UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	IN 54000/IN 9400 — Machine Learning for Image Analysis
Day of examination:	June 3 2020
Examination hours:	9.00 June 3−9.00 June 10

This exercise set consists of 7 pages.

Appendices:	None
Permitted aids:	All

Read the entire exercise text before you start solving the exercises. Please check that the exam paper is complete. If you lack information in the exam text or think that some information is missing, you may make your own assumptions, as long as they are not contradictory to the "spirit" of the exercise. In such a case, you should make it clear what assumptions you have made.

Read the following link for important exam information: https://www.mn.uio.no/english/about/hse/corona/examination-2020.html

You will submit a single PDF file. In this file, the exercises should be answered in given order. Any figures or drawings MUST be included at the correct place in the document. If you make sketches or computations on paper, include photo of them in good quality and with good resolution, and place the photo under the appropriate subtask.
If you use any source different from the course material to answer a question, include full reference to the source in the answer of the corresponding subtask. Use proper styles for citing wepages or papers.
Most of your answers should include a discussion, typically a few sentences.

Every subtask has equal weight in the evaluation.

(Continued on page 2.)

Exercise 1 Computations for a simple network

You are given a simple feedforward network with one hidden layer with 2 nodes, activation function sin(z), and a binary classification problem. The output layer uses sigmoid activations. The input vector to the network is $x = [x_1, x_2, x_3]$.

1a

Make a sketch of a computational graph for the network.

1b

Let $x_1 = 2$, $x_2 = 1$, $x_3 = 3$
We assume the initial values of the weights are: $W_{11}^{[1]} = 1$, $W_{12}^{[1]} = 2$, $W_{13}^{[1]} = 1$, $W_{11}^{[2]} = 3$, $W_{12}^{[2]} = 4$.

Compute the predicted value $\hat{y}$. Show your computations.

1c

If we use a logistic cost functon, and the true output for the single sample is $y = 1$, compute the first update of $W_{13}^{[1]}$ if we use gradient descent with a learning rate of 0.2.

Include all your computations.

1d

Discuss if sin(z) would be a good activation function. Give your views on some properties that an activation function should have.

1e

Explain why the range or standard deviation of weight in a layer in a neural network should depend on the size of the input to the layer.

Exercise 2 Convolutional neural networks - 2020

2a

Explain briefly why fully conntected networks do not work well for image classification.

(Continued on page 3.)

2b

Argue whether the effective receptive field (field of view) is larger or smaller than the theoretical receptive field.

We can count the number of paths (connections) from the input data to a neuron. Input data located right below a neuron have more paths compared to input data located at the edge of the receptive field. The number of paths can be viewed as a weighting function, hence the effective receptive field will be smaller than the theoretical receptive field.

2c

Discuss whether a CNN is efficient for detecting long range dependencies in an image.

We typically need many CNN layers to be able to connect distant information within an image. We can use pooling/stride and dilation to faster increase the receptive field, but in general CNN's are inefficient for detecting long range dependencies.

2d

Consider the model architecture consisting of a CNN followed by two dense (fully-connected) layers. Discuss whether the model is invariant to translation.

Anne: use either c or d, but not both?

CNNs are translational invariant assuming no padding. Dense layers are however not transnational invariant, and will break the models translational invariance.

2e

Consider the model architecture consisting of a CNN followed by averaging over the spatial dimensions (global average pooling). Briefly discuss if this model is invariant to translation.

Yes

Exercise 3 Recurrent neural networks - 2020

Anne: cut a?

3a

Why are RNNs slow to train, and why might it be faster to use a CPU compared to a GPU for training/inference?

(Continued on page 4.)

RNNs uses sequential computation. GPUs are powerful for parallelized computation, but runs with a slow clock speed compared to CPUs. For this reason, sequential computation may be faster om a CPU.

Anne: work on better formulation?

3b

How could you structure a CNN to be used for time series (1d) with a stream of incoming data?

We can use temporal convolutions (causal convolutions). The output is computed from past data only. No future information used.

3c

Discuss when you would consider to use a bi-directional RNN.

Bi-directional RNNs can be useful when you have access to the past and the future, and the prediction depends on both.

Exercise 4 Statistical properties of classification

4a

Discuss briefly challenges associated with a binary classification problem where you have labelled data from 9900 samples for class 1 and 100 samples for class 2.

4b

Explain what a minibatch is, and discuss any potential drawbacks of using either a small or a large minibatch size.

4c

Suggest a measure to compute similarities between two distributions, and give two examples of how we have used this in the course.

