

i General information about the exam.

This exam has 8 questions (You can see 9 questions in Inspira but Question 9 is a placeholder for points from projects). Each question is worth 4-10 points and the total number of points from the exam is 55 points.

No examination support material is allowed on this exam.

Sheets for handwriting/drawing

On this exam it will be possible to attach hand-drawn sketches/illustrations or handwritten text to your digital exam answer. This is recommended for Question 8.

An exam question code will be available under each of the questions in the exam set. Ask the invigilator for drawing paper. The exam question code is unique for each question per student, so be sure to mark the sheet you have written or drawn on with the exam question number and the question code for the question you have answered on the sheet during examination.

In the 15 minutes after exam end time, you can fill out other requested information at the top of the page: date, your candidate number, course code, number of pages etc. Your candidate number can be found in the exam system.

Please ask an invigilator if you have trouble finding questions codes or your candidate number. When you have finished your exam, the sheets are to be submitted together, in the order they will be added to your answer paper, to the head invigilator in the venue.

¹ Embeddings

Dagny has a problem. Read the description below and help her:

Dagny:

I am training my own small language model. As a preprocessing step, I wanted to train a word embedding model.

I have a vocabulary of size 20. I trained embeddings using CBOW with context size 2. Training loss was small and everything looked fine.

I heard that good embeddings should capture semantic meanings of the words. For example, one could do linear algebra with them and get results like **King** - **Man** + **Woman** \approx **Queen**.

I got the following embeddings:

[illegible][illegible]

Queen = [0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0], and

[illegible]

However, when we do the computations, we get

King – **Man** + **Woman** = $[0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ -1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0]$
which is far from the embedding for the word **Queen**.

What is the problem? Is this a sign of overfitting or could it be something else?

Fill in your answer here

Words: 0

Maximum marks: 4

2 AE vs VAE

What are autoencoders (AE) and variational autoencoders (VAE)? How do they differ? How does one choose which one is more appropriate for the task at hand?

Fill in your answer here

Format ▾ | **B** *I* U \times_2 \times^2 | \mathcal{I}_x | | | | Ω | Σ |

Words: 0

Maximum marks: 5

3 Adam

You are working as a group leader at INF265. You are approached by a student named Holger. He shows the formulas for Adam from the course book:

$$\begin{aligned}\mathbf{v}_t &\leftarrow \beta_1 \mathbf{v}_{t-1} + (1 - \beta_1) \mathbf{g}_t \\ \mathbf{s}_t &\leftarrow \beta_2 \mathbf{s}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{v}}_t &= \frac{\mathbf{v}_t}{1-\beta_1^t} \\ \hat{\mathbf{s}}_t &= \frac{\mathbf{s}_t}{1-\beta_2^t}\end{aligned}$$

$$\mathbf{g}'_t = \frac{\alpha \hat{\mathbf{v}}_t}{\sqrt{\hat{\mathbf{s}}_t + \epsilon}}$$

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \mathbf{g}'_t$$

Holger is slightly frustrated and asks the following question.

Holger: What is this mess? I do not see any point of doing things in such a complicated way. Why can't we just use regular gradient descent?

Explain Holger the intuition behind different parts of the formulation and what are potential benefits of using it.

Fill in your answer here

Maximum marks: 6

4 Initialisation strategies

Laura, Merethe and Nina are training deep neural networks for a classification task. They have the following discussion.

Laura: I initialised the weights in my network by setting all of them to zeros.

Merethe: No, that is not a good strategy. I initialised the weights by setting all of them to ones.

Nina: No, you both are wrong. The best way to initialise weights is to set them all to $1/m$ where m is the number of neurons in the hidden layer.

Merethe: Did you normalise your data?

Nina: No, I didn't.

Merethe: Oh, I see. If you had normalised your data, you wouldn't have to have the $1/m$ term in your initialisation.

Laura: Merethe, what you say is true for regression tasks. But you forgot that we are solving a classification task. You do not need normalise the input because the softmax at the output layer normalises the output, so it is between zero and one regardless of the input.

Comment the discussion and the strategies that they have chosen. What do you recommend them to do?

Fill in your answer here

Format
B
I
U
 x_2
 x^2
 I_x

Words: 0

Maximum marks: 6











5 Attention


Answer briefly the following questions:

1. What is the attention mechanism and how does it work?
2. What are the main benefits of the attention mechanism compared to recurrent neural networks?

