Predicting TSS using Deep Learning

Evgeny Kurbatsky

April 6, 2016

Data

Were to find TSS data

	mm9	hg18
Coding genes	34988	53901
All genes	88435	178586
CAGE peaks near coding genes	74392	87939
All CAGE peaks	158429	201177

Not TSS is random part of sequence at least 2000 far from all CAGE peaks.

Clustering

1. TSS can overlap
2. To handle this we merge nearby TSS in clusters

Data division

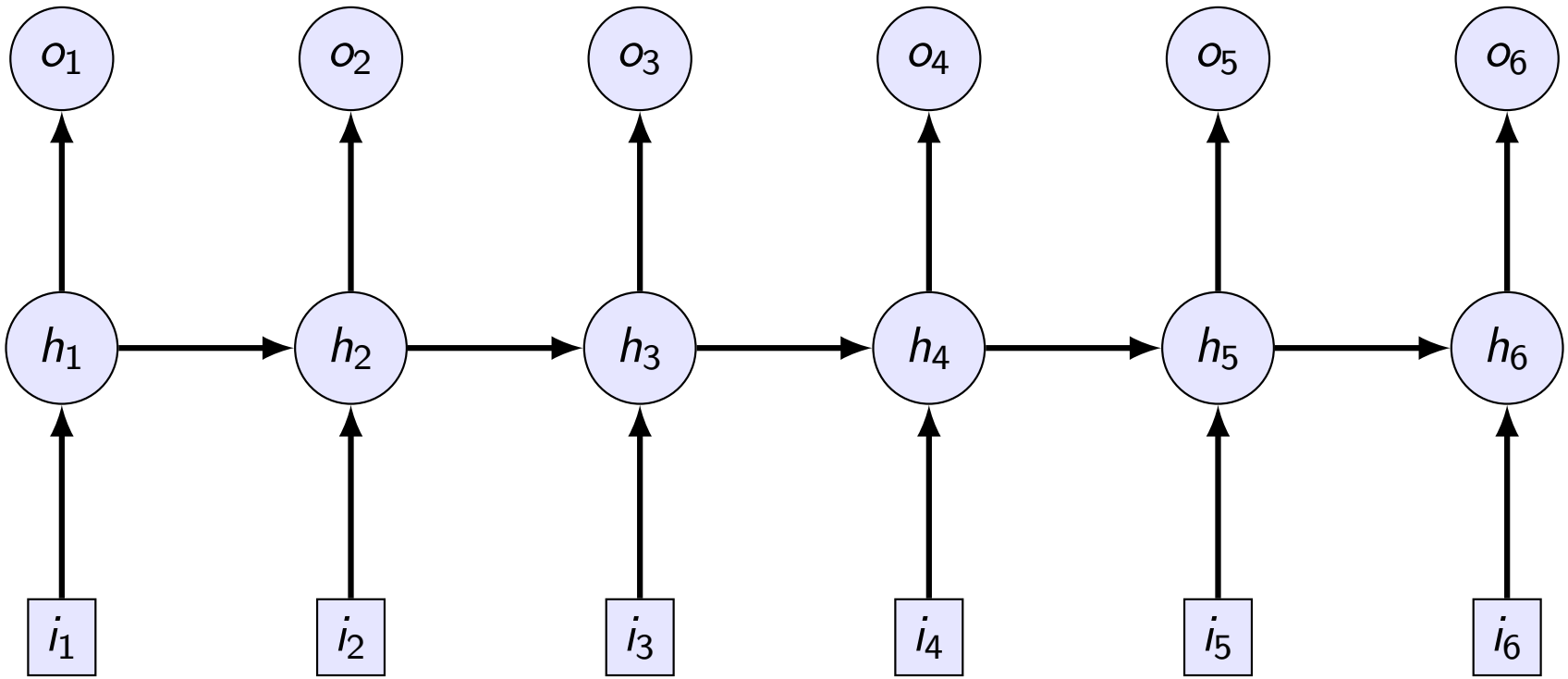
1. Training - 80
2. Validation - 10
3. Testing - 10

