions in the same feature

E. Don't know

Question 22

What does the SegNet and the U-Net have in common?

A. They both have skip connections

B. They cannot be applied to multi-class segmentation

C. They both have dimensionality reduction built in through the encoder-decoder structure

D. They both carry location information from each layer of the encoder to the corresponding layer of the decoder

E. Don't know

Question 23

Which segmentation validation score is given by the formula

$$\frac{2|A \cap B|}{|A| + |B|}?$$

A. Intersection over Union

B. Dice score

C. Jaccard index

D. None of the above

E. Don't know

Question 24

For which of the following segmentation problems are the Dice and IoU scores not very well suited?

A. Segmenting flies from background

B. Segmenting birds from background

C. Segmenting tape worms from background

D. Segmenting bumblebees from background

E. Don't know

Question 25

Which of the following dilation techniques is referred to as dilated convolutions?

A. Dilating the kernel before convolving with the image

B. Dilating the image before convolving with the kernel

C. Stacking multiple convolutions with the same kernel to get a larger receptive field

D. Upsampling the image before applying a convolution

E. Don't know

Question 26

Given the transpose convolution shown in Fig. 26.1, what is the receptive field, in the original image, of pixel $(2, 3)$? Indices are given in Python zero indexing convention.

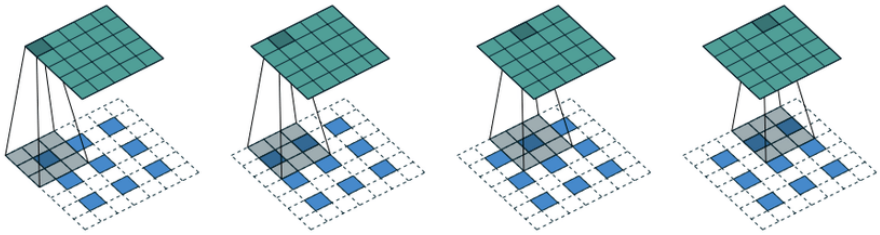

Figure 26.1: A transpose convolution.

A. $(0:1, 1:1)$

B. $\mathbf{(1:1, 1:2)}$

C. $(2:4, 3:5)$

D. $(3:5, 4:6)$

E. Don't know

Question 27

Which of the following is not considered weak annotation for image segmentation?

A. Image level annotations

B. Bounding boxes

C. Scribbles

D. Segmentation masks

E. Don't know

Question 28

Which of the following tools could be used to learn segmentation from image level annotations?

A. Segmentation uncertainty quantification

B. Class activation mappings

C. MC Dropout

D. Bounding boxes

E. Don't know

Question 29

Which loss function is given by the formula below?

$$\mathcal{L}(y, \hat{y}) = -\sum_i \left[(1 - \sigma(\hat{y}_i))^\gamma \, y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i)) \right].$$

A. Focal loss

B. Dice loss

C. Binary Cross Entropy

D. Root Mean Square Error

E. Don't know

Question 30

Why would you apply dropout when making predictions on your test set?

A. You only apply dropout when training the model

B. **To quantify model (epistemic) uncertainty**

C. To make sure the predictions are made under the same conditions as the training

D. To quantify data (aleatoric) uncertainty

E. Don't know

Question 31

A Faster R-CNN object detection model is trained with 4 losses: (i) RPN classification loss, (ii) RPN regression loss, (iii) Object classification loss, (iv) Object regression loss. Which of these losses are also used in a Fast R-CNN model?

A. RPN classification and RPN regression

B. Object classification and Object regression

C. RPN classification and Object classification

D. Only the Object classification loss

E. Don't know

Question 32

A Faster R-CNN object detection model is trained with 4 losses: (i) RPN classification loss, (ii) RPN regression loss, (iii) Object classification loss, (iv) Object regression loss. All losses contribute to the training and the final performance of the detector. However, two of these losses are essential for having a Faster R-CNN model and two of them could potentially be omitted and still be able to train a Faster R-CNN model but with inferior performance. Which losses could be omitted?

A. RPN classification and RPN regression

B. Object classification and Object regression

C. RPN classification and Object classification

D. RPN regression and Object regression

E. Don't know

Question 33

Assume that you have a YOLO object detector that operates on a 7x7 grid with 4 anchors per cell and aims to detect objects for 20 categories. The size of the model prediction on a test image is:

A. variable and it is based on how many target objects appear in the image

B. always 98

C. always 1960

D. always 3920

E. Don't know

Question 34

We have a YOLO object detector that operate on a 3x3 grid with 1 anchor per cell and aims to detect objects for 20 categories. What is the maximum number of object instances from the same class that this model can detect in a single image?

