**CM225 Project Paper**

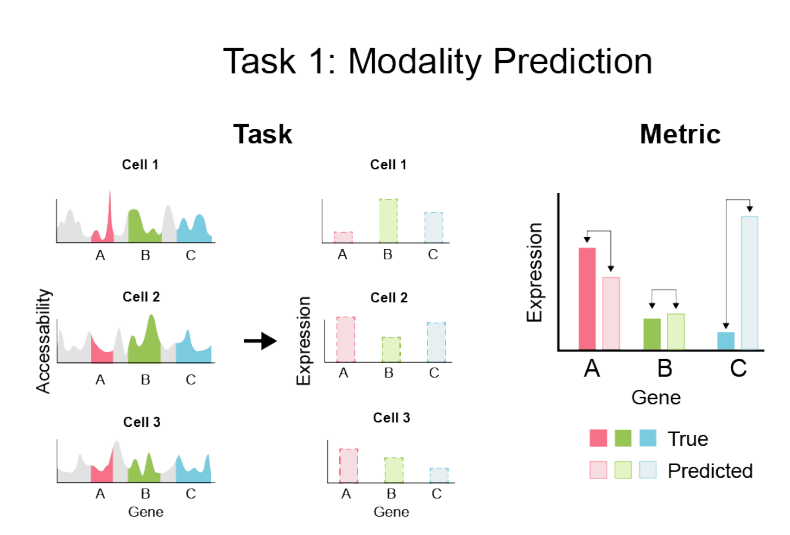
Ashish Kumar Singh (UID: 105479019)

**Predicting one modality from another of single cell data**

**Introduction**

Genetic information propagates through DNA to RNA to proteins. In this project, we test if it is possible to predict all the features of a modality from the features of another model of a single cell data. Specifically, we focus on protein data to RNA gene expression prediction. This was one of the tasks in Neurips 2021 competition [1]. Current literature uses regression methods and simple neural networks for prediction [2,3,4,5].

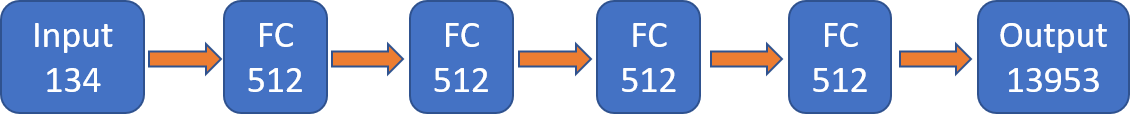
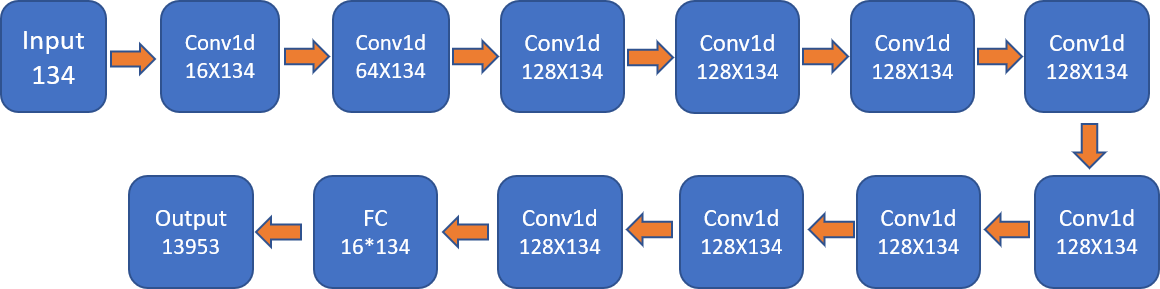
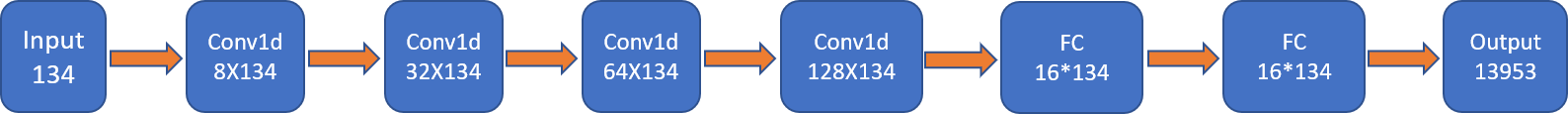
In this project we trained two deep neural networks with convolution layers and benchmarked against one of the competition winner model. We used ADT2GEX modality prediction data from the Neurips competition. All the models were trained on a Nvidia RTX 2060 GPU using Pytorch framework. We observe that the convolution neural network performs better than the benchmark model on ADT2GEX test data. Moreover, deeper network performs slightly better than shallow network. Details of the method and results are discussed below.

