UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	IN3310/IN4310 — Deep Learning for Image Analysis
Day of examination:	June 4, 2024
Examination hours:	9:00 – 13:00
This problem set consists of 15 pages.	
Appendices:	None
Permitted aids:	None

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

General information:

- Read the entire exercise text before you start solving the exercises. Please check that the exam paper is complete.
- Remember that your exam answers must be anonymous; do not state either your name or that of fellow students.
- Most of your text answers should include a discussion and be brief, typically at most a few sentences.
- When you do calculations to answer an exercise, include the intermediate steps in your answer.
- 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 justifiable within the context of the exercise. In such a case, you should make it clear what assumptions you have made.
- Plan your time so that you can try to answer as many subtasks as possible. Each subtask is normally weighted equally.

Cheating on school exams

"For school exams, it is considered as cheating to use support materials, unless it is explicitly stated in the exam question that this is permitted. Having access to illegal support materials may also be considered cheating, even if they are not used.

All communication between candidates in the exam room is prohibited. Contact with other candidates or other persons during trips to the lavatory and breaks is also prohibited. Mobile phones and other electronic equipment should be turned off and packed away."

`https://www.uio.no/english/studies/examinations/cheating/index.html`

(Continued on page 2.)

Problem 1

Suppose you want to predict the value of a house as a function of the following inputs: 1) Size of the house x_1; 2) Number of bedrooms x_2; Age of the house x_3; 4) Monthly energy usage of the house x_4. Let y denote the true value and $\hat{y}$ denote the prediction. Suppose you use a neural network that predicts $\hat{y}$ as follows:

$$\hat{y} = \sum_{j=1}^{5} v_j g\Big(\sum_{i=1}^{4} w_{i,j} x_i + b_j \Big) + b$$

where $v_j, w_{i,j}, b_j, b$ are trainable parameters and g is an activation function.

1a

Is this problem classification or regression? Why?

1b

How many hidden layers does the neural network have? How many hidden nodes (neurons in hidden layers) does it have?

1c

Let z_1 denote the output (activation) of neuron 1 in hidden layer 1. What is the expression of z_1 in terms of inputs $x_1, \ldots, x_4$?

1d

Suppose you want to slightly modify the final layer of the neural network so the output becomes non-negative regardless of the value of trainable parameters and without changing the number of parameters. Note that the desired output can be any and possibly large positive number but cannot be a negative number. How do you modify the final (output) layer?

(Continued on page 3.)

Problem 2

2a

A two-dimensional convolutional layer applies filters with spatial size $(5, 5)$, stride 2, and standard padding to an input of size (32,32). What will be the spatial output size?

2b

Suppose that a number of above filters with spatial size $(5, 5)$, stride 2, and standard padding, 10 input channels and 100 output channels, and no bias terms are applied in a layer. What is the number of trainable parameters in this layer?

2c IN4310 students only

Suppose we use a second two-dimensional convolutional layer having filters with spatial size $(3, 3)$, stride 3, and standard padding on top of the first convolutional layer described above and a third convolutional layer having filters with spatial size $(3, 3)$, stride 1, and standard padding on top of the second convolutional layer.

What is the spatial size of the region of the input to the first layer that will be used to compute the output of a single neuron in the third layer? You need to answer only for neurons in the center of the feature maps and ignore padding effects at the boundaries.

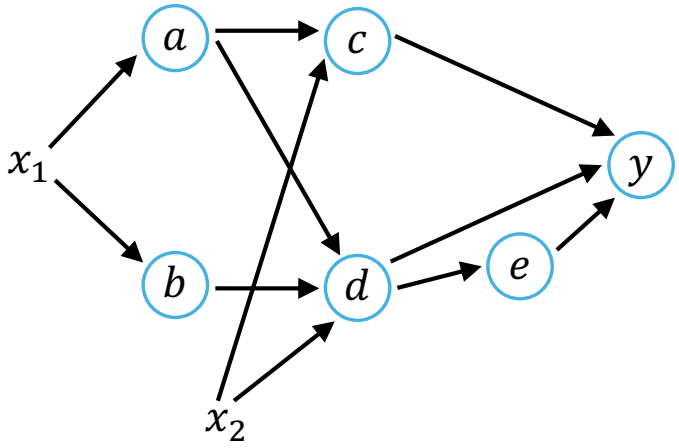
Include your calculations in your answer even if you also include drawings

*Hint $R^{[0]} = 1$ and $R^{[1]} = R^{[0]} + (k^{[1]} - 1)$ where $k^{[1]}$ is the spatial size of the first layer filter

(Continued on page 4.)