A. 9

B. 18

C. 180

D. as many as appear in the image

E. Don't know

Question 35

An object detection output (i.e., object bounding boxes on a test image) can be obtained directly by

A. a semantic segmentation prediction

B. an instance segmentation prediction

C. a multi-label multi-class image classification prediction

D. all of the above

E. Don't know

Question 36

The goal of an object proposal algorithm or a Region Proposal Network is to generate window proposals:

A. with high precision

B. with high recall

C. that do not overlap with each other

D. that cover all the pixels of the image

E. Don't know

Question 37

Assuming that you have already pre-computed the object proposals using Selective Search for a whole dataset including both the training and the test set. Which of the following models is faster?

A. RCNN model

B. Fast RCNN model

C. Faster RCNN model

D. The training and inference time of all of them is the same

E. Don't know

Question 38

An object detector predicts two bounding boxes of the class C_1 for a test image. The coordinates of these bounding boxes are [10,50,20,20] and [40,10,40,30]. The image contains 2 objects of the class C_1 and the ground-truth bounding boxes have coordinates [40,60,40,20] and [15,60,15,10]. *Note that the bounding box coordinates are given in the form of [x_0,y_0,w,h].* The IoU values of the predicted boxes are:

A. 0.000 and 0.375

B. 0.000 and 0.000

C. 0.375 and 0.000

D. 0.375 and 0.375

E. Don't know

Question 39

We have trained an object detector to predict 2 categories C_0 and C_1. We apply this detector on a test image and we obtain the following bounding boxes: [10,20,10,20], [60,60,60,60]. The first bounding box belong to C_0 and the second one to C_1.
The image contains two C_0 object with ground-truth coordinates [10,30,10,10] and [60,60,60,60].
The IoU values of the predicted boxes are:

A. 0.5, 1.0

B. 0.5, 0.0

C. 0.0, 0.0

D. 0.0, 1.0

E. Don't know

Question 40

We apply an object detector on a test image and we obtain the following bounding boxes: B_0=[65,65,50,50], B_1=[125,120,25,30], B_2=[60,60,60,60], B_3=[120,120,30,30]. All boxes belong to the same object class and the confidence scores boxes are 0.4, 0.6, 0.9, 0.7.
After applying non-maximum suppression (with IoU=0.7) to this output, which set of bounding boxes do we obtain?

A. $\{B_2\}$

B. $\{B_2, B_3\}$

C. $\{B_0, B_1, B_2\}$

D. $\boldsymbol{\{B_0, B_2, B_3\}}$

E. Don't know

Question 41

The Non-maximum suppresion (NMS) is a common part of an object detection pipeline. Which of the following sentences are true?

A. NMS is *only* applied at test time for post-processing the object detection output

B. NMS can *optionally* be applied during training of the object detection model before applying the classification and bounding box regression loss.

C. NMS should *necessarily* be applied during training of the object detection model before applying the classification and bounding box regression loss.

D. Note of the above.

E. Don't know

Question 42

Even though that the IoU metric is by far the most popular approach to quantify an object detection prediction, an object detector is not trained to optimize the IoU loss of the prediction with respect to the ground-truth box. Why?

A. IoU is not differentiable

B. It is computationally expensive to compute IoU during training

C. It can be used but it does not lead to good results

D. It is hard to compute it during training

E. Don't know

Question 43

You applied an object detection model on the following 2 test images and you obtained the bounding box predictions as shown in the images. Assuming that your whole test set contains only these two images, that is the final Average Precision (AP) for the class *dog* (We assume that a bounding box is correct if its IoU with the corresponding ground-truth is greater than 0.5).

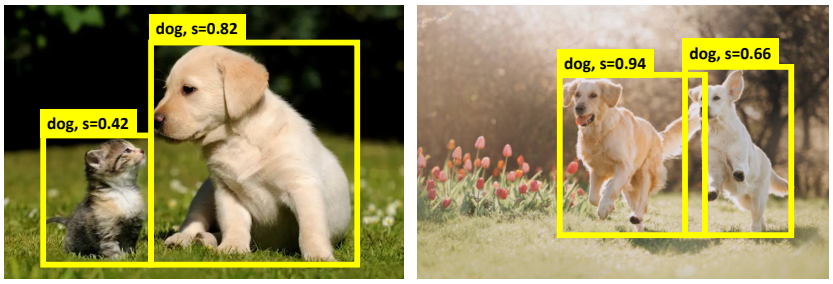

Figure 43.1: Object detection output.

