Protein Secondary Structure Prediction

Project Report

The goal of this project is to develop and evaluate deep learning models for predicting protein secondary structure from amino-acid sequences. The task is formulated as a sequence-to-sequence prediction problem, where each residue in the input sequence is assigned a corresponding secondary structure label in both Q8 and Q3 annotation schemes. Bidirectional deep learning architectures were implemented and analyzed to capture long-range contextual dependencies within protein sequences and to assess their effectiveness in accurately modeling structural patterns.

● Approach 1: Bidirectional Vanilla RNN

In the first approach, a Bidirectional Vanilla Recurrent Neural Network (BiRNN) was implemented and trained from scratch for protein secondary structure prediction.

Architecture :

Input: Amino-acid sequence encoded as integer indices of length.

Embedding Layer:

- Vocabulary size: 23
- Embedding dimension: 128
- Masking**:** Padding index is masked to ignore padded residues.

Recurrent Layer:

- Model type: Vanilla RNN
- Number of layers: 2
- Hidden dimension: 256 per direction
- Bidirectional: Yes (forward + backward)
- Output feature size per residue: 512 (256 × 2)
- Processing: Batch-first enabled

Regularization:

- Dropout probability: 0.5 (applied between RNN layers and before classification).

Output Layer:

- Type: Fully connected linear layer
- Input features: 512
- Output classes: 8 (Q8 secondary structure labels)

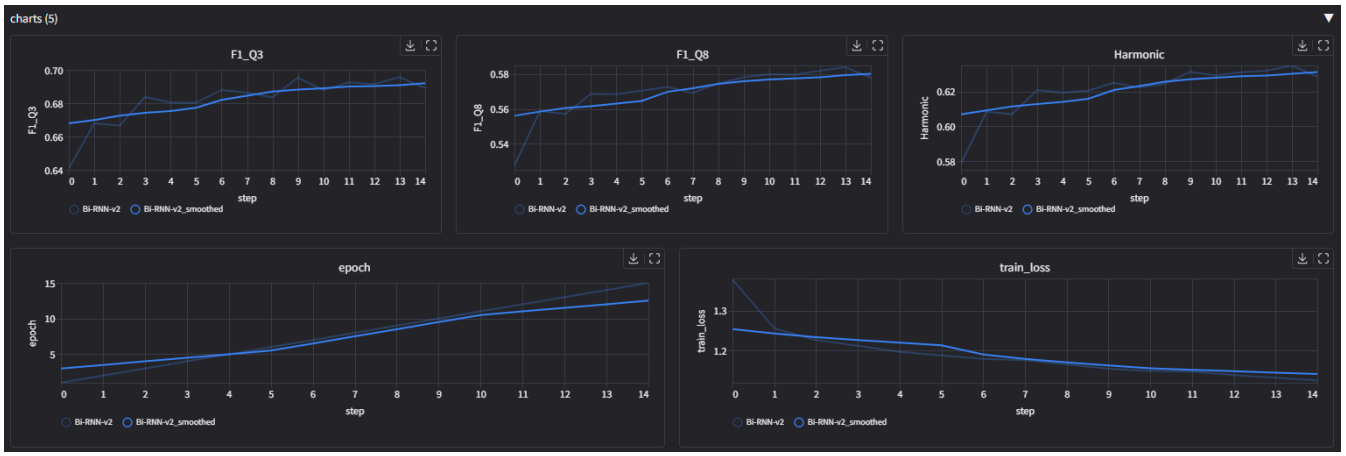
Output Shape: (Batch size, Sequence length, 8)

Loss Function:
Q8 prediction: CrossEntropyLoss applied at the residue level.

Optimizer:
Adam optimizer with a standard learning rate.

Regularization:
Dropout is used between RNN layers and before the final classification layer to mitigate overfitting.

Result:
The model achieved a public leaderboard score of **0.449**, demonstrating its ability to learn basic sequential dependencies in protein sequences, but highlighting limitations in capturing complex long-range interactions compared to more advanced architectures.



● Approach 2: Bidirectional GRU (BiGRU):

In the second approach, a Bidirectional Gated Recurrent Unit (BiGRU) network was implemented to improve sequence modeling and capture longer-range dependencies in protein sequences more effectively than a vanilla RNN.

Architecture :

The model is designed to process an amino-acid sequence to predict its Q8 secondary structure.

1. Input:

- Amino-acid sequence is encoded as integer indices.
- Sequence length is denoted by *L*.

2. Embedding Layer:

- Vocabulary Size: `len(AA_VOCAB) = 23` (Amino Acid Vocabulary size).
- Embedding Dimension: 128.
- Function: Masks the padding index to ignore padded residues.

3. Recurrent Layer:

- **Model Type:** Gated Recurrent Unit (GRU).
- Configuration: Bidirectional (forward and backward directions).
- Number of Layers: 2.
- Hidden Dimension: 256 per direction.
- Dropout (Between layers): 0.6
- Processing: Batch-first processing is enabled.
- Output Feature Size (per residue): 256 (forward) + 256 (backward) = 512

4. Regularization:

- **Dropout Probability:** 0.6, applied after the Bidirectional GRU layer.

5. Output Layer (Prediction Head):

- Type: Fully connected linear layer.
- Input Features: 512.
- Output Classes: 8 (for Q8 secondary structure labels).

