Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2019/2020

Exercise Sheet 8

Deadline: 19.01.2021, 23:59

Instructions

For a deeper understanding of the material:

- Posts by Roan Gylberth on medium.com
- Practical Recommendations for Gradient-Based Training of Deep Architectures, Yoshua Bengio, 2012

Exercises

Exercise 8.1 - Possible Problems (1.5 + 2 + 1 = 4.5 points)

a) One of the optimization challenges is *ill-conditioning*. For answering the following questions read this article (6 min read). Answer the questions in your own words.

- Read part 8.2.1 Ill-Conditioning and explain why very small steps increase cost function when the Hessian matrix is ill-conditioned. Start from the equation of the second-order Taylor series expansion of the cost function.
- In practice, how can we spot ill-conditioning?
- What can we do to solve the problem of ill-conditioning?

b) Another challenge discussed in the lecture is the so-called Exploding gradient problem. To be more specific, this problem happens only during the backward pass in training (very deep) Neural Networks. Assume that you have a 100-layer Feed Forward Neural Network with ReLU activation function as non-linearities. Explain the phenomenon of exploding gradient with the formula of the backward pass. How can we avoid the problem of gradient explosion?

c) In the lecture you learned that neural networks have a large number of local minima. Are all local minima problematic? Explain in 2-3 sentences. How can you test if your network is stuck in a local minimum (or another critical point which is not global minimum)? For answering this question consult Chapter 8 of Deep Learning book.

Exercise 8.2 - Batch Size (1.5 + 0.5 + 1 = 3 points)

a) Discuss pros and cons of (1) stochastic (m=1), (2) batch (m = size of dataset) and (3) minibatch gradient descent (m is the number of points passed at a time). For each point provide one supporting sentence.

b) Why is it important to shuffle data before applying minibatch gradient descent?

c) When deciding on batch size it is important to think about the technical aspects. Read this article (7 min read) about GPU and CPU and discuss the relation between batch size and the choice of processor.
* Note that when using GPUs, it is common that batch sizes of power of 2 (e.g. 32 or 128) offer better runtime. Here you can read why.

Exercise 8.3 - SGD with Momentum (2.5 points)

Familiarize yourself with the SGD with momentum from the lecture slides (slides 16-20 in Chapter 8) and Deep Learning book and understand how it works. It is known that the cost function of NNs usually has many saddle points (DL book chapter 8.2.3). How does SGD with momentum help to alleviate the problem of getting stuck at these saddle points when compared to vanilla SGD (Chapter 8 slide 15)? Describe a situation in which vanilla SGD will get stuck at one saddle point, while SGD with momentum won't.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You have to submit the solutions of this assignment sheet as a team of 2-3 students.
- Hand in a **single** PDF file with your solutions.
- Make sure to write the student Teams ID and the name of each member of your team on your submission.
- Your assignment solution must be uploaded by only **one** of your team members to the course website.
- If you have any trouble with the submission, contact your tutor **before** the deadline.