DanQ: a hybrid convolutional and recurrent neural network for quantifying the function of DNA sequences

Adi, Armin, Behrooz, Karen, Philip, Wendy

CS273B Presentation

Oct. 24, 2016

Central Dogma of Molecular Biology

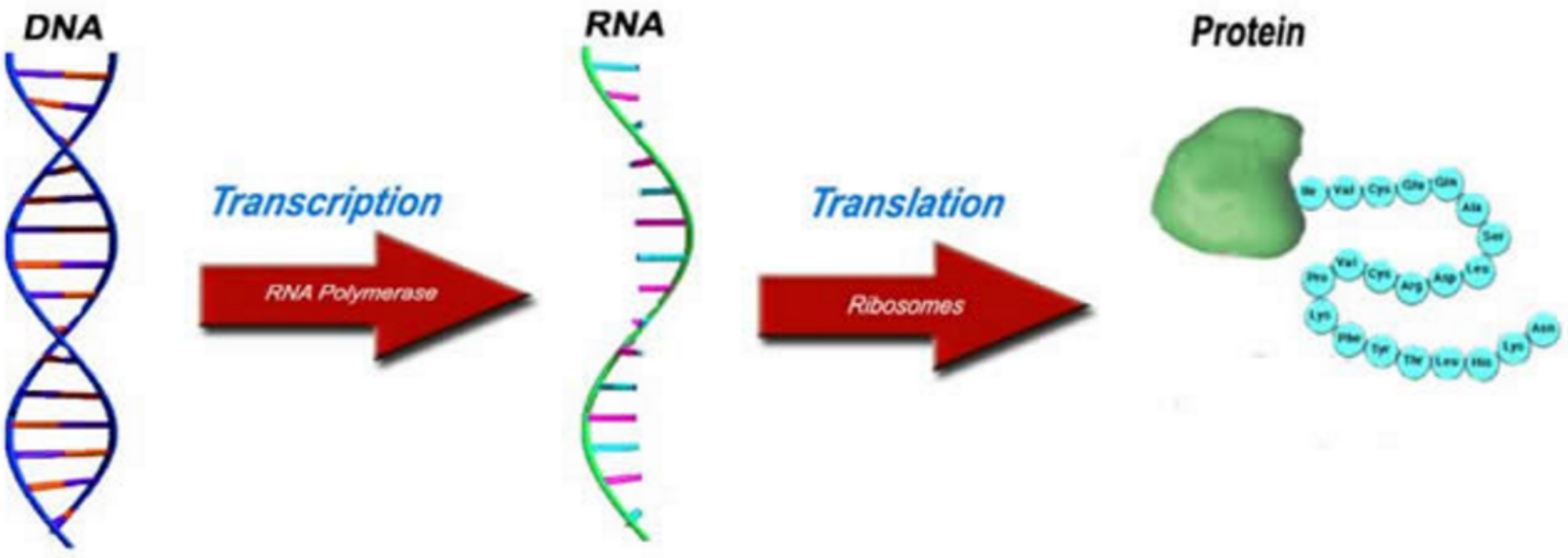
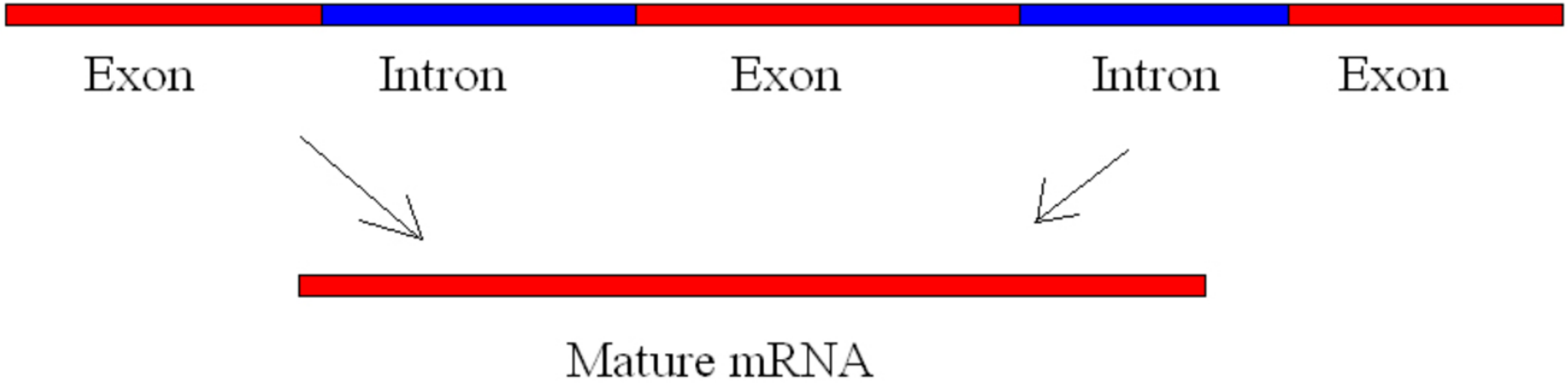


