Exam module 5 : Visual recognition
Date: May 3rd, 2021
Time: 16-18h

Name _____

NIU _____

- Each multiple choice question (MCQ) is worth 0.25 points.
- Wrong answers to MCQs will not subtract any point.
- Short answer questions (SAQ) are worth *up to* 0.5 points : 0, 0.25 or 0.5 points.
- Please, answer
  - MCQs on the exam sheets (except students doing the exam online)
  - SAQs into blank sheets
- Write your name on this and all sheets

Object detection

1. Figure 1 illustrates the R-CNN object detection system (Girshick et al. 2014). Explain why the image warping technique was necessary and which are the negative implications of this design in terms of training/inference time. What was the solution proposed by later methods?

   It was necessary to warp images to a fixed input size because the region-wise feature extractor (a CNN model such as AlexNet or VGG) requires a fixed-size input of 227×227 pixels. This design makes both training and inference very slow as the CNN features for each one of 2k proposals from the first stage must be computed. The solution to this problem comes from the Spatial Pyramid Pooling (SPP) and/or ROI Pooling layers, that apply a pooling operation to a region of the last feature map of a CNN providing a fixed-size output independently of the region geometry. With these layers the feature extraction is made much faster as a single computation of the CNN layers is shared for all the regions.

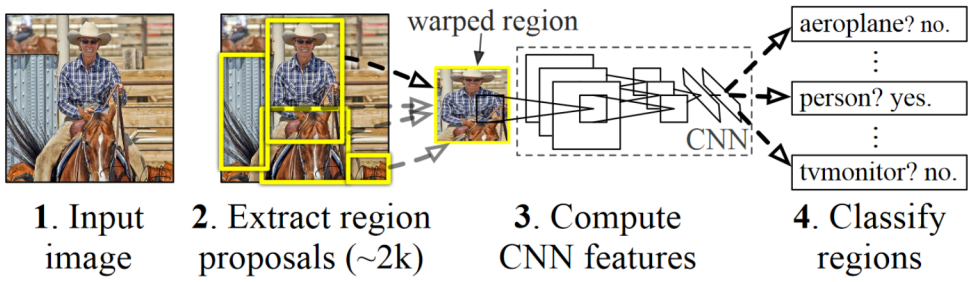

Figure 1: R-CNN object detection system overview.

2. Explain the concept of "anchor boxes" and why they are used in deep learning models for object detection. Name two models, one two-stage model and one single-stage model, that use anchor boxes.

   Anchor boxes (or prior boxes) are a predefined set of boxes with their geometry (size and aspect ratio) chosen to match the ones of objects in a dataset. They are usually obtained by some clustering algorithm over the bounding box annotations of the training set. They are used as proposals in various spatial locations of a conv. feature map, as it is easier for the model to learn the offsets to an anchor box than the actual bounding box geometry. The Faster RCNN (two-stage) model uses anchor boxes, and also the single stage model SSD among many others.

3. Explain the You Only Look Once (YOLO) object detection model (Redmon et al. 2016). Describe briefly the key idea, how is the architecture and output of the model, and which are its main strength and limitation.

   YOLO is a one-stage object detection model. The key idea of the model is to divide the input image into a grid of 7x7 regions and, for each cell of the grid, make 2 bounding box predictions and one class label prediction. The network has 24 convolutional layers followed by 2 fully connected layers. The final output of the model is a 7x7x30 tensor. (30 = 2 bounding boxes * 5 values (coordinates + confidence) + 20 classes). Its main strength is its speed. The base algorithm runs at 45 frames per second (FPS). The main limitation is the number of objects that can be detected, since each cell predicts only two boxes and one class, this limits the model in detecting groups of overlapping objects.

4. Which of the following statements about the evaluation of object detection methods are true. Indicate all the true ones.

   (a) The intersection over union (IoU) is a function that measures the overlap of two bounding boxes and its output values range from -1 to 1.

