Soft vs Hard Multitask Models for Chromatin Accessibility

Ana Istrate, Sebastian Le Bras, Michael Painter
Stanford University

Task

The goal of this project was to systematically compare soft and hard multi-task models for chromatin accessibility. Variations of the state of the art BASSET architecture are used to explore the validity and effectiveness of multi-task models.

Biological Background/Motivation

Chromatin Accessibility refers to the availability of a certain chromatin sequence and can be best understood, in this context, by its implications. Chromatin Accessibility coincides with chromatin activity, which gives us a glimpse into a cells regulation and provides us with useful information about cell differentiation, environmental signaling and disease development[1].

Dataset

We use the DNaseI-seq peak data provided by the ENCODE and Epigenetics Roadmap projects[2]. The datasets provide accessible regions for **164 different cell types**. For each cell type, peak positions were identified using the **HotSpot algorithm**, as described by John et al. Peak. Locations are combined across reference genomes by stacking peak positions and then considering full width at half maximum. Simply put, the data identifies regions of accessible chromatin --- per cell type --- that were found to be accessible in more than half of the reference genomes considered.

There are a total of **~1,200,000** entries in the dataset, each consisting of a **600bp** long one-hot encoded DNA sequence. The corresponding label to the dataset is a length 164 binary vector denoting whether a certain cell type passes the threshold for chromatin accessibility. The dataset is quite imbalanced consisting of significantly more class 0: inaccessible chromatin than class1: accessible chromatin.

Preprocessing: Peaks within 200bp of one another are greedily merged. The 164 binary vector of labels for the merged peak is taken to be the union of of the two vectors being merged. Moreover, the new peak position is taken to be the weighted average of the old peak positions, weighted by the number of cell types the peak was found in.

Metrics & Oracle

Accuracy: A measure of correct prediction over the total predictions. Highly sensitive to class imbalance.

AUC(area under the ROC curve): the area under the ROC curve measure the True positive Rate against the False Positive rate. It is a good measure against Class Imbalance

Basset: AUC: 0.892

Methods

Basset Architecture: All models were based on the architecture of the Basset model [3], described in Table 1. The model defines each task as a binary prediction of a certain cell types chromatin accessibility given a 600bp sequence of data. The loss functions used in the networks are described below where $\hat{\boldsymbol{y}}$ refers to the prediction, $\mathbf{y}$ refers to true value and $\mathcal{B}$ refers to a minibatch of the entire dataset $\mathscr{D}$.

$$\mathrm{CE}_i(\hat{y}, y) = y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

$$\mathcal{L}_i^u(\mathcal{B}) = \sum_{(x,y)\in\mathcal{B}} \mathrm{CE}_i(f_i(x), y)$$

Since our dataset is imbalanced we make sure to weigh the underrepresented class more heavily.

$$\rho_i = \frac{1}{|\mathcal{D}|}\sum_{(x,y)\in\mathcal{D}} \mathbb{I}[y_i = 1]$$

$$r_i(y) = \begin{cases} 1 - \rho_i & \text{if } y_i = 1 \\ \rho_i & \text{if } y_i = 0 \end{cases}$$

$$\mathcal{L}_i(\mathcal{B}) = \sum_{(x,y)\in\mathcal{B}} r_i(y)\,\mathrm{CE}_i(f_i(x), y)$$

Soft multi-task model: Soft multi-task learning refers to a constrained model that still maintains separate parameters for each task. Each task within the network, in this case the binary prediction for each cell type, has its own full network. We achieve the soft sharing through the use of **L2 normalization** across the weights of layers 1-9. The loss function is described in the equations below:

$$d(W_i, W_k) = \ell_p(W_i, W_k) = \left(\sum_{j\in J} |(W_i)_j - (W_k)_j|^p\right)^{1/p}$$

$$\mathcal{L}(\mathcal{B}) = \sum_{i=1}^{164} \mathcal{L}_i(\mathcal{B}) + \lambda \sum_{0<i<k\leq 164} d(W_i, W_k)$$

A visualization of the soft multi-task model can be seen in Figure 1A.

Layer	Layer Description	Output Shape
0	In:	(600,4)
1	Conv1D, [19, 300, 1]:	(582,300)
2	BN:	(582,300)
3	MP, [4]:	(146,300)
4	Conv1D, [11,200,1]:	(136,200)
5	BN:	(136,200)
6	MP, [4]	(34,200)
7	Conv1D, [7,200,1]:	(28,200)
8	BN:	(28,200)
9	MP, [4]:	(7,200)
10	FC:	(1000)
11	FC:	(1000)
12	Softmax:	(2)

