Mechanistic Interpretability

Housekeeping
WT 2024/25

Frederick Riemenschneider

17.10.2024

Overview

1 Preliminaries

2 Organization

3 What now?

Preliminaries

About me

- **mail**: riemenschneider@cl.uni-heidelberg.de

What you should already know about

- Neural Networks and how to ~~tame~~ train them
- common NLP architectures and concepts
  - LSTM
  - seq2seq
  - attention
  - Transformer

Organization

This seminar

- rest of today: organisation and paper overview
- next week: fundamental concepts
- after that: presentations by students
- literature list with schedule on the course page

How to get points

- **participation**
- **presentation**
- **implementation project**

How to get points

- **active participation**
  - no more than one unexcused absence
  - active participation in classroom discussion
- **preparation**
  - Read all scheduled papers.
  - Hand in two questions or comments about each paper via mail.
  - deadline: each tuesday before the seminar, 3pm
  - part of your participation grade

Presentation

- preparatory seminar
  - usually one paper
  - $\sim$ 30 minutes
- main seminar
  - usually two papers
  - $\sim$ 60 minutes
- **Discuss the presentation with me before the seminar.**

Presentation grading

- presentation content
  - explain methods and results
  - point out strengths and weaknesses
- presentation style
  - structure
  - clarity of the presentation
  - design of the slides, use of illustrations, etc.

Presentation grading

- A good presentation adds to the paper.
  - Put it in context with related work/what we have seen in the seminar.
  - If there is code, look at it to explain unclear passages.
- source for all illustrations you use
- **text is evil**

Implementation project

- max. 8 pages (both PS and MS)
- possibilities
  - clean reimplementation of one of the approaches
  - exploration of one of your own ideas

Submission and final grade

- submission of all final projects and papers by 31st of March
- If you do a second presentation, you are done by the end of the semester.
- final grade is made up of
    - participation (30%)
    - presentation (40%)
    - implementation project or second presentation (30%)

What now?

What now?

- write a mail to riemenschneider@cl.uni-heidelberg.de containing...
  - ...three topics that you would like to present, ranked by your preferences.
  - ...at most one date on which you can absolutely not present (current dates might change).
  - ...whether you want to take it as a preparatory or main seminar
  - ...your name.