A. AP(dog)=100%

B. AP(dog)=86%

C. AP(dog)=80%

D. AP(dog)=72%

E. Don't know

Question 44

Assuming that you have a test set that contains only one image. You have 4 object detectors that give you the following predictions on this image. Which of the object detectors is the best according to the AP metric?

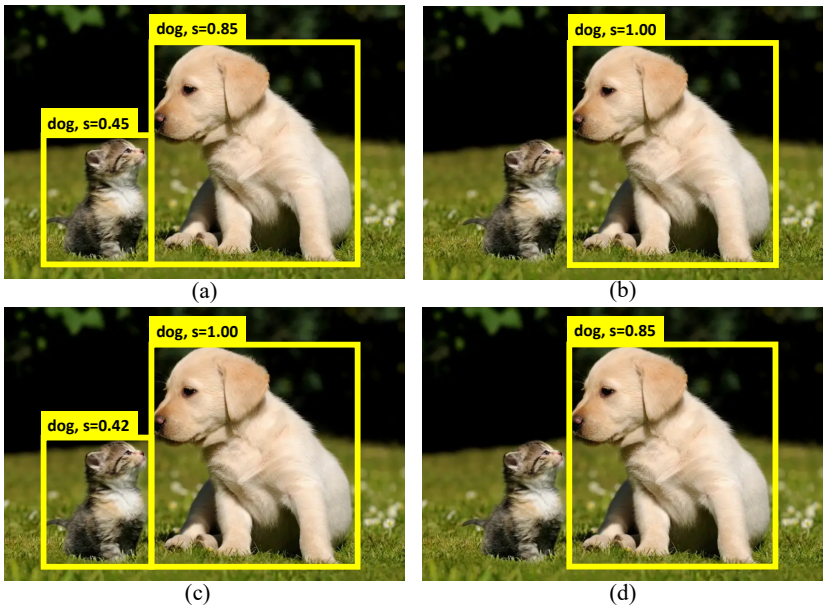

Figure 44.1: Four different object detection outputs.

A. b

B. b and d

C. c

D. all 4 detectors have the same AP

E. Don't know

Question 45

When training an two-stage object detection model to detect N classes, the size of the object classification output has:

A. N outputs

B. $\boldsymbol{N + 1}$ **outputs**

C. 2 outputs

D. 4 outputs

E. Don't know

Question 46

When training a Faster RCNN object detection model to detect N classes, the size of the object classification and the RPN classification output have:

A. N outputs and 2 outputs

B. $\boldsymbol{N+1}$ **outputs and 2 outputs**

C. N outputs and N outputs

D. $N+1$ outputs and $N+1$ outputs

E. Don't know

Question 47

Why do we prefer an object proposal paradigm instead of a sliding window paradigm for an object detection model?

A. Sliding window is computationally very expensive

B. Sliding window can not handle variable sized outputs

C. The performance of sliding window detectors are always worse

D. All of the above

E. Don't know

Question 48

The idea to grow both the generator and discriminator by starting from a low resolution and progressively add new layers to the model was first introduced in the Progressive GAN (ICLR 2018). Which of the following GAN models use this key idea to generate high resolution images?

A. Only the Progressive GAN (ICLR 2018)

B. only the Progressive GAN (ICLR 2018) and the StyleGAN (CVPR 2019)

C. only the Progressive GAN (ICLR 2018), the StyleGAN (CVPR 2019) and the StyleGAN2 (CVPR 2020)

D. the Progressive GAN and all its successors (StyleGAN, StyleGAN2, StyleGAN2-ADA, StyleGAN3)

E. Don't know

Question 49

You are training a StyleGAN model (CVPR 2019) following the settings of the original implementation and you aim to generate images that have a resolution of 256x256. How many convolutional layers does the synthesis network g has besides the 1×1 convolution at the output?