All experiments were replicated 3 times

Models

1. Recurrent neural networks
2. Convolutional neural networks

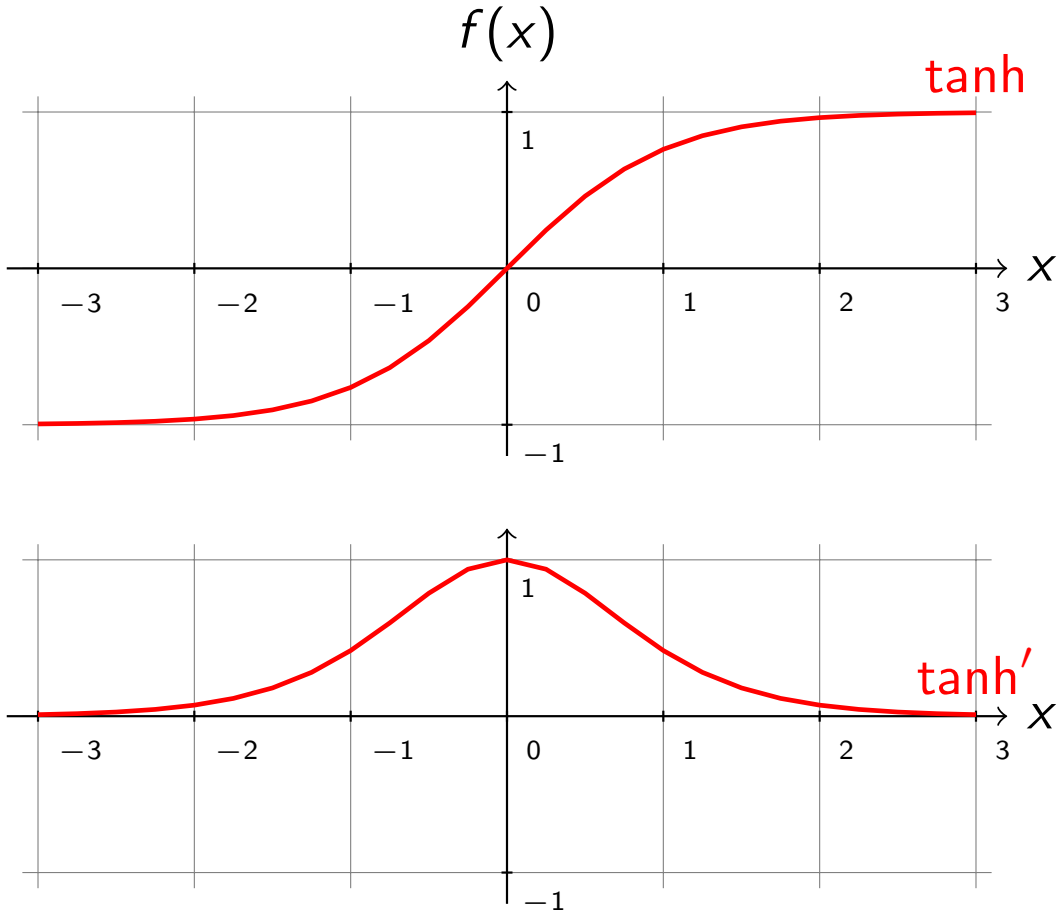
RNN



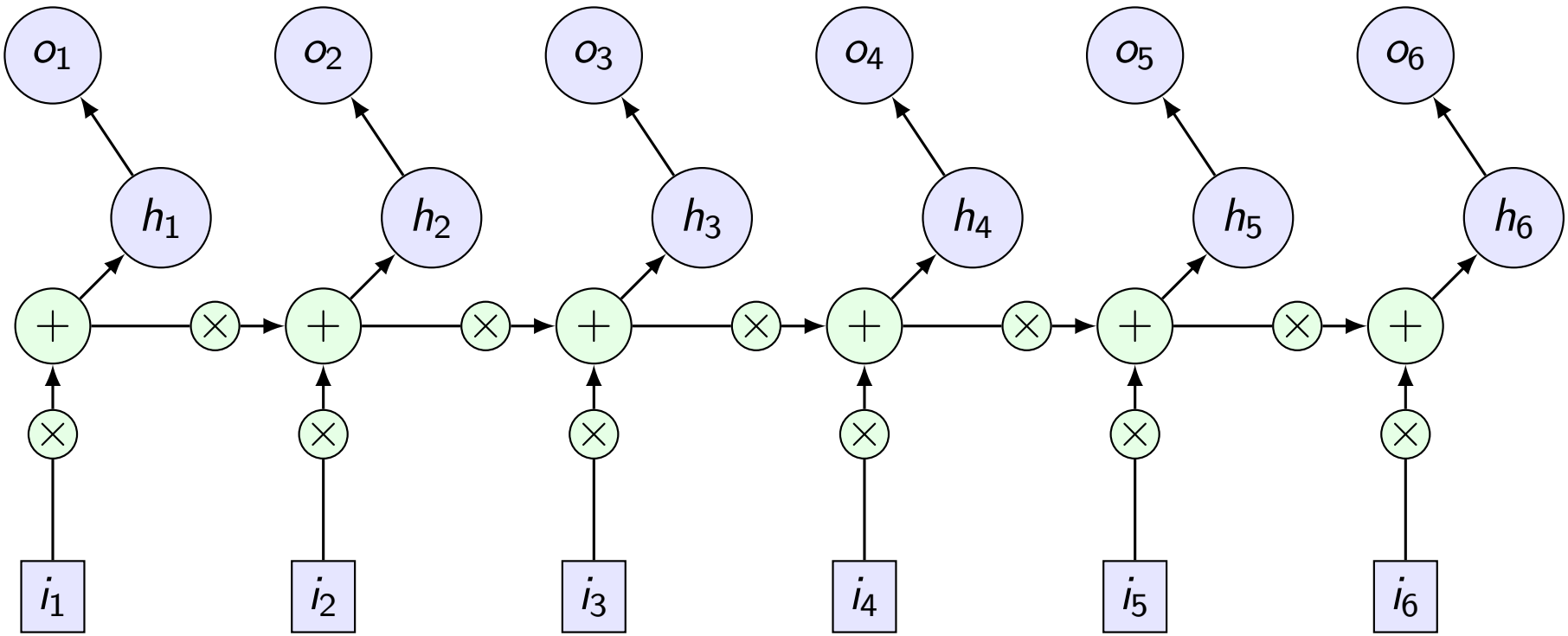
Gradiend wanishing / exploding

