ℹ Front page

UNIVERSITY OF OSLO
The Faculty of Mathematics and Natural Sciences

Written examination in TEK5040/9040
2022 Fall
Duration: 30.11.2022, 09:00 to 30.11.2022, 13:00
Permitted aids: all printed and written aids, all kinds of calculators

It is important that you read this cover page carefully before you start.

Information about hand drawings (Scantron)
In this question set you have the opportunity to answer with hand drawings (all questions). You use the handed out sketch sheets. It is possible to use several sheets per assignment. See instructions on how to fill in the sketch sheets in the link below the assignment overview.

Below the taskbar you will find a digital calculator, an instruction for use of sketch sheets, and the following lecture slides for support during your examination:
Lecture_slides_intro_cnn_tflow_rnn_attention,
lecture_slides_Reinforcement_learning,
lecture_slides_learning_concepts,
lecture_slides_3D_processing_and_tracking,
lecture_slides_bayesian_and_sequence_modeling,
lecture_slides_control_gan_diffusion_models

1 Field of View

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Assume that you apply two convolutional layers consecutively to an input image x. The first convolutional layer has a kernel size 5 and stride 1 in both height and width dimension. The second convolutional layer has a kernel size of k in both height and width dimensions. We define the *field of view* of a neuron to be the pixels in the input image x that may affect the output of the neuron. The field of view for a neuron in the first convolutional layer forms a rectangle of size 5×5.

The size of the (rectangular-shaped) field of view for a neuron in the second convolutional layer is required to be 8 in both height and width dimensions. What is the value of the kernel size k of the second convolutional layer?

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

Maximum marks: 12

2 Self-Attention

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Self-attention based networks are widely used in sequence modelling.

A self-attention layer is fed with a sequence of three word vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ in the given order. The corresponding output sequence is $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. Then the input order is changed to $\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_1$. Write down the output sequence corresponding to the new input sequence. Give a brief explanation to your answer.

One disadvantage of self-attention is the high computational complexity for long sequences. How can you remedy this disadvantage? State a disadvantage of the remedy itself.