   (b) By convention in all object detection datasets it is considered that a bounding box prediction is correct when its intersection over union (IoU) with one of the ground-truth bounding box annotations is greater than 0.95.

   (c) When for a given ground-truth bounding box annotation there is no bounding box prediction with an intersection over union (IoU) greater than the predefined IoU threshold it counts as a False Negative (FN).

   (d) When a bounding box prediction has an intersection over union (IoU) smaller than the predefined IoU threshold it counts as a True Negative (TN).

   (e) To calculate the Average Precision (AP) metric of a model the first step is to rank all the predictions in descending order according to their predicted confidence score.

   (c),(e)

5. Which of the following statements are true. Indicate all the true ones.

(a) The Single Shot Detector (SSD) model predicts bounding boxes after multiple convolutional layers. Since every convolutional layer functions at a diverse scale, this makes easier to detect objects with different sizes.

(b) The You Only Look Once (YOLO) model is a two-stage object detector with a custom CNN backbone.

(c) The Focal Loss of the RetinaNet model tries to mitigate class imbalance within the classification loss by down-weighting loss contributions of easily classifiable examples.

(d) Training a Faster R-CNN object detection model involves training per-class linear SVMs binary classifiers with hard negative mining. This helps the model to be more discriminative for rare classes.

(a),(c)

Image and instance segmentation

6. Compute the output of the max unpooling layer of Figure 2. The input is a 4x4 matrix where a max pooling operation is applied. After a sequence of layer operations, an max unpooling layer linked to max pooling layer is added.

   (a) Fill the output 4x4 matrix of Figure 2. **0.25 points**

   (b) What popular semantic segmentation network uses the max unpooling layer? **0.25 points**

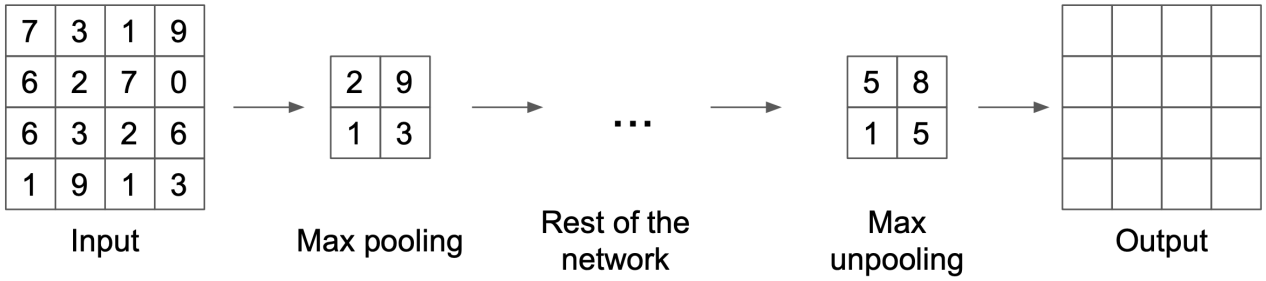

Figure 2: Unpooling

(a) The solution is the result in Figure 3.

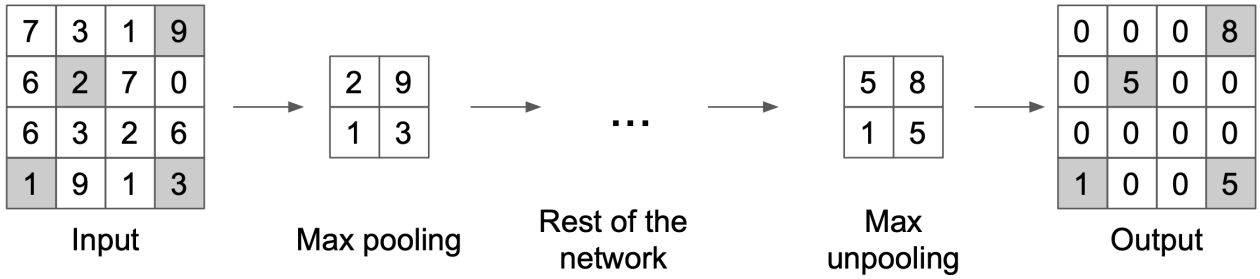

Figure 3: Unpooling solution

Note: The max pooling operation was wrong in the exam, here is the corrected version in Figure 4.

