DREAM Challenge 2022

Predicting gene expression using millions of random promotor sequences

By
MadLab

Abstract

In this report we present a transformer based model to predict gene expression from promotor sequences. The model is composed of 3 building blocks, namely a convolutional network, a transformer and a recurrent network. The convolutional net is implemented with multiple receptive fields, to capture features at different base pair scales. The transformer is implemented following the enformer (Avsec, Ž. et al. 2021) architecture.

1. Description of data usage

The training data was randomly divided into training (n = 4.703.150) and validation (n = 95.982) using a 98-2 percent split. Before splitting the data, ~ 50% of the training sequences that were only observed once (assumed by y value being discrete) were filtered out to reduce the number of low confidence observations, yielding 4.799.133 sequences.
Before training the data was encoded using one-hot (with N's encoded as [0,0,0,0]).

2. Description of model

The model takes as input one-hot encoded sequences and performs 4 1D convolutions separately, with increasing kernel sizes (15 - 30). The idea behind this operation is to learn sequence features acting on multiple scales.

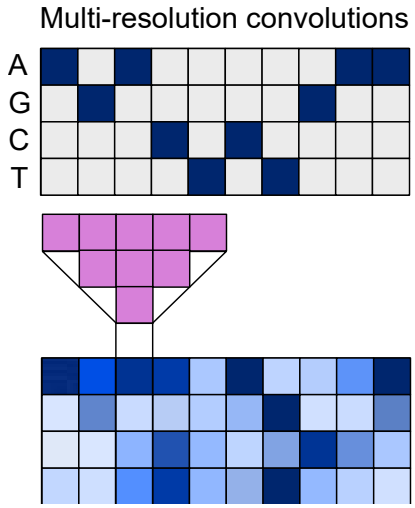

Figure 1: Illustration of multi-scale convolutions

The output of the convolutional operations is added together and fed into a enformer-style (Avsec, Ž. et al. 2021) transformer with relative positional embeddings. The output of the transformer is fed into a single bidirectional LSTM layer followed by a output layer with linear activation.

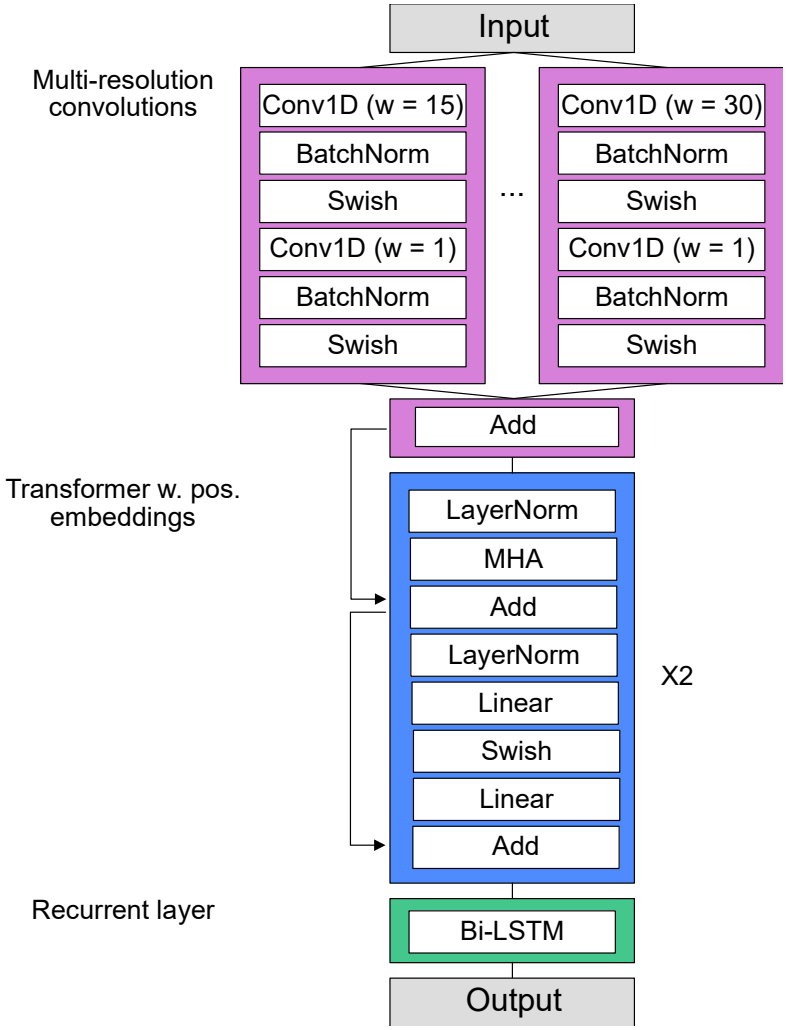

Figure 2: Illustration of network architecture

The network was trained with parameters found in table 1.

Table 1: Table describing network parameters

Parameter	**Value**
Epochs	10
Batch size	1024
Activation function	Swish
CNN	
Convolution sizes	15, 20, 25 and 30
Convolution kernels	192 and 96
Kernel initializer	He Normal
Transformer	
Value size	96
Key sise	96
N heads	1
RNN	
Bidirectional LSTM units	4

3. Training procedure

The model was trained with the parameters listed in table 2, to minimize mean squared error on the training set. The model with the lowest Pearson correlation on the validation set, after all epochs was saved for evaluation on the test set. Training scores for that model, can be found on table 3.

Table 2: Table describing training parameters

Parameter	Value
Loss function	Mean squared error
Optimizer	RMSProp
Learning rate	0.0001
L2 regularization weight	0.001

Table 3: Model scores

Dataset	MSE	R^2	Pearson correlation
Training	0.4404	0.5710	0.5807
Validaiton	0.4494	0.5608	0.5719

4. Other important features

5. Contributions and acknowledgement

a. Contributions

Name	Affiliation	Email
Andreas Møller	University of Southern Denmark	andreasfm@bmb.sdu.dk
Gabija Kavaliauskaite	University of Southern Denmark	gkav@sdu.dk
Jesper Madsen	University of Southern Denmark	jgsm@imada.sdu.dk

b. Acknowledgements

6. References

Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 18, 1196–1203 (2021). https://doi.org/10.1038/s41592-021-01252-x