Endterm Exam SS2024

Introduction to Deep Learning (Technische Universität München)

Chair of Visual Computing
School of Computation, Information and Technology
Technical University of Munich

E0349
Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

2169-01-1384-E0349-01

Introduction to Deep Learning

Exam:	IN2346 / endterm	**Date:**	Friday 26th July, 2024
Examiner:	Prof. Dr. Matthias Nießner	**Time:**	08:30 – 10:00

	P 1	P 2	P 3	P 4	P 5	P 6	P 7
I							

2169-01-1384-E0349-01

Working instructions

- This exam consists of **16 pages** with a total of **7 problems**.
  Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 90 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
  - none
- **Answers are only accepted if the solution approach is documented.** Give a reason for each answer unless explicitly stated otherwise in the respective subproblem.
- Do not write with red or green colors nor use pencils.
- Physically turn off all electronic devices, put them into your bag and close the bag.

Left room from ________ to ________ / Early submission at ________

☐ Exam empty

2169-01-1384-E0349-01

Problem 1 Multiple Choice (18 credits)

Mark correct answers with a cross ☒
To undo a cross, completely fill out the answer option ■
To re-mark an option, use a human-readable marking ×■

Please note:

- For all multiple choice questions any number of answers, i.e. either zero (!), one or multiple answers can be correct.
- **For each question, you'll receive 2 points if all boxes are answered correctly (i.e. correct answers are checked, wrong answers are not checked) and 0 otherwise.**

1.1 Which of the following statements on Convolutional Neural Networks are true?

- ☐ Early layers typically capture high-level features.
- ☐ Deep layers typically capture high-level features.
- ☐ Pooling layers can be used to reduce the spatial dimension of the feature maps.
- ☐ All layers in a CNN have the same receptive field size.

1.2 Which of the following statements on activation functions are true?

- ☐ The Softmax activation function is invariant to scale, i.e. $Softmax(cx) = Softmax(x)$.
- ☐ When designing a neural network, (Linear → Dropout → ReLU) is equivalent to (Linear → ReLU → Dropout).
- ☐ Skip connection adds non-linearity to the model; therefore, it is a type of activation function.
- ☐ The output of Leaky ReLU is always non-negative.

1.3 Which of the following layers are used in the original Transformer model?

- ☐ Convolutional layers
- ☐ Recurrent layers
- ☐ Fully connected layers
- ☐ Softmax layers

1.4 Which of the following statements correctly describes the relationship between loss curves, overfitting, and underfitting?

- ☐ A model is underfitting if the training loss is low and the validation loss is high.
- ☐ Overfitting occurs when both training and validation loss are high.
- ☐ Underfitting is indicated by high training loss and high validation loss.
- ☐ A loss curve showing a consistently decreasing training loss and a decreasing validation loss indicates overfitting.

1.5 Which of the following statements about Dropout are true?

- ☐ It increases the gap between validation loss and training loss in general.
- ☐ It can be seen as an ensemble of networks.
- ☐ During evaluation, it activates all nodes and scales up the output value.
- ☐ It can not be applied to CNN.

1.6 Your model for classifying different cat species is getting a low training error with a high validation error. Which of the following options are promising things to try to improve the validation performance?

- ☐ Transfer learning from a pre-trained large model
- ☐ Add more linear layers
- ☐ Decrease the learning rate
- ☐ Add weight regularization

1.7 Which of the following propositions about a Convolutional layer are true?

- ☐ The total number of parameters depends on padding.
- ☐ The total number of parameters depends on the width and height of the kernel.
- ☐ The number of channels of the input image and the number of filters can be different.
- ☐ The size of the Convolutional layer's output depends on the stride.

1.8 Which of the following functions are **NOT** suitable as activation functions to add non-linearity to a network?

- ☐ The floor function $f(x) = \lfloor x \rfloor$
- ☐ $f(x) = x$
- ☐ $f(x) = \sqrt{x}$
- ☐ $f(x) = |x|$

1.9 Which of the following statements on Generative Adversarial Networks for image generation are true?

- ☐ The Generator decodes a latent vector into an image.
- ☐ When training the Generator, we use a frozen pre-trained Discriminator as supervision.
- ☐ Training a GAN does not require manual "real/fake" labeling of the image.
- ☐ A reduced generator loss generally means an increased discriminator loss.

Problem 2 Short Questions (16 credits)

0 1

2.1 Explain why you would use 1x1 convolutions.

0 1 2

2.2 Explain the concept of receptive field. What happens if the receptive field of your model's output is too small?

0 1 2

2.3 Consider two different models for image classification of the MNIST dataset. The models are (i) a fully connected network with three hidden layers of size 16, (ii) VGGNet. Which of the two models is more robust to the translation of the digits in the images? Give a short explanation of why.

