183.663 Deep Learning for Visual Computing

Exam Questions and Information

Christopher Pramerdorfer

Computer Vision Lab, TU Wien

January 21, 2017

This document lists all exam questions. Every exam consists of a different subset of these questions. The exam questions will be very similar to those stated here, but not necessarily identical. You can answer the questions in English or German, but please use the English terminology in any case. Questions should be answered in your own words and in sufficient detail; do not just write down keywords or phrases from the lecture slides. The overall style should be similar to the style of the assignment reports, but of course briefer. You have 60 minutes to answer the questions. No documents are allowed during the exam.

What is the task definition of image classification? Explain at least 5 challenges and give examples. What is object detection and how does it differ from classification?

Why do we need datasets? What are the challenges of dataset collection and annotation in deep learning? Explain the purpose of the different subsets. What is the rule of thumb about how many images per classes are needed for CNNs to perform well?

Assume a company asks you to develop an application that is able to predict which kind of bird is depicted in a given image. List and explain the individual steps you'd follow to solve this problem using deep learning.

What is the motivation for solving vision tasks via machine learning? What is a machine learning algorithm and how are they used for solving image classification problems?

Explain the differences between classification and regression as well as supervised learning and unsupervised learning. What do discriminative and generative models learn, and how can generative models be used for classification?

Why are machine learning algorithms tested on data unseen during training? How can such algorithms perform well on unseen data? What does dataset bias mean?

How is the performance of a classifier measured? Which factors determine the performance, and how is underfitting and overfitting related to these factors? Draw sketches that illustrate underfitting and overfitting.

Explain the terms algorithm capacity, bias, and variance. Draw a sketch that illustrates how the capacity, training error, and test error are related. Why does the training set size affect the optimal capacity of a model?

What is a hyperparameter? Name at least 3 hyperparameters in the context of deep learning using convolutional neural networks. What is the purpose of hyperparameter selection, which search strategies exist, and how do they work?

How does the k nearest neighbor classifier work? Create a sketch for illustration, assuming a two-dimensional feature space and two different classes. Draw at least three training samples per class (must not lie on a line) as well as (roughly) the resulting decision boundaries. What are the limitations of this classifier?

Why do general machine learning algorithms (those expecting vector input) perform poorly on images? What is a feature, and what is the purpose of feature extraction? Explain the terms low-level feature and high-level feature.

What is the definition of a parametric model? What do the parameters of such models control (what effect do they have?), and how are they set?

What is a linear model, which types of parameters does it have, and what do they specify? Draw a sketch assuming two-dimensional feature space and three different classes. Draw a few samples per class so that the classes are linearly separable. Draw the decision boundaries a linear classifier might learn and explain how the individual boundaries are related to the classifier output (no need to calculate anything).

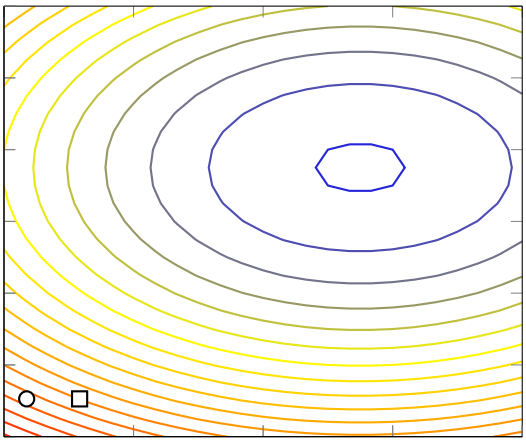
What is the purpose of a loss function? What does the cross-entropy loss on a dataset $\mathcal{D}$ measure? Which criteria must the ground-truth labels and predicted class-scores fulfill to support the cross-entropy loss, and how is this ensured?

What is the purpose of optimization in the context of machine learning? How does the

gradient descent algorithm work? What is the gradient of a function?

