Exam module 5 : Visual recognition
Date: May 3rd, 2021
Time: 1.5 hours

Note: questions 3-6 and 26-29 belong to subjects not covered in course 2019-20.

Answer all questions right into these sheets. Each multiple choice question is worth 0.25 points. Wrong answers will not subtract any point. Short answer questions are worth up to 0.5 points (0, 0.25 or 0.5). **Write your name on this and all the answer sheets**.

1. In the context of metric learning, $L = \max\{0, m + d(a,p) - d(a,n)\}$ is the loss for a triplet network, where $d(x_1, x_2) = ||f(x_1) - f(x_2)||_2$ (Euclidean distance) and $f(x_1), f(x_2)$ denote the output of the network for the two inputs x_1, x_2. a) What's the meaning of a, p, n and consequently of this loss ? Please make also a drawing to explain the different cases of zero and non-zero loss. b) What's the difference between this loss and the contrastive loss of Siamese networks in terms of the samples ?

   see slides number 28-30

2. Again in metric learning, a) what is mining and why do it ? b) what is a hard positive pair for a Siamese network ?

   see slides number 44,45

3. Choose the false sentence referred to CUDA programming:

   (a) A CUDA program should distinguish between host and device memory resources.

   (b) The keyword `__global__` indicates that the next procedure is going to be executed on the device and it will be called from the host.

   (c) Memory allocated using the `__shared__` keyword can be shared equally among host and device code.

   (d) Host and device memory are separate entities. One special case is `__shared__` memory, which is fast on-chip device memory allocated per block and not visible to threads in other blocks.

   (e) The `__device__` keyword indicates that the following procedure or function is going to be executed on the device.

   the right choice is: Memory allocated using the `__shared__` keyword can be shared equally among host and device code.

4. Select the sentence making a false statement:

   (a) Convolutions in the Fast Fourier transformed domain (FFT) are more efficient when the filter dimensions are large enough with respect to the input data dimensions.

(b) cuDNN is a GPU-accelerated library of primitives for Deep Neural Networks providing highly tuned implementations of routines arising frequently in DNN applications.

(c) Winograd algorithms for convolutions are always faster than FFT, GEMM and Direct Convolution methods because they require less memory resources.

(d) The cuDNN context is associated with only one GPU device. However multiple contexts can be created on the same GPU device.

(e) Lowering convolutions to matrix multiplications (GEMM) becomes less worth as the product of the input and kernel tensor dimensions becomes smaller.

the right choice is: Winograd algorithms for convolutions are always faster than FFT, GEMM and Direct Convolution methods because they require less memory resources.

5. A common approach to implement DNNs is to recast the most computationally expensive operations as general matrix multiplications (GEMM). Select the right answer that completes the missing elements for the following GEMM transformation :

`Image data:` $\begin{pmatrix} d_0 & d_1 & d_2 \\ d_3 & d_4 & d_5 \\ d_6 & d_7 & d_8 \end{pmatrix}$

`Filter data:` $\begin{pmatrix} f_0 & f_1 \\ f_2 & f_3 \end{pmatrix}, \begin{pmatrix} g_0 & g_1 \\ g_2 & g_3 \end{pmatrix}$

`GEMM transformation:` $\begin{pmatrix} f_0 & f_1 & \star & \circ \\ g_0 & g_1 & \bullet & \triangle \end{pmatrix} \times \begin{pmatrix} d_4 & d_5 & d_7 & d_8 \\ d_3 & d_4 & d_6 & d_7 \\ d_1 & \diamond & d_4 & d_5 \\ d_0 & d_1 & d_3 & d_4 \end{pmatrix} = \begin{pmatrix} o_0 & o_1 & o_2 & o_3 \\ o_4 & o_5 & o_6 & o_7 \end{pmatrix}$

(a) $\star = g_0, \quad \circ = g_1, \quad \bullet = g_2, \quad \triangle = g_3, \quad \diamond = d_3.$

(b) $\star = f_2, \quad \circ = f_3, \quad \bullet = g_2, \quad \triangle = g_3, \quad \diamond = d_3.$

(c) $\star = f_2, \quad \circ = f_3, \quad \bullet = g_2, \quad \triangle = g_3, \quad \diamond = d_2.$

(d) $\star = g_2, \quad \circ = g_3, \quad \bullet = f_2, \quad \triangle = f_3, \quad \diamond = d_2.$

(e) None of the previous answers are correct.

the right choice is: $\star = f_2, \quad \circ = f_3, \quad \bullet = g_2, \quad \triangle = g_3, \quad \diamond = d_2.$

6. The basic difference between correlation and convolution is that correlation measures the similarity between two signals, while the convolution measures the effect that one signal has on the other signal. Said this, choose the best answer (more complete) from the following options:

(a) The mathematical calculation of correlation is the same as for the convolution in the time domain, except that the signal is not reversed before the multiplication process.

(b) 6a is correct. And thus, if the filter is symmetric then the output of both the expressions would be the same.