Fill in your answer here

Format

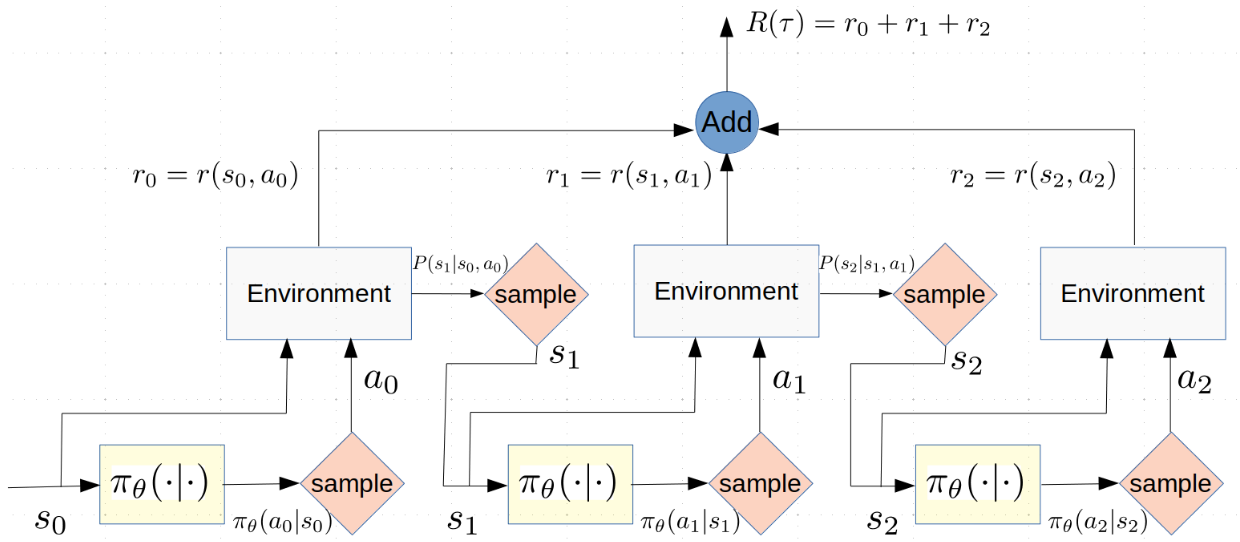
Words: 0

Maximum marks: 12

3 Policy Gradients

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

The following figure shows a roll-out of a reinforcement learning system for an episode of three time steps, $t = 0, 1, 2$. At each time step, the policy $\pi_\theta(\cdot|\cdot)$ generates the action probability distribution $\pi_\theta(a_t|s_t)$ and an action a_t is sampled. We assume that a policy is implemented using a neural network (policy network) with parameters θ. The environment generates state transition distribution $P(s_{t+1}|s_t, a_t)$ from which the next state s_{t+1} is sampled. At the same time, the environment generates a reward $r_t = r(s_t, a_t)$. Assume both the state space and the action space are discrete. For a given trajectory $\tau = \{s_0, a_0, s_1, a_1, s_2, a_2\}$, we can calculate the total reward $R(\tau) = r_1 + r_2 + r_3$.



A student tries to estimate the policy network parameters by optimizing $R(\tau)$ with respect to θ through gradient descent. Therefore he tries to back-propagate gradients of $R(\tau)$ through the above network. Explain why this approach is unlikely to be successful.

A better approach would be to maximize the expected reward $J(\theta) = E_{\tau\sim\theta}[R(\tau)]$, where $\tau \sim \theta$ means that the trajectory τ is sampled with a policy π_θ parameterized by θ . It can be shown that $\nabla_\theta J(\theta) = E_{\tau\sim\theta} \sum_{t=1}^{T} [\nabla_\theta \log \pi_\theta(a_t|s_t) R(\tau)]$ where T is the length of the trajectory τ. In this case, what is the effective loss we need to back-propagate through the policy network?

How can you interpret the gradients calculated in this way ?
(Hint: compare this with the cross-entropy loss in supervised learning.)

Fill in your answer here

Format

Words: 0

Maximum marks: 14

4 Generative Adverserial Networks

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Assume that we are given a data-set where each data example consists of an image and the corresponding single word description. For example, a data example can be an image of a car together with the word vector representing the word "car". There can be several examples containing the same description word, but paired with different images. For example, there can be several different images of cars but always coupled with the word vector "car".

Describe a suitable approach for training a generative adversarial network (GAN) which can be used to generate an image corresponding to a given description word. Draw a diagram to support your description.

In using the trained GAN above, how can you generate a set of diverse images corresponding to a given description word.

What can be an advantage of a diffusion model (more precisely denoising diffusion model - DDM) over a GAN in the above task?

Fill in your answer here

Format

Words: 0

Maximum marks: 12

5(a) Meta-Learning

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Explain what is meant by k-class n-shot learning.

Give an example of a 5-class one shot learning problem. Describe the support set and the query set in both meta-training and meta-testing phases of your example.

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

Maximum marks: 10

5(b) Active Learning

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Briefly describe how the most "valuable" samples are predicted in active learning.

Fill in your answer here

Format

Σ

Words: 0

Maximum marks: 2

6(a) Bayesian Deep Learning

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Let $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$ be a training data-set where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are respectively a sequence of input vectors and the corresponding outputs of a neural network parameterized by $\boldsymbol{w}$. In Bayesian deep learning we try to estimate the variance of $p(\hat{y}|\hat{x}, \boldsymbol{w})$ for a given input $\hat{x}$ and an output $\hat{y}$.

How can we use this variance in safety-critical applications?