0 1

2.4 AlexNet is using an 11×11 convolutional filter in the first layer. Name 1 disadvantage of using such a large filter.

0 1 2

2.5 You are trying to solve a binary classification problem, but the positive class is very underrepresented (e.g., 8 negatives for every positive). Describe a technique that you can use during training to help alleviate the class imbalance problem. Would you apply this technique at test time as well? Why or why not?

0 1

2.6 Give one situation where it would make sense to overfit the model on purpose.

2.7 You want to train a voice-to-text model and have collected some audio files of people speaking and their transcripts. Write down 4 data augmentation methods that can be applied.

0 1 2

2.8 If your Convolutional layer's input has shape [10×20×30×40] (batch size × number of channels × height × width), how many parameters are there in a **single** 3x3 convolution filter operating on this input with stride 2, including bias?

0 1

2.9 The Adam optimizer is described with the following formulae:

0 1 2

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)\nabla L(x,\theta), \quad v_k = \beta_2 v_{k-1} + (1-\beta_2)[\nabla L(x,\theta) \circ \nabla L(x,\theta)], \tag{1}$$

$$\hat{m}_k = \frac{m_k}{1-\beta_1^k}, \quad \hat{v}_k = \frac{v_k}{1-\beta_2^k}, \tag{2}$$

$$\theta_k = \theta_{k-1} - \frac{\eta}{\sqrt{\hat{v}_k} + \epsilon}\hat{m}_k. \tag{3}$$

Please explain why we need the operations in (2).

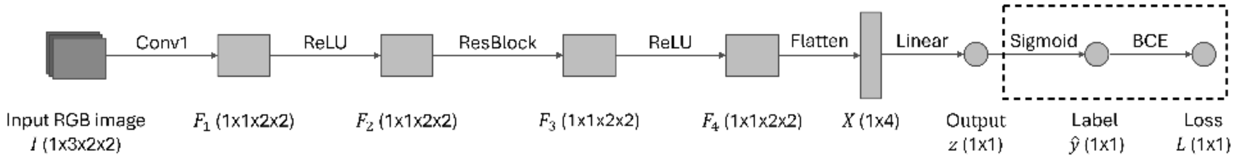
2.10 Assume the input of batch normalization in a CNN has shape [BxCxHxW], where B, C, H, W represent batch size, number of channels, height, and width, respectively. Please:

0 1 2

1. Write down the variables of batch normalization, which are updated during training, and their shapes.
2. Explain why these variables need to be stored together with the model and the purpose of using them.

Problem 3 Back Propagation (14 credits)

Consider the following network for the task of Binary Classification:



0 1 2 3 4

3.1 Let's focus on the fully connected layer (Linear), $z = XW + b$. Note that X is the flattened output of a ReLU function.

$$X_{1\times 4} = \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix}, W_{4\times 1} = \begin{bmatrix} 1 \\ -2 \\ 3 \\ 3 \end{bmatrix}, b_{1\times 1} = [0]$$

We've calculated the forward pass for you. Our RGB image's ground-truth label is $y = 1$.

- $z = XW + b = 1 * 1 + 2 * (-2) + 0 * 3 + 1 * 3 + 0 = 0$
- $\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} = 0.5$
- $L = BCE(y, \hat{y}) = -\frac{1}{1}\sum_{i=1}^{1} 1 \cdot \ln(0.5) = 0.693$

Calculate the following gradients using the chain rule:

1. $\frac{\partial L}{\partial \hat{y}}$, Hint: $\frac{\partial \ln(x)}{\partial x} = \frac{1}{x}$
2. $\frac{\partial L}{\partial z}$,
3. $\frac{\partial L}{\partial W}$
4. $\frac{\partial L}{\partial b}$

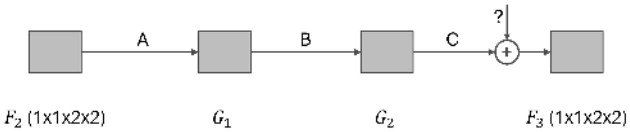
0 1 2

3.2 With the same context of 3.1, write down the variables that should be cached during the forward pass in order to compute the backward pass efficiently.

3.3 With the same context of 3.1, explain why in the given network, the gradient $\frac{\partial L}{\partial W}$ will always be non-positive (negative or zero).

3.4 The structure of the ResBlock is shown below. Please:

1. Assign operations “ReLU”, “Conv1”, and “Conv2” to A, B, and C in the figure below. Each operation can be assigned only once.
2. The addition operator is missing one input. Write down which one of F_2, G_1, and G_2 should connect to the addition operator.
3. Conv1 has kernel size 1, stride 1, and padding 0. Conv2 has kernel size 3, stride 1. Write down the padding of Conv2.



