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 (Lahnemann et al., 2020; Vitak et al., 2017). 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 (Wu et al., 2017; Bost et al., 2020; Kinker et al., 2020). 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 (Navin, 2015; Mannarapu et al., 2021). 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 (Wauters et al., 2021).

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 (Lahnemann et al., 2020). 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 intricate 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.

–>

Chapter 2

Overview of scRNA-seq processing pipeline

Various scRNA-seq techniques (like SMART-seq, Drop-seq and 10X genomics sequencing [21, 22] are available nowadays with their sets of advantages and disadvantages. However, the data content and processing steps of different scRNA-seq techniques are quite standard and conventional. A typical scRNA-seq dataset consists of three files: genes quantified (gene IDs), cells quantified (cellular barcode) and a count matrix (number of cells $\times$ number of genes), irrespective of the technology or pipeline used. Conventionall, a series of essential steps are applied to these matrices for different analysis objectives as illustrated Fig. 1. These steps are:

of the technology or pipeline used. A series of essential steps in scRNA-seq data processing pipeline and optional tools for each step with both ML and DL approaches are illustrated in Fig. 1.

With the advantage of identifying each cell and unique molecular identifiers (UMIs) for expressions of each gene in a single cell, scRNA-seq data are embedded with increased technical noise and biases [23]. Quality control (QC) is the first and the key step to filter out dead cells, double-cells, or cells with failed chemistry or other technical artifacts. The most commonly adopted three QC covariates include the number of counts (count depth) per barcode identifying each cell, the number of genes per barcode, and the fraction of counts from mitochondrial genes per barcode [24].

Normalization is designed to eliminate imbalanced sampling, cell differentiation, viability, and many other factors. Approaches tailored for scRNA-seq have been developed including the Bayesian-based method coupled with spike-in, or BASiCS [25], deconvolution approach, scran [26], and sctransfrom in Seurat where regularized Negative Binomial regression was proposed [27]. Two

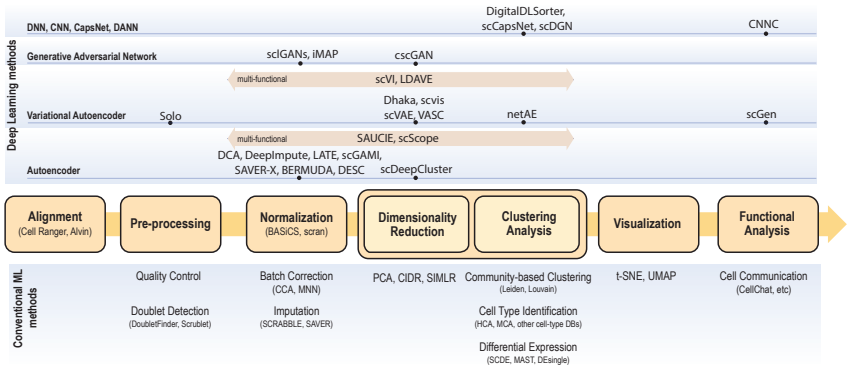

Figure 1

Figure 2.1: Single cell data analysis steps for both conventional ML methods (bottom) and DL methods (top). Depending on the input data and analysis objectives, major scRNA-se analysis steps are illustrated in the center flow chart. The convential ML approaches along with optional analysis modules are presented below each analysis step. Deep learning approaches are categorized as neural network models (DNN, CNN, CapsNet, and DANN), Generalive Adversarial Network (GAN), Variational Autoencoder, and Autoencoder. For each DL approach, optional algorithms are listed on top of each step in the pipeline.

important steps, batch correction and imputation, will be carried out if required by the analysis: • Batch Correction is a common source of technical variation in high-throughput sequencing experiments due to variant experimental conditions such as technicians and experimental time, imposing a major challenge in scRNA-seq data analysis. Batch effect correction algorithms include detection of mutual nearest neighbors (MNNs) [28], canonical correlation analysis (CCA) with Seurat [29], and Hormony algorithm through cell-type representation [30]. • Imputation step is necessary to handle high sparsity data matrix, due to missing value or dropout in scRNA-seq data analysis. Several tools have been developed to "impute" zero values in scRNA-seq data, such as SCRABBLE [31], SAVER [32] and scImpute [33]. Dimensionality reduction and visualization are essential steps to represent biological meaningful variation and high dimensionality with significantly reduced computational cost. Dimensionality reduction methods, such as PCA, are widely used in scRNA-seq data analysis to achieve that purpose. More advanced nonlinear approaches that preserve the topological structure and avoid overcrowding in lower dimension representation, such as LLE [34] (used in SLICER [35]), tSNE [36], and UMAP [37] have also been developed and adopted as a standard in single-cell data visualization.