What is the difference between a local and a global optimum? Draw a sketch that illustrates the difference. Are local minima a problem in deep learning? Why (not)?

Consider the following contour plot of a function with two parameters. How might gradient descent proceed in this case, assuming the circle in the bottom-left corner as the starting point? Mark the individual steps and connect them using lines. Give a brief explanation of momentum. How would momentum affect the training progress in the example case? Mark the individual steps gradient descent with momentum might take, assuming the bottom-left square as the starting point?



Explain the differences between batch, minibatch, and stochastic gradient descent. Which version is most commonly used in deep learning and why? What effects has the minibatch size? Write pseudo-code that illustrates the overall structure of minibatch-based training and validation with early stopping, continuing below the following line. A high-level overview is sufficient, no need to use math.

```
while epoch <= MAX_EPOCHS:
```

How do weight and bias parameters affect the input $\mathbf{x}$? What must be considered when initializing these parameters? What happens if the weights are set too large or too small? We discussed a heuristic for controlling the magnitude of weights. What is the intention behind this heuristic (no math required)?

Explain the purpose of preprocessing. How do per-sample normalization and per-training-set normalization differ in terms of operation and purpose? In the latter case, which (if any) preprocessing is applied during validation and testing?

What are the goals of optimization and machine learning? Why do they differ? Create two sketches with each showing the training progress over time in terms of both training and test error; one that is good from an optimization perspective but bad from a machine learning perspective, and one that is worse from an optimization perspective but better from a machine learning perspective. Explain both sketches.

What is the purpose of regularization? What is weight decay and what is its purpose? What is early stopping and how does it work? In deep learning, is it better to increase regularization or to decrease the model capacity by other means, and why?

What is the definition of a feedforward neural network? Which types of units do such networks have? Draw a graph of such a network.

What is the definition of a multilayer perceptron? What operation do (non-input) units perform? What is an activation function and which functions are common? Draw a sketch that shows the layers of such networks and how the individual units are connected.

Why are multilayer perceptrons not suitable for deep learning for image analysis?

What is the motivation and purpose for representation learning? How is deep learning related to representation learning, and what is its definition?

What is the motivation for using locally connected layers for image analysis? What does sparse connectivity mean? Assume input data of dimension $W \times H \times D$. How many many weight and bias parameters does a locally connected (but not convolutional) layer have, assuming 3×3 connectivity and $W \times H \times 2$ output data dimension, and why?

What is the purpose of a convolutional layer? How do such layers differ from locally connected layers, and why are they almost always preferred? Assume input data of

dimension $W \times H \times D$. How many weight and bias parameters does a convolutional layer have, assuming 3×3 connectivity and $W \times H \times F$ output data dimension, and why? Why are such layers called convolutional layers? What are feature maps?

What is the receptive field of a neuron? Assume a network consisting of two convolutional layers with 3×3 connectivity followed by a 2×2 max-pooling with stride 2 and again two convolutional layers with 3×3 connectivity. What is the receptive field of neurons in the final convolutional layer? How does the receptive field affect feature extraction? Why do convolutional layers usually use 3×3 connectivity?

What is the purpose of pooling layers? Calculate the output of a 2×2 max-pooling layer with stride 2 assuming the following input. How many pooling layers with stride 2 should a CNN have assuming an input resolution of 64×64 and convolutional layers that do not change the resolution, and why?

$$\begin{bmatrix} 1 & 1 & 2 & 4 \\ 5 & 6 & 7 & 8 \\ 3 & 2 & 1 & 0 \\ 1 & 2 & 3 & 4 \end{bmatrix} \quad \Longrightarrow$$

Give a general overview of convolutional neural networks and their purpose, and draw a sketch that illustrates their overall structure (typical layer types and their arrangement). What are the two overall stages of such networks?