(c) 6a is correct. However, even if the filter is symmetric the output of both expressions might be different.

(d) 6a is incorrect. In fact, the GEMM transformation from the previous exercise (Ex. 5) represents a correlation operation, and not a convolution.

(e) None of the previous answers are correct.

the right choice is: 6a is correct. And thus, if the filter is symmetric then the output of both the expressions would be the same.

7. The loss function in a typical single shot object detection model consists in a combination of two independent loss functions, each one optimizing a particular task. Which are these two tasks? Mention one specific example of loss function for each one of them.

   (1) Classification (e.g. Cross-entropy loss), and (2) Localization / Bounding box regression (e.g. L1, smooth L1, L2, etc.)

8. Explain how the Region of Interest pooling (ROI-pooling) layer works, and why is it useful in object detection pipelines.

   (1) Divide the region proposal into equal-sized 2D sections (the number of which is the same as the desired dimension of the output). (2) Find the max value in each section. (3) Copy those max values to the output (fixed-size) buffer. It it useful in object detection pipelines because it allows to reuse the visual features extracted from the CNN backbone for classifying regions proposals independently of their dimensions.

9. Which one(s) do we use as loss function for semantic segmentation?

   (a) Mean Intersection over Union (MIoU) or Jaccard Index
   (b) Global pixel accuracy
   (c) Pixel-wise cross entropy
   (d) Average per-class accuracy
   (e) All of the above

   c

10. What is a dilated convolution? Which of the studied works did introduce? Why is it useful?

    see slide number 49

    Dilated convolutions are equivalent to normal convolutions but with dilated filters (like expanding their size while filling the gaps with zeroes). DeepLab introduced it and it is useful as a way to avoid the use of pooling operations. Pooling operations are needed because if they are not used the receptive field becomes too small; however, pooling loses spatial information. Dilated convolutions, thanks to dilated filters, widen the receptive field while avoiding spatial resolution coarsening

11. The goal of weakly supervised segmentation is to...

    (a) integrate global context with local information to improve accuracy
    (b) take advantage of temporal information to improve consistency
    (c) skip certain training stages to improve efficiency
    (d) exploit coarse labels (e.g., bounding boxes) to get around heavy labeling efforts
    (e) none of the above

    d

12. Select all correct sentences about multimodal representations.

    (a) Joint representations use fusion mechanisms to combine modalities.
    (b) Coordinated representations share a common representation space for all modalities.
    (c) Joint representations enable the combination of heterogeneous modalities (e.g. image, text).
    (d) Coordinated representations are more efficient than joint representations.

(e) Multimodal representations cannot represent sequential data.

a and c

13. Select all correct sentences.

    (a) The dimension of a word embedding representation is the same as the corresponding one-hot representation.

    (b) The space of word embeddings is continuous.

    (c) Distances between one-hot representations can encode rich semantic relations.

    (d) Word embeddings are sparse.

    (e) One-hot representations are sparse.

    b and e

14. Describe the typical architecture of an image captioning system. Specify the type of networks used.

    An image encoder, typically implemented as a convolutional neural network, followed by a language/text decoder, typically implemented as a recurrent neural network.

15. Which (possibly several) sentences are correct about conditional GANs and unconditional GANs?

    (a) Conditional GANs can be used to control the class and attributes of the generated images.

    (b) Conditions are limited to labels and attributes.

    (c) The condition is applied only to the generator.

    (d) Text-to-image synthesis can be implemented as a conditional GAN.

    (e) None of the above.

    a and d

16. What is the idea of cycle consistency for unpaired image-to-image translation? What loss uses CycleGAN?

    The image translated from one domain to the other and then translated back to the original domain (i.e. a cycle) should ideally map to the input image. In CycleGAN and similar methods each direction is implemented with a separate generator/translator (e.g. $y = G(x)$ and $x = F(y)$).

    The cycle consistency loss penalizes the reconstruction error between reconstructed image $\hat{x} = F(G(x))$ and input image x (analogously for $\hat{y} = G(F(y))$ and y).

    Note: not necessary to explicitly include more specific details (e.g. that the loss in CycleGAN is $\|F(G(x) - x)\|_1 + \|G(F(x) - x)\|_1$)

17. Which (possibly several) sentences about diversity in image-to-image translation are correct?

    (a) pix2pix can directly generate diverse outputs.

    (b) Adding noise to the input image and using pix2pix generates diverse images.

    (c) Diversity can be achieved by forcing invertibility to reconstruct the random latent code.

    (d) MUNIT disentangles code and style latent codes, where sampling the style latent code provides diversity.

