# Outer Product-based Neural Collaborative Filtering
## Collaborative Filtering Project Report

Akhil Goel[1][2015126], Anirudh Singh[1][2015015], and Mohit Agarwal[1][2015060]

IIIT Delhi

**Abstract.** Single cell RNA sequencing is emerging as a powerful tool in biomedical research. Developed by Tang et al. [10], the method can be used to study heterogeneity and inner workings of the cell. However there are few challenges associated with this method and one such is the presence of dropout events in the data generated through it. In our project we leverage the outer product-based neural collaborative filtering, proposed by He et al.[4] to impute dropouts in single cell expression data.

**Keywords:** scRNA-seq · ConvNCF · Outer Product · NDCG, Hit Ratio

## 1 Introduction

Transcriptomics is the study of the transcriptome which in simple terms refers to the complete set of RNA transcripts that are produced by the genome under specific circumstances. Comparison of these transcriptomes allows us to identify genes that are expressed differently in distinct cell populations. These transcriptomes can be studied in significant detail using RNA sequencing methods. These studies have provided several valuable insights into complex biological systems and have contributed notably towards stem cells and cancer cells research among others.

Major advancements in RNA-sequencing in the last couple of decades have revolutionized the field of transcriptomics. RNA-sequencing has been mostly done using ”Bulk” RNA-seq since its breakthrough introduction in 2000s. However, as bulk RNA-seq measures the ”average” expression level of a gene across a large (bulk) population of input cells, it is insufficient in case of heterogenous systems and might obscure the relevant biological differences across the heterogenous population.

Single-cell RNA-seq (scRNA-seq), a new technology first published in 2009 by Tang et al [10] represents an approach to overcome this problem. scRNA-seq can provide us with the expression profile of ”individual” cells and thus allows the assessment of fundamental properties of cell populations at an unprecedented resolution. Considering the huge impact of scRNA-seq, contributing to its research is of even more importance.

Despite its benefits, scRNA-seq has its own share of challenges. The most prominent of them being ”gene dropouts” in which a gene is observed at a moderate expression level in one cell but is not even detected in another cell. In simple terms, some of the non-zero entries in the cell-gene data matrix is missed and interpreted as zero. This is majorly due to the low number of RNA transcriptomes (i.e. expression level data of a gene is not available for all the cells) and the stochastic and variable nature of the gene expression across cells. Not including the missed entries in the analysis can have a drastic impact in such a high stake scenario where any negligence can lead to severe changes in the results of cells like cancer. This calls for a need of coming up with efficient methods to impute (replacing missing data with substituted values) scRNA-seq datasets.

The problem of imputing the scRNA-seq datasets is closely related to that of collaborative filtering where we have a user-item matrix and a rating corresponding to some user-item pairs. This user-item matrix is also partially filled and the objective is to estimate the complete matrix with the values we already have. Analogous to this, we have cells (users) and genes (items) in our scRNA-seq datasets, the expression level (rating) of a gene across cells and the objective here again is to estimate partially filled (because of dropouts) cell-gene matrix with the expression values we already have. Using this idea, various techniques have been devised to impute scRNA-seq datasets like McImpute, DrImpute, scImpute. McImpute [8], is a low-rank matrix based completion technique that uses nuclear norm minimization(NNM) for imputing the dropouts in single cell expression data. Another method scImpute developed by Li and Li[7], imputes dropouts by first calculating for each cell the probability of dropout of a gene using a mixture model and then taking the information of same gene from other similar cells. These techniques have shown encouraging results and thus validates our claim of modelling the imputation problem of cell-gene matrix using a collaborative filtering approach.

Recently a new technique Outer product-based Neural Collaborative Filtering (ONCF), proposed by He et al. [4] has been introduced in the collaborative filtering literature. It essentially computes outer product of the embedding dimensions of user and item, feeds it into a ConvNet, to finally generate an item recommendation list. It has proven to be more practical than traditional model-based, matrix factorization-based methods. It harnesses the benefits of both outer product (for modelling expressive correlations between embeddings) and convolutional networks (for effective feature learning). We believe that using this method on the scRNA-seq datasets can bring us closer to resolving the dropout problem associated with these datasets.