$$f(g(x))' = f'(g(x))g'(x)$$

$$f(g(h(x)))' = f'(g(h(x)))g'(h(x))h'(x)$$



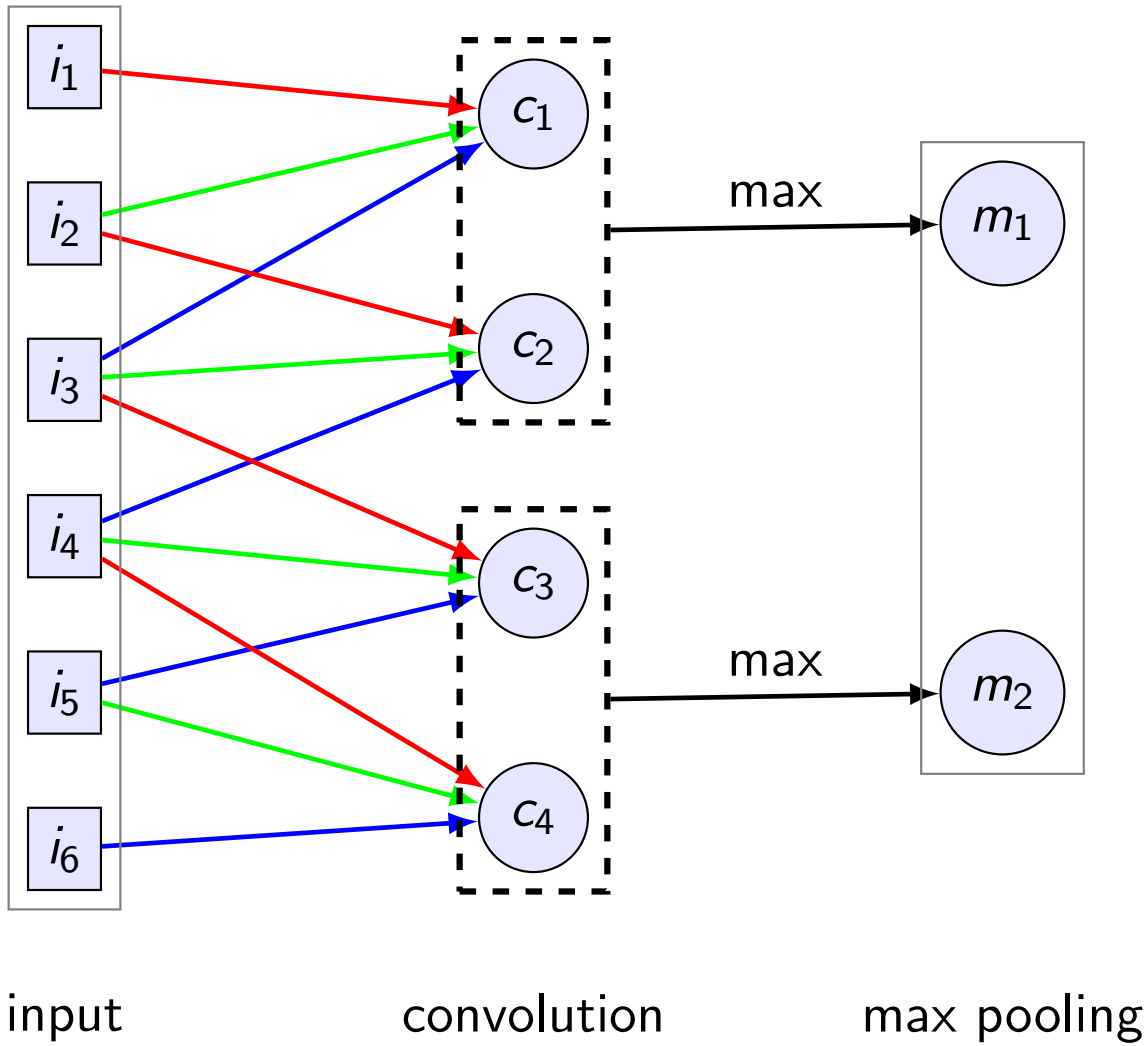
Long short-term memory



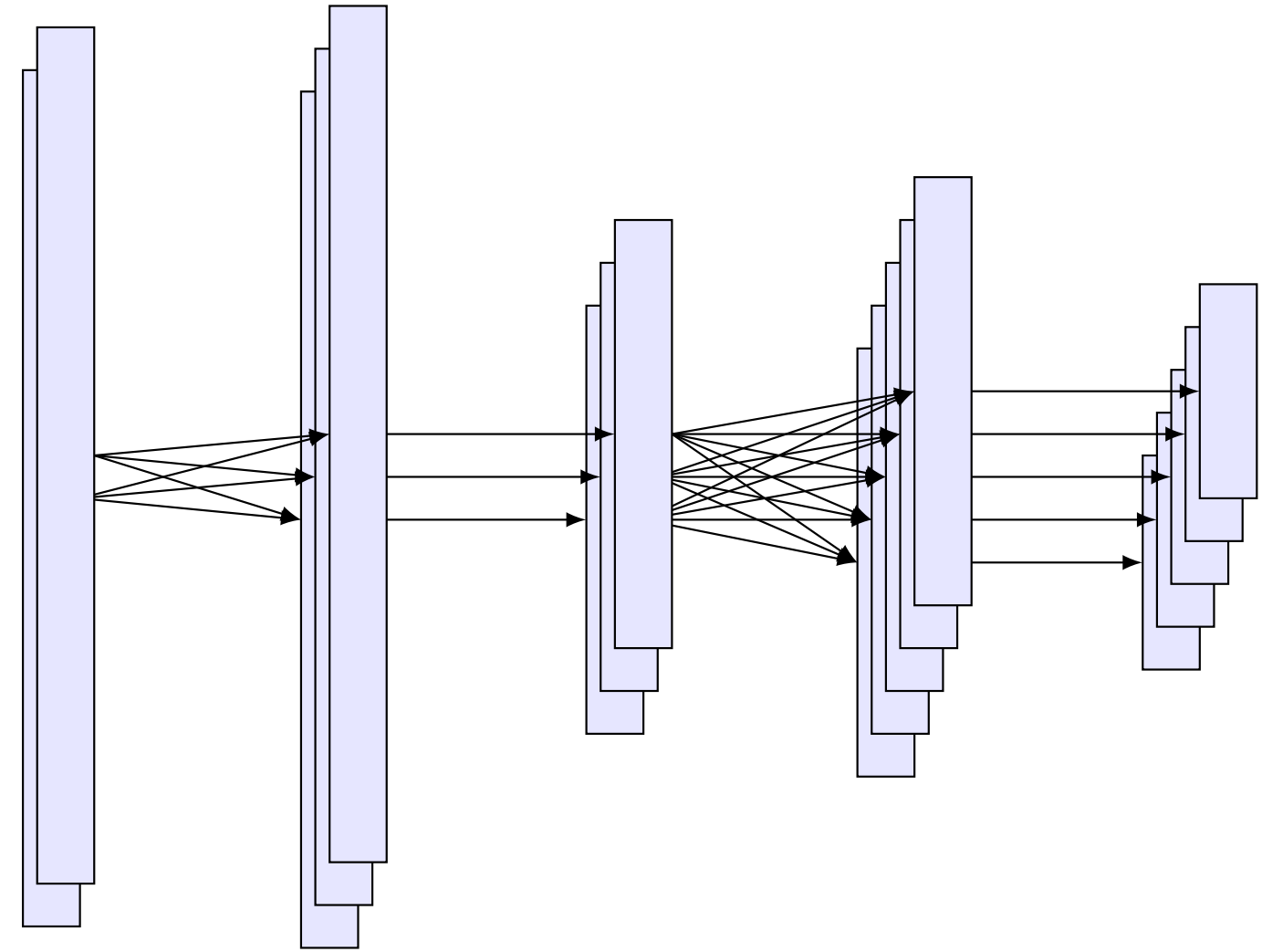
LSTM Results

coding genes error: 0.131
no further development

