Dilated Convolutions for Modeling Long-Distance Genomic Dependencies

Ankit Gupta, Alexander M. Rush
Harvard University

Summary

- Use dilated convolutional neural network to model long-term dependencies in DNA
- Appoximately match LSTM performance on small-context baseline for predicting regulatory markers
- Using new long-term dependency dataset, achieve best performance using dilated convolutions for predicting regulatory markers

Genetic Regulation Overview

Key Attributes:

- DNA has a complex three-dimensional conformation that is not captured by its 1D sequence
- Distal elements in 1D space can be adjacent in 3D space, and thus able to interact
- Capturing long-term dependencies (in 1D space) may allow network to learn motifs from spatially close regions

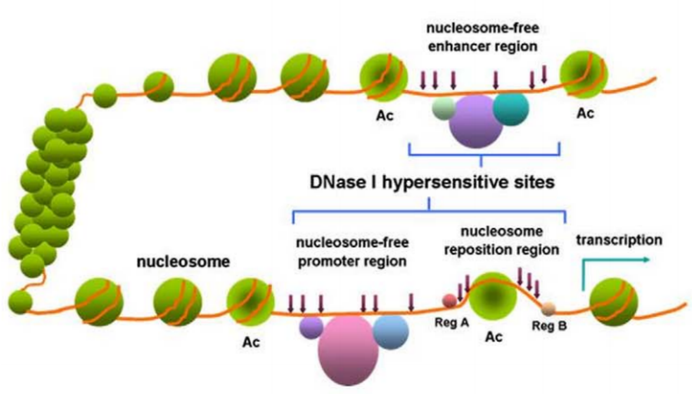


Figure: From Wang et al. (2012). A gene regulatory network. A transcription factor that binds at the location "nucleosome-free enhancer region" is spatially close to the transcription start site.

High-level overview of task: Given DNA region, predict whether each regulatory marker is present in region.

- Transcription Factor Binding Sites (TFBSs): TFs are proteins that bind to DNA, and either promote or repress gene transcription.
- Histone Modifications: Histones are proteins that DNA is wound around. Chemical modifications to histones can change how tightly wound DNA is, thus making regions more or less accessible.
- DNAse hypersensitivity sites: These regions correspond with more accessible regions of the genome, where we expect regulatory activity to occur.

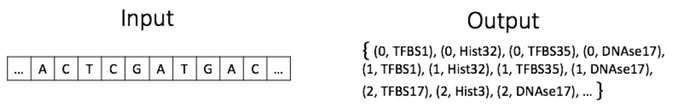


Figure: An overview of the inputs and outputs we use for Task 2. The input is a DNA sequence of length 25000. The output is a set of tuples, representing where in the input each regulatory factor is present.

References

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Quang, D. and Xie, X. (2016). Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, page gkw226.

Wang, Y.-M., Zhou, P., Wang, L.-Y., Li, Z.-H., Zhang, Y.-N., and Zhang, Y.-X. (2012). Correlation between dnase i hypersensitive site distribution and gene expression in hela s3 cells. *PloS one*, 7(8):e42414.

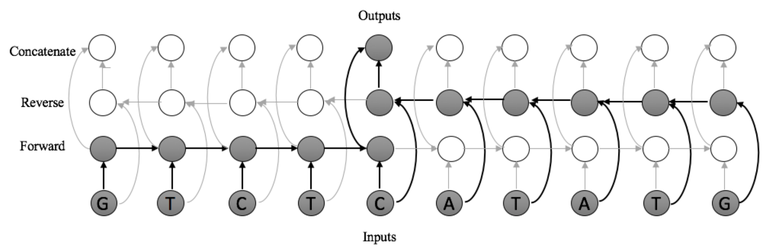
Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zhou, J. and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10):931–934.

harvardnlp

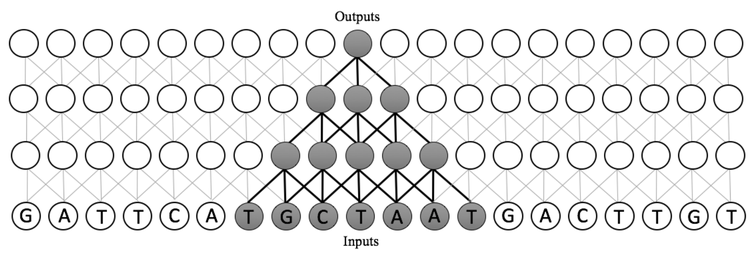
Model Comparison

Bidirectional LSTM



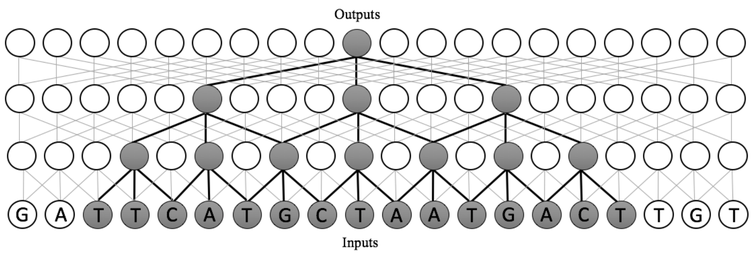
- Receptive field (in bold) of each output contains every input
- Backpropagation distance is proportional to sequence length
- Overview: Large receptive field, but long backprop distance