## 2 Outer product-based Neural Collaborative Filtering

Outer Product-based Neural Collaborative Filtering or ONCF (Figure 1) uses the outer product of the embedding of the one hot encoded user and item features, unlike other methods that either concatenate or do element-wise dot product of the embedding dimensions for the recommendation. The 2D interaction map of pairwise correlations of embedding dimensions generated through outer product has following advantages:

- It subsumes matrix factorization.
- It consists of more encoded interaction signal as compared to correlation generated through element-wise product of the embedding dimensions
- It is useful for a deep learning model to generalize better on sparse data in comparison with the features generated through simple concatenation of the user and item embedding dimensions

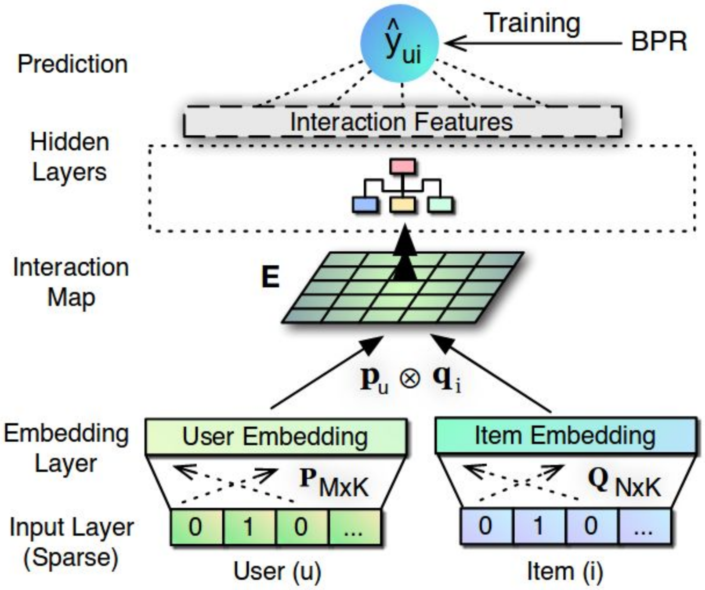

**Fig. 1.** Illustration of the working of ONCF model. (Image taken from paper[4])

### 2.1 ConvNCF

Various methods such as NCF[5] employ MLP (Multi-Layer Perceptron) classifier to extract high-level features from feature vector for the recommendation purpose. However the MLP classifier has three major drawbacks which makes it less useful to extract feature from the 2D interaction map:

- Since MLP classifier model has a large number of parameters, it requires machines that are powerful and have large memories to store it.
- It requires huge training data.
- It needs to be carefully tuned on the regularization of each layer to ensure good generalization performance.

ONCF, therefore uses convolutional neural collaborative filtering (ConvNCF) (Figure 2) model for extracting high-level features from 2D interaction map of pairwise correlations. The ConvNCF model is comparatively more stable and generalizable than the MLP classifier as it uses CNN which has few parameters having smaller magnitudes.
The ConvNCF in our experiment uses 6 convolutional layers with rectified linear unit(ReLU) activation function. The final feature map generated is used for the prediction score.

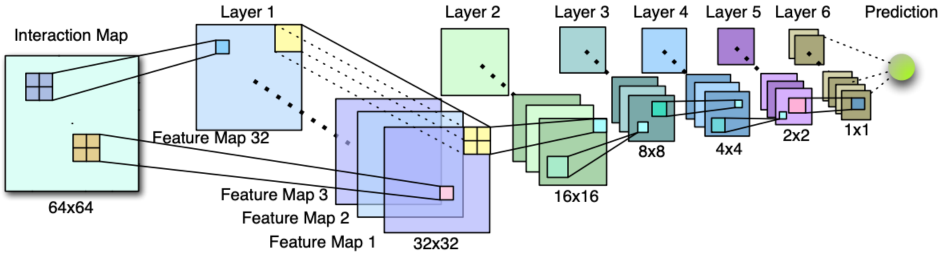

**Fig. 2.** Illustration of the working of ConvNCF model. (Image taken from paper[4])

### 2.2 Objective function

The objective function used to optimize ConvNCF model is Bayesian Personalized Ranking (BPR)[9] objective function.

$$L(\varDelta) = \sum_{(u,i,j)\in D} -ln\sigma(y_{u,i} - y_{u,j}) + \lambda_{\varDelta}||\varDelta||^2 \tag{1}$$

where $y_{u,i}$ and $y_{u,j}$ are the prediction scores on $i^{th}$ and $j^{th}$ item respectively, $\lambda_{\varDelta}$ are parameter specific regularization parameters and D is the set of training instances such that D = {(u,i,j) — $i \in Y_u^+ \cap j \notin Y_u^+$}. Here $Y_u^+$ denotes the items that user u has consumed. It improves the ranking of observed items by maximizing the score of observed items and minimizing the score of unobserved items. Since ONCF here is used for personalized ranking task, the objective function can customize it to correctly predict relative order between the interactions as opposed to pointwise loss objective function which only minimizes the error in actual score prediction.