Exercise 5 Generalization - 2020

5a

Discuss what to expect if the training and test data is drawn from different distributions (processes).

You should expect the model to not generalize well, or at least you have no guarantee of on the generalization error.

(Continued on page 5.)

5b

Discuss the use of the data sets (training, validation, test).

We use the training set to adjust model parameters such as the weights in the neural network. The validation set is used for hyper-parameter tuning, and the test set is used to get an estimate of the out-of-sample error (real world error).

5c

What can be a drawback of extreme hyper-parameter tuning?

Extreme hyper-parameter tuning can result in over-fitting on the validation data set.

5d

Discuss whether multitask learning is a reasonable regularizer compared to weight decay (L2).

Both multitask learning and weight decay limits the number of network weights configurations. The reduction in likely hypothesis can thus reduce the generalization error. Weight decay has the effect of lowering the values of the network weights and making the decision boundary smoother. Forcing the network weights to be small may not always be beneficial. Training the model simultaneously on related tasks can infer more general features which hence reduce the generalization error.

Exercise 6 Selected topics

6a

PhD students only Find a paper that compares ReLU and LeakyRelu. Describe briefly their findings, the data set, and the difference in performance the paper reports.

6b

Describe an example from the course on how we can use upsampling in convolutional networks.

6c

Select a method from the course for visualizing or interpreting what a trained network has learned, and describe the method briefly.

(Continued on page 6.)

6d

Describe briefly the difference between an autoencoder and a variational autoencoder.

6e

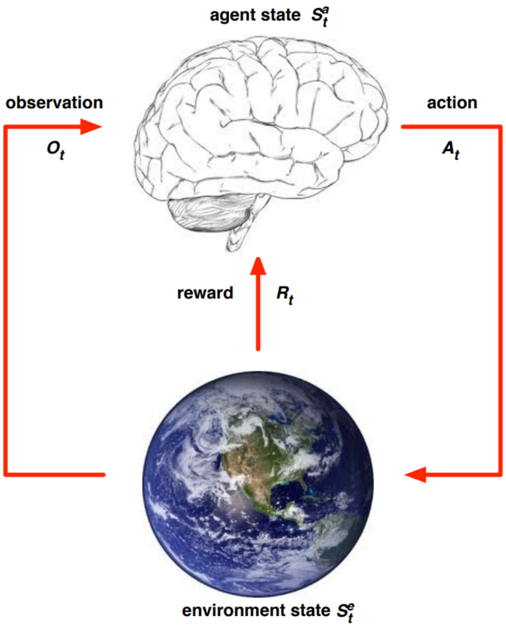
Find an application of deep learning that in your opinion raises some ethical aspects. Describe the application and discuss the ethical concerns.

Exercise 7 Reinforcement learning- 2020

Anne: cut 1-2 of these

7a

Explain the quantities, S_t^a, A_t, S_t^e, R_t, and O_t in the figure:



- S_t^a - is the agents representation/state/understanding of the world (system). The subscript, t, indicate the current time step t. The agents state can be e.g. a table, a neural network, or the environment state in full observatory systems. The agent base its actions on the agent state.
- A_t - is the action selected by the agent at time step t.
- S_t^e - is the environment state. The environment state is the true configuration of the world/system.

(Continued on page 7.)

- R_t - is the reward sampled by the environment state given by the agents action.
- O_t - is the observation done by the agent of the environment state. The observation can include full knowledge of environment state, but can also be a partical observation.

7b

What can be the benefit of approximating the action-value (q-value) function by a neural network compared to storing the action-values in a table?

If a table were to be used for storing predicted action-values, the model (agent) would not be able to generalize between similar states and actions. Similar state and action pairs would be stored in separate cells within the table. A neural network however can learn that two states are close (similar) and that the agent should perform e.g. the same action.

7c

What does the action-value (q-value) function represent?

The action-value (q-value) represent the expected accumulative reward given a state and an action.

7d

Explain the trade-off between exploration and exploitation.

Reinforcement learning is different than supervised learning as the agent affects the training data itself. The agent selects actions and based on these actions the training data is generated.

In the Bellman (optimality) equation it is evident that the q-value is defined by selecting the future actions with the highest q-value. For an untrained agent, the action with the highest q-value can be wrong and result in the agent only visiting a part of the state-space. For this reason, early in the training phase (of the agent) it is common to "explore" with selecting a random action with probability ϵ.

For a trained agent it is inefficient to keep exploring sub-optimal regions of the state-space. We can exploit the knowledge of the agent to search in already promising regions of the state-space by selecting promising actions.