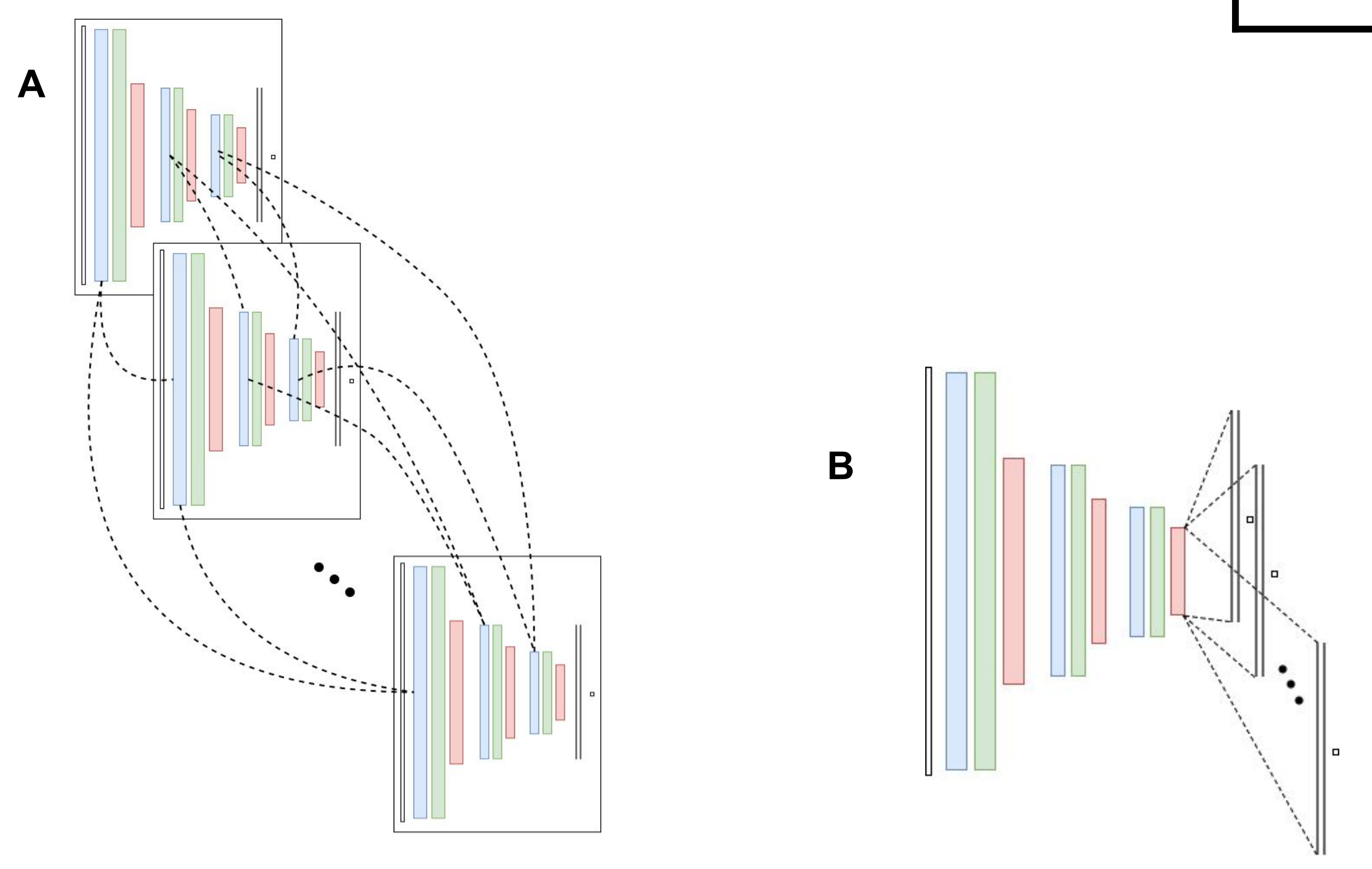

Figure 1: Blue layers -> convolutional layer, green layer -> batch normalization, red layer -> max pooling **A**: a visualization of the soft multi-task model. The dotted lines refer to the L2 normalization soft sharing. **B**: a visualization of the hard multi-task model

Table 1: A table description of the Basset Architecture. In the "Layer Description" column, the layers labeled Conv1D have format: Conv1D, [filter size,number of filters, stride], while the layers labeled MP (max pooling) have format: MP, [pool size].

Methods (continued)

Hard multi-task model: Hard multi-task learning refers to parameter sharing across different tasks in the beginning of the network. For our network, layers 1-9, seen in table 1, are shared. Each task has its own fully connected layers and loss function which are then combined into the final loss function as follows:

$$\mathcal{L}(\mathcal{B}) = \sum_{i=1}^{164} \mathcal{L}_i(\mathcal{B})$$

The visualization of the hard multi-task model can be seen in Figure 1B

Results & Analysis

A summary of our results can be seen in Table 2.

- So far we have unfortunately been unable to achieve similar results to what was achieved in BASSET.
- Interestingly, our models tend to hit close to 1.0 recall, suggesting that our models are overfitting to the positive examples.
- The soft model shown was run with very low weight decays which would have affected performance by significantly reducing regularization in the network.

Table 2	Accuracy	AUC
BASSET	-	0.892
Soft multi-task	79%	0.604*
Hard multi-task	92%	0.253*

Next Steps

- Variations of hard & soft models:
  - Hard models that share fewer/more layers
  - soft/hard model hybrids
  - Different Initializations
  - Transfer Learning: hard -> soft
- Debug poor performance of models compared to BASSET.

References

1. Tsompana, Maria, and Michael J. Buck. "Chromatin accessibility: a window into the genome." *Epigenetics & chromatin* 7.1 (2014): 33.
2. ENCODE Project Consortium. "The ENCODE (ENCyclopedia of DNA elements) project." Science 306.5696 (2004): 636-640.
3. Kelley, David R., Jasper Snoek, and John L. Rinn. "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks." Genome research 26.7 (2016): 990-999.