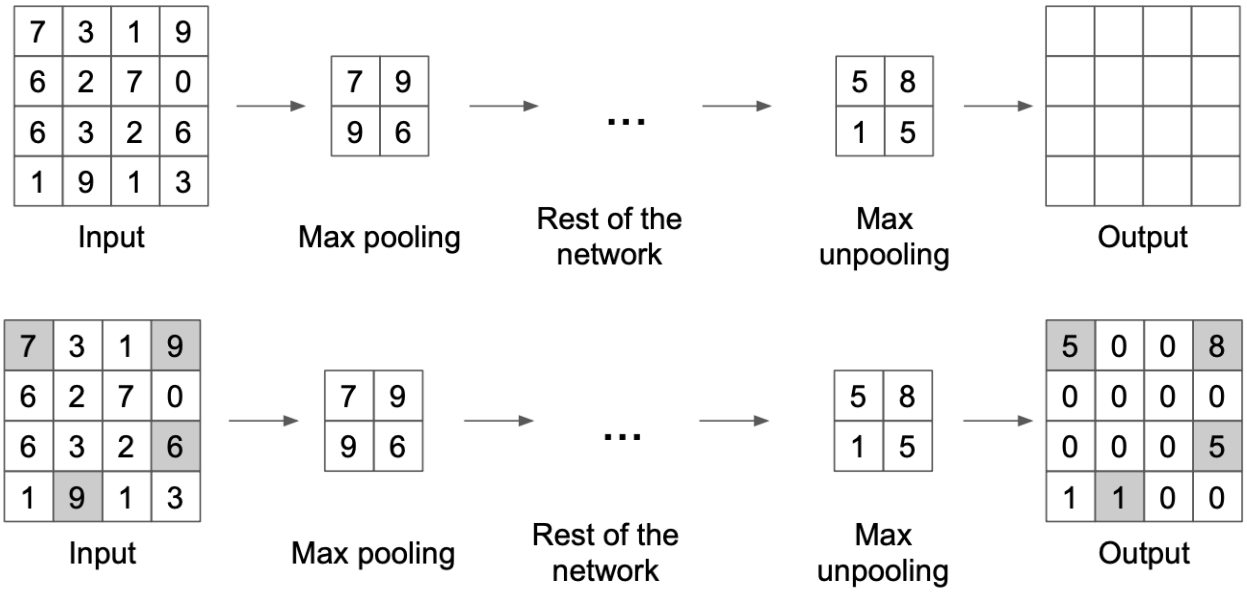

Figure 4: Unpooling exercise corrected with its solution

(b) SegNet uses the max unpooling technique.

7. What is an skip connection and why is it used in semantic segmentation? **0.25 points** Name two networks that use skip connections. **0.25 points**

   A skip connection is a connection between two layers that are not consecutive. I.e., it skips some layers in between. While in ResNets are used to reduce the problem of vanishing gradients, in semantic segmentation these skip connections are used for enhancing the resolution of the output. Since the input size and the output size should be the same and the encoding network has max poolings, there is need of adding upsamplings. The skip connections usually connect the layers from the encoder with their counterpart of the same resolution in the decoder. FCN, UNet use skip connections. Networks based on ResNet use skip connections which are residual connections.

8. Explain Mask RCNN. Include in the explanation what problems can be solved with MaskRCNN, the used backbone, the RPN network, the ROI alignment, the output heads, and a sketch of the network.

   Mask RCNN is a deep neural network aimed to solve instance segmentation problem in machine learning or computer vision. In other words, it can separate different objects in a image or a video. You give it a image, it gives you the object bounding boxes, classes and masks.