## 3 Experiments

### 3.1 Datasets

For the experiment the following datasets were used:

- **Yelp**: This dataset is obtained from Yelp dataset challenge having user ratings on various businesses. The dataset consists of 730,791 ratings done by 25,815 users on 25,677 items.
- **MovieLens 100K**: MovieLens 100K [3] dataset has 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 items.
- **Blakeley**: The dataset, developed by Blakeley et al.[1], contains single RNA sequence data of 30 cells from human embryo. The cells have been annotated on the basis of their association with three cell lineages of human blastocyst: epiblast, trophectoderm and PE(Primitive Endoderm).
- **Jurkat-293T**: This dataset, developed by Zheng et al.[14], contains expression profiles of 3,388 cells. The cells are annotated Jurkat or 293-T if they expressed CD3D or XIST respectively.
- **Kolodziejczyk**: This dataset, developed by Kolodziejczyk et al. [6], contains single cell RNA sequence data of 704 mouse embryonic stem cells. The cells in the dataset have been annotated based on the conditions they were cultured: 2i, ai or LIF.
- **Preimplantation**: This dataset[12] consists single cell RNA sequence data of 317 cells from mouse preimplantation embryos. Since these cells are from zygote at different stages they have been annotated into 13 categories: 16cell, 4cell, 8cell, BXC, C57twocell, early2cell, earlyblast, late2cell, lateblast, mid2cell, midblast, fibroblast and zygote
- **Quake**: This dataset, developded by Darmanis et al.[2], consists of expression profiles of 461 healthy human brain cells. The cells have been categorized into 9 major neuronal, glial and vascular cell types in the brain.
- **Usoskin**: This dataset, developed by Usoskin et al.[11], consists single cell RNA sequence data of 622 mouse neurons. PCA classification was used to derive and assign one of the 4 labels: NP, PEP, NF and TH to them.
- **Zeisel**: This dataset, developed by Zeisel et al.[13], consists of data of 3005 cells from the mouse somatosensory cortex (S1) and hippocampal CA1 region. The molecular distinctiveness of these cells, discovered using divisive biclustering method, was used to annotate them into 9 different categories.

### 3.2 Metrics

The following metrics were used to evaluate the performance of ONCF on the above datasets:

- **Hit Ratio (HR)**: It is a recall based metric that measures the occurence of the test item in the top k items in the recommendation list generated by the recommended system.
- **Normalized Discounted Cumulative Gain (NDCG)**: It is a ranking based metric that is used to measure the performance of the recommended system based on the position of the items in its recommended list. The items in top k ranks are given high scores.

### 3.3 Protocol

Before feeding to the network, the cell-genes data is pre-processed using the process.m procedure described in [8][1]. Training-Testing splits are designed in the following fashion:

- All but one rated item by a user is added to the training set.
- The remaining one rated item is added to the testing set.
- $n$ entries out of the total number of un-rated items by the user are added to the negative set[2].

ConvNCF model is trained and tested on the above splits using Bayesian Personalized Ranking (BPR) as the objective function and HR@10 and NDCG@10 as evaluation metrics.

For our experiments we have compared the results of ConvNCF with an existing approach called MF-BPR [9]. In MF-BPR Rendle et al. formulate the item recommendation problem in a standard Matrix Factorization fashion. They proceed to optimize the formulation using their proposed BPR optimization which is derived from the maximum posterior estimator for optimal personalized ranking.

