Collaborative Text Filtering

Peter J. Bakke, Daniel Horvath, Christian Hansen and Thomas Brenner

DTU Compute, Technical University of Denmark

Introduction

Collaborative text filtering is one of the most popular and effective approaches for recommender systems. Recommender systems are based on the idea, that given previously collected data about users and their interactions with items, you can predict whether a given user wants to have an interaction with a given item. This is widely used for platforms like Netflix, Amazon, Youtube and news websites. These platforms can increase their profits by being able to predict their consumers interests and showing content relevant for the user.
The purpose of this project is to match two text descriptions of varied lengths. More concretely job applications with job descriptions. The hope is of course, to have a model that matches job applicants with open job positions.

Data

- We use the publicly available **MovieLens** dataset from `https://grouplens.org/datasets/movielens/` for the first part of our project
- We use the publicly available **CiteULike** from `http://www.citeulike.org/faq/data.adp` for the second part of our project
- We use a private dataset called **TalentFox** for the final part of our project

Key points

- We construct a baseline model using **Matrix Factorization** on the MovieLens, CiteUlike, and TalentFox data
- We construct a **Collaborative Text Filtering** model on the same data using
  - Feed Forward Networks
  - LSTM Networks

  And compare the results to the baseline model
- We implement the models using the **Pytorch** deep learning framework and **TorchText**
- We train the models on the **Google Colab** GPU cloud

MovieLens

The MoveieLens dataset consist of 20 million ratings on a scale from 1 to 5, of 27,000 different movies by 138,000 users. Taking outset in the MovieLens dataset the objective is to predict how a specific user will rate a specific movie.

Table 1: Results

Model	Best accuracy	Best Epoch	Something else
MF	0.1337	2	0.1337
FNN	0.1337	12	0.1337

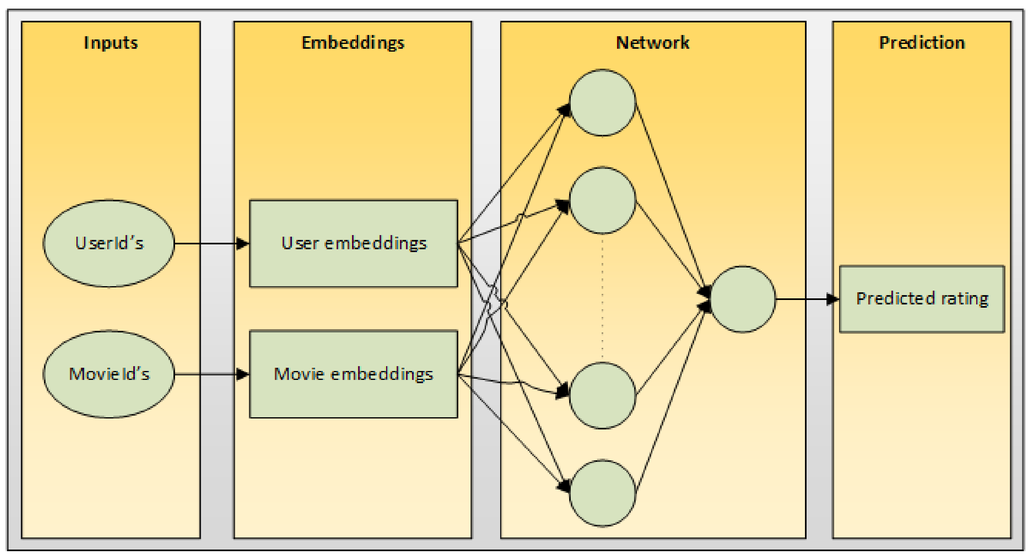

Figure 1: MovieLens neural net representation

CiteULike

The CiteULike dataset consist of users that have marked that they have read different scientific articles. In total we have approximately 200k articles spread across approximately 5500 users Taking outset in the CiteULike dataset the objective is to predict whether a specific reader is interested in an article.