1D convolutional neural network

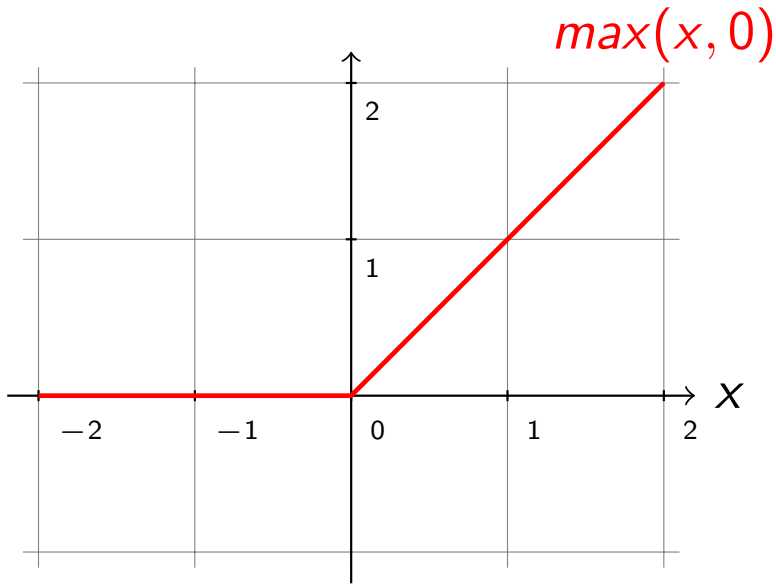


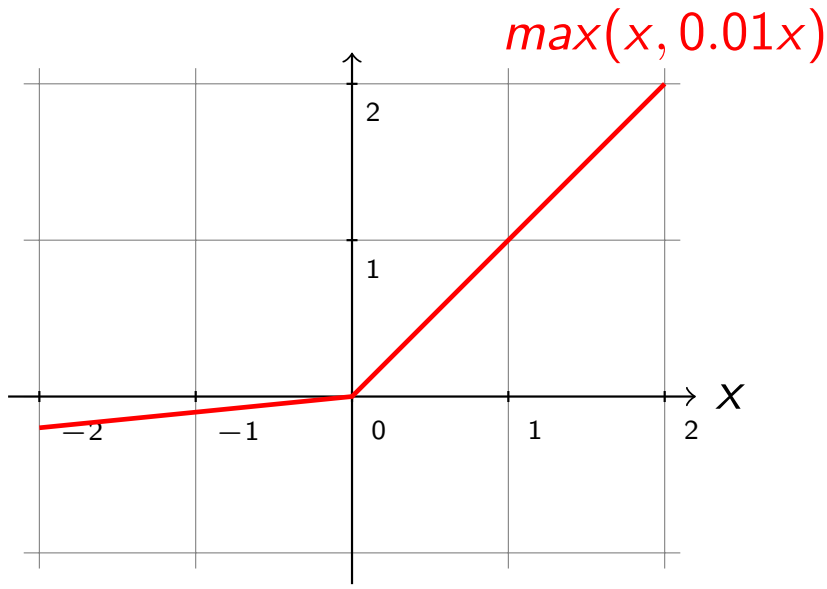
1D convolutional neural network



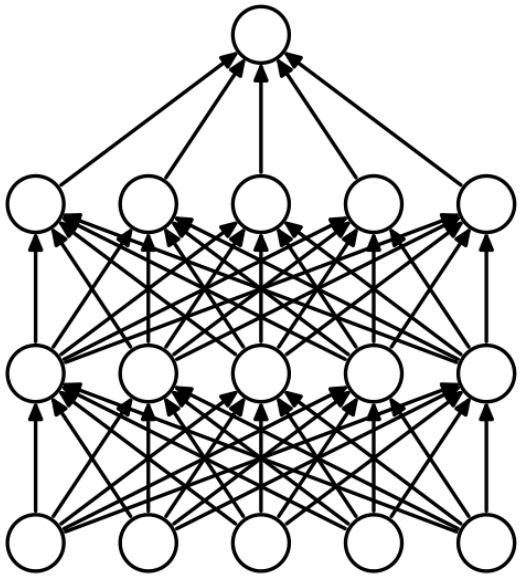
input convolution max pooling convolution max pooling