 [2]

**Method**

The task was to convert ADT protein data to GEX data. We used train/test data from the Neurips competition available here [1]. The train set contains ~50K samples and test set contains 1K samples. The input ADT data has 134 features while the output GEX data has 13953 features. We directly fed the input data to the network and trained it for predict the output features. We trained two deep convolutional neural network and benchmark them to one of the competition winner model ‘Novel’.

The architecture of all three models is depicted in the figure below. The architecture of Novel method contains 4 hidden fully connected layers with 512 dimensions. The architecture of ConvNet6 contains 4 Conv1d layers with kernel size 3. It is followed by 2 fully connected layers to predict the output features. We use convolution in 1-d as the input data is a 1d vector.



Novel

ConvNet6

ConvNet12

The architecture of ConvNet12 contains 10 Conv1d layers followed by 1 fully connected layer. We also use residual connections after every two layers in ConvNet12. Residual connections are helpful in training deeper networks as they avoid the vanishing gradient problem.

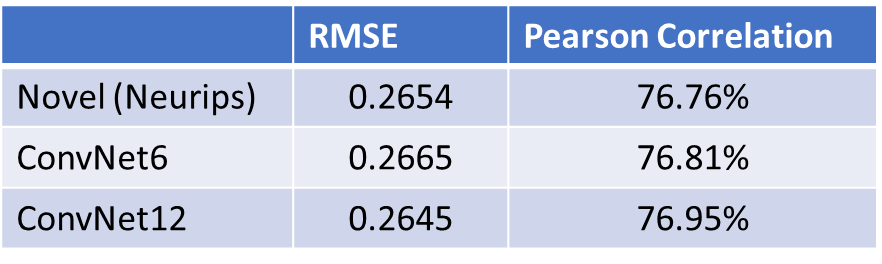
We use MSELoss as the loss for training. Training is performed using Adam optimizer for 100 epochs. We did hyperparameter tuning for learning rate and weight decay. We compared the model on test data using RMSE and Pearson Correlation metrics.

All the three networks are trained on Nvidia RTX 2060 GPU using Pytorch framework. The code for the project is available at

<https://github.com/2ashish/UCLA-Courses/tree/main/CM225/project/code>

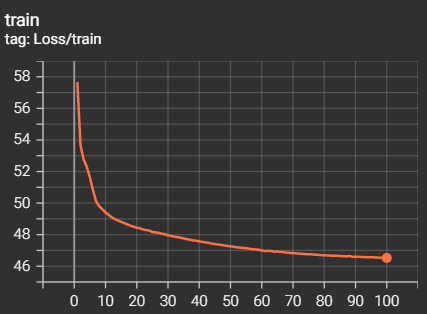
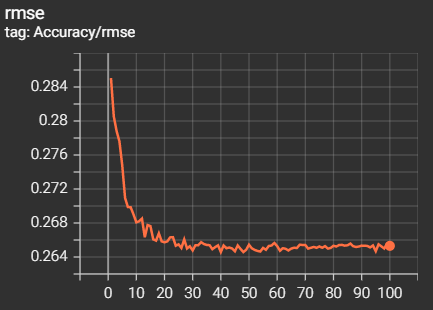
**Results**

The test metrics results are shown below.



We see that Convolutional Neural network performs slightly better that the baseline Novel method. Also, the deeper model is better that shallow network. In terms of RMSE, ConvNet12 achieves the best metric with value 0.2645. In terms of Pearson Correlation also we see that ConvNet12 achieves the best metric with 76.95% correlation.

Graph below on left shows the training loss curve for ConvNet12 training. We see that the training is properly done and stopped early. We can train a little bit more. The graph on right shows the test RMSE values which plateaus showing that the current model is the best model possible.



**Conclusion and Future Work**

We found that Convolution neural networks perform better in modality prediction task than vanilla fully connected networks. Moreover, deeper network performs better than shallow network.

Future work can include optimizing the hyperparameters for the convnets, training with deep supervision can help training faster. We can also experiment with different kernel size and architecture for better performance. Neural architecture search is also an option for search better architecture.

**References**

[1] https://openproblems.bio/neurips\_2021/

[2] Luecken et. al. “A sandbox for prediction and integration of DNA, RNA, and protein data in single cells”

[3] Luke Zappia and Fabian J Theis. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. bioRxiv, page 2021.08.13.456196, aug 2021.

[4]Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular Systems Biology, 15(6):e8746, jun 2019.

[5] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biology, 21(1):218, dec 2020