Table 2: Results

Model	Best accuracy	Best Epoch	Something else
MF	0.1337	2	0.1337
FNN	0.1337	12	0.1337
LSTM	0.1337	24	0.1337

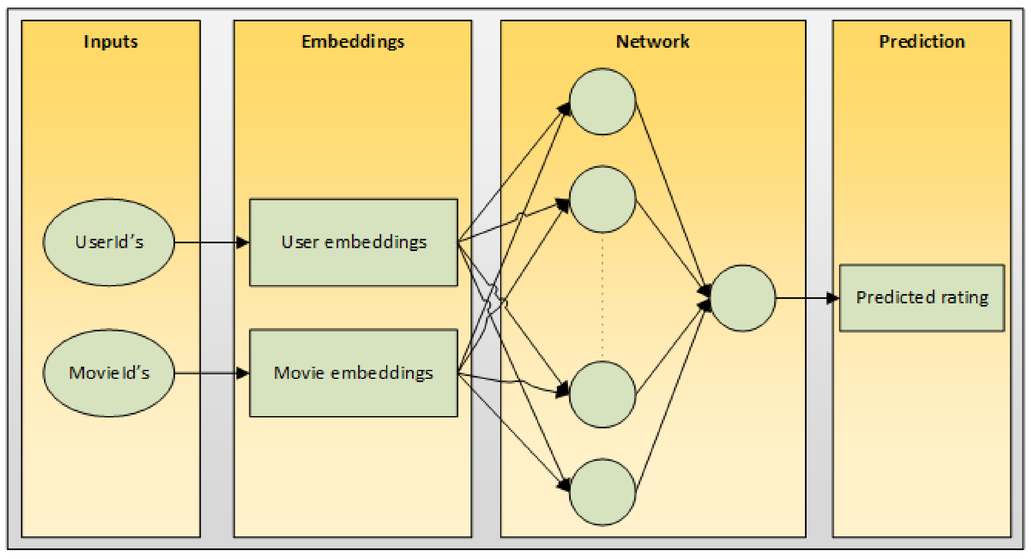

Figure 2: Need to draw the LSTM net

TalentFox

The TalentFox dataset consist of job applicants and job descriptions. The modelling aim is to find the best suited candidate for a specific job description

Table 3: Results

Model	Best accuracy	Best Epoch	Something else
MF	0.1337	2	0.1337
FNN	0.1337	12	0.1337
LSTM	0.1337	24	0.1337

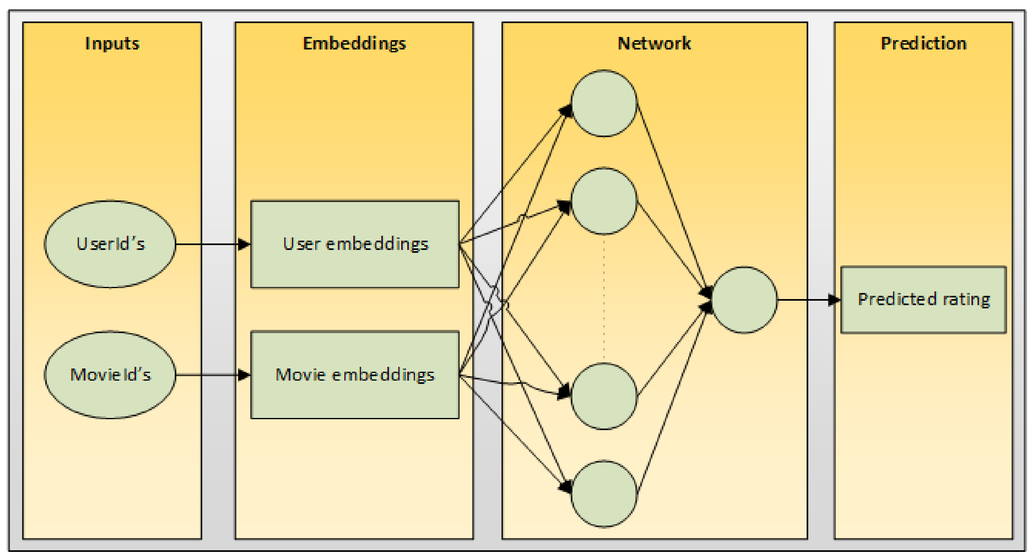

Figure 3: Need to draw another diagram here

Comparison of results

Table 4: Confusion matrix of the test set on the final DeepLoc model using profiles encoding. Sens. = Sensitivity, MCC = Matthews Correlation Coefficient