ReLU, Leaky ReLU

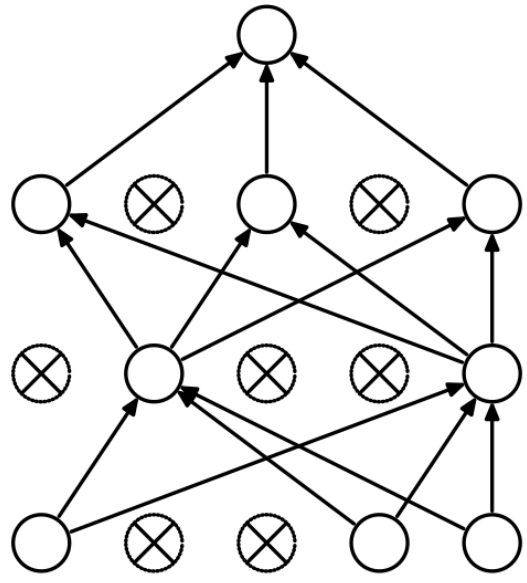




Dropout

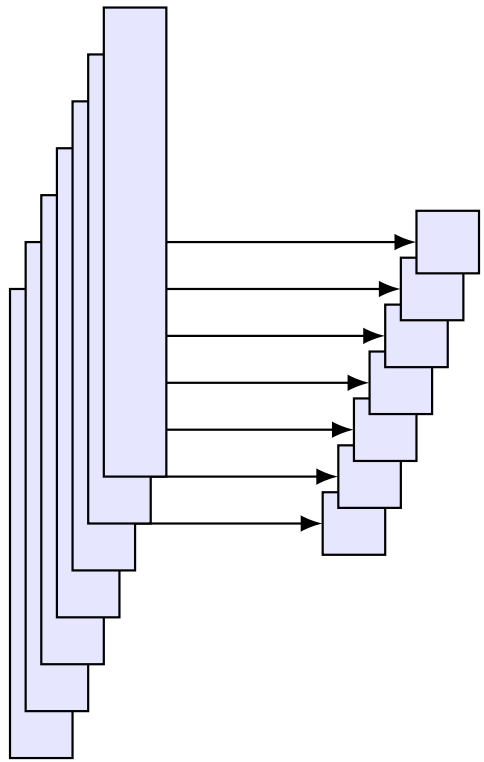


(a) Standard Neural Net

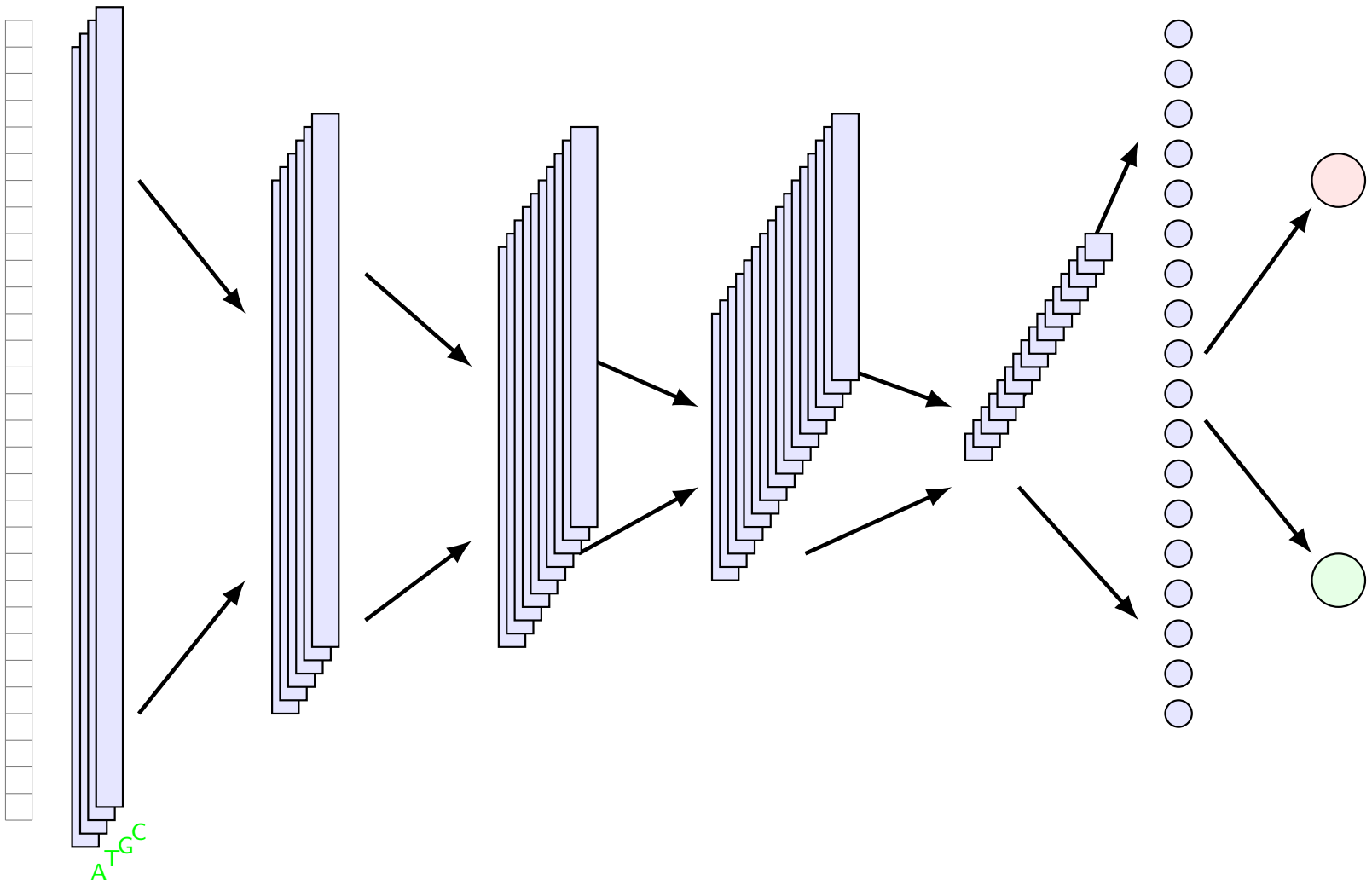


(b) After applying dropout.

Feature layer



Network structure



Network structure

Input TSS[-1000, 500]

- dropout: 0.1
- shape: 4×1500

Convolution 1

- features: 30
- conv size: 4
- max pooling: 2
- dropout: 0.2
- shape: 30×748

Convolution 2

- features: 60
- conv size: 6
- max pooling: 2
- dropout: 0.2
- shape: 60×371

Network structure

Convolution 3

- features: 60
- conv size: 6