Calculation of the variance above involves evaluating integrals of the form $I = \int F(\boldsymbol{w}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}$, where $F(\boldsymbol{w})$ is a function of $\boldsymbol{w}$. Explain why it is hard to evaluate this integral in practical deep learning. (Hint: Focus on $p(\boldsymbol{w}|\mathcal{D})$)

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | Ω | Σ |

Words: 0

Maximum marks: 6

6(b) Variational Inference

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Outline how variational inference can be applied to estimate the integral described in Question 6(a), $I = \int F(\boldsymbol{w}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}$. (Hint: Avoid details, focus on the main steps)

Fill in your answer here

Format

Σ

Words: 0

Maximum marks: 6

7(a) Inverse Reinforcement Learning

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Inverse Reinforcement Learning (IRL) has been proposed as a remedy for the difficulty in manually designing complex reward functions. Assume that a set of expert demonstrations $\mathcal{D} = \{\tau_1, \tau_2, \cdots, \tau_M\}$ has been collected where each τ_i for $i = 1, 2, \cdots, M$ is a sequence of state-action pairs. The reward function $r_\psi(\cdot, \cdot)$ is implemented as a neural network with parameters ψ, and hence each trajectory τ_i can be assigned a total reward $R_\psi(\tau_i)$.

What is the objective function of maximum entropy inverse reinforcement learning?

Fill in your answer here

Format B I U x₂ x² Iₓ Ω Σ

Words: 0

Maximum marks: 4

7(b) Behaviour Cloning

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

State two weaknesses of behavior cloning applied to control learning.

A student claims that reinforcement learning is the best approach to control learning. Briefly discuss two arguments against and two arguments in favor of her claim.

Fill in your answer here

Format

Σ

Words: 0

Maximum marks: 8

8(a) 3D Processing

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

A student is given a set of point clouds and each point cloud in the set consists of N four-dimensional points. Each point is represented by the spatial co-ordinates x, y, z and the reflectivity r. The student is asked to design a network to perform segmentation on any given point cloud from the set. He designs a graph neural network based on a series of graph convolution layers.
What is the likely dimension of the feature vectors of the initial nodes?

Fill in your answer here

Format | B I U x_2 x^2 | I_x | | | Ω | |
Σ |

Words: 0

Maximum marks: 3

8(b) 3D Processing

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Each layer of the graph neural network in Question 8(a) can be described by the following general operation:

$$\boldsymbol{x}_i' = \square_{j \in \mathcal{E}_i} \boldsymbol{h}_{\boldsymbol{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where $\boldsymbol{x}_i$ is the feature vector of the concerned node, $\boldsymbol{x}_j$ is the feature vector of a neighbor node, $\mathcal{E}_i$ is the set of neighbor nodes of node i, $\boldsymbol{h}_{\boldsymbol{\Theta}}$ is the edge function with parameters $\boldsymbol{\Theta}$ and $\boldsymbol{x}_i'$ is the updated feature vector of the concerned node. The aggregation operation $\square$ is assumed to be symmetrical.

If the student uses an edge function $\boldsymbol{h}_{\boldsymbol{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{h}_{\boldsymbol{\Theta}}(\boldsymbol{x}_i)$ and hence drops the aggregation operation, what disadvantage would the resulting network have?

Another student uses the edge function $\boldsymbol{h}_{\boldsymbol{\Theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{h}_{\boldsymbol{\Theta}}(\boldsymbol{x}_j)$ instead. State one advantage and one disadvantage of this choice.

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

8(c) Multi-object Tracking

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

A graph neural network $\mathcal{G}_t$ can be used in multi-object tracking to solve the association problem. In this case, each detection is represented by a node and possible associations are represented by edges.

Describe briefly what initial node features and initial edge features represent.

The graph neural network $\mathcal{G}_s$ in Questions 8(a) and 8(b) uses only node features (i.e. no edge features). Explain why explicit edge features are not essential for $\mathcal{G}_s$, whereas they are desirable for $\mathcal{G}_t$.

Fill in your answer here

Format B I U x_2 x^2 I_x Ω Σ

Words: 0

Maximum marks: 4

8(d) Single Object Tracking

In this exercise you can answer with digital hand drawing. Use your own sketch sheet (distributed). See instructions for filling in the sketch sheet in the link below the task bar.

Explain why a fully connected siamese neural network used for single object tracking is faster than the MDNet.

Fill in your answer here

Format

Words: 0

Maximum marks: 3