Regulatory Activity Prediction with Attention-based Models

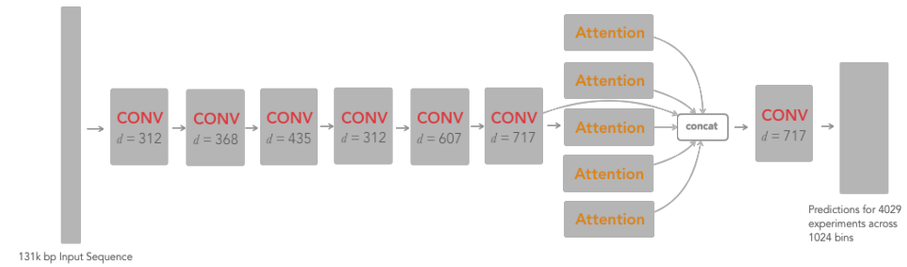
Sabri Eyuboglu and Geoffrey Angus

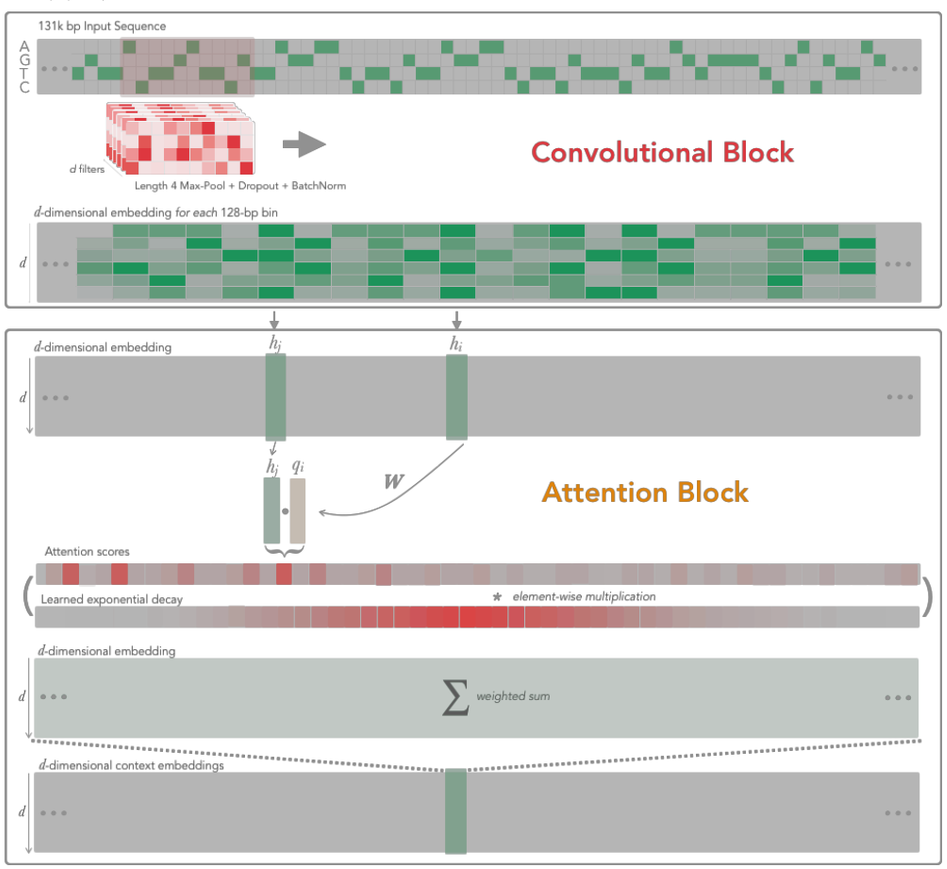
CS273B
Deep Learning in Genomics
Stanford University

Model Architecture

Computational models that predict gene expression from genome sequence rely on variations in non-coding regions (e.g. promoters and distal regulatory elements) to inform their predictions. Recent works, including the FANTOM and ENCODE projects, have produced large datasets of epigenetic and transcriptional experiments on variety of cell-types. These datasets have fueled the development of deep learning models for genome expression prediction.

We present an attention-based neural network architecture designed to capture the influence of distal regulatory elements on gene expression.