figure from http://aisbiology.pbworks.com/w/page/49012795/DNA,%20RNA,%20and%20Proteins

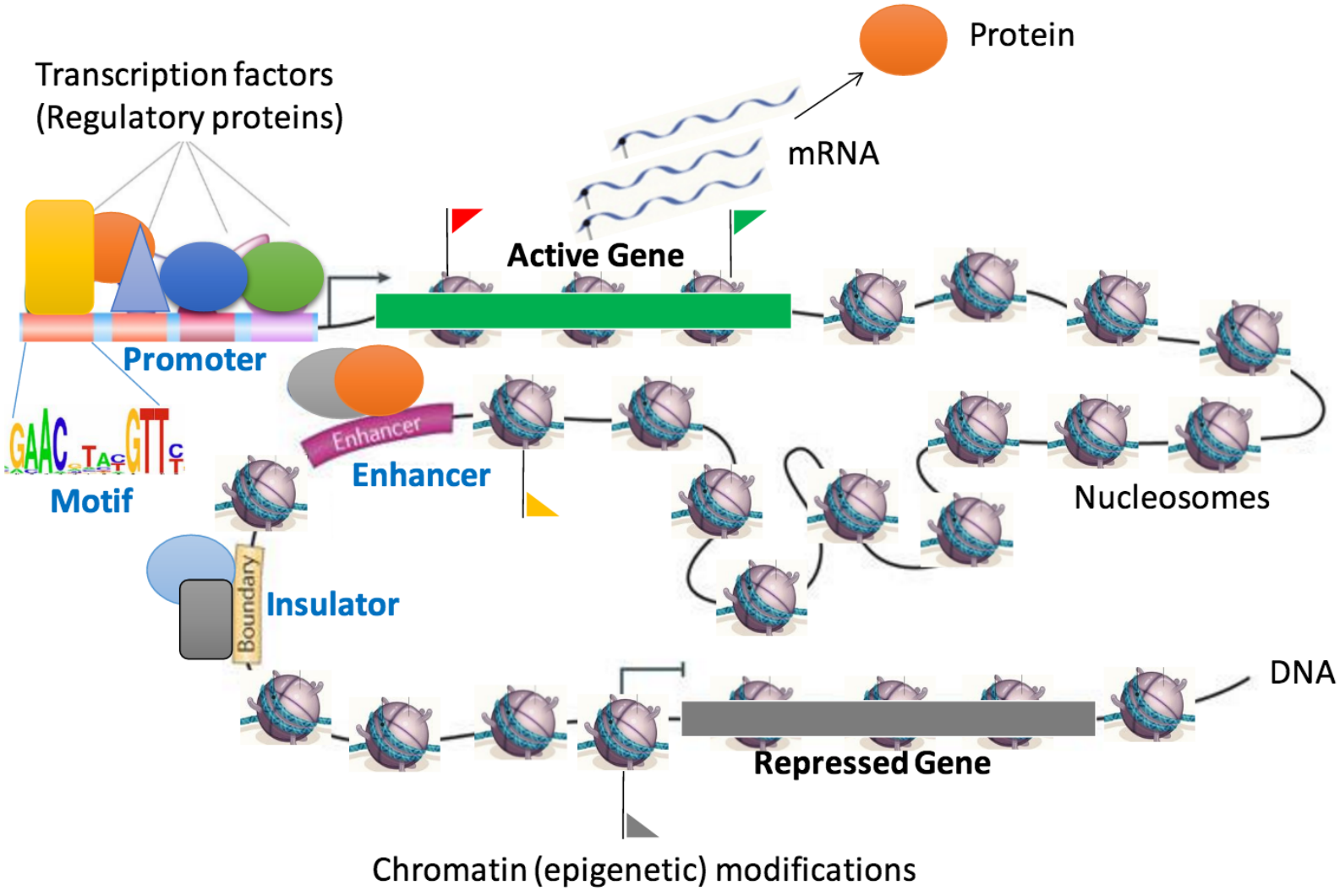
Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.



Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.



Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.
- Mutations in the coding parts can change the protein produced, however these are not the only mutations that could be "important".
  - 93% of disease associated variants are in non-coding regions.
- What could non-coding SNPs affect?

Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.
- Mutations in the coding parts can change the protein produced, however these are not the only mutations that could be "important".
  - 93% of disease associated variants are in non-coding regions.
- What could non-coding SNPs affect?
  - Transcription factor binding

Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.
- Mutations in the coding parts can change the protein produced, however these are not the only mutations that could be "important".
  - 93% of disease associated variants are in non-coding regions.
- What could non-coding SNPs affect?
  - Transcription factor binding
  - DNA accessibility

Coding vs. Noncoding

- Most of the genome (98%) of the genome is non-coding.
- Mutations in the coding parts can change the protein produced, however these are not the only mutations that could be "important".
  - 93% of disease associated variants are in non-coding regions.
- What could non-coding SNPs affect?
  - Transcription factor binding
  - DNA accessibility
  - Epigenome (histones, methylation)

Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional

Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional
- A variant that changes chromatin profile may imply a change in biology

Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional
- A variant that changes chromatin profile may imply a change in biology
  - TF binding may get affected

Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional
- A variant that changes chromatin profile may imply a change in biology
  - TF binding may get affected
  - Change in methylation

Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional
- A variant that changes chromatin profile may imply a change in biology
  - TF binding may get affected
  - Change in methylation
  - Increase in DNase accessibility

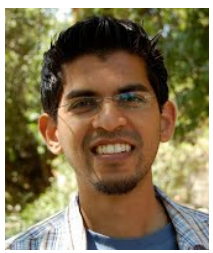
Data

- Labels (from ENCODE[1] and Roadmap Epigenomics)

[1] Thanks Anshul!

Data

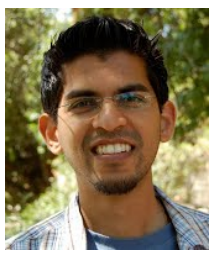
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix



[1] Thanks Anshul!

Data

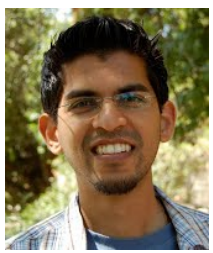
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq



[1] Thanks Anshul!

Data

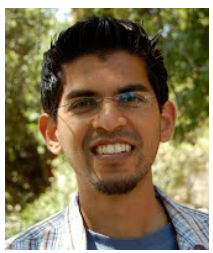
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq
    - DNase-seq



[1] Thanks Anshul!

Data

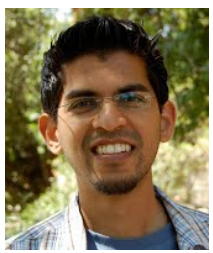
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq
    - DNase-seq
- Labels for Regression



[1] Thanks Anshul!

Data

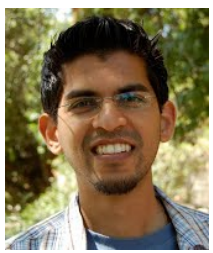
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq
    - DNase-seq
- Labels for Regression
  - Functional Labels



[1] Thanks Anshul!

Data

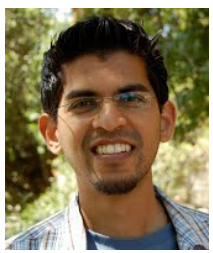
- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq
    - DNase-seq
- Labels for Regression
  - Functional Labels
- One-hot encoded input sequence



[1] Thanks Anshul!