- max pooling: 2
- dropout: 0.2
- shape: 60×183

Feature layer

- neurons number: 60
- dropout: 0.5

Fully connected

- neurons number: 60
- dropout: 0.1

L1 = 0.00001

L2 = 0.00001

Results

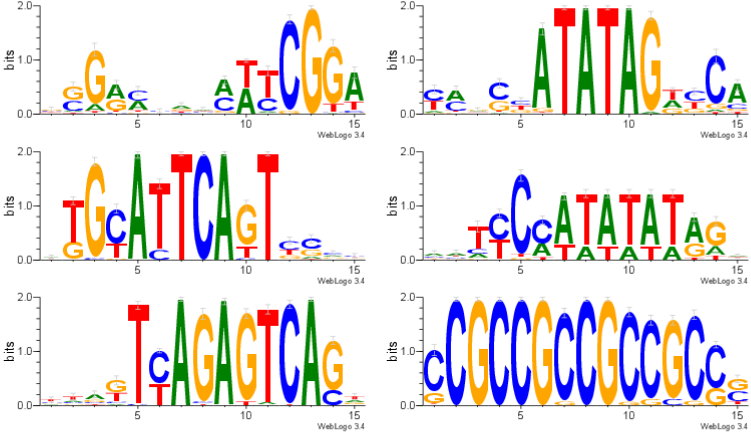
		error rate	ROC
mm9	genes coding	0.075	0.976
	genes all	0.129	0.944
	cage near coding	0.062	0.985
	cage all	0.096	0.965
hg19	genes coding	0.077	0.975
	genes all	0.144	0.930
	cage near coding	0.059	0.985
	cage all	0.088	0.970

So what?

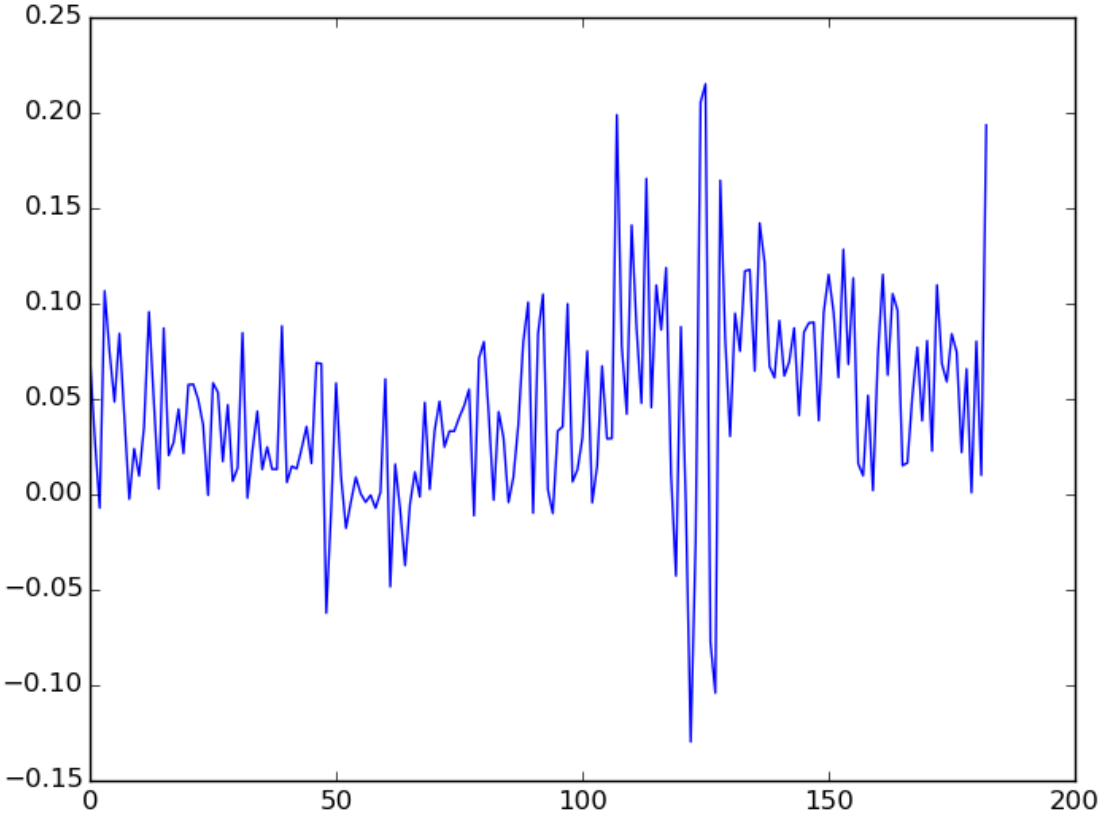
1. Full genome prediction
2. Patterns extraction
3. Relevance