Standard Convolution



- Small receptive field (in bold) for every output: O(nlayers)
- Backpropagation distance is short
- Overview: Short backprop distance, but very small receptive field

Dilated Convolution



- Introduced for image segmentation by Yu and Koltun (2015)
- Large receptive field (in bold) for every output: O(2^{nlayers})
- Overview: Short backprop distance **and** very large receptive field

Main Takeaway: Dilated convolutions allow for large receptive fields like LSTMs, and short backpropagations, like convolutions. This makes them promising for modeling problems with very long-term dependencies.

Task 1: Short Inputs, Existing Data

Use the dataset from Zhou and Troyanskaya (2015). Given a short sequence of d DNA nucleotides, predict whether each of m regulatory factors is present anywhere in that sequence.

- $\mathcal{V} = \{A, C, T, G\}$, $d = 1000$, $m = 919$
- $\boldsymbol{x} = \mathcal{V}^d$, $\boldsymbol{y} = \{0, 1\}^m$
- Task: maximize $p(\boldsymbol{y}|\boldsymbol{x})$.
- Goal: Match SOTA LSTM performance using Dilated Convolutions

Model	Hidden	Type	Parameters
Baseline: LR	0	-	3,676,919
Baseline: MLP	1	Fully	4,551,919
Baseline: CNN3	3	Conv	155,159,839
Baseline: LSTM	2	LSTM	46,926,479
Dilated3	3	Dilated	37,056,519
Dilated6	6	Dilated	25,758,079

CNN3 is the baseline from Zhou and Troyanskaya (2015), and LSTM is the baseline from Quang and Xie (2016). Parameter counts are from the best-case hyperparameter configuration.

Results

Model	PR AUC		
	TFBS	Hist	DNAse
Baseline: LR	0.042	0.143	0.097
Baseline: FF	0.046	0.181	0.106
Baseline: CNN3	0.205	0.273	0.319
Baseline: LSTM	**0.305**	**0.340**	**0.407**
Dilated3	0.190	0.271	0.299
Dilated6	0.285	0.320	0.396

- Dilated convolutions allow for significant improvements over simple convolutional models
- Dilated6 performs better than standard convolutions on all three metrics, and only slightly underperforms the LSTM-based model

Task 2: Long Inputs, New Dataset

Construct new dataset with inputs with larger contexts.
Dataset properties:

- Longer input sequences: $d = 25000$
- Total of 93880 non-overlapping sequences
- Comprises 2.3 billion nucleotides
- Excludes sequences with large percentage of unknown nucleotides or multimapped regions
- Constructed from ENCODE genome regulatory data (Consortium et al., 2012)

With large context, each output is likely to be present in very large number of inputs. Thus, predict whether each output is at *each location* in the input.

- $d = 25000$, $m = 919$
- $\boldsymbol{x} = \mathcal{V}^d$, $\boldsymbol{y} = \{0, 1\}^{d \times m}$
- Task: maximize $p(\boldsymbol{y}|\boldsymbol{x})$.
- Goal: Demonstrate that with longer inputs, dilated convolutions are better able to predict the locations of regulatory markers than LSTMs.

Loss: Multilabel Binary Cross Entropy Loss: if x_i is the prediction for the ith label, and z_i is the true value:

$$\frac{1}{m} \Sigma \left(-z_i \log(x_i) - (1 - z_i) \log(1 - x_i) \right)$$

Model Descriptions and Results

Model	Layers	Conv Type	Parameters
CNN1	1	Conv	137,187
CNN3	3	Conv	341,803
CNN7	7	Conv	656,363
Dilated	6	Dilated Conv	635,739
Bi-LSTM	4	Conv, LSTM	764,395
ID-CNN	15	Iterated Dilated	631,263

Model	Validation PR AUC			Test PR AUC		
	TFBS	Hist	DNAse	TFBS	Hist	DNAse
CNN3	0.013	0.053	0.035	-	-	-
CNN3	0.059	0.115	0.100	-	-	-
CNN7	0.167	0.166	0.180	0.167	0.165	**0.186**
Dilated	0.274	0.279	0.178	**0.274**	**0.273**	0.179
Bi-LSTM	0.104	0.288	0.116	0.107	0.264	0.113
ID-CNN	0.166	0.247	0.147	-	-	-

- Substantially higher performance using dilated convolutions on predicting transcription factor binding sites and histone modifications
- No improvement using dilated convolutions on predicting DNAse hypersensitivity sites

Conclusions

- With small input context (Task 1), dilated convolutions do better than standard convolutions, but not LSTMs.
- With larger input contexts (Task 2), dilated convolutions do much better than standard convolutions **and** LSTMs
- LSTMs appear less capable of scaling to long backpropagations.
- Suggests that dilated convolutions may be an important model for studying complex genetic phenomena

{ankitgupta@college,srush@seas}.harvard.edu