Matthew Uryga, Steven Wang
CSCI-4969
Spring 2022

04/19/2022

Capstone 1 Writeup

1 Distribution of Work

1.1 Yu-Kai (Steven) Wang

Steven implemented the multihead self attention, outer product mean, and triangular self attention. He also worked on parallelization and tinkering with the model to allow it to work on the DCS cluster with multiple GPUs.

1.2 Matthew Uryga

Matthew implemented the triangular multiplication, as well as constructing the overall structure of the evoformer trunk from the modules that were constructed above. He also implemented the dataset for training/testing, as well as the training loop and evaluation.

1.3 Repository Link

The code for our implementation of the evoformer trunk can be found here:
https://github.com/mnuryga/MLBinfCapstone.

2 Results

2.1 Alphafold1 Results

Accuracies for top $\frac{L}{k}$ predictions:

k	Short	Medium	Long
1	0.230443	0.221818	0.230140
2	0.347676	0.307315	0.296237
5	0.511692	0.431034	0.376984
10	0.615373	0.506860	0.438546
20	0.693295	0.586715	0.480660
50	0.762319	0.630435	0.526087
100	0.818116	0.660507	0.556522

Accuracies for top $\frac{L}{k}$ predictions with thresholding (>0.5):

k	Short	Medium	Long
1	0.736420	0.759364	0.784539
2	0.736420	0.759364	0.784539
5	0.736397	0.760673	0.785875
10	0.742424	0.762611	0.794528
20	0.763906	0.774161	0.829209
50	0.800072	0.797717	0.847826
100	0.835290	0.825797	0.849275

2.2 Alphafold2 Results

Accuracies for top $\frac{L}{k}$ predictions:

k	Short	Medium	Long
1	0.230443	0.221818	0.230140
2	0.347676	0.307315	0.296237
5	0.511692	0.431034	0.376984
10	0.615373	0.506860	0.438546
20	0.693295	0.586715	0.480660
50	0.762319	0.630435	0.526087
100	0.818116	0.660507	0.556522

Accuracies for top $\frac{L}{k}$ predictions with thresholding (>0.5):

k	Short	Medium	Long
1	0.736420	0.759364	0.784539
2	0.736420	0.759364	0.784539
5	0.736397	0.760673	0.785875
10	0.742424	0.762611	0.794528
20	0.763906	0.774161	0.829209
50	0.800072	0.797717	0.847826
100	0.835290	0.825797	0.849275

3 Conclusion

3.1 Key Findings

As shown above, the contact prediction accuracy is relatively good when thresholding, but the accuracy falls off significantly when the predictions are not thresholded and when greater than $\frac{L}{10}$ of the predictions are considered. From this, it can be concluded that the model is not predicting enough contacts with high enough confidence to effectively estimate the protein structure. However, it is worth noting that while training, the prediction accuracy has only increased. Given more time, and perhaps more dilation blocks, the models may be able to achieve much better accuracy.

3.2 Comparison

Based on the test accuracies, it is evident that Alphafold1 performs better than Alphafold2, which contradicts the expected output. This may be due to several factors:

(1) The Alphafold2 evoformer trunk was trained for significantly less time than the Alphafold1 model.

(2) The evoformer‘s parameters may not be optimal.

(3) The reduced dimensionality model parameters (including N_{res}) have a significant impact on the potential performance of the model.

(4) There is something fundamentally wrong with the implementation of the evoformer trunk.

4 Script Output

4.1 Alphafold1

Note that not all of the training output was recorded, as it was done over several days.

```
Epoch 22, 70,880 crops:
        Train loss per crop = 0.049809
        Valid loss per crop = 0.052133
Epoch 23, 73,640 crops:
        Train loss per crop = 0.047115
        Valid loss per crop = 0.050202
Epoch 24, 69,980 crops:
        Train loss per crop = 0.046847
        Valid loss per crop = 0.050519

Test loss per crop: 0.022741

---Accuracies for L/k sequences--
        short       med      long
1    0.230443  0.221818  0.230140
2    0.347676  0.307315  0.296237
5    0.511692  0.431034  0.376984
10   0.615373  0.506860  0.438546
20   0.693295  0.586715  0.480660
50   0.762319  0.630435  0.526087
100  0.818116  0.660507  0.556522
```

4.2 Alphafold2

```
Epoch 22, 70,880 crops:
        Train loss per crop = 0.049809
        Valid loss per crop = 0.052133
Epoch 23, 73,640 crops:
        Train loss per crop = 0.047115
        Valid loss per crop = 0.050202
Epoch 24, 69,980 crops:
        Train loss per crop = 0.046847
        Valid loss per crop = 0.050519

Test loss per crop: 0.022741

---Accuracies for L/k sequences--
        short       med      long
1    0.230443  0.221818  0.230140
2    0.347676  0.307315  0.296237
5    0.511692  0.431034  0.376984
10   0.615373  0.506860  0.438546
20   0.693295  0.586715  0.480660
50   0.762319  0.630435  0.526087
100  0.818116  0.660507  0.556522
```