Unsupervised transfer learning on PBMC single cell gene expression dataset

Sanket Suhas Deshpande, Dr. Debarka Sengupta

Centre for Computational Biology, Indraprastha Institute of Information Technology

Abstract

Transfer learning is domain of machine learning where a trained model for a task can be repurposed to do different tasks without starting from scratch. Transfer Learning reduces the vast compute and time resources required. Here we use Doc2Vec - Natural Language Processing method to train and create embeddings on PBMC single cell gene expression dataset. Goal is to train the model on a large size dataset and put this trained model and embeddings in public domain for other researchers to use it for classification or clustering.

Dataset

We are using PBMC single cell gene expression dataset for prototyping. The expression data are log counts per 10,000 based on UMI counts for all methods except for Smart-seq2, which is based on read counts. Data is very sparse and unbalanced to use it as it is.

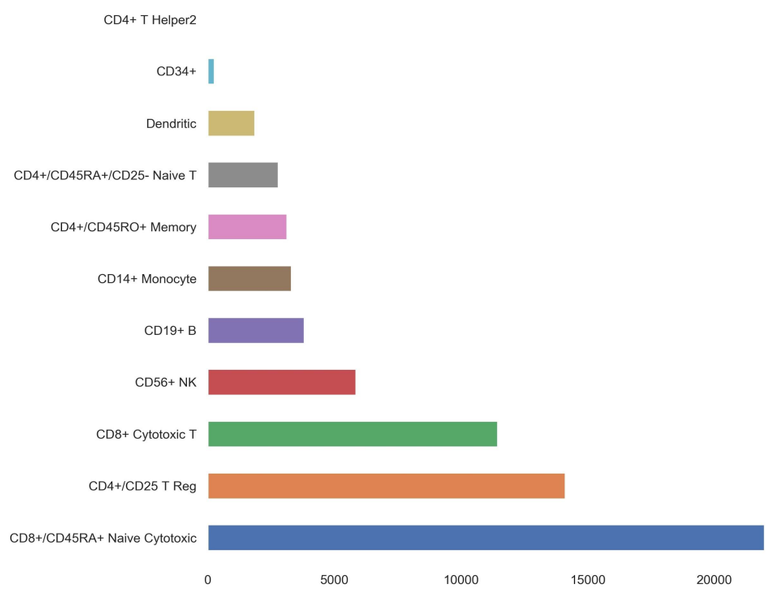

Fig: bar graph of number of cells per cell type

Introduction

Biological datasets are huge and using these whole datasets for statistical modeling/ machine learning is computationally expensive and time consuming. More efficient method is to train and create a generic model beforehand and then use it for specific tasks later on. These datasets exists in form of large sparse matrices. For any downstream analysis we need to represent data in fewer dimensions. Embeddings are fixed length vector representation of a sample. To create embeddings we are using doc2vec, each sample is considered a document and genes are the words in the document which are selected on the basis of different cutoffs on the z scores. The intuition behind using doc2vec is similar to its use in NLP, two documents are similar if they contain similar words in similar order. Here two samples are close to each other if their genes are expressed in almost similar range.

Methodology

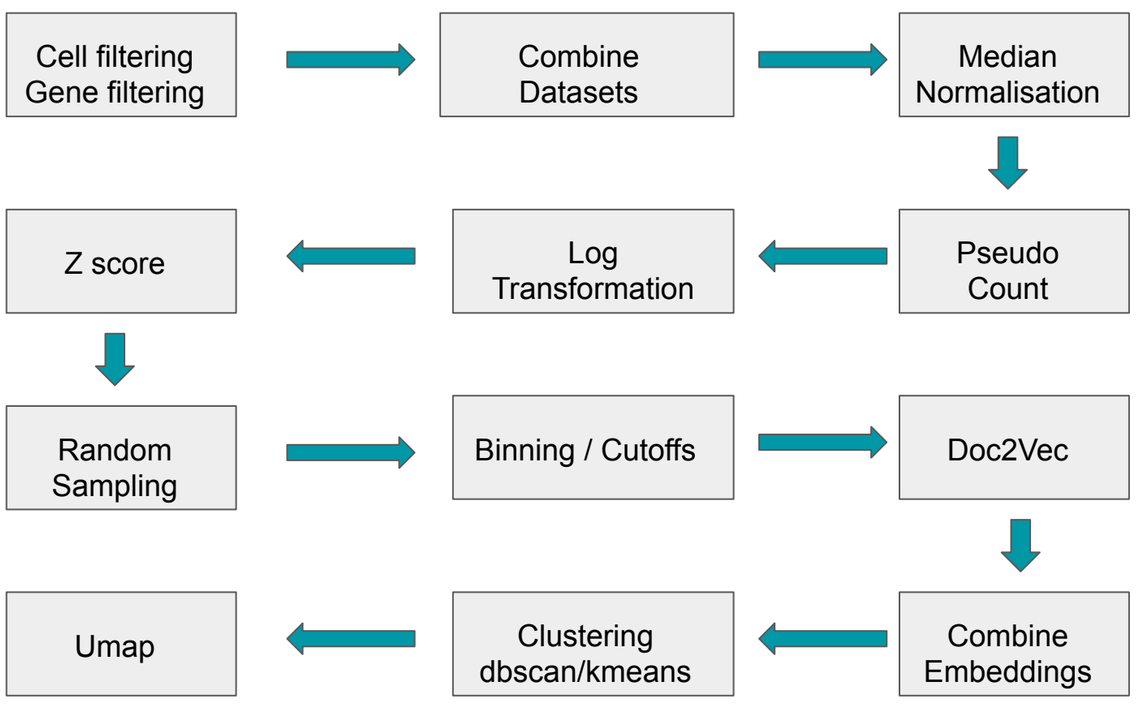

Fig: Pipeline followed in this project

Matrix2Dov2vec Model

Matrix to documents transformation by binarizing genes between 0 and 1.