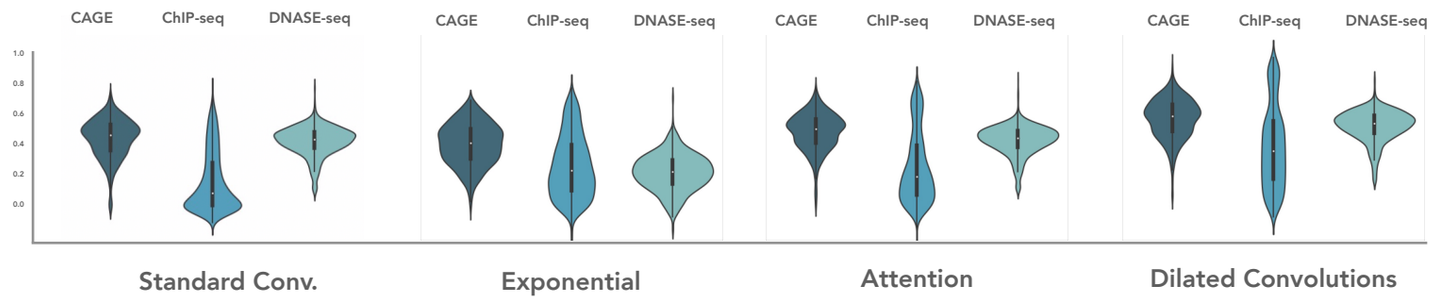


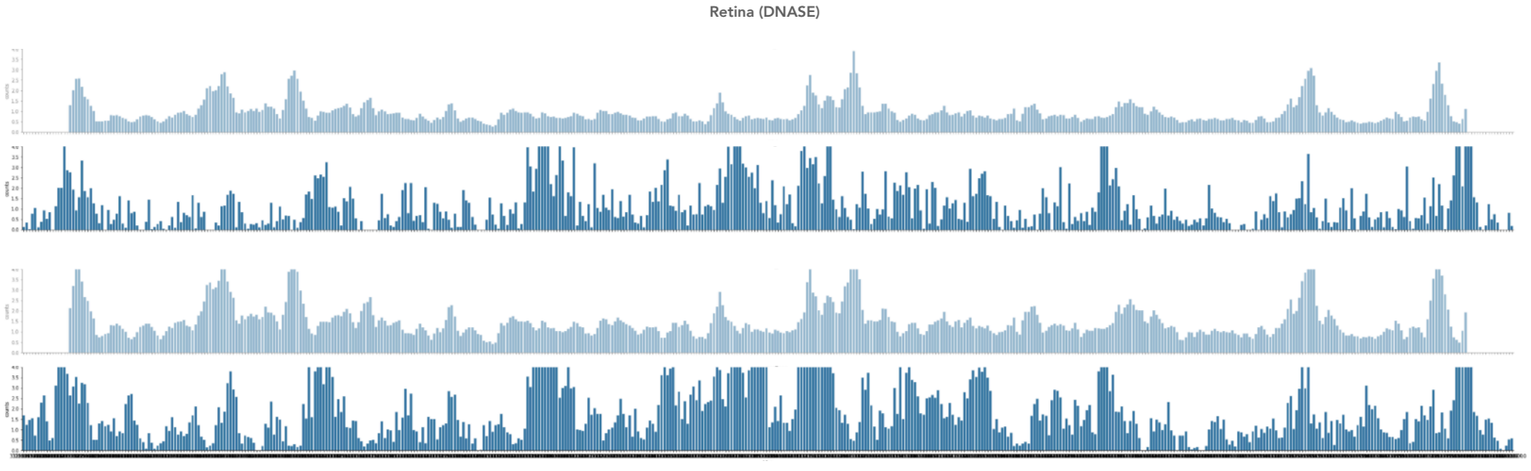
Results

We use the same datasets and preprocessing techniques that Kelley et al. presented. The dataset includes a total of 973 CAGE experiments, 949 DNase-seq experiments and 2,307 histone modification ChIP-seq experiments spanning a diverse set of cell-types (CAGE experiments measure gene expression levels in a given cell-type or tissue, ChIP-seq assay identify protein binding sites on DNA sequences and DNase-seq experiments locate regulatory regions in a sequence). These experiments were carried out by the FANTOM5, ENCODE and Roadmap Epigenomics projects. The output of these experiments is a set of reads (short DNA sequences) mapped to experimental counts.

Model	r^2	r
Standard Conv.	0.277	0.506
Exponential	0.281	0.538
Multi-Head Attention	0.334	0.562
Dilated Conv.	0.405	0.628

Here we compare the Pearson Correlation scores of four different models– the manuscript and baseline models, followed by two of our own models– partitioned across the three different datasets. The weights in each experiment are initialized as a pre-trained subset of the weights found in the manuscript model. We see here that performance is most similar on the CHIP-seq data, which shows very low score density across all four figures.





Here we compare the output of the proposed Dilation w/ Attention model to the target values on a DNASE experiment run on different donors on a region of the CHR2 chromosome. We see that many of the peaks in the predicted sequence (dark blue) align with those of peaks in the experimental reads (light blue).

Future Work

The work we have out forth in this project is among some of the first attempts to apply attention to the specific task of predicting regulatory activity across chromosomes. At time of writing, this approach seems to be at least on par with models utilizing exponential decay, but underperforms dilated convolutions.