6. Final Output Shape: `(Batch size, Sequence length, 8)`.

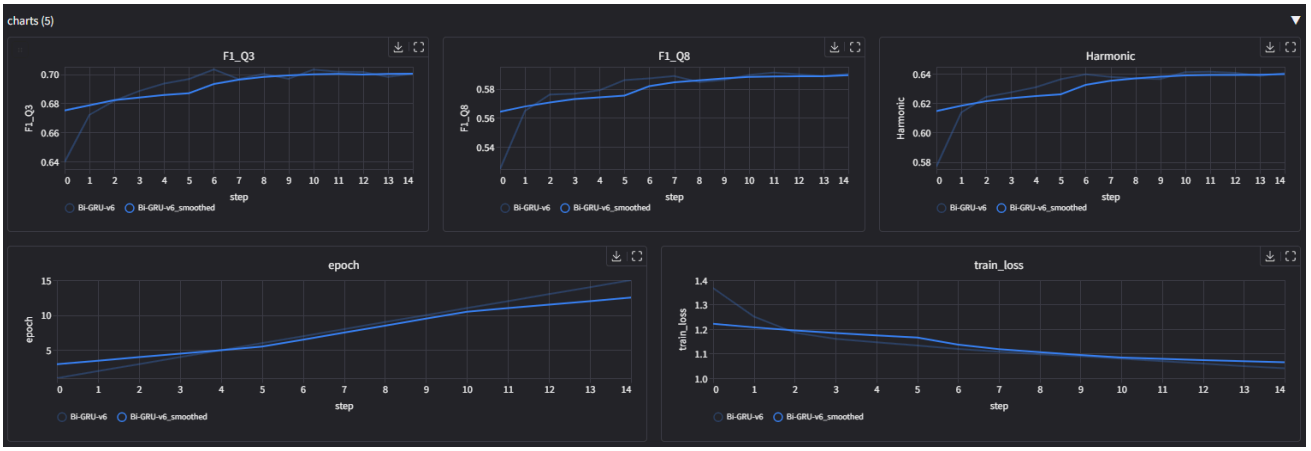
Loss Function:
Q8 prediction: CrossEntropyLoss applied at the residue level.

Optimizer:
Adam optimizer with a standard learning rate.

Regularization:
Dropout is used between RNN layers and before the final classification layer to mitigate overfitting.

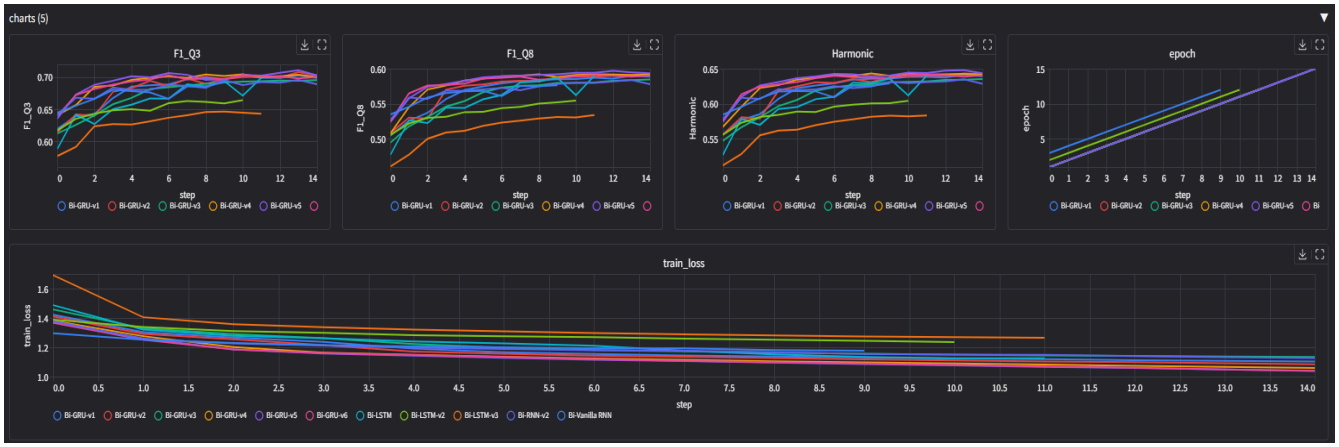
Result:

The Bidirectional GRU model achieved a public leaderboard score of **0.47**, showing a clear improvement over the vanilla BiRNN baseline. This gain highlights the effectiveness of GRU gating mechanisms in modeling long-range dependencies and reducing vanishing gradient issues in protein sequence prediction tasks.



- **Comparison and Results:**

Model	Architecture	Loss Function	Training Epochs	Public Score	Remarks
BiRNN (Vanilla)	2-layer Bidirectional RNN	CrossEntropyLoss	15	0.449	Baseline sequential model. it captures local context but struggles within long-range dependencies.
BiGRU	2-layer Bidirectional GRU	CrossEntropyLoss	15	0.470	Improved performance due to gating mechanism enabling better modeling of long-range interactions



- **Key Learnings and Takeaways:**

  This project provided valuable hands-on experience in designing and implementing sequence-to-sequence deep learning models for protein secondary structure prediction. Through practical experimentation with bidirectional RNN and GRU architectures, I developed a deeper understanding of how sequential models capture contextual dependencies across long protein sequences. A key takeaway from this work was the critical role of regularization when working with limited datasets. By systematically tuning the dropout rate, I observed that a higher dropout value of **0.6** consistently produced better generalization performance, effectively reducing overfitting. This reinforced the insight that model capacity, dataset size, and regularization strength must be carefully balanced, especially in biological sequence modeling tasks where labeled data is often scarce.

- **Links :**
  - **Huggingface-space:** click here
  - **Huggingface-datasets :** click here