   There are two stages of Mask RCNN. First, it generates proposals about the regions where there might be an object based on the input image using the a region proposal network (RPN). Second, it predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposal. Both stages are connected to the backbone structure which is a Feature Pyramid Network (FPN) style deep neural network. It consists of a bottom-up pathway, a top-bottom pathway and lateral connections.

   The ROIAlign is a trick to locate the relevant areas of feature map in the FPN that correspond to each ROI.

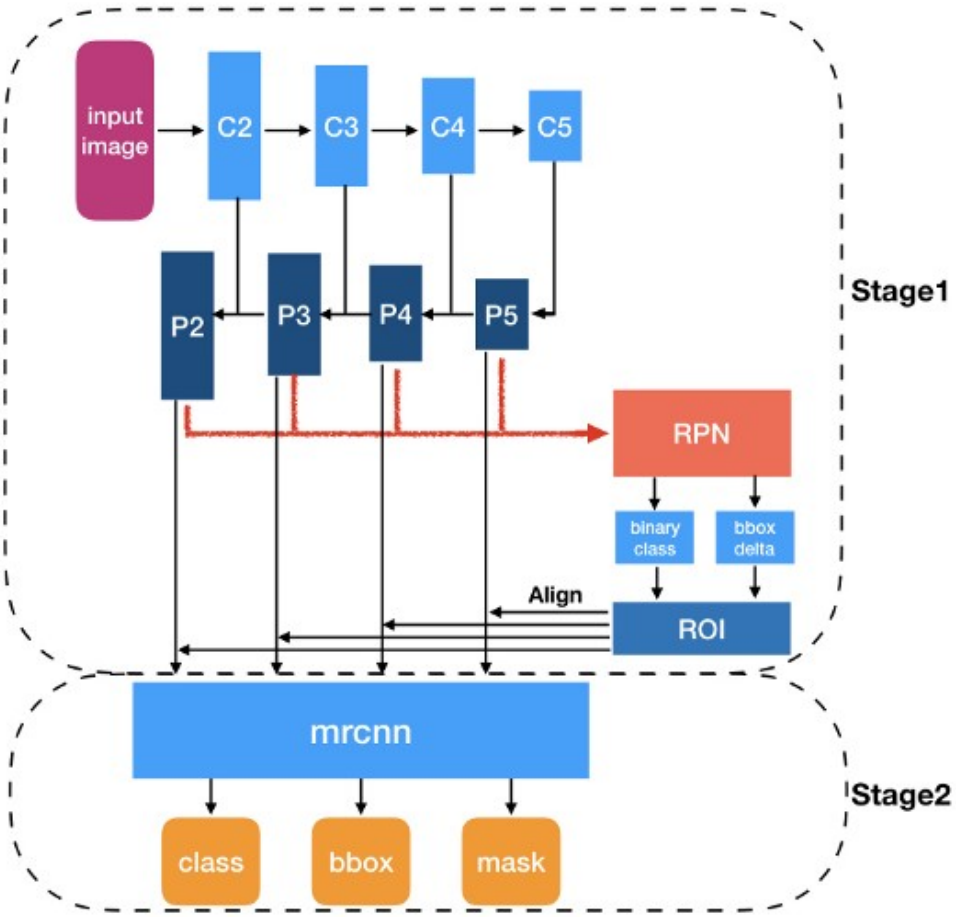

Figure 5: Sketch of the MaskRCNN architecture

9. Which of the following statements are true. Indicate all the true ones.

   (a) Semantic segmentation consists of labelling every pixel of the input image with its class label.

   (b) Instance segmentation consists of labelling every pixel of the input image with its class label and its object id.

   (c) Panoptic segmentation mixes semantic and instance segmentation. It labels with classes the staff and with labels and object ids the things.

   (d) Amodal segmentation mixes natural language processing and instance segmentation. In amodal segmentation the object to segment is expressed by a sentence.

   (e) Referring image segmentation segments the visible and not visible pixels of an image that correspond to an object.