### 3.4 Github submission

The code of our project can be found at[3]

## 4 Results

### 4.1 Results Reproducibility

We were able to reproduce the results produced by He et al. in their experiments on Yelp dataset[4]. Table 4.1 shows the highest value of HR@10 and NDCG@10 and Figure 3 shows the HR@10 and NDCG@10 plots in He et al.'s experiments and our experiments on Yelp dataset.

| Metric | **Highest value in He et al.'s experiments** | **Highest value in our Experiments** |
|---|---|---|
| **HR@10** | 0.31 | 0.31 |
| **NDCG@10** | 0.162 | 0.16 |

**Table 1.** Analysis of highest value of HR@10 and NDCG@10 in He et al.'s experiments and our experiments on Yelp dataset

[1] https://github.com/aanchalMongia/McImpute_scRNAseq

[2] $n$ is a heuristic which depends on the dataset architecture and size

[3] https://github.com/akhil15126/CF-Project

[4] https://github.com/duxy-me/ConvNCF

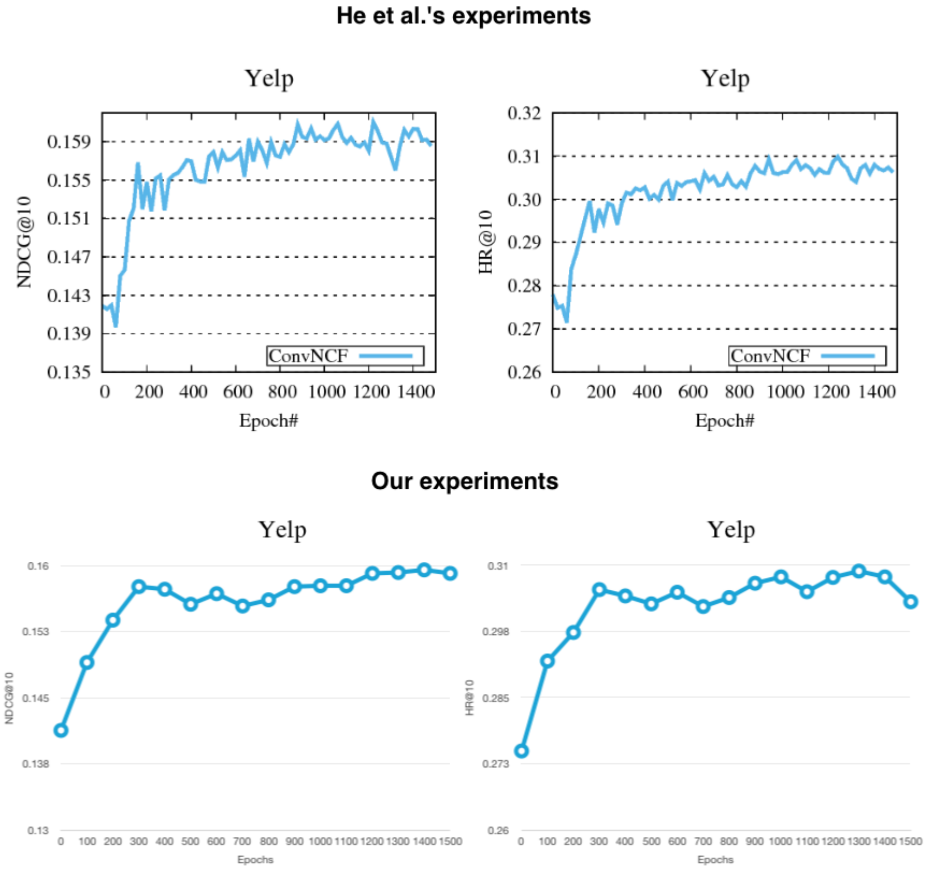

**Fig. 3.** HR@10 and NDCG@10 on Yelp dataset by He et al. and our experiments.

| | **Best ConvNCF** | | **Best MF-BPR** | |
|---|---|---|---|---|
| **Dataset** | **HR@10** | **NDCG@10** | **HR@10** | **NDCG@10** |
| Blakeley | **0.9** | **0.8667** | 0.7667 | 0.6802 |
| Movie Lens | **0.08** | **0.047** | 0.03 | 0.015 |
| Jurkat | **0.502** | **0.3562** | 0.4906 | 0.3462 |
| Kolodziejczyk | **0.6477** | **0.512** | 0.6449 | 0.5119 |
| Preimplantation | **0.703** | **0.615** | 0.5836 | 0.4455 |
| Quake | **0.41** | **0.2975** | 0.2885 | 0.1752 |
| Usoskin | **0.4224** | **0.2973** | 0.4019 | 0.2663 |
| Yelp | **0.30** | **0.159** | 0.2880 | 0.1506 |
| Zeisel | **0.306** | **0.206** | 0.2825 | 0.1814 |