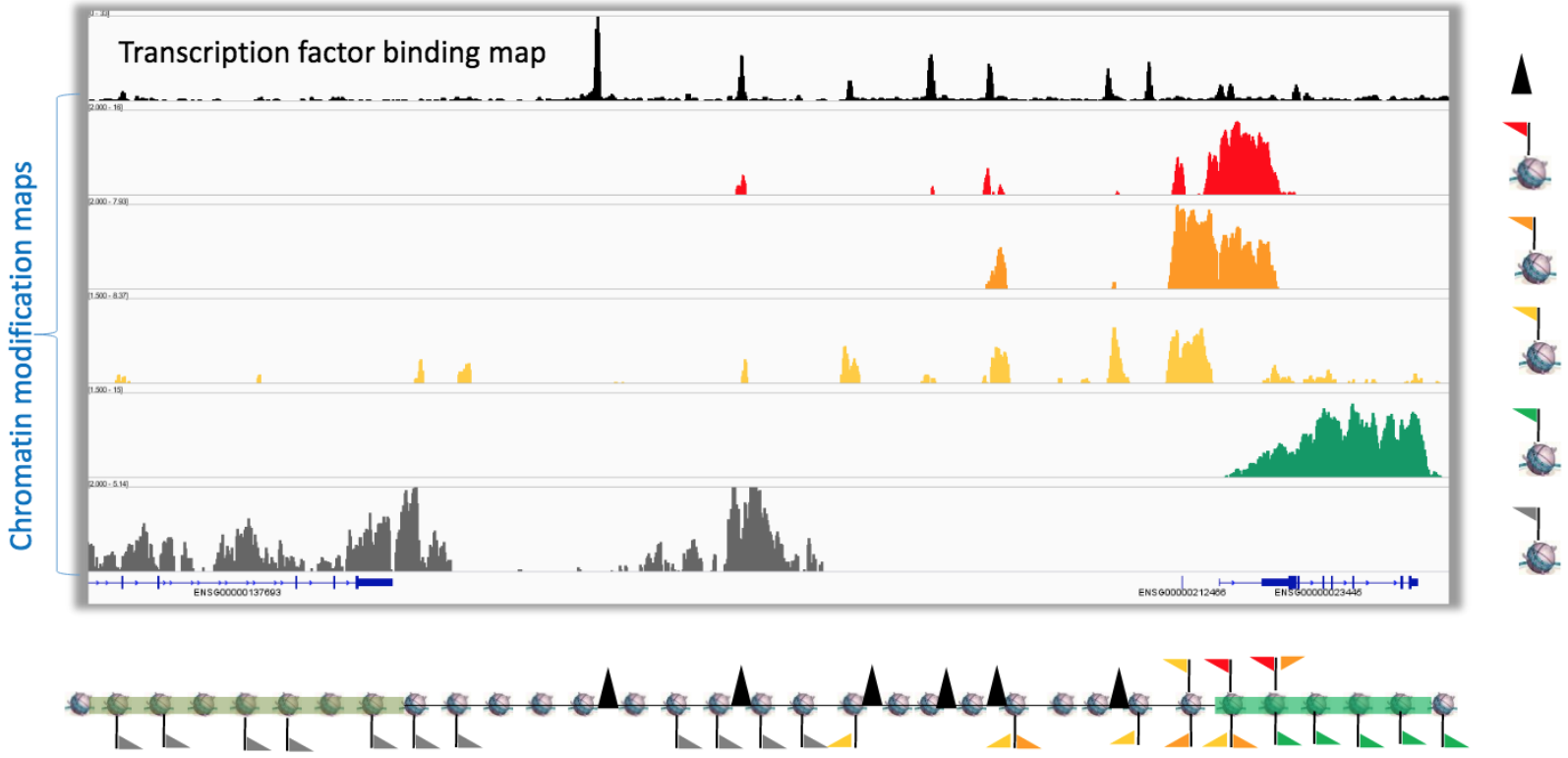
Data

- Labels (from ENCODE[1] and Roadmap Epigenomics)
  - Multi-task Label Matrix
    - ChIP-seq
    - DNase-seq
- Labels for Regression
  - Functional Labels
- One-hot encoded input sequence
  - 1000bp sequence centered on a 200bp bin



[1] Thanks Anshul!

Data (ChIP-seq)