   (a)(b)(c)

10. Match the methods from the first list (a-e) with its description from the second list(1-5). Please, answer like a-1, b-2, c-3 . . . not like 1-a, 2-b . . . , and in a separate sheet, not here.

    (a) Path aggregation

    (b) Mask Scoring RCNN

    (c) TensorMask

    (d) RetinaMask

(e) YOLACT

(1) proposes to use dense sliding window instead of region proposal networks.

(2) is a realtime instance segmentation network which combine the prototypes with the mask coefficients.

(3) calibrates the misalignment between mask quality and mask score including a network block to learn the quality of the predicted instance masks.

(4) improves MaskRCNN by boosting information flow from lower layers.

(5) extends RetinaNet single-shot detector for instance segmentation.

- (a)(4) - **Path aggregation network** improves MaskRCNN by boosting information flow from lower layers.
- (b)(3) - **Mask Scoring RCNN** calibrates the misalignment between mask quality and mask score including a network block to learn the quality of the predicted instance masks.
- (c)(1) - **TensorMask** proposes to use dense sliding window instead of region proposal networks.
- (d)(5) - **RetinaMask** extends RetinaNet single-shot detector for instance segmentation.
- (e)(2) - **YOLACT** is a realtime instance segmentation network which combine the prototypes with the mask coefficients.

Deep metric learning

11. Consider a Siamese network implemented as two identical branches with shared weights. Let be

- $f(x)$ the d-dimensional vector produced by a branch of the network
- x_1, x_2 inputs to branches
- $D = ||f(x_1) - f(x_2)||_2$
- $y = 0$ if x_1 is considered similar to x_2 and 1 if not
- m a certain given value

(a) What is the meaning of the following loss ? Explain it. How is it called ? **0.25 points**

$$L(x_1, x_2, y) = (1 - y)\, D^2 + y\, (\max(0, m - D))^2$$

(b) Draw a figure showing three cases: $L > 0, y = 0$, $L > 0, y = 1$ and $L = 0, y = 1$. **0.25 points**

a) Contrastive loss.

If x_1 and x_2 are considered similar (for instance, belong to the same class) then the loss is their squared Euclidean distance: the more apart they are, the higher the loss, so this loss tends to pull together samples which are considered similar.

If x_1 and x_2 are dissimilar and their Euclidean distance is less than m, the margin, then the loss is the positive difference between the margin and the distance. If distance is larger than the margin the loss is zero. Therefore this loss pushes away pairs of samples of different class.

b) figure of slides, page 20. Figure 6.

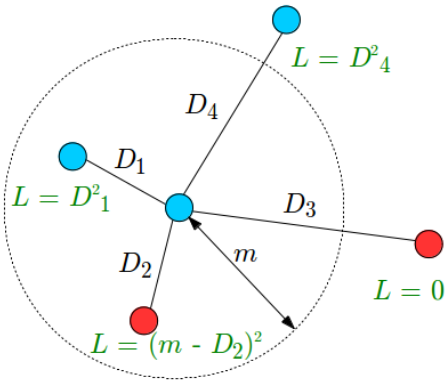

Figure 6: Contrastive loss.

12. Consider again a Siamese network made of two branches with shared weights. For a minibatch of n samples we obtain $n/2$ pairs, being each one either positive or negative ($y = 0$ or 1). Each pair is one term of the loss. Is there a way to obtain more pairs from the same minibatch ? How ? Why is it better to do it this way ?

    Instead of two branches we have just one. Pairs are made at the embedding space, that is, we first process all the n samples (pass them through the network) to obtain $f(x_1) \dots f(x_n)$ and *then* make the pairs of transformed samples. In total we can obtain $n(n-1)/2$ pairs. Each pair is a term of the loss. More terms is better, in principle, to estimate the gradient. This is called online contrastive loss. Explained in slide 48.

Multimodal deep learning

13. How can we increase the number of classes of a classifier using a word embedding? What architecture and type of loss would you use?