**Table 2.** Comparison of ConvNCF results with MF-BPR on different datasets

### 4.2 Movie Lens

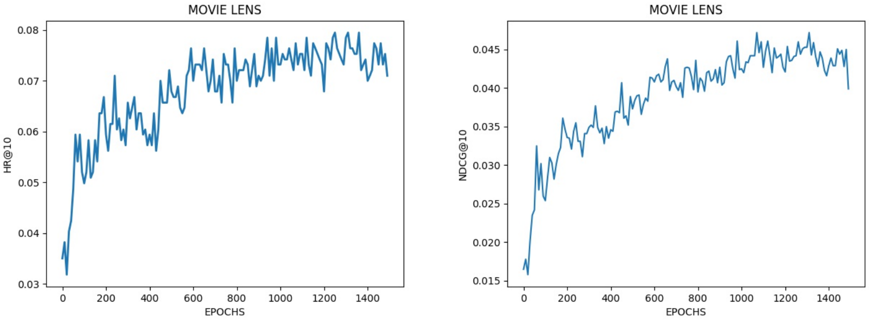

**Fig. 4.** HR@10 and NDCG@10 on Movie Lens dataset

HR value starts at around 0.03 and converges at around 0.075 whereas NDCG value starts at around 0.015 and converges at around 0.045. Refer Fig. 4

### 4.3 Blakeley

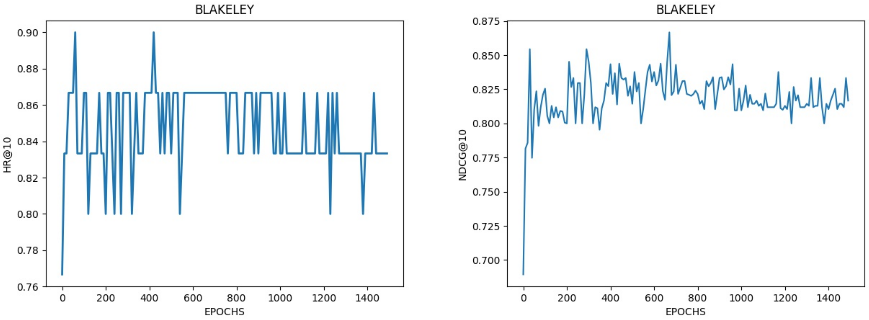

**Fig. 5.** HR@10 and NDCG@10 on Blakeley dataset

HR value starts at around 0.76 and converges at around 0.85 whereas NDCG value starts at around 0.6 and converges at around 0.825. Refer Fig. 5

### 4.4 Jurkat

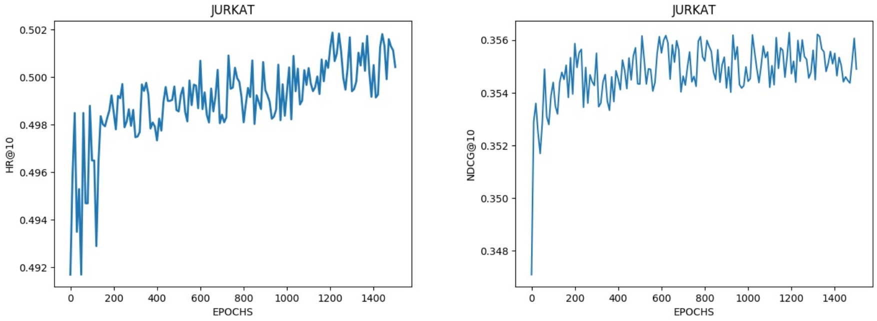

**Fig. 6.** HR@10 and NDCG@10 on Jurkat dataset

HR value starts at around 0.492 and converges at around 0.5 whereas NDCG value starts at around 0.34 and converges at around 0.356. Refer Fig. 6

### 4.5 Kolodziejczyk

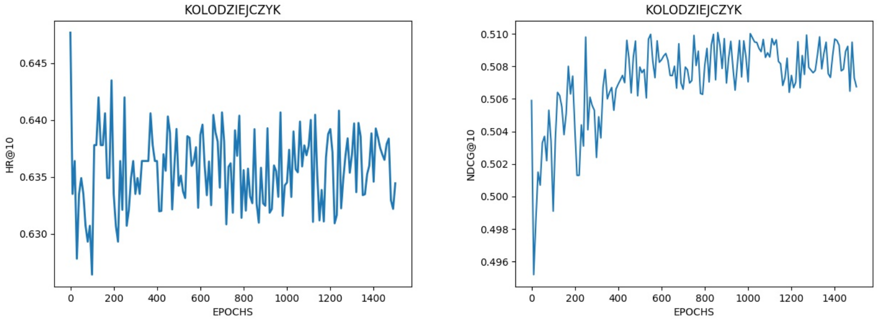

**Fig. 7.** HR@10 and NDCG@10 on Kolodziejczyk dataset

HR value starts at around 0.629 and converges at around 0.635 whereas NDCG value starts at around 0.48 and converges at around 0.508. Refer Fig. 6