Fill in your answer here

Format

B *I* U x_2 x^2 I_x          

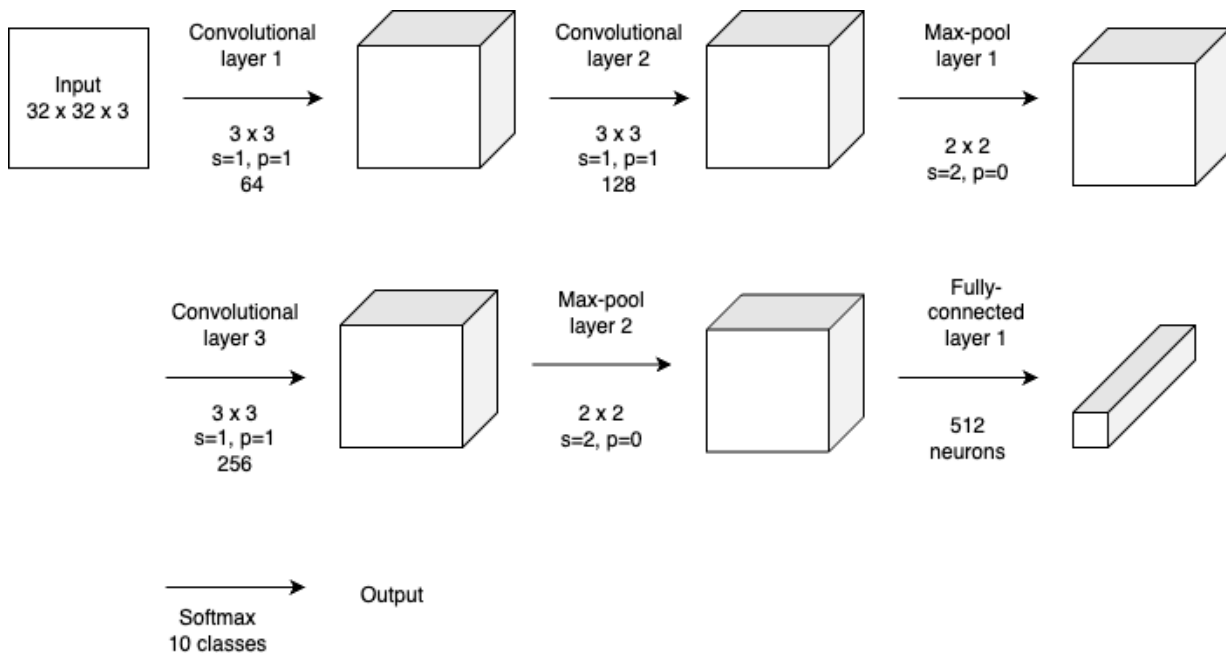


Words: 0

Maximum marks: 8

6 Convolutions

Suppose we have the following convolutional network architecture:



The input is a 32x32 color image (3 channels). For each convolutional layer, the first row tells the size of the filters. The second row tells stride (s) and padding (p) and the third row the number of filters. The full-connected layer has 512 neurons. Activation functions in each convolutional layer as well as the fully-connected layer is ReLU. The output layer has a softmax activation function.

Consider the following modification to the architecture:

1. Replace the 3x3 filters in convolutional layer 1 by 5x5 filters.
2. Increase the number of filters in convolutional layer 2 to 256.
3. Remove max-pool layer 1.
4. Change the filters of the max-pool layer 2 to 4x4 and use stride=4.
5. Replace ReLU activation functions with sigmoid functions

Note that you are supposed to consider each modification separately. That is, what happens if you make the modification 1 to the original architecture (and nothing else). Then consider what happens if you make modification 2 to the original architecture (and nothing else). And so on.

For each modification, answer the following:

- How does the change affect the number of parameters of the network? You do not need to compute the exact number of parameters. Just tell which layers are affected and how.
- Which adjustments do you have to make to other layers in order to make the network work?

Suppose that the original model was underfitting. Which of the changes are most likely to remedy this problem? Justify your answer.

Fill in your answer here

[illegible]

Words: 0

Maximum marks: 8

7 Modelling help

You have started a new job as a machine learning engineer at a hospital. The department of digital pathology wants to develop a system to assist medical doctors to make diagnoses. Specifically, they want to predict a diagnosis (skin condition) based on an image and some metadata. Currently, their "automatic" solution is a manually constructed rule-based system (some if-then rules). You have been given the task to train a machine learning-based diagnosis system.

The data is collected from 10,015 patients. For each patient, there is a digital image of a skin lesion (Dictionary definition: A lesion is a region in an organ or tissue which has suffered damage through injury or disease, such as a wound, ulcer, abscess, or tumour.). In addition, they have collected age (numerical), sex (categorical) and anatomical site of the lesion (text) for each patient. The class label is one of eight classes of skin conditions.

To get started, your boss asks you to write a short plan how to begin. The plan should include:

- How would you preprocess the data?
- What type of architecture would you use?
- What would be the loss function?
- How would you figure out whether your model performs better than the existing solution? How would you evaluate the models? What performance measure would you use? Why?

Your boss is also slightly worried that the data set is too small to train a decent model. Suggest him some possible remedies for this problem.

Fill in your answer here

Format
B
I
U
 x_2
 x^2
 I_x
 $\frac{1}{2}$
 $\frac{3}{4}$
 Ω

--	--	--

 Σ

Words: 0

Maximum marks: 8

⁸ Forward and backward propagation

Consider the following neural network with one hidden layer for a regression task. The network

has weight matrices $\mathbf{W}^{(1)} = \begin{bmatrix} 1 & 3 & -1 \\ 2 & 1 & 1 \end{bmatrix}$, $\mathbf{W}^{(2)} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, $\mathbf{b}^{(1)} = [0 \quad 1 \quad -3]$ and $\mathbf{b}^{(2)} = 0$

where $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$ are the parameters of the hidden layer and $\mathbf{W}^{(2)}$ and $\mathbf{b}^{(2)}$ are the parameters of the output layer. The hidden layer has a ReLU activation function.

Answer the following questions:

A) Forward pass. Suppose we perform a forward pass with the input $\mathbf{x} = [-1 \quad 1]$. What is the output of the network? Show intermediate results.

B) Loss. Consider the squared loss $L = \frac{1}{2}(\hat{y} - y)^2$ where \hat{y} is the prediction of the network. Suppose the true label is $y = 5$. What is the loss?

C) Backward pass. Compute the partial derivatives of the parameters using back-propagation. Show your work.

D) Gradient descent update. Update the parameters using gradient descent with learning rate 0.1.

E) Here we did a gradient descent update using a mini-batch with size 1. Why do we usually prefer larger mini-batches?

It is recommended to solve this task on paper. See instructions under "General information".

Fill in your answer here

Words: 0

Maximum marks: 10

9 Points from projects

This is a placeholder for points from the projects.

Maximum marks: 45