    We can map the class labels and visual features to the same shared space and align the representations with pairs using contrastive loss, ranking loss, classification loss or canonical correlation analysis. For example, using the word embedding space as shared space, we can simply use a trainable linear layer to map visual features to the word embedding space, while keeping the (pretrained) visual feature extractor and word embedding mapping frozen.

14. Figure 7 shows the multimodal attention mechanism of a visual question answering system. Generating the attention maps requires looping over the spatial locations of the visual feature. How would you modify the architecture to avoid the loop operation and extract the attention map in a single feed forward pass?

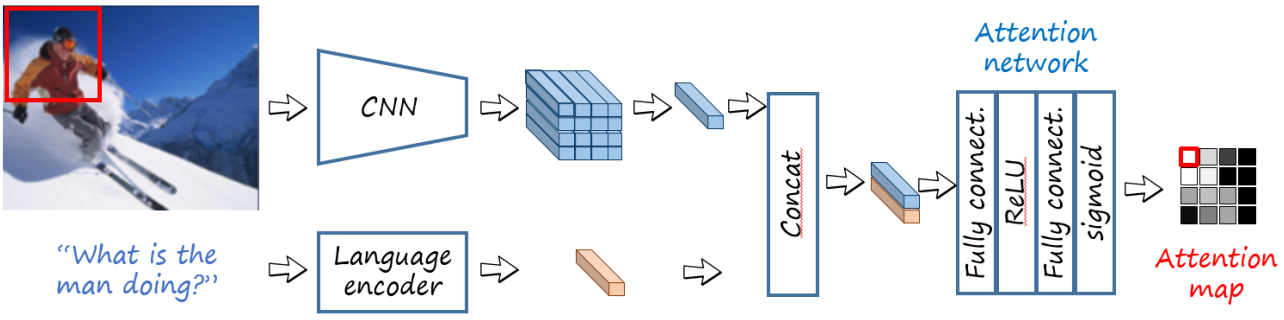

Figure 7: R-CNN object detection system overview.

Use tiling over the spatial locations to replicate the text feature vector into a text feature tensor with the same spatial dimensions of the visual feature. In this way we can concatenate the visual and text feature tensors over the channel dimension. Reshape the fully connected layers of the multilayer perceptron as 1×1 convolutions.

Transfer learning

15. In Multi-task learning (MTL):

    (a) We train a model to solve several tasks on the same image domain

    (b) We train a model trained on one task adapting it to another task(s)

    (c) Each of the multiple tasks has different domain

    (d) MTL means we have a multi-label classification problem.

    (e) MTL performs inductive knowledge transfer. It achieves high generalization by sharing the domain information between different tasks.

    a, e

16. Self-taught Clustering is a kind of Transfer Learning where:

    (a) We have a lot of unlabeled data in the source domain and a few unlabeled data in the target domain

    (b) We train through labeled data in the source domain how to cluster data in the target domain

    (c) The source and target domain share some common features to help the clustering in the target domain

    (d) The purpose is to extend the information theoretic co-clustering algorithm for transfer learning

    (e) We apply the hierarchical clustering but on the output of the layer before the softmax of the neural network model.

    a, c, d

17. What is Negative Transfer?

    Most approaches to transfer learning assume transferring knowledge across domains be always positive. However, in some cases, when two tasks are too dissimilar, brute-force transfer may even hurt the performance of the target task, which is called negative transfer. The negative transfer field studies how to measure relatedness among tasks. Still how to design a mechanism to avoid negative transfer is open problem to be deeply studied theoretically and practically.

Graph neural networks

18. What is the key idea behind graph attention networks? Write how the message (aggregate function) is computed and contrast with GCN's message definition.

    The key idea behind GAT is to weigh the neighbors influence at aggregation time. GAT does so by learning attention coefficients. The message is defined as $\sum_{v \in \mathcal{N}(u) \cup \{u\}} \alpha_{uv} \mathbf{H}_v^{(k-1)}$, where α_{uv} is a learnable weight computed by an attention model. By contrast, GCNs learn a weight to be applied to all neighbors, and only control the neighbors influence by applying a handcrafted normalization that down-weighs information coming from neighbors with higher degrees.

