

i General information about the exam

This exam has 6 questions (You can see 7 questions in Inspira but Question 7 is a placeholder for points from projects). Each question is worth 5-15 points and the total number of points from the exam is 55 points.

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 5.

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 must shade the exam question code in the boxes under each digit in the question code and also fill out other requested information at the top of the page. 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 Network architectures

Amanda and Bertha are constructing image classifiers that are supposed to recognise cows and horses. Their input are images with 20x20 pixels and 3 channels for colours.

Amanda uses a standard feedforward network. She has one hidden layer with 20 units with ReLU activation. As the network is solving a binary classification task, the output layer consists of a single sigmoid unit.

Bertha uses a convolutional neural network. The first layer is a convolutional layer with 10 filters. The filters are 3x3 filters with padding=1 and stride=1. The activation function is ReLU. The convolutional layer is followed by a 2x2 max pooling layer with stride=2. After the pooling the network has a fully-connected layer with 20 units and ReLU activation. Finally, the output layer consists of a single sigmoid unit.

Answer the following questions:

A) Which network has less parameters? Justify your answer.

B) Amanda and Bertha have a large training set with thousands of labelled images of cows and horses. However, both models have a rather disappointing performance in practice. Give Amanda and Bertha advice how to build a model that is likely to outperform their current models. Justify your recommendation.

C) Cecilie wants to learn a classifier to distinguish between hippos and rhinos. She explains: "Unfortunately, I have only 12 labelled images of hippos and 17 images of rhinos which is not enough to construct a classifier. It is really frustrating. If I would be interested in classifying between cows and horses like Amanda and Bertha, I would have enough labelled images, but my data set is way too small."

Help Cecilie to classify hippos and rhinos. Is her problem solvable? What kind of techniques would you recommend to her?

Fill in your answer here

Format
B
I
U
 x_2
 x^2
 I_x

Words: 0

Maximum marks: 10

Frode: $f_F(z) = \min(0, z)$

What are pros and cons of each suggestion?

Fill in your answer here

Maximum marks: 5

3 Bias-variance tradeoff

This task is about the bias-variance tradeoff. **Answer the following questions:**

A) Consider a feedforward neural network with one hidden layer. What effect does increasing the number of hidden neurons have for bias and variance? What effect does increasing the number of hidden layers have for bias and variance? Justify your answer.

B) What is the relationship between bias and variance and overfitting?

C) What is regularisation and how is it related to overfitting?

Fill in your answer here

Maximum marks: 10

4 Modules

Researchers have been developing different modules that combine several neural network layers. In this task, we take a look at two commonly used modules.

A) **LSTM**. You can see the structure of an LSTM module below. What are common challenges in learning recurrent neural networks? What are the roles of different parts of LSTM and how do they help to mitigate the problems?

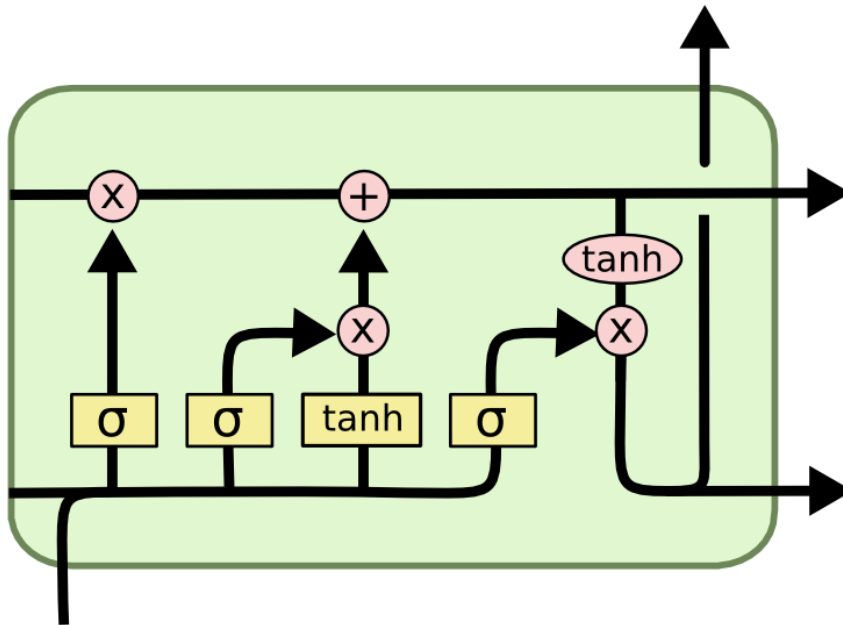


Image source: Chris Olah

B) **Inception**. You can see an illustration of an Inception module below. What is the motivation for using such a module? Why are there 1x1 convolutions in this module?

