DeepCellType: Refining Bulk Data with Deep Learning

Juan Irizarry-Cole

The Problem

"Bulk" biological data contains mixtures of numerous cell types.

Modern sequencing and expression analysis techniques are capable of single cell resolution.

Differential gene expression across cell types confounds usefulness of conclusions reached with bulk data.

Mountains of data remain convoluted within that bulk data

The Goal

Use a machine learning approach to automatically classify different coarse grain cell types.

Extracting a clear image from convoluted data requires a deconvolution approach

Robust deconvolution models already exist for images

The DREAM challenge specified 8 cell types to be labeled

Target Cell Types
Fibroblasts
B Cells
Endothelial Cells
NK Cells
Neutrophils
Myeloid Lineage Cells
CD4+ Cells
CD8 T Cells

The Caveats

Whereas image data is usually structured by pixel distance, allowing convolutional approaches to work, microarray data is not

- Possible future solutions: force the data indices to conform to actual genomic order

Likewise, microarray data is not time series data so they are not easily fed properly into a recurrent neural network

The Multilayer Perceptron is always available to us

However, an MLP can only do so much with data…

The Data (Oh brother)

- However, several problems exist with the data
  - Labels are extracted from reports automatically using regex, which may cause incorrect labeling of certain samples
  - The samples use different probe sets with different indices
  - Many of the samples are outside the 8 detected classes or ambiguously labeled
  - Even within the same probe sets, the length of the probe lists differ.
  - Even within probe sets of the same lengths, the indices are reordered.
  - Some samples simply don't exist

Cumulative Data Subset	Sample Size	Percentage of total
Total	49,320	100%
Data that exist and use the [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array probe set and which fall into one of the 8 listed cell types	9,400	19.1%
Data that exists with length 54,675	7,607	15.4%

Data Sparseness

7,607 samples is small for a complex network such as the genome

Without a defined order to the data, only a few deep learning model archetypes are easily usable

Is it even possible to train a model on a subset of this subset?

Yes.