A. 12 convolutional layers

B. 14 convolutional layers

C. 16 convolutional layers

D. 18 convolutional layers

E. Don't know

Question 50

Imagine that you have learned two latent directions w_1 and w_2 in a GAN model, where w_1 is smiling and w_2 is adding bearding to faces. You would now like to apply both directions with m_1 and m_2 to your favorite image in order to generate a face smiling and with a beard at the same time.
You follow two strategies: (i) you first apply the direction w_1 with m_1 and then you apply the direction w_2 with m_2 and (ii) you first apply w_2 with m_2 and then apply w_1 with m_1.
The final output images of the two strategies would be:

A. always identical

B. different

C. identical only if $m_1 = m_2$

D. identical only if $m_1 > 0$ and $m_2 > 0$

E. Don't know

Question 51

Imagine that you have a GAN model trained to generate faces and you want to learn a single latent direction that can add beard to faces and at the same time add sunglasses. To do that you use two sets of images: one set with men that have beard and wear sunglasses and another set with men without beard and without sunglasses. Would it be possible to learn this direction?

A. No, because every direction in the latent space affects only on face characteristic.

B. No, because this direction has to change both the upper (sunglasses) and the bottom (beard) part of the image

C. No, the latent direction that you learned can change only one of the two face attributes in the image.

D. Yes

E. Don't know

Question 52

When training a GAN model a good strategy to initialize the generator (G) and discriminator (D) is:

A. Initialize both G and D randomly (train them from scratch).

B. Initialize the weights of D using a strong pre-trained model and use transfer learning to finetune it and initialize the weights of G randomly.

C. Initialize the weights of G using a strong pre-trained model and use transfer learning to finetune it and initialize the weights of D randomly.

D. Initialize both G and D using a strong pre-trained model and finetune them both on your dataset using transfer learning.

E. Don't know

Question 53

In a GAN model, which is trained using the minimax loss, the loss of the generator becomes minimum when:

A. the discriminator predicts that all real images are real with score 1.0.

B. the discriminator predicts that all real images are fake with score 1.0.

C. the discriminator predicts that all generated images are real with score 1.0.

D. the discriminator predicts that all generated images are fake with score 1.0.

E. Don't know

Question 54

In a GAN model, which is trained using the minimax loss, the loss of the discriminator becomes maximum when:

A. the discriminator predicts that all real images are real with score 1.0 and all generated images are fake with score 1.0.

B. the discriminator predicts that all real images are fake with score 1.0 and all generated images are real with score 1.0

C. the discriminator predicts that all generated and all real images are real with score 1.0.

D. the discriminator predicts that all generated and all real images are fake with score 1.0.

E. Don't know

Question 55

You are given a StyleGAN model that is trained to generate 512x512 images from an initial latent of 512 dimensions. What is the shape of the intermediate latent code w?

A. 18x512

B. 16x512

C. 14x512

D. 1x512

E. Don't know

Question 56

You are training a vanilla GAN model with a minimax loss and you observe that your discriminator becomes too good and the generator training fails due to vanishing gradients. What can you do so that your GAN is not suspectible to this problem?

A. Increase the capacity of the discriminator

B. Change the loss and use WGAN

C. Try to train the model again

D. Use data augmentation on the generated images

E. Don't know

Question 57

Consider a conditional GAN model on a class label y. Which of the following sentences are true?

A. The label y is used as an additional input only to the generator together with the noise vector Z.

B. Similar to an unconditional GAN, the model tries to learn a probability distribution $p(x)$.

C. Given a certain latent code z, the conditional GAN generates the same image regardless of the label y.

D. None of the above.

E. Don't know

Question 58

Which of the following GAN models are conditional GANs?

A. BigGAN

B. Pix2pix

C. CycleGAN

D. All of the above

E. Don't know

Question 59

We have a StyleGAN model trained on the FFHQ dataset and we want to use the CLIP model to optimize the latent code of the GAN and generate an image with the input text "an image of a purple car". What we should expect from the result of the iterative optimization?

A. It can work reasonably well if we use a good initialization for the GAN's latent code.

B. It will not work well because the CLIP model can not recognize well car images.

C. We will probably get an image with a car but it will not be purple since it's a rare color for a car.

D. It will not generate a car since the pre-trained GAN model can not generate cars.

E. Don't know

Question 60

We have a pre-trained GAN model and we want to use the CLIP model to optimize the latent code of the GAN and generate an image given an input text. During the iterative optimization process:

A. We finetune the weights of the GAN generator and the weights of the image and text encoder of the CLIP.

B. We freeze the generator and the CLIP encoders and we do not update their weights at each iteration.

C. We freeze the weights of the CLIP encoders but we finetune the weights of the GAN generator.

D. We freeze the weights of the GAN generator but we finetune the weights of the CLIP encoders.

E. Don't know