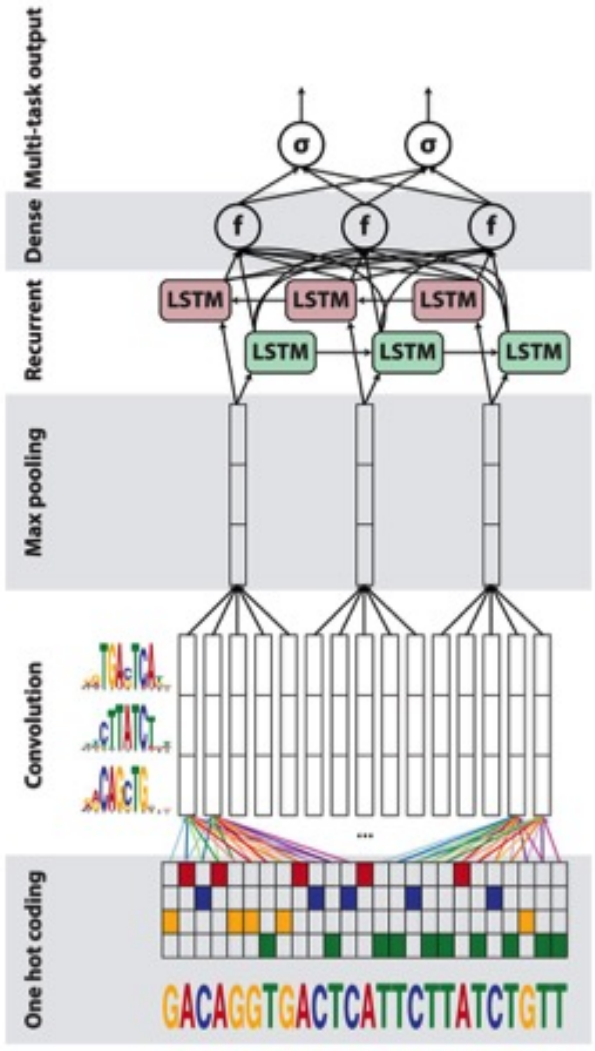


Data (DNase-seq)

Architecture Overview

- Goal: To predict chromatin profile from sequence in order to determine whether non-coding variants are functional
- Hybrid convolutional and recurrent deep neural network
- Convolution layer captures regulatory motifs
- Adds recurrent layer to capture long term (extension of DeepSEA approach)

Architecture



Daniel Quang, and Xiaohui Xie Nucl. Acids Res. 2016;44:e107

Example of hybrid CNN / RNN architecture

- "Deep Visual-Semantic Alignments for Generating Image Descriptions" (Karpathy and Li, CVPR 2015)

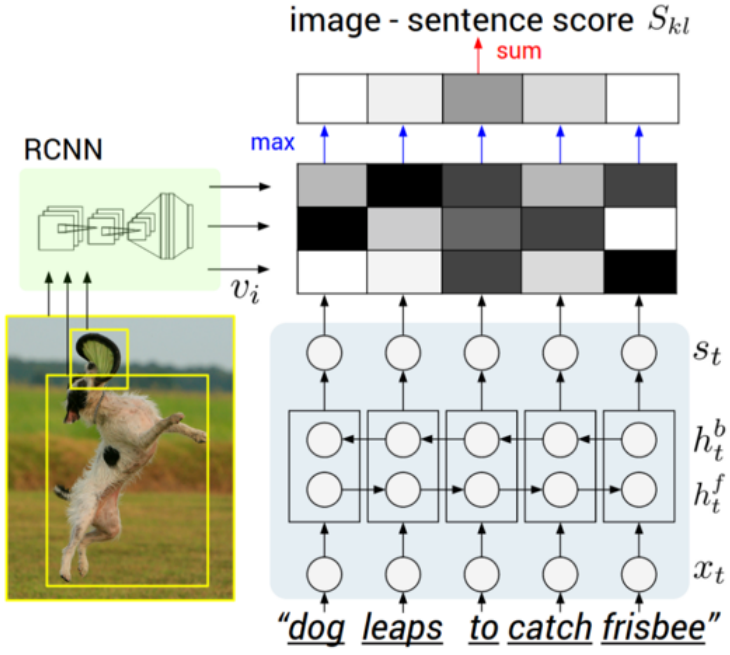
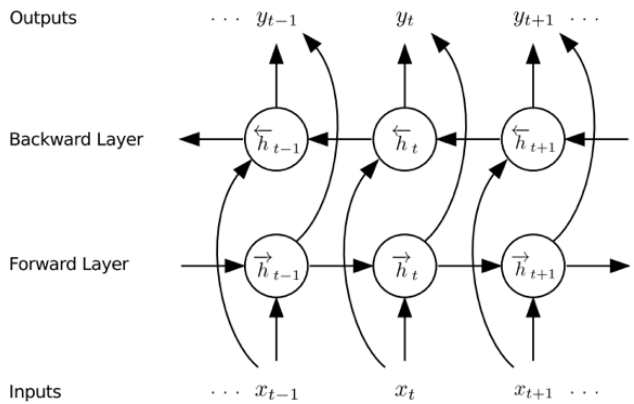


Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

Bidirectional LSTM - example

- "Translation Modeling with Bidirectional