Problem 3

Consider the following neural network:



The output (activation) for any neuron $q \in \{a, b, c, \ldots\}$ is given by:

$$g\left(b + \sum_l v_l x_l + \sum_u w_u u\right) \tag{1}$$

where $g(\cdot)$ is some activation function, b is the bias term, $v_l = 0$ for neurons that have no direct inputs x_l, and $w_u = 0$ for neurons that have no inputs from neuron u.

This task asks you to write down expressions for gradients of this network. Please consider the following advices before doing so:

- Write the expression in terms of:
  - $\frac{\partial q_1}{\partial q_2}$, where q_2 is direct input to q_1, and
  - $\frac{\partial q}{\partial x_l}$, where the network input x_l is direct input to neuron q.
- This task is about the sequences of partial derivatives along the edges.
- You **do not need** to plug in or compute how $\frac{\partial q_1}{\partial q_2}$ or $\frac{\partial q}{\partial x_l}$ looks like.
- You **do not need** to multiply out terms in parentheses, so $(a + b)c$ or $((a + b)c + (d + e)f)g$ is fine to keep like that.

Write down an expression for the following gradients of this network:

3a

$\frac{\partial y}{\partial x_2}$

(Continued on page 5.)

3b

$\frac{\partial y}{\partial x_1}$

(Continued on page 6.)

Problem 4

Let $h_\theta : \mathcal{X} \to \mathcal{Y}$ denote a predictor parameterized by $\theta \in \mathbb{R}^d$. Suppose we have collected N independently and identically distributed samples (x_n, y_n) to train our predictor and use some loss function ℓ so $\ell(h_\theta(x_n), y_n)$ is the loss on sample n.

4a

Suppose the optimization problem for empirical risk minimization (ERM) is denoted by $\min_\theta R_N(\theta)$. What is $R_N(\theta)$ in terms of training samples?

4b

Suppose you use Gradient Descent with learning rate schedule α_t for solving the optimization problem and training your predictor. What is the update rule for parameter θ_t at iteration t?

4c

Instead of training the predictor with Gradient Descent (GD), your friend suggests you to use Stochastic Gradient Descent (SGD). Why does your friend think SGD is better than GD?

4d

When you implement SGD by drawing one sample from training samples uniformly at random and computing stochastic gradients, you note that stochastic gradients are too noisy and variance is huge. How can you control the variance?

(Continued on page 7.)

Problem 5

5a

What is true for "Dropout Regularization"? Briefly explain your selection(s)!

1. Dropout removes neurons out of the network at inference time.
2. Dropout is proposed to improve generalization.
3. If a neuron is active and its activation dominates the output of the network, dropout ensures its activation throughout training.

5b

Why are Residual Connections popular along with convolutions? Provide one advantage of using residual connections in backward pass and provide one advantage of using residual connections in forward pass?

5c

Consider the following code:

```
class SimpleCNN(nn.Module):
    def __init__(self, num_classes=10):
        super(SimpleCNN, self).__init__()
        self.conv1 = nn.Conv2d(...)
        self.conv2 = nn.Conv2d(...)
        self.conv3 = nn.Conv2d(...)
        self.conv4 = nn.Conv2d(...)
        self.conv5 = nn.Conv2d(...)
        self.fc1 = nn.Linear(...)

    def forward(self, x):
        x = F.relu(self.conv1(x))
        x = F.relu(self.conv2(x))
        x = F.relu(self.conv3(x))
        x = F.relu(self.conv4(x))
        x = F.relu(self.conv5(x))
        x = self.fc1(x)

        return x
```

By modifying, adding, removing **at most three lines of code** in total, how do you change the code such there there exists a residual connection between the "input to self.conv3" and the "output of self.conv5"?

(Continued on page 8.)

Problem 6

For a model $f : \mathcal{X} \to \mathcal{Y}$ and a transformation $T : \mathcal{X} \to \mathcal{X}$, we want the following property to hold:

$$f(T(x)) = f(x), \quad \forall x \in \mathcal{X}.$$

6a

Explain how this property is used for training neural networks with data augmentation in a *supervised* context. Discuss how this affects a model's ability to generalize.

6b

Explain some of the assumptions involved in the choice of T for data augmentation, and name three examples of T commonly used for image classification tasks.

6c

Discuss how this is applied for training models without supervised labels.

(Continued on page 9.)

Problem 7

7a

Discuss the impact of image compression on adverserial attacks.

7b

Explain the concept of *gradient masking* as an adverserial defence strategy, and relate this to the effect of image compression on adverserial attacks.

(Continued on page 10.)

Problem 8

Suppose you are developing a deep learning model to classify emails as either "spam" or "not spam" based on various features such as the content of the email, the sender's address, and the presence of certain keywords. Within your training date, you notice that 40% of emails are classified as spam. Over time, the behavior of email users and spammers changes. For the test data, you notice that the proportion of spam emails has decreased significantly to 10% because email providers have implemented more stringent spam filters and security measures. Meanwhile, the distribution of features given spam and not spam labels does not change.