### 4.6 Preimplantation

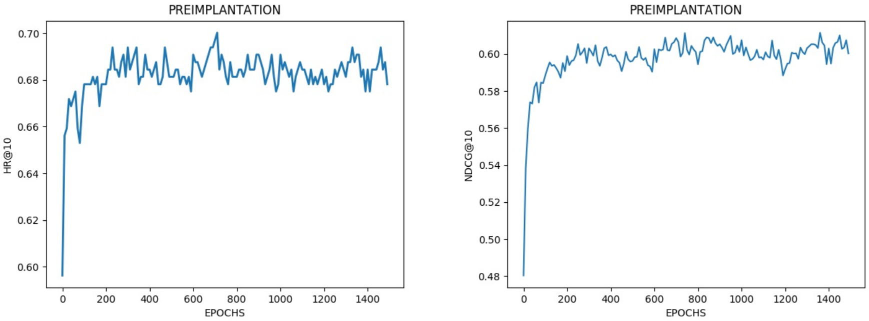

**Fig. 8.** HR@10 and NDCG@10 on Preimplantation dataset

HR value starts at around 0.6 and converges at around 0.69 whereas NDCG value starts at around 0.48 and converges at around 0.6. Refer Fig. 8

### 4.7 Quake

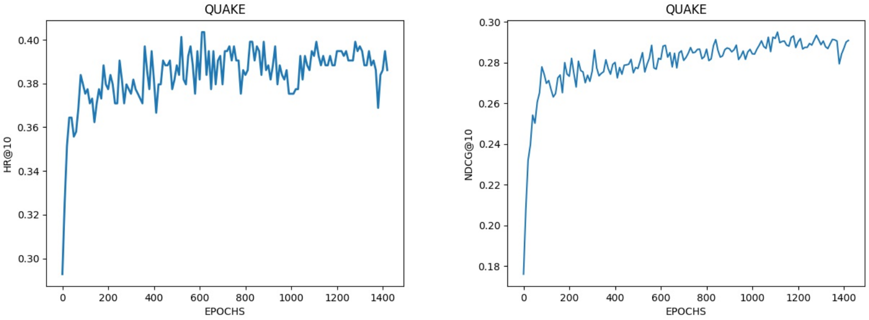

**Fig. 9.** HR@10 and NDCG@10 on Quake dataset

HR value starts at around 0.28 and converges at around 0.4 whereas NDCG value starts at around 0.17 and converges at around 0.3. Refer Fig. 9

### 4.8 Usoskin

HR value starts at around 0.35 and converges at around 0.41 whereas NDCG value starts at around 0.24 and converges at around 0.29. Refer Fig. 10

