Predicting RNA 3D Folding

Elizabeth Adeoye
Modestor Tembo
Viviana Yamell Mejia

Project 4

Stanford Kaggle Competition

Competition Link

The Ask

The Stanford Kaggle competition asks participants to develop machine learning models to predict the 3D structure of RNA molecules from its sequence.

The Goal

To improve our understanding of biological processes and drive new advancements in medicine and biotechnology.

The Dataset

The dataset was taken from the Kaggle competition itself.

The input data provides RNA sequences with details like their nucleotide makeup.

Our Approach

Used supervised learning to predict RNA structures. The models were trained to improve their performance as they are run.

What is RNA?

- Single strand structures
- Involved in the process of protein synthesis and gene regulation
- Carry genetic information for some viruses
- Nitrogenous bases: Adenine (A), Guanine (G), Cytosine (C), and Uracil (U)
- Understanding RNA's structure can help in areas like drug discovery and understanding diseases.

The Code

jupyter

python™

Definitions in Our Code

- Epoch: one complete pass.
- Early stopping: performance decrease.
- R^2 value: perfect fit from 0 to 1.
- Mean Absolute Error (MAE): average difference between predicted and actual values.
- Training Loss: difference between predictions and actual model.
- Validation Loss: understanding of new data.
- Dropout rate: prevents overfitting.
- Learning rate: determines step size.
- Hidden dimension: amount of information allowed to process.

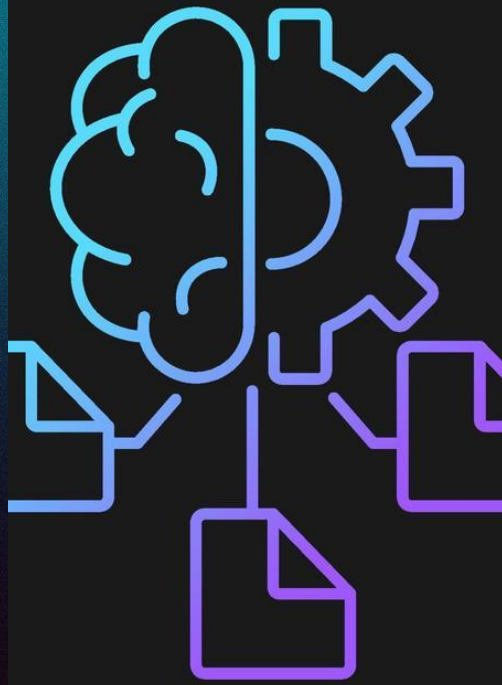
“

Frameworks to Install

- TensorFlow
- PyTorch
- SeaBorn
- Sklearn
- Torch

Conclusion

Given the complexity surrounding rna data, we predicted a 42% R^2 based on the model. This means that the average R^2 score across all folds, indicating the overall proportion of variance explained by the model. In conclusion, rna research is still vastly unknown and there is room for more research for better machine learning models and predictions from data and biology experts.



THANK YOU