Deep learning tackles single-cell analysis - A survey of deep learning for scRNA-seq analysis

Mario Flores, Zhentao Liu, Tinghe Zhang, Md Musaddaqui Hasib, Yu-Chiao Chiu, Z

2021-09-13

Contents

Front Matter

Authors

Mario Flores[1§], PhD, is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Texas at San Antonio, and joint program Faculty of Biomedical Engineering at the University of Texas Health San Antonio. Before joined ECE, he was a postdoctoral fellow at the National Center for Biotechnology Information of the National Institutes of Health from 2015 to 2019. His research focuses on DNA and RNA sequence methods, transcriptomics analysis, epigenetics, comparative genomics, and deep learning to study mechanisms of gene regulation, single-cell RNA-seq, and Natural Language Processing.

Zhentao Liu1 is a PhD student in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. His research focuses on deep learning for cancer genomics and drug response prediction.

Tinghe Zhang1 is a PhD student in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. His research focuses on deep learning for cancer genomics and drug response prediction.

Md Musaddaqui Hasib1 is a PhD student in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. His research focuses on interpretable deep learning for cancer genomics.

Yu-Chiao Chiu2, PhD, is a postdoctoral fellow at the Greehey Children's Cancer Research Institute at the University of Texas Health San Antonio. His postdoctoral research is focused on developing deep learning models for pharmacogenomic studies.

Zhenqing Ye2,3, PhD, is an assistant professor in the Department of Population Health Sciences and the director of Computational Biology and Bioinformatics at Greehey Children's Cancer Research Institute at the University of Texas Health San Antonio. His research focuses on computational methods on next generation sequencing and single-cell RNA-seq data analysis.

Karla Paniagua1* is a PhD student in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. Her research

focuses on ~

Sumin Jo1 is a PhD student in the Department of Electrical and Computer Engineering, the University of Texas at San Antonio. Her research focuses on m6A mRNA methylation and deep learning for biomedical applications.

Jianqiu Zhang1, PhD, is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Texas at San Antonio. Her current research focuses on deep learning for biomedical applications such as m6A mRNA methylation.

Shou-Jiang Gao4,6, PhD, is a Professor in UPMC Hillman Cancer Center and Department of Microbiology and Molecular Genetics, University of Pittsburgh. His current research interests include Kaposi's sarcoma-associate herpesvirus (KSHV), AIDS-related malignancies, translational and cancer therapeutics, and systems biology.

Yufang Jin1, PhD, is a Professor in the Department of Electrical and Computer Engineering at the University of Texas at San Antonio. Her research focuses on mathematical modeling of cellular responses in immune systems, data-driven modeling and analysis of macrophage activations, and deep learning applications.

Yidong Chen2,3, PhD, is a Professor in the Department of Population Health Sciences and the director of Computational Biology and Bioinformatics at Greehey Children's Cancer Research Institute at the University of Texas Health San Antonio. His research interests include bioinformatics methods in next-generation sequencing technologies, integrative genomic data analysis, genetic data visualization and management, and machine learning in translational cancer research

Yufei Huang5,6, PhD, is a Professor in UPMC Hillman Cancer Center and Department of Medicine, School of Medicine, University of Pittsburgh. His current research interests include uncovering the functions of m6A mRNA methylation, cancer virology, and medical AI & deep learning.

§Corresponding authors: Mario Flores (mario.flores@utsa.edu), Yidong Chen (cheny8@uthscsa.edu), and Yufei Huan (yuh119@pitt.edu)

Author Affiliations

1Department of Electrical and Computer Engineering, the University of Texas at San Antonio, San Antonio, TX 78249, USA

2Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA

3Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA

4Department of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, PA 15232, USA

5Department of Medicine, School of Medicine, University of Pittsburgh, PA 15232, USA

6UPMC Hillman Cancer Center, University of Pittsburgh, PA 15232, USA

Book Maintainer

Hello there :D

Any feedback and contributions will be appreciated.

Mail: sumin.jo@utsa.edu

Website: Huang-AI4Medicine-Lab

About this book

This book is full version of our research paper [Journal-Info-Here].

Keywords deep learning; single-cell RNA-seq; imputation; dimention reduction; clustering; batch correction; cell type identification; functional prediction; visualization

Abstract

Since its selection as the method of the year in 2013, single-cell technologies have become mature enough to provide answers to complex research questions. However, together with the growth of single-cell profiling technologies, there has also been an increase of computational challenges to process the generated datasets. It's here that by effectively leveraging large data sets, Deep Learning (DL) is positioning as the first option for single-cell analyses. Here we provide a unified mathematical description of the DL methods used in single cell RNA sequencing (scRNA-Seq) followed with the survey of the most representative published DL algorithms for scRNA-Seq in the field.