8a

What is the type of distribution shift you are facing in this example?

8b

What is the expression for importance weight (density ratio) in terms of train and test distributions?

8c

Suppose the Empirical Risk Minimization (ERM) for training the model parameters θ is given by

$$\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} \ell(h_\theta(x_n), y_n)$$

How do you modify the above optimization problem with importance weights to solve Importance-weighted ERM (assuming you have access to perfect estimates of density ratios)?

(Continued on page 11.)

Problem 9

Consider a dataset for a classification task. It consists of 1000 samples. 970 samples have label A, and the remaining 30 have label B.

9a

Is accuracy a good metric to evaluate the performance of a classifier on this dataset? If not, then which other metric would be a good choice? Explain briefly.

9b

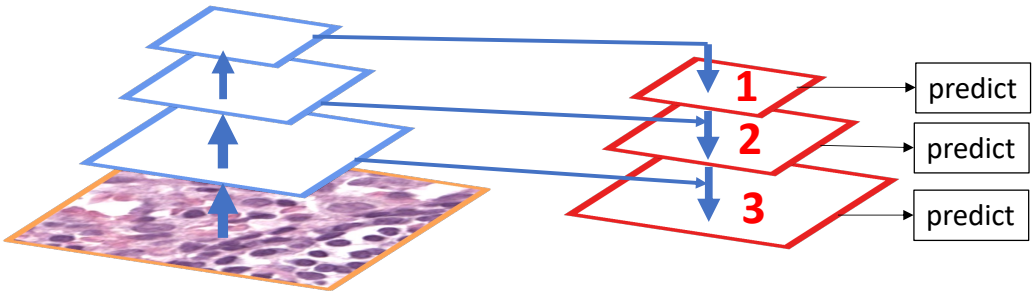
Suppose we care more about class B than class A. We don't want our classifier to misclassify class B as A, but classifying some instances of class A as class B is tolerable. Which of the following metrics would you look at for class B? Briefly justify your choice in at most 3 sentences. **Note:** class B is the positive class for these metrics.

1. Precision
2. Recall / Sensitivity
3. Specificity

(Continued on page 12.)

Problem 10

Consider the Feature Pyramid Network shown in the image below:



This network's predictions are made using anchor boxes on the 3 red feature maps numbered 1, 2, and 3.

10a

As seen in the image, the 3 feature maps have different resolutions. Furthermore, each feature map uses a different set of anchor boxes to make predictions. The resolution of the feature map and the number of anchor boxes used by it are listed below:

- Feature map 1: Resolution = 2×2. Anchor boxes = 5.
- Feature map 2: Resolution = 3×3. Anchor boxes = 4.
- Feature map 3: Resolution = 4×4. Anchor boxes = 2.

How many predictions will the network make in total? Show your calculations.

10b

To train the network, we need to match the predictions to the ground truth boxes. Given a prediction box and a list of ground truth boxes, how will you decide which ground truth box (or none of them) should be used as the ground truth for the prediction box?

(Continued on page 13.)

Problem 11

You are creating a Convolutional Network for image segmentation to segment out tumorous regions in the given input image. Your task is to not just identify the pixels corresponding to the tumour but also further grade the tumour as either grade 1, grade 2, grade 3, or grade 4.

11a

How many output channels will your network have?

11b

To reduce the number of parameters in your network, you would like to replace transposed convolution with nearest neighbour upsampling. If we want to upsample the matrix $\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix}$ by $2\times$ (make it twice as large), what would be the output of using nearest neighbour upsampling for that?

(Continued on page 14.)

Problem 12

12a

The cell state in LSTM is updated using the following equation:

$$\mathbf{C_t} = \mathbf{f_t} \odot \mathbf{C_{t-1}} + \mathbf{i_t} \odot \hat{\mathbf{C}}$$

Where f_t is the forget gate. Why is the forget gate's bias initialised to a high positive value? Explain briefly in 2-3 sentences.

12b

If a bidirectional LSTM layer has 10 LSTMs for each direction separately, what would be its output size? Justify your answer in 1-2 sentences.

(Continued on page 15.)

Problem 13

13a

Self-attention computes dot products between keys and queries. If a given self-attention has 4 attention heads and its one-dimensional input size is 3, how many dot products will it compute? Show your calculations.

13b

Swin Transformer limits the computation self-attention to a local window. Patches within a window do not interact with patches from other windows. This prevents the layer from capturing the global context. What does Swin Transformer do to solve this problem? Briefly explain in 2-3 sentences.