Clustering analysis is a key step to identify cell subpopulations or distinct cell types to unravel the extent of heterogeneity and their associated cell-type-specific markers. Unsupervised clustering is frequently used here to categorize cells into clusters by their similarity often taken the aforementioned dimensionality-reduced representations as input, such as community detection algorithm Louvain [38] and Leiden [39], or data-driven dimensionality reduction followed with k-Means cluster by SIMLR [40].

Feature selection is another important step in single-cell RNA-seq analysis is to select a subset of genes, or features, for cell-type identification and functional enrichment of each cluster. This step is achieved by differential expression analysis designed for scRNA-seq, such as MAST that used linear model fitting and likelihood ratio testing [41]; SCDE that adopted a Bayesian approach with a Negative Binomial model for gene expression and Poisson process for dropouts [42], or DEsingle that utilized a Zero-Inflated Negative Binomial model to estimate the dropouts [43].

Besides these key steps, downstream analysis can include cell type identification, coexpression analysis, prediction of perturbation response, where DL has also been applied. Other advanced analyses including trajectory inference and velocity and pseudotime analysis are not discussed here because most of the approaches on these topics are non-DL based.

Bibliography

Bost, P., Giladi, A., Liu, Y., Bendjelal, Y., Xu, G., David, E., Blecher-Gonen, R., Cohen, M., Medaglia, C., Li, H., Deczkowska, A., Zhang, S., Schwikowski, B., Zhang, Z., and Amit, I. (2020). Host-viral infection maps reveal signatures of severe covid-19 patients. *Cell*, 181(7):1475–1488 e12.

Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M. S., McFarland, J. M., Warren, A., Rodman, C., Roth, J. A., Bender, S. A., Kumar, B., Rocco, J. W., Fernandes, P., Mader, C. C., Keren-Shaul, H., Plotnikov, A., Barr, H., Tsherniak, A., Rozenblatt-Rosen, O., Krizhanovsky, V., Puram, S. V., Regev, A., and Tirosh, I. (2020). Pan-cancer single-cell rna-seq identifies recurring programs of cellular heterogeneity. *Nat Genet*, 52(11):1208–1218.

Lahnemann, D., Koster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T. H., Lelieveldt, B. P. F., Mandoiu, I., Marioni, J. C., Marschall, T., Molder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J., Saliba, A. E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schonhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biol*, 21(1):31.

Mannarapu, M., Dariya, B., and Bandapalli, O. R. (2021). Application of single-cell sequencing technologies in pancreatic cancer. *Mol Cell Biochem*, 476(6):2429–2437.

Navin, N. E. (2015). The first five years of single-cell cancer genomics and beyond. *Genome Res*, 25(10):1499–507.

Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., Carbone, L., Steemers, F. J., and Adey, A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods*, 14(3):302–308.

Wauters, E., Van Mol, P., Garg, A. D., Jansen, S., Van Herck, Y., Vanderbeke, L., Bassez, A., Boeckx, B., Malengier-Devlies, B., Timmerman, A., Van Brussel, T., Van Buyten, T., Schepers, R., Heylen, E., Dauwe, D., Dooms, C., Gunst, J., Hermans, G., Meersseman, P., Testelmans, D., Yserbyt, J., Tejpar, S., De Wever, W., Matthys, P., collaborators, C., Neyts, J., Wauters, J., Qian, J., and Lambrechts, D. (2021). Discriminating mild from critical covid-19 by innate and adaptive immune single-cell profiling of bronchoalveolar lavages. *Cell Res*, 31(3):272–290.

Wu, H., Wang, C., and Wu, S. (2017). Single-cell sequencing for drug discovery and drug development. *Curr Top Med Chem*, 17(15):1769–1777.