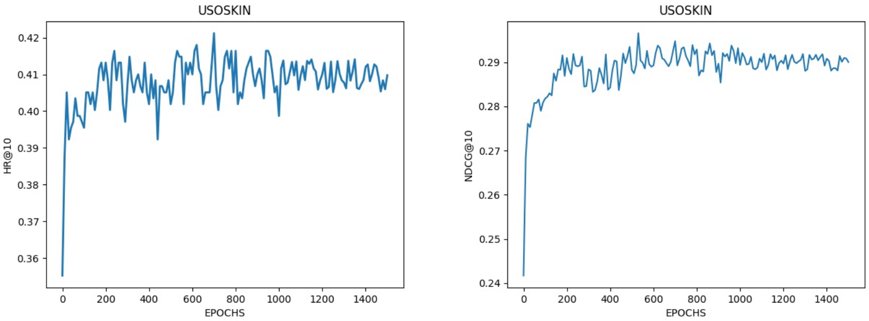

**Fig. 10.** HR@10 and NDCG@10 on Usoskin dataset

### 4.9 Zeisel

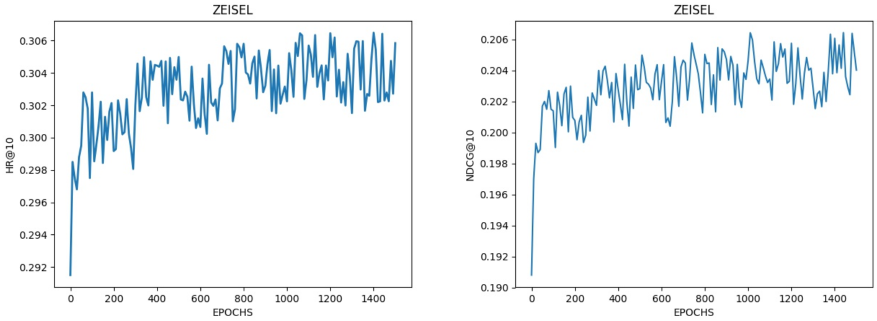

**Fig. 11.** HR@10 and NDCG@10 on Zeisel dataset

HR value starts at around 0.29 and converges at around 0.304 whereas NDCG value starts at around 0.185 and converges at around 0.206. Refer Fig. 11

### 4.10 Comparison with MF-BPR

Table 2 compiles the result of ConvNCF model and MF-BPR model on different datasets. It is clearly evident that ConvNCF provides a higher Hit Ratio and a higher Normalized Discounted Cumulative Gain on all the tested datasets.

## 5 Challenges

Following challenges were faced while working through the project:

- **Time** : Dimensionalities of the datasets were huge. Operations like convolution on the outer product space and optimization over a number of trainable parameters were very time consuming.
- **Limited Hardware Capacity** : With the limited capabilities of a standard i-5 processor, handling such expensive operations was hard.

The first challenge was tackled by starting early with the project and the second challenge was handled by running our codes on Google Collaboratory (a cloud based Jupyter notebook which provides free access to a Tesla K80 GPU on a 12 hour lease)

## 6 Conclusion

scRNA sequencing has proved to be groundbreaking in measuring gene expression levels for different cells. However, the presence of gene dropouts in the data can hinder this progress of discovering new properties of the cell that may have brought significant advances in the field of biomedical research. We have presented a new approach that leverages the advantages of both outer product and convolutional networks to solve one of the major challenges associated with single cell RNA sequencing technique. The proposed method showed an incredible performance in imputing the dropout events in terms of Hit ratio and NDCG in comparison with the benchmark method MF-BPR. The application of outer product and convolutional network for imputing the dropouts is a stepping stone to solve the problem in a much more efficient way and can advance this field further to a point where we can solve this problem much more reliably and practically.

## 7 Acknowledgement

This work was done with support from our professor Dr. Angshul Majumdar. We thank him and Ms. Aanchal Mongia for providing us the dataset.

## References

1. Paul Blakeley, Norah ME Fogarty, Ignacio Del Valle, Sissy E Wamaitha, Tim Xiaoming Hu, Kay Elder, Philip Snell, Leila Christie, Paul Robson, and Kathy K Niakan. Defining the three cell lineages of the human blastocyst by single-cell rna-seq. *Development*, pages dev–123547, 2015.
2. Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290, 2015.
3. F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.
4. Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua. Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*, 2018.
5. Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.
6. Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Jason CH Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N Natarajan, Alex C Tuck, Xuefei Gao, Marc Bühler, Pentao Liu, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell stem cell*, 17(4):471–485, 2015.
7. Wei Vivian Li and Jingyi Jessica Li. scimpute: accurate and robust imputation for single cell rna-seq data. *BioRxiv*, page 141598, 2017.
8. Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Mcimpute: Matrix completion based imputation for single cell rna-seq data. *bioRxiv*, page 361980, 2018.
9. Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 452–461. AUAI Press, 2009.
10. Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377, 2009.
11. Dmitry Usoskin, Alessandro Furlan, Saiful Islam, Hind Abdo, Peter Lönnerberg, Daohua Lou, Jens Hjerling-Leffler, Jesper Haeggström, Olga Kharchenko, Peter V Kharchenko, et al. Unbiased classification of sensory neuron types by large-scale single-cell rna sequencing. *Nature neuroscience*, 18(1):145, 2015.
12. Liying Yan, Mingyu Yang, Hongshan Guo, Lu Yang, Jun Wu, Rong Li, Ping Liu, Ying Lian, Xiaoying Zheng, Jie Yan, et al. Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology*, 20(9):1131, 2013.
13. Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.
14. Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott,

Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.