Location	Number of predicted proteins										Sens.	MCC
Nucleus	680	103	4	5	2	8	1	2	2	1	0.842	0.784
Cytoplasm	94	361	7	18	5	4	3	8	1	7	0.711	0.608
Extracellular	3	5	365	5	5	4	2	0	4	0	0.929	0.907
Mitochondrion	9	21	0	247	0	5	14	2	1	3	0.818	0.812
Cell membrane	5	15	6	1	203	20	1	4	18	0	0.744	0.732
Endoplasmic ret.	3	6	6	3	18	120	1	7	8	1	0.694	0.654
Plastid	1	2	0	8	0	0	140	0	1	0	0.921	0.883
Golgi apparatus	4	17	1	0	9	8	1	26	4	0	0.371	0.414
Lysosome/Vacuole	0	7	11	1	20	9	0	4	12	0	0.188	0.194
Peroxisome	0	13	0	4	1	4	0	0	0	8	0.267	0.321

More results

We are able to represent what regions in the sequence are relevant for each subcellular localization to perform the prediction.

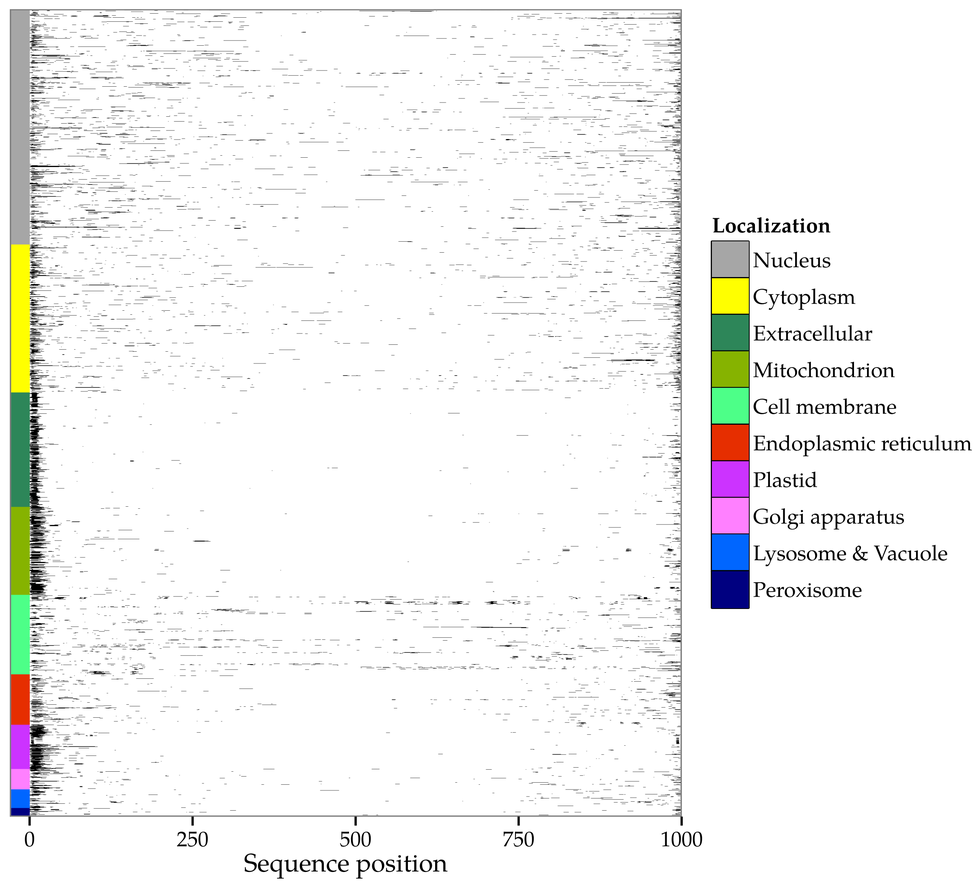

Figure 4: Sequence importance across the protein sequence of DeepLoc test set when making the prediction.

Acknowledgements

The authors wish to thank Alexander R. Johansen and Jose Juan Almagro Armenteros from DTU Lyngby for their constructive feedback and fruitful discussions during the process of the project.

References

[1] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5):367–374, 2004.

s183268@student.dtu.dk