3.5 With the same context (**including your solution**) of 3.4, write down the gradient $\frac{\partial L}{\partial F_2}$ and explain why the ResBlock introduces a “highway” for the gradient flow. You may use the term $\frac{\partial L}{\partial F_3}$ in your answer.

Problem 4 Dataset and Transfer Learning (10 credits)

You want to train a classifier to classify images of dogs, cats, and birds. You downloaded 1M images of these classes from the Internet with various resolutions and start labeling them by hand. After a while, you have collected a dataset with 1000 images of dogs, cats, and birds.

0 1

4.1 To train and test the model, you first split your dataset into train, validation, and test subsets. Write down a meaningful percentage of data you would assign to each subset.

0 1

4.2 Before training, you apply normalization to the images. Explain the benefit of doing this.

0 1

4.3 After splitting, you realize that the train set only contains pictures taken during the day, whereas the others only have pictures taken at night. What will happen if you continue with these splits? How can you correct it?

0 1 2

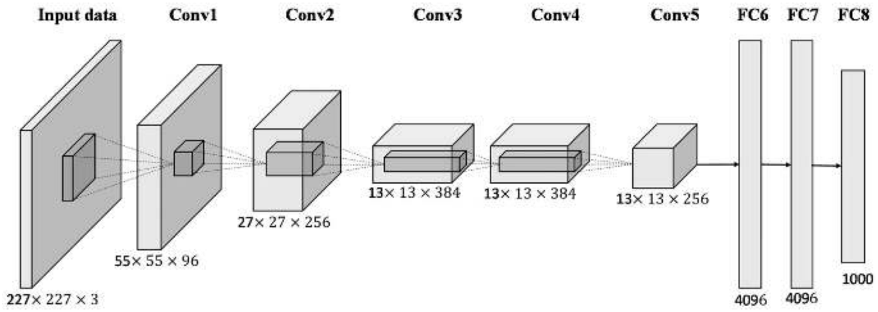
4.4 With the same context of 4.1, you initialized a classification network and trained it with cross entropy loss using the train set, then ran the network to predict images in both the train and validation set. What result do you expect? (Note: the dataset contains 1000 images.) Write down two approaches **(besides transfer learning)** to further improve your model.

0 1

4.5 You decide to use transfer learning. Instead of downloading a pre-trained model from the Internet, you can train one yourself. Explain how you can do it.

4.6 To save time, you downloaded a pre-trained model of AlexNet that outputs 1000 class labels. The model structure is shown in the figure below (activations are hidden). Explain step by step how you use this model for transfer learning to train your classifier **efficiently**.

0
1
2



4.7 One of your classmates found a classifier trained to classify brands from car images with high accuracy. Would you use this model for your task? Explain why.

0
1

Problem 5 Activation and Regularization (10 credits)

After you told your friend about what you learned in I2DL, they are excited to try those Neural Networks. Because they did not attend I2DL, they made some mistakes.
They try the following model architecture on a 10-class classification problem:

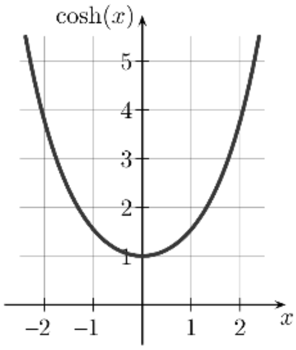
- Linear(784 to 400) → ReLU → Linear(400 to 200) → ReLU → Linear(200 to 10) → output $\boldsymbol{z}$
- Instead of the regular Multi-class cross entropy, they use the Categorical cross entropy $CCE(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_i y_i \log(\hat{y}_i)$, where $\hat{\boldsymbol{y}} = \text{Sigmoid}(\boldsymbol{z})$, and $\boldsymbol{y}$ is the ground truth label.
- The weights are initialized using Xavier initialization.
- They additionally applies a regularization loss $reg = \sum_{w \in \theta} \exp(-w)$ where θ contains all weights of the network excluding biases.

0 1 2

5.1 What weight values does the used regularization encourage? Name two problems the resulting weights can cause.

0 1

5.2 After you explained the problem to your friend, they want to try $reg = \sum_{w \in \theta} \cosh(w) - 1$. They insist that the "$-1$" is necessary so zero-valued weights are not punished. Will the "-1" change the weights resulting from training? Explain why.



0 1

5.3 After training, the network achieves a CCE loss close to zero but only around 10% accuracy. Explain how this could happen.

5.4 Your friend has read that shuffling the data during training can be beneficial. They use the following code to generate training and validation sets in each epoch.

```
def generate_data_for_epoch(data):
    np.random.shuffle(data) # inplace shuffling of data
    split_index = int(0.8 * len(data))
    train = data[:split_index]
    val = data[split_index:]
    return train, val
```

Indicate the problem with their approach and explain what the consequence of using this code is.