Consistent Ordering

Throwing random numbers won't work

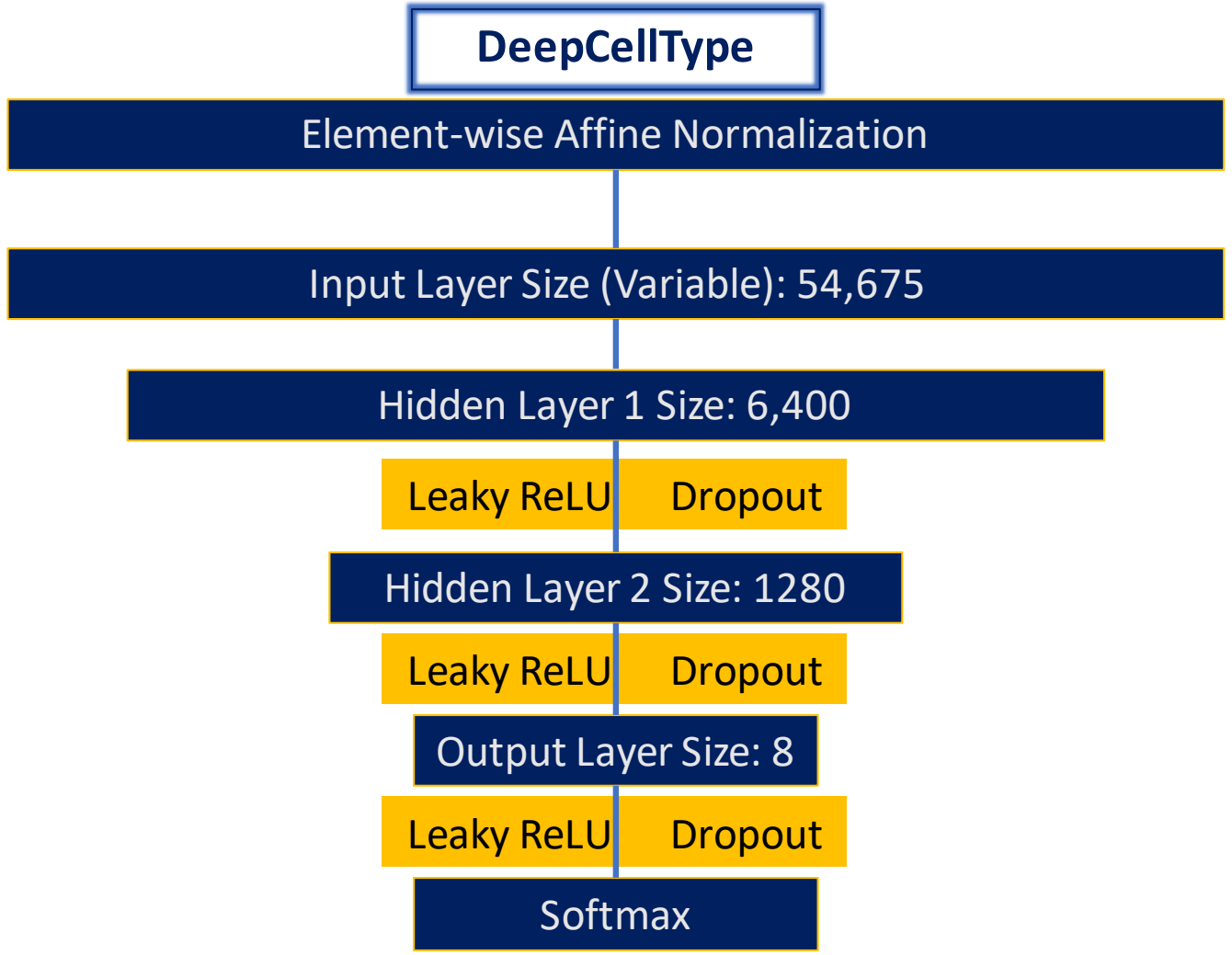
The order can be arbitrary as long as it is consistent

Using Python's Pandas package, one sample's index can act as the standard index across all samples of the same length.

- Maximum length may be unnecessary; future testing can involve minimal index and cross-platform sorting.

Remember: the main goal is to see if a model can learn at all.

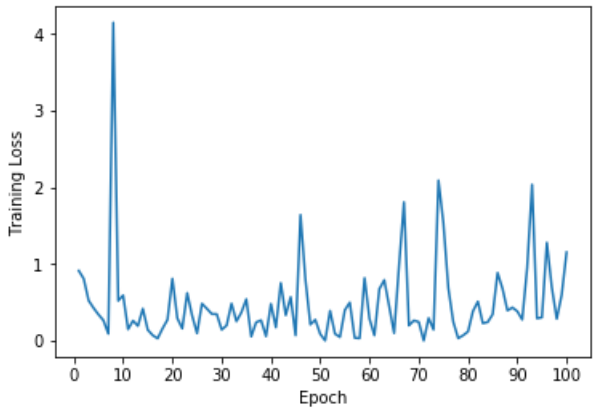
Finally, the Model



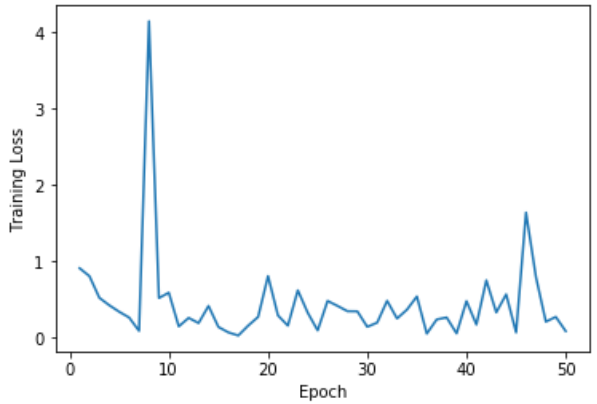
Training Attributes and Info

Attribute	**Info**
Model Parameters	35,932,718
Loss Function	Cross Entropy Loss
Optimizer	ADAM
Learning Rate	0.001
Dropout Probability	0.7
Epochs	100
Train/Test Ratio	0.9:0.1
Batch Size	50