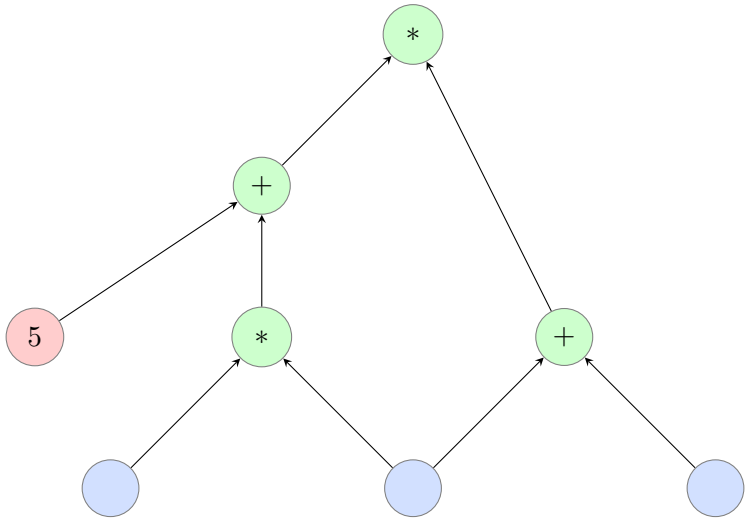
How is the depth of a CNN defined? What effect does increasing the depth have? How does one choose a suitable network depth to solve a given image classification problem?

What is the backend of a CNN? Discuss the backends of VGGNet and GoogLeNet/ResNet and their pros and cons.

What are residual networks (ResNets) and which problem do they overcome? Explain what a residual block computes and create a sketch of such a block.

What is the purpose of the backpropagation algorithm? Assume the following computational graph. First insert digits from your Matrikelnummer into the empty input nodes, going from right to left both in terms of nodes and digits. (Assuming Matrikelnummer 0123456, the values of the rightmost node would be 6, that of the node left of it would be 5, and so on.) Then compute the partial derivative of the topmost node with respect to all input nodes via backpropagation. Write computation node values after the forward pass left of the nodes, local gradients left of the edge connecting the corresponding nodes,

and “cached” partial derivatives of the topmost node right to the nodes. Explain the steps of the algorithm at a given node.



What is the purpose of data augmentation? Assume that the task is to train a digit classifier. Think of and explain data transformations that are applicable in this case, and at least one that is not.

What is the purpose of dropout, how does it work, and why is it effective? Why do “dropped” neurons have no effect on the output of the next layer? To which CNN layers is dropout commonly applied?

What is the purpose and aim of batch normalization? Why does it have a regularizing effect? To which CNN layers is batch normalization applied and where?

What does the learning rate hyperparameter specify? Draw a sketch that shows how training progresses over time (epochs) when setting the learning rate much too high, too high, properly, and too low, and explain the sketch. Explain a heuristic for adapting the learning rate during training and why doing can be beneficial.

What is the purpose of oversampling and how does it work? What is ten-crop oversampling? Assume the task of dog breed classification. Think of and explain transformations that are applicable in this case. What about ten-crop oversampling?

What is the purpose and intuition behind using model ensembles? How might the individual models differ from each other? How is the output of these models combined?

What are the challenges in using medical imaging data? Can deep networks be used effectively for medical tasks (give 3 examples)? How can we use the training data most efficiently?

What is the difference between 2D, 2.5D and 3D deep CNN networks?

What approaches are key components to use deep CNNs in medical imaging applications, especially when data for training is not or sparsely available? Describe two approaches and the benefit of using them in medical imaging.

Explain the challenges, benefits and drawbacks of using expert vs. non-experts to obtain labels of medical image data? Give an example of an alternative method to expert-based manual labeling of medical data for deep learning approaches.

Describe the main purpose of RNNs and give an example application. Draw and explain a sketch illustrating the overall architecture of such networks. What is the main limitation of traditional RNNs (as opposed to LSTMs)?

What is the main advantage of LSTMs over traditional RNNs? Draw a sketch that illustrates the dataflow within a LSTM block. List and explain the purpose of the different LSTM gates.