0 1 2

5.5 You noticed that the Xavier initialization is not the best choice, why? What do you recommend to use instead?

0 1

5.6 You had a debate with your friend about Linear and Convolutional layers. They made the following statement:

> A convolutional layer with fixed input and output sizes can be replaced by three layers (Flatten → Linear → Reshape). By carefully setting the hyperparameters and parameters (input and output sizes of Linear, output shape of Reshape, weights and biases of Linear), these three layers can fully replicate the convolutional layer's output.

Do you agree with their statement? Explain why.

0 1 2

Problem 6 Recurrent Neural Networks (12 credits)

You are a big fan of *The Lord of the Rings* and want to build a model to translate English into Elvish (a constructed language used in the novel).

0 1 2 3

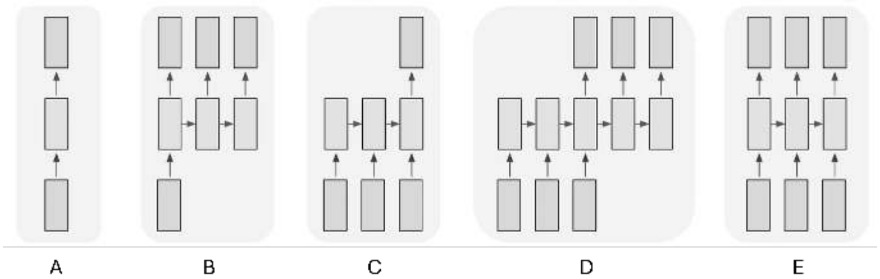
6.1 You first collected the vocabulary of Elvish and built an embedding dictionary. Please answer:

1. What is an embedding dictionary?
2. Is an embedding dictionary of Elvish enough for your project? If not, what else do you need?
3. Besides the words, what else do you need as embeddings?

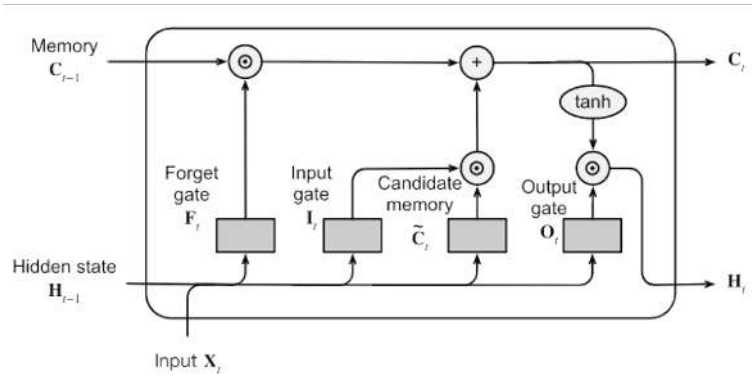
0 1 2 3

6.2 You started with a simple Recurrent Neural Network (RNN). Please explain:

1. What are the input and the output of the RNN **block**?
2. Which of the following pipelines is best for your task? Explain why?



6.3 RNNs can struggle with long sentences. Now you want to try LSTMs. Here is an illustration of an LSTM. Please answer:



1. What are the activation functions for the Forget gate, Input gate, Candidate memory, and the Output gate, respectively?
2. Is it a good idea to use ReLU for the Forget gate? Explain why.
3. Why are LSTMs better at processing long sequences?

0 1 2 3

6.4 Recently, many large language models use Transformers with multi-head attention instead of LSTMs. Please answer:

1. How to inform the multi-head attention layer about the ordering of the words in a sentence?
2. The attention layer takes Key K, Query Q, and Value V as inputs. How are the attention weights computed (in formula or words), and what are their characteristics?
3. How is the output of attention layers computed (in formula or words)? Why is this process called "attention"?

0 1 2 3

Problem 7 Autoencoders (10 credits)

0 1 2 3

7.1 Please answer:

1. What are the main purposes of the encoder and the decoder of an autoencoder?
2. In a **fully convolutional autoencoder**, what kinds of layers are used in the encoder and the decoder, respectively (name 2 for each)?

0 1 2

7.2 For the task of image semantic segmentation, you want to use the U-net architecture. Explain two differences between the architectures of an image autoencoder and a U-net. Is it reasonable to use a U-net architecture for autoencoding? Explain why.

0 1 2

7.3 What are the differences between the autoencoder and the variational autoencoder in terms of the goal and loss?

7.4 The diffusion model learns a denoising process to generate samples in a distribution. It has been shown to be able to learn complicated distributions such as images. The latent diffusion model, on the other hand, learns the distribution of the latent feature encoded by a VAE. With your knowledge of VAEs in mind, write down two advantages in terms of **speed** and **image quality** of the latent diffusion model compared to the diffusion model that operates directly on the RGB domain and explain why.

0
1
2
3

Additional space for solutions–clearly mark the (sub)problem your answers are related to and strike out invalid solutions.