Single-cell multimodal integration

Predicting single-cell gene expression from chromatin accessibility with multi-layer perceptron and truncated singular value decomposition.

About dataset

Input: chromatin accessibility (row: cell, column: chromosome locations)

Target: gene expression (row: cell, column: genes), library-size normalized log1p

Extreme sparsity: 2% non-zero values;

High dimensionality: 229k for input, 23k for output;

(105942, 228942) -> (105942, 23418)

Preprocessing (input)

80-20 fixed training-validation split

Compression: convert to sparse matrix data type to save memory

Binarization: non-zero entries = 1, improves performance

tSVD dimensionality reduction: remove sparsity, concentrate information

Normalization: divide by row-wise mean

Non-negative matrix factorization and autoencoder: computationally expensive

Preprocessing (target)

80-20 fixed training-validation split

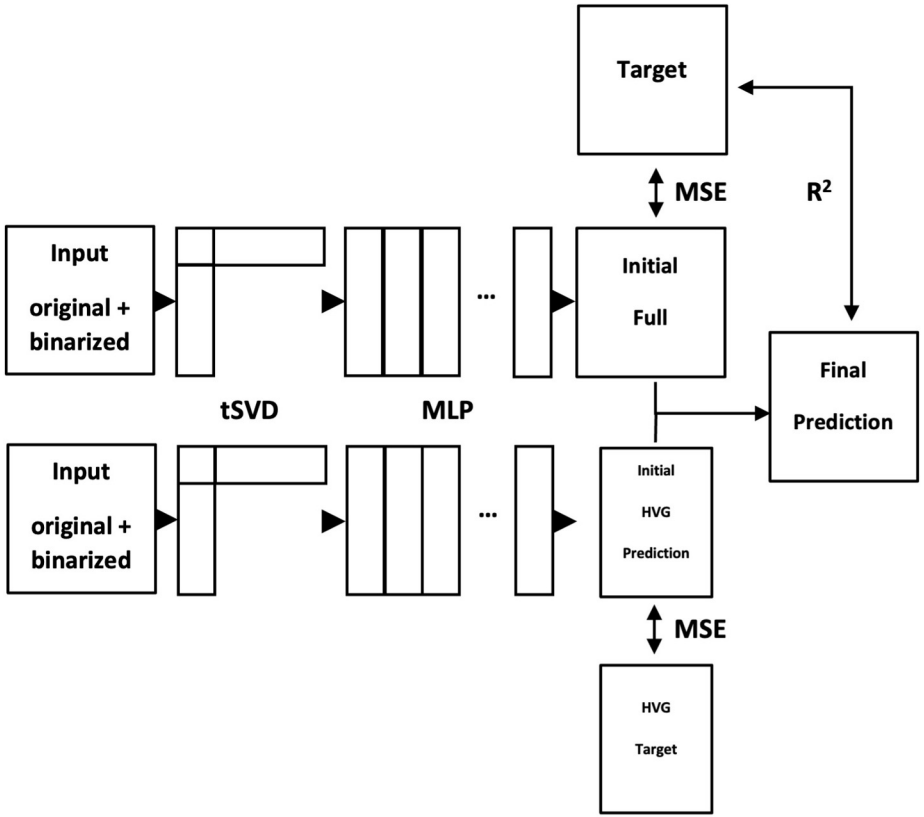
Compression: convert to sparse matrix data type to save memory

Scanpy highly variable gene (HVG) selection: compensatory model, 2048 in total

Normalization: divide by non-zero row-wise mean

tSVD on target data limits the ceiling of performance

Model architecture



Training

MLP1: 6-layer network with ReLU, BatchNorm and Dropout, hidden size: 2048

MLP2: 6-layer network with ReLU, BatchNorm and Dropout, hidden size: 2048

Batch size: 256, 100 epochs

AdamW optimizer, ReduceLROnPlateau scheduler, MSELoss

Software: PyTorch + Sklearn, trained on Google Colab

Current Result

Metric: Pearson correlation coefficient (self-implemented vectorized version)

MLP1: 0.66785, MLP2: 0.70221

My final score: 0.66838

Reference score: 0.670 (5th solution, not directly comparable)

Thank you.