19. Select all true statements (if any). When dealing with graph-structured data:

    (a) Node embedding methods learn structural information, which is subsequently used for tasks such as node classification.

    (b) Structural information is not useful for any node classification task.

    (c) Structural information may be learnt by e.g. pushing node representations to be predictive of the neighboring nodes.

    (d) Flattening the adjacency matrix of a graph and feeding it to an MLP is an effective way to leverage structural information which ensures permutation equivariance.

    (e) In order to leverage graph structural information, one should use convolutional neural networks out of the box.

    a,c

20. Select all true statements (if any). GNN message passing updates ...

    (a) address over-smoothing by leveraging attention.

    (b) never contain information from reaches beyond the first-hop neighbor, even as iterations progress.

    (c) are only defined in node embedding methods.

    (d) are composed of a function aggregating neighborhood information and an update function transforming the node features.

    (e) leverage node features but do not aggregate any neighborhood information.

    d

Reinforcement learning

21. Define MDP and contrast with POMDP. What is the main difference between the two?

    MDPs provide a framework to formally describe an environment. MDPs are defined by a tupe composed of:

    - a state space
    - an action space
    - a transition operator (probability of landing on a state given the previous state and action)
    - a reward function mapping state-action pairs to a real number
    - a discount factor determining how much to weigh future rewards.

    PoMDPs are a generalization to MDPs, they are defined by a tuple containing the same information as MDPs but also require defining

    - an observation space
    - an emission function

    In POMDPs, the agent only indirectly observes the environment.

22. Select all true statements (if any). In reinforcement learning (RL):

    (a) We face the exploration vs. exploitation fundamental dilemma.

(b) Exploration allows us to gather new information by taking non-greedy actions.

(c) Exploration consists on always choosing the best action to perform.

(d) Exploitation is not part of the fundamental dilemma in reinforcement learning.

(e) Exploitation always chooses the best action to perform given the current information.

a,b,e

23. Select all true statements about RL (if any).

(a) In RL, the environment is always known by the agent.

(b) RL problems can be easily posed as supervised problems because decisions are isolated and do not affect future decisions.

(c) RL is concerned with the problem of sequential decision making, the goal is to pick the sequence of actions to maximize future reward.

(d) RL agents are always myopic.

(e) In RL, the agent performs trial and error learning.

c, e

Generative models: GANs and VAEs

24. Which of the following are latent variable models:

(a) Variational Autoencoders

(b) Autoregressive models

(c) Generative Adversarial Networks

(d) AutoEncoders

(e) None

a, c, d

25. Which of the models can compute exact likelihood:

(a) Variational Autoencoders

(b) Autoregressive models

(c) Generative Adversarial Networks

(d) PixelCNN

(e) None

b, d

26. Select the correct statements about the reparametrization trick in Variational Autoencoders:

(a) It is used at inference time

(b) It enables gradient propagation towards the inference model

(c) It shifts the output of the inference model by performing one multiplication and one addition

(d) It is used to incorporate stochasticity in the model

(e) None of the above is correct

b, c, d

27. Which of the following approaches are valid ones to evaluate Generative Adversarial Networks:

    (a) Likelihood

    (b) Frechet Inception Distance

    (c) Visual inspections of the model samples

    (d) Mean Squared Error

    (e) Inception Score

    b, c, e

28. Explain mode collapse issue in Generative Adversarial Networks.

    Mode collapse happens when the generator learns to model only a sub-space of data distribution. For example, the model can learn to generate only particular instances of given classes, failing to model the full data diversity.

29. When talking about Variational Autoencoders we introduced Evidence Lower Bound (ELBO). Explain in your own words what the concept of tightness of ELBO means.

    The tightness of ELBO is a concept that measures how well the variational distribution approximates the true posterior.