Introduction to Deep Learning
Exam SS18, 16.07.2018

Multiple-/Single-Choice questions

- ❑ Means: several answers could be correct, tick all correct ones

- ● Means: only one answer is correct

And Regression

Given is the following and-operation between inputs x_1 and x_2.

X_1	x_2	and(x_1, x_2)
1	1	1
1	0	0
0	1	0
0	0	0

Which of the following weights and biases represents a valid linear regression for that task.

- ● W = (1,1), b = 2
- ● W = (1,1), b = -0.5
- ● W = (1, 1), b = -1.5
- ● None of the above

Which of the following introduce nonlinearities?

- ❑ Convolutions
- ❑ Pooling
- ❑ Batch Normalization
- ❑ ReLUs

Why do we use multiples of 2 for the mini batch size?

- ❑ GPUs are optimized for that
- ❑ Makes computation faster
- ❑
- ❑

Short Questions

What kind of architecture would you choose for a network to translate english to german?

Why is minimizing the cross entropy between desired output distribution q and and actual estimation distribution p (H(q,p) the same as minimizing their KL divergence KL(q,p) = H(q,p) - H(p)?

Batch Normalization

Given is the Batch-Normalization-operation

$$y = \frac{(x-\mu)}{\sigma}\xi + \phi$$

1. Compute the gradients of the loss function with respect to the parameters ξ, ϕ given the gradient $\partial L / \partial y_{ij}$.
2. How can you accumulate the backward gradients? ?? Very unsure about that one

Convolution

Given is a simple convolutional operation. You have an input of dimension 2x2x2 and apply a convolutional kernel with filter size 1 and one output channel. This convolutional layer is followed by a max-pooling layer of size 2x2 with stride 2 and a padding of 1.

Given are the following values:

$$x = (\lceil 1 -1.5; 1 -2 \rceil, \lceil -21; -1.51 \rceil), ytarget = [01; 10]$$

1. Compute the output.
2. Compute the binary cross-entropy loss.
3. Compute the weight update given a gradient with respect to the outputs consisting of ones and a learning rate of 1.

You have an input of size 50x50x200 and a convolutional layer with kernel 10x10, your output should be size 10x10x10, how many multiplications does one forwards pass involve?
What would you do, to reduce this number?

Basic network stuff

Given is a deep fully-connected network with 15 layers that uses tanh as an activation function.

1. What do large inputs mean for the gradients?
2. Do any problems arise from that?
3. Now the weights are initialized with small random weights. Do you foresee any problems?
4. Name a method that makes the network robust against initialization. (1 point)
5. If one now uses a ReLU as activation-function, which problems are solved? Which problems do still exist?

Optimization

1. Explain difference between vanilla SGD and RMSProp.
2. Explain difference between Newton and quasi-Newton-methods, like BFGS.
3. Explain difference between Newton methods and the Gauss-Newton method.
4. Explain beta_1 and beta_2, the parameters of the Adam optimizer.

You have two Models, *Model 1* and *Model 2* and train them the same way. You find out, that *Model 1* has too few parameters, to perform well. What is this called. Draw test and validation loss.
You now severely increase the model complexity of *Model 1*, but *Model 2* still performs better. What happened now? Draw again training and validation loss.

Other questions:

1. Name two techniques used in Deep learning where you use a dataset separate from the training dataset
2. Name two technique for learning rate scheduling.

Formulas for a Batch normalization layer was given, calculate the partial derivative with respect to the scaling and shifting parameter.

You decrease you mini batch size, what do you do with your learning rate, why?

In general:
-It was 17 pages of mostly short questions (90 points in total, mostly 2 per question, sometimes 4, sometimes 1)
-time was tight, but not undable