Learning the Language of the Genome Using RNNs

Jesse Zhang and Govinda Kamath

Background

Epigenetics is the study of how the genome is regulated by external mechanisms. Biological experiments have shown that subsequences of the human genome are regulated by specific proteins. The purpose of this project is to explore how an RNN architecture can be used to learn sequential patterns in genomic sequences. A robust method for modeling the genome can offer insights on genetic patterns related to health and disease.

Previous deep learning approaches[1,3,6,7]

- Process the input sequence using a convolutional layer
- Incorporate extra information about each input sequence, and the information is obtained from tedious experiments.

This project explores how RNNs can predict information about a sequence without convolutional layers and side information.

Problem Statement

We explore a sequence classification problem. For each of many features, we assess an RNN's ability to predict whether a given genomic feature will be present based solely on the sequence. Different RNN architectures will be compared based on perplexity. The chosen RNN architecture will be evaluated against a logistic regression baseline using the F1 score.

Dataset

The dataset was collected from experiments reported in the **ENCODE**[2] and **Roadmap Epigenomics**[5] data releases.

- 80000 sequence-label pairs for the training dataset,
- 8000 pairs for the testing dataset
- 2000 pairs for the validation dataset.

Each sequence has a length of exactly 100 and is a subsequence of the length-$3*10^9$ human genome. The set of labels describes 919 binary prediction tasks, and each indicates whether part of a particular sequence was observed as accessible for a particular experiment. For evaluation, we look only at tasks where at least 1% of the training examples were positive (498 of the 919 tasks).

Evaluation Metric

When designing the RNN, we minimize average **perplexity** across all tasks and examples:

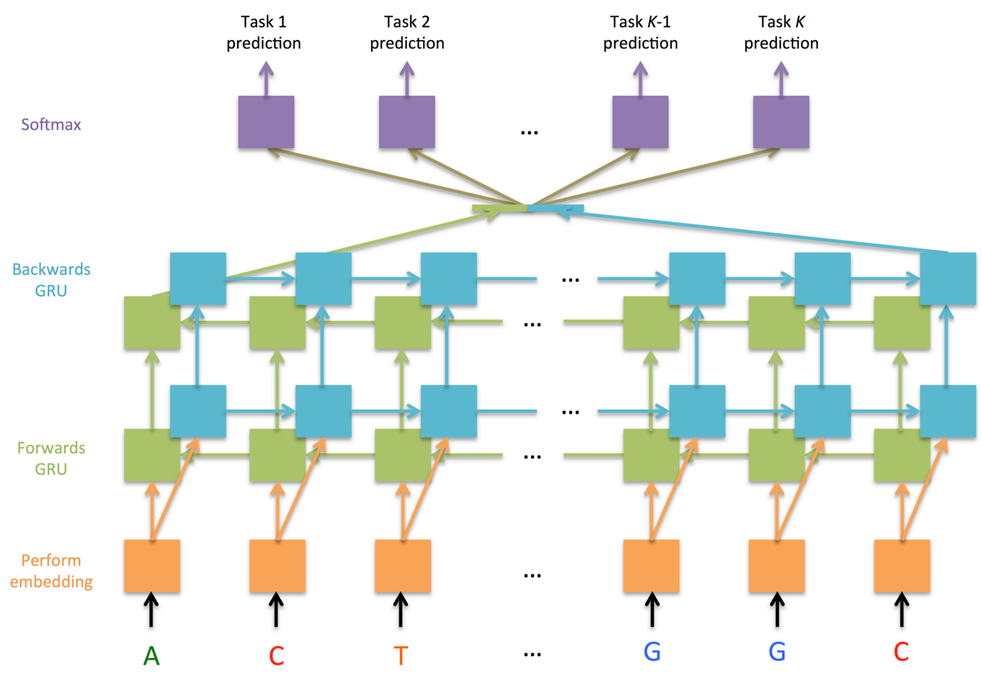
$$PP(\mathbf{y},\hat{\mathbf{y}}) = \exp\left\{-\frac{1}{n-1}\sum_{t=1}^{n-1}\sum_{i=1}^{|V|} y_i^{(t)} \log \hat{y}_i^{(t)}\right\}$$

When evaluating the final RNN, we compare the F1 score achieved by the model to the F1 score achieved by the **baseline:** a logistic regression (LR) model for each of the 498 prediction tasks. For the LR model, an input sequence is mapped to a length-1364 feature vector where each feature was the count of a particular *k*-mer (for k = 1, …, 5).

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Precision: proportion of correct positive predictions
Recall: proportion of positive examples that were predicted to be positive

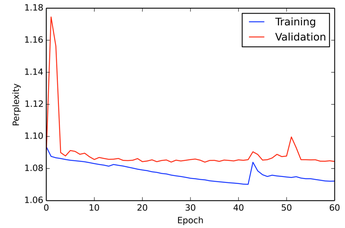
Model Architecture



Bidirectionality captures how the genome can potentially be read in either direction. Gated recurrent units (GRUs) memorize long-range information. Hidden representations are shared across tasks, but each task uses its own set of softmax weights. Character-level prediction alleviates some computational workload.

Results

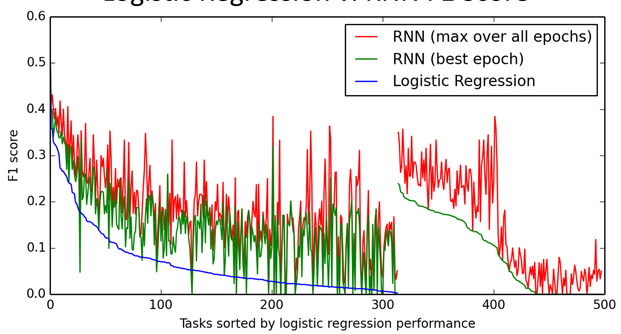
Training & Validation Loss



For the 498 tasks with at least 800 training examples:

- RNN outperforms LR on 408 tasks (81.9%)
- RNN F1 scores are 0.091 points higher on average

Logistic Regression v. RNN F1 Score



Conclusion

We empirically showed that an RNN can predict biological features about a genetic sequence without outside information. The RNN model outperforms a LR model on a significant majority of 498 binary labeling tasks. This confirms how RNNs can leverage 1) correlation between tasks and 2) large amounts of data to make more robust predictions. Unlike LR, RNNs require significantly more training time.

Extensions: Train on more data (lots available), test an architecture better suited for very long-range information such as a clockwork RNN[4], use genomic word-level RNN

References

1. B. Alipanahi *et al. Nature biotechnology,* 2015.
2. ENCODE Project Consortium *et al. Nature*, 489(7414):57–74, 2012.
3. D. R. Kelley *et al. bioRxiv*, page 028399, 2015.
4. J. Koutnik *et al. arXiv preprint arXiv:1402.3511*, 2014.
5. A. Kundaje *et al. Nature*, 518(7539):317–330, 2015.
6. D. Quang *et al. bioRxiv*, page 032821, 2015.
7. K. Zhou *et al. Nature methods*, 12(10):931–934, 2015.