(e) CycleGAN can directly generate diverse outputs.

c and d

18. Image generation falls within a category of:

    (a) Supervised Learning
    (b) Unsupervised Learning
    (c) Reinforcement Learning
    (d) All of the above
    (e) None of the above

    b

19. All methods discussed in the generative models class:

    (a) Model dataset density explicitly.
    (b) Use latent variables.
    (c) Provide us with sampling mechanisms that allow us to sample from the model.
    (d) b and c.
    (e) None of the above.

    c

20. Autoregressive models for image generation:

    (a) Assume sequential ordering among images in dataset.
    (b) Assume sequential ordering of pixel in the image.
    (c) Assume sequential ordering of channels in the pixel.
    (d) All of the above.
    (e) b and c.

    e

21. Select all correct statements. Variational autoencoders:

    (a) Optimize exact likelihood.
    (b) Use ELBO as model sampling mechanism.
    (c) Use amortized inference to estimate the parameters of the posterior.
    (d) Assume sequential ordering among images in dataset.
    (e) None of the above

    c

22. Select all correct statements. Generative Adversarial Network (GAN):

    (a) Is a latent variable model.
    (b) During training, it uses reparametrization trick.
    (c) Computes lower bound on image likelihood.
    (d) Provides an efficient model sampling mechanism.
    (e) None of the above

a, d

23. Select all correct statements. When evaluating GANs:

    (a) We can compute test set likelihood.
    (b) We can visually analyze the samples.
    (c) We can compute Inception Score.
    (d) We can compute Fréchet Inception Distance.
    (e) We can compute Intersection over Union.

    b, c, d

24. Given two distributions (e. g. two Normal distributions), which of the following allows us to measure the similarity score between the two?

    (a) Fréchet Inception Distance
    (b) Inception Score
    (c) KL divergence
    (d) a and c
    (e) All of the above.

    d

25. We use reparametrization trick to:

    (a) Sample from the model.
    (b) Speed up the model training.
    (c) Remove stochasticity in the forward pass through the model.
    (d) Remove stochasticity from the backward pass through the model (backpropagation).
    (e) None of the above.

    d

26. Recurrent Neural Networks

    (a) are well suited to deal with sequential data.
    (b) contain loops, which allow information to persist.
    (c) may suffer from vanishing/exploding gradients because weights are shared across different time steps.
    (d) All of the above.
    (e) None of the above.

    d

27. LSTM and GRU's gates:

    (a) do not apply any non-linearity to their output.
    (b) apply rectifier (ReLU) non-linearity to their output.
    (c) apply a sigmoid non-linearity to their output.
    (d) apply leaky ReLU non-linearity to their output to ensure gradient flow.

(e) LSTMs and GRUs do not have any gating system.

c

28. HyperNetworks

(a) endow RNNs with a soft attention mechanism to make them fully differentiable.

(b) relax the RNNs' weight sharing by using a second network to predict their parameters.

(c) endow RNNs with an external memory bank, overcoming the problem of modeling long term dependencies.

(d) reduce the number of parameters in RNNs by applying weight sharing among all gates.

(e) None of the above.

b

29. Residual Networks can be reformulated as

(a) shallow RNNs by applying weight sharing across different layers, leading to orders of magnitude less parameters.

(b) deep bidirectional HyperNetworks by applying weight sharing across different layers.

(c) deep RNNs by applying weight sharing across different layers, significantly increasing the number of parameters.

(d) RNNs without any modification, but this usually leads to significant performance drops.

(e) transformer networks, by endowing them with external memory banks.

a

30. Select all true statements. In reinforcement learning:

(a) We care about single and isolated decisions that do not affect the future.

(b) We learn by trial and error, without direct supervision.

(c) Feedback is delayed.

(d) Time does not matter and thus, data follows the i.i.d. assumption.

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

b, c and e

31. A Markov Decision Process (MDP) is defined as:

(a) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}$ the transition operator, r the reward, and $\gamma \in [0, 1]$ the discount factor.

(b) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mathcal{O}, \mathcal{E}\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}$ the transition operator, r the reward, $\gamma \in [0, 1]$ the discount factor, $\mathcal{O}$ the observation space and $\mathcal{E}$ the emission probability.

(c) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}$ the transition operator, r the reward, $\gamma \in \mathbb{R}$ the discount factor.

(d) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mathcal{O}\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{P}$ the transition operator, r the reward, $\gamma \in \mathbb{R}^+$ the discount factor and $\mathcal{O}$ the observation space.

(e) None of the above.

a

32. Which of the following statements about rewards, returns and value functions is correct?

    (a) The reward is a scalar feedback telling us how well we are doing.

    (b) The return is the total (discounted) reward.

    (c) The action-value (Q) is the expected return from taking a given action from a given state.

    (d) By comparing the action-value (Q) to the state-value (V), we can assess which actions are better than average.

    (e) All of the above.

e

33. Reinforcement learning algorithms:

    (a) In value-based algorithms, we directly differentiate the RL objective.

    (b) In policy gradient, we learn a policy by taking gradient steps to improve the return.

    (c) In actor-critic, we have two neural networks: one to learn the policy and a second one to fit a value function.

    (d) DQN overcome the problem posed by non-stationary targets by introducing a replay buffer.

    (e) b and c.

e