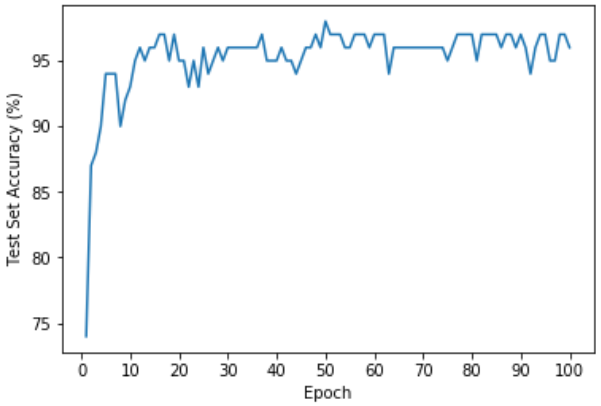
Loss over 100 Epochs



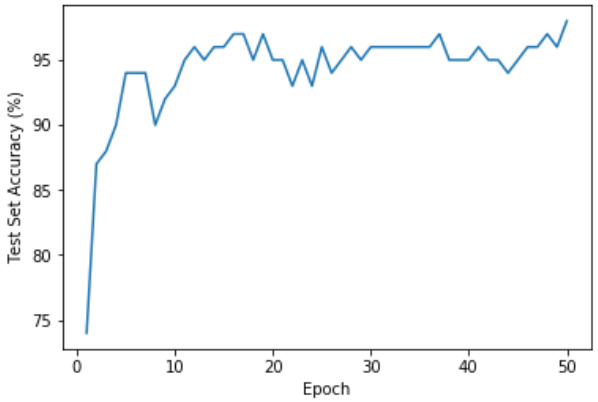
Loss over 50 Epochs



Accuracy over 100 Epochs



Accuracy over 50 Epochs

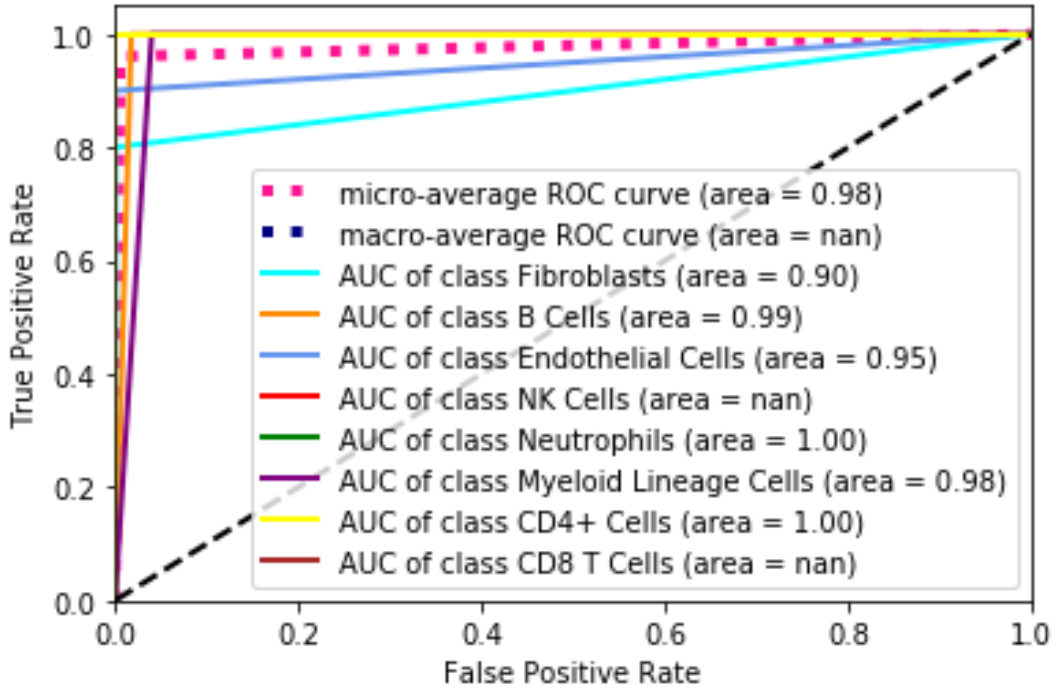


Training Results

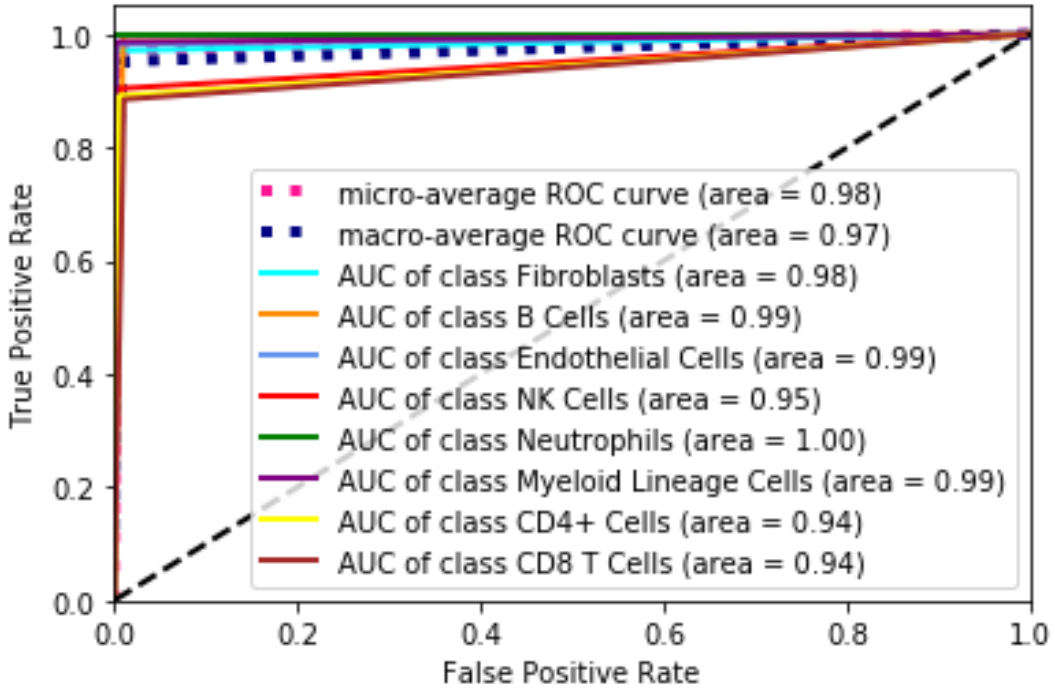
Peak Test Set Accuracy: ***98%***

The Moment of Truth

Area Under the Curve on 5% of the Training Set



Area Under the Curve on the Full Training Set



Conclusions, Discussions, and Next Steps

- A simple deep learning model is capable of learning deep connections in microarray data to predict broad stroke cell types with high confidence.
- Fine grain cell types may require more data and a more robust model. Reordering the data by genomic position may allow convolutions to improve results.
- Normalizing the data and implementing dropout layers made for drastically better results
- The data loader can be tweaked to give a more universally usable model using the Entrez gene IDs to standardize probe sets across many platforms and sample sizes.
- Additionally, the model can easily be adapted to bulk RNA-Seq data.

Thank you!

Questions?