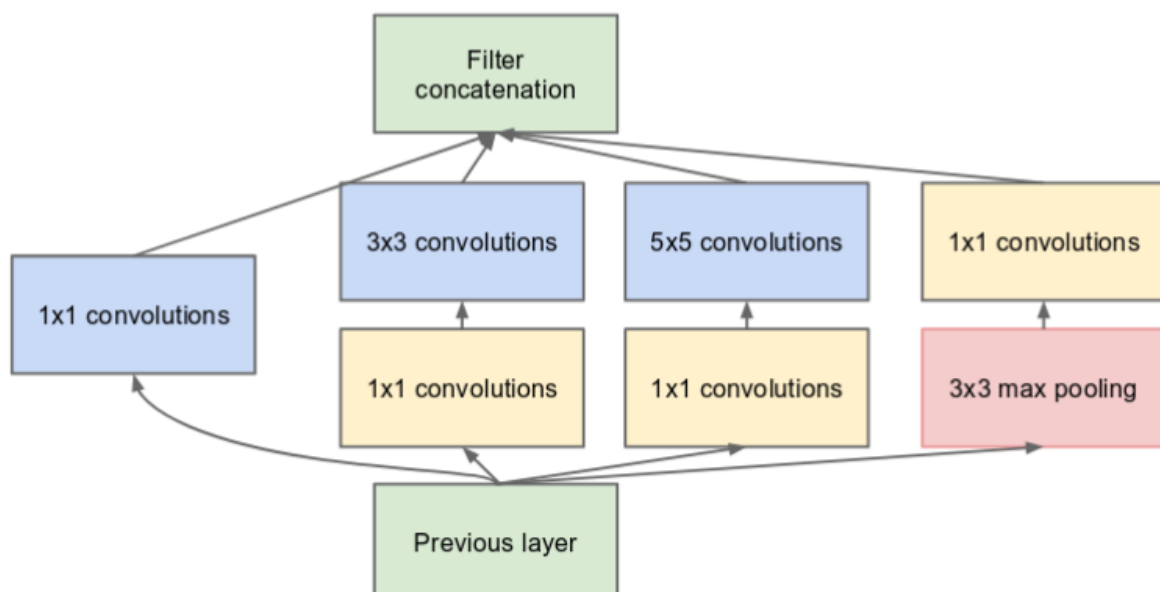


Image source: Szegedy et al.

Fill in your answer here

Maximum marks: 10

⁵ Forward and backward propagation

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

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

Format

▼

B *I* U x_2 x^2 I_x           



Words: 0

Maximum marks: 15

6 TSP

Tuva is solving an instance of the Traveling Salesperson Problem (TSP). That is, she is given a set of cities and distances between them and the goal is to find the shortest route such that the salesperson visits every city once. In other words, the solution is a sequence of cities. Tuva has heard that the problem is NP-hard and thus it is unlikely that there exists efficient exact algorithms.

Tuva has decided to resort to machine learning. She has trained an attention-based transformer model that, given the coordinates of the all cities and the route so far, can compute the probability for any of the remaining cities to be the next in the sequence given the cities visited so far. In other words, the probability for a city is high if the model "thinks" that going to that particular city is likely to lead a short route. Note that the model tries to minimise the length of the remaining route and going to the nearest city is not always the best option.

Unfortunately, Tuva has a problem. While her transformer is performing well in its task, it cannot make predictions for the whole route at once. It can only predict how good each city would be if they were placed next in the route.

Help Tuva. How can she find a short route given that she has the trained transformer at her disposal? Tuva also hopes to have some flexibility. Sometimes she needs a fast algorithm that finds some route but sometimes she is willing to spend more time to find a better route.

Fill in your answer here

Maximum marks: 5

7 Obligatory projects

This is a placeholder for points from the projects.

Maximum marks: 45