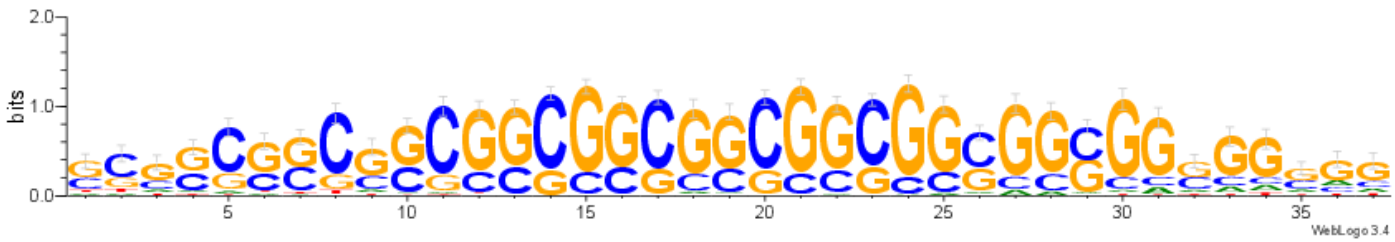
Full genome prediction

Convolution 2 patterns

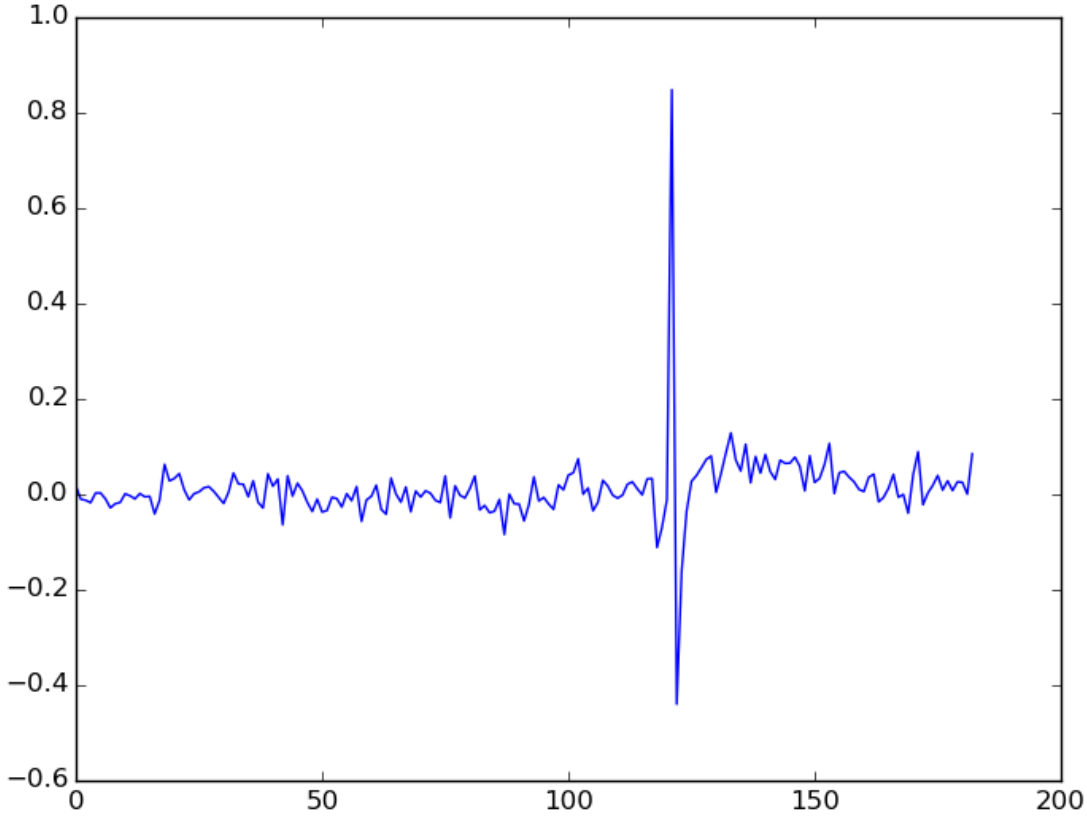


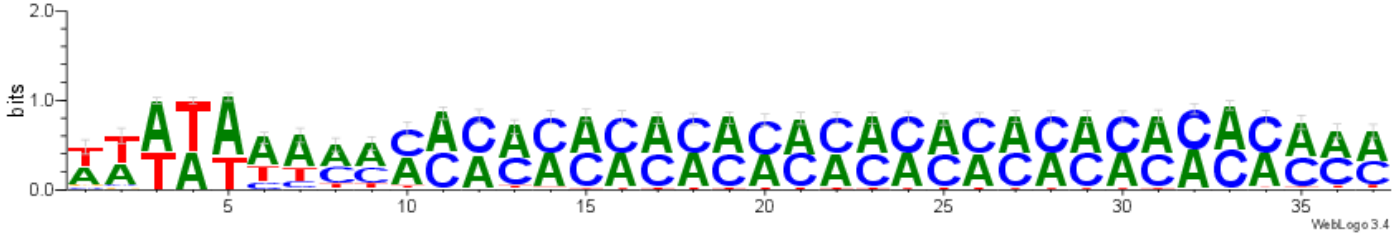
Convolution 3 pattern



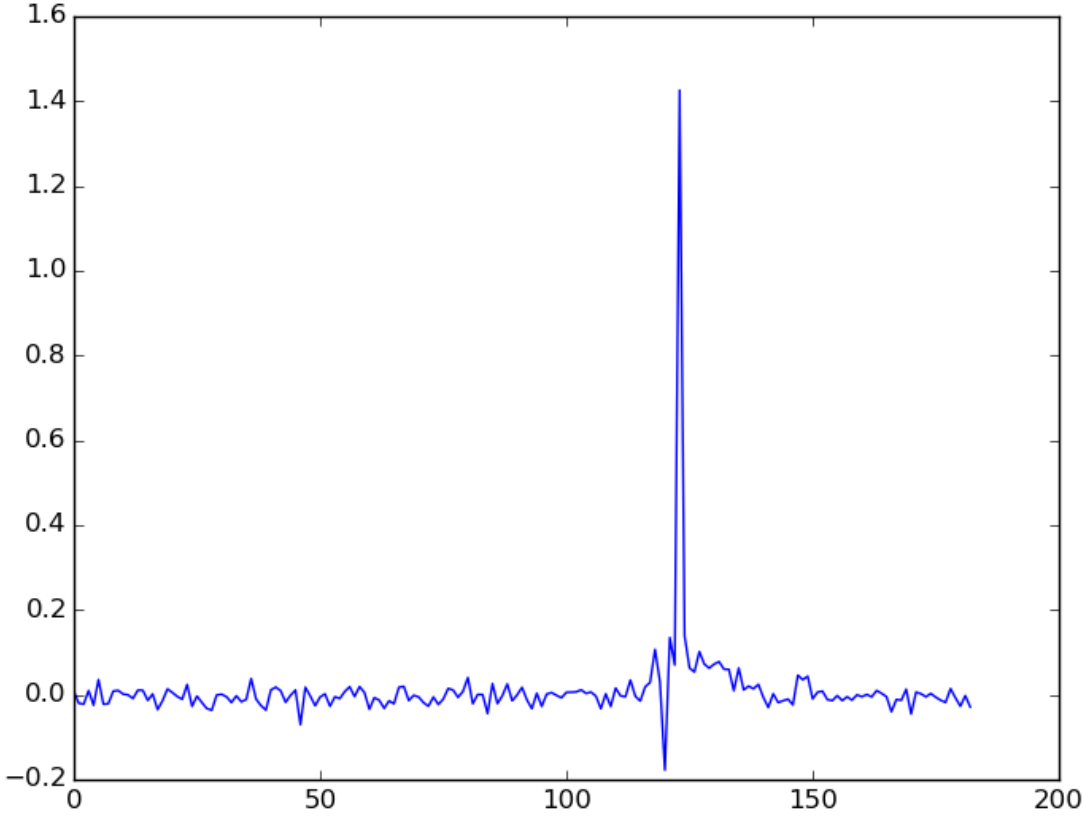


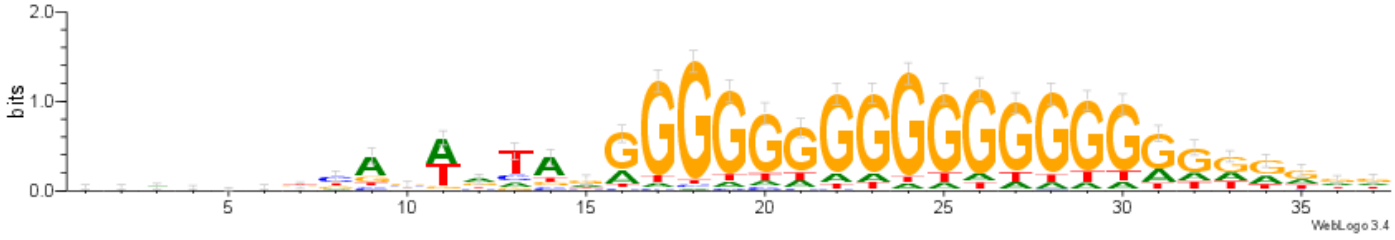
Convolution 3 pattern





Convolution 3 pattern





Relevance

Try to find parts of input that give more contribution

Plans

1. Comparison with others
2. Get better data
3. There still a lot Deep Learning tricks
4. Biological results

To be continued

Questions?