Key Points

- Single cell RNA sequencing technology generate large collection of transcriptomic profiles of up to millions of cells, enabling biological investigation of hidden structures or cell types, predicting their effects or responses to treatment more precisely, or utilizing subpopulation to address unanswered hypotheses.
- Current Deep Learning-based analysis approaches for single cell RNA seq data is systematically reviewed in this paper according to the challenge they address and their roles in the analysis pipeline.
- A unified mathematical description of the surveyed DL models is presented and the specific model features were discussed when reviewing

each approach.

- A comprehensive summary of the evaluation metrics, comparision algorithms, and datasets by each approaches is presented.

Chapter 1

Introduction

Single cell sequencing technology has been a rapidly developing area to study genomics, transcriptomics, proteomics, metabolomics, and cellular interactions at the single cell level for cell-type identification, tissue composition and reprogramming (**??**). Specifically, sequencing of the transcriptome of single cells, or single-cell RNA-sequencing (scRNA-seq), has become the dominant technology in many frontier research areas such as disease progression and drug discovery (**???**). One particular area where scRNA-seq has made a tangible impact is cancer, where scRNA-seq is becoming a powerful tool for understanding invasion, intratumor heterogeneity, metastasis, epigenetic alterations, detecting rare cancer stem cells, and assessing therapeutic response . Currently, scRNA-seq is applied to develop personalized therapeutic strategies that are potentially useful in cancer diagnosis, therapy resistance during cancer progression, and the survival of patients (**??**). The scRNA-seq has also been adopted to combat COVID-19 to elucidate how the innate and adaptive host immune system miscommunicates resulting in worsening immunopathology produced during this viral infection (**?**).

These studies have led to a massive amount of scRNA-seq data deposited to public databases such as Single Cell PORTAL, Single Cell Expression Atlas, PanglaoDB, and scRNASeqDB. Expressions of millions of cells from 18 species have been collected and deposited, waiting for further analysis. On the other hand, due to biological and technical factors, scRNA-seq data presents several analytical challenges related to its complex characteristics like missing expression values, high technical and biological variance, noise and sparse gene coverage, and elusive cell identities (**?**). These characteristics make it difficult to directly apply commonly used bulk RNA-seq data analysis techniques and have called for novel statistical approaches for scRNA-seq data cleaning and computational algorithms for data analysis and interpretation. To this end, specialized scRNA-seq analysis pipelines such as Seurat [9] and Scanpy [10], along with a large collection of task-specific tools, have been developed to address the intri-

cate technical and biological complexity of scRNA-seq data.

Recently, deep learning has demonstrated its significant advantages in natural language processing and speech and facial recognition with massive data. Such advantages have initiated the application of DL in scRNA-seq data analysis as a competitive alternative to conventional machine learning approaches for uncovering cell clustering [11, 12], cell type identification [11, 13], gene imputation [14-16], and batch correction [17] in scRNA-seq analysis. Compared to conventional machine learning (ML) approaches, DL is more powerful in capturing complex features of high-dimensional scRNA-seq data. It is also more versatile, where a single model can be trained to address multiple tasks or adapted and transferred to different tasks. Moreover, the DL training scales more favorably with the number of cells in scRNA-seq data size, making it particularly attractive for handling the ever-increasing volume of single cell data. Indeed, the growing body of DL-based tools has demonstrated DL's exciting potential as a learning paradigm to significantly advance the tools we use to interrogate scRNA-seq data.

In this paper, we present a comprehensive review of the recent advances of DL methods for solving the present challenges in scRNA-seq data analysis (Table 1) from the quality control, normalization/batch effect reduction, dimension reduction, visualization, feature selection, and data interpretation by surveying deep learning papers published up to April 2021. In order to maintain high quality for this review, we choose not to include any (bio)archival papers, although a proportion of these manuscripts contain important new findings that would be published after completing their peer-reviewed process. Previous efforts to review the recent advances in machine learning methods focused on efficient integration of single cell data [18, 19]. A recent review of DL applications on single cell data has summarized 21 DL algorithms that might be deployed in single cell studies [20]. It also evaluated the clustering and data correction effect of these DL algorithms using 11 datasets.

In this review, we focus more on the DL algorithms with a much detailed explanation and comparison. Further, to better understand the relationship of each surveyed DL model with the overall scRNA-seq analysis pipeline, we organize the surveys according to the challenge they address and discuss these DL models following the analysis pipeline. A unified mathematical description of the surveyed DL models is presented and the specific model features are discussed when reviewing each method. This will also shed light on the modeling connections among the surveyed DL methods and the recognization of the uniqueness of each model. Besides the models, we also summarize the evaluation matrics of these DL algorithms and compare the tools that integrate these DL algorithms. Access to these DL algorithms with the original research results, available datasets used by these methods are also listed to demonstrate the advantages and utility of the DL algorithms. We envision that this survey will serve as an important information portal for learning the application of DL for scRNA-seq analysis and inspire innovative use of DL to address a broader range of new challenges

in emerging multi-omics and spatial single-cell sequencing.

–>

Bibliography