SRM	NECA	SDHB	CAPZ	MINO	LYX	ID2	RAC1	CSK
1.3	-0.3	0.23	3.45	0.75	0.9	1.6	8.2	3.44
2.3	1.40	0.56	1.56	0.23	0..33	0.45	2.56	6.12
0.7	2.3	0.2	2.89	0.9	0.16	3.12	0.45	0.56
3.14	0.56	3.01	0.11	2.46	0.13	0.89	1.86	0.98

Doc1: “SDHB MINO LYX ”
Doc2: “SDHB MINO LYX ID2”
Doc3: “SRM SDHB MINO LYX RAC1 CSK”
Doc4: “NECA CAPZ LYX ID2 CSK”

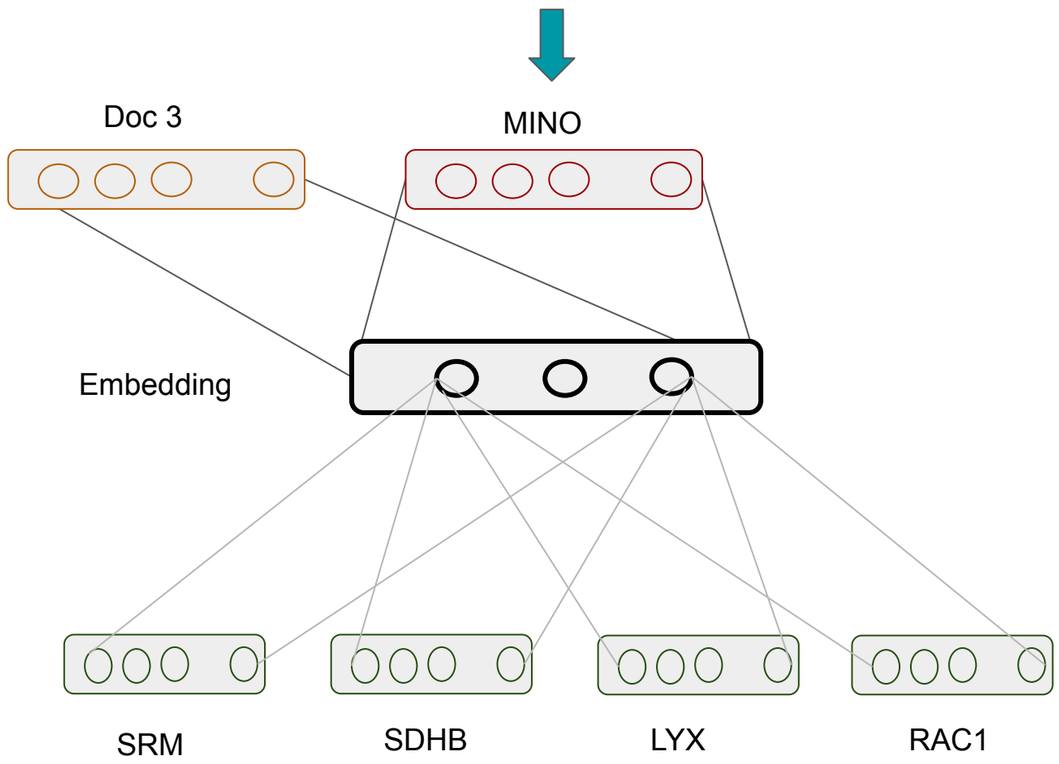

Fig: Neural network architecture of doc2vec

Clustering

Dbscan and kmeans clustering and umap on combined embeddings of vectors generated in each of the 5 bins.

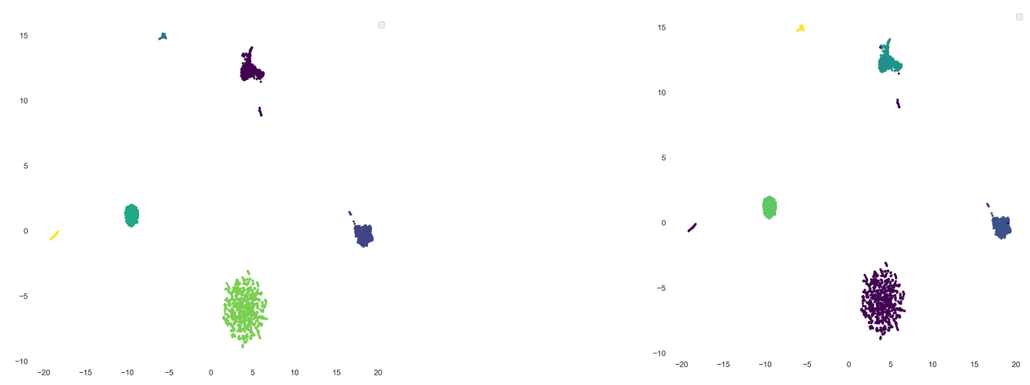

Fig: clusters formed using dbscan and kmeans

Future work

- Skip gram method for training by ordering genes on the basis of their location in DNA.
- Validation of clusters and embeddings.
- Increase training capacity of neural network.

References

1] Sinha D, Kumar A, Kumar H, Bandyopadhyay S, SenguptaD, ‘dropClust: efficient clustering of ultra-largescRNA-seq data.’ Nucleic Acids Res. 2018

2] Quoc Le, Tomas Mikolov, ‘Distributed Representations of Sentences and documents’ In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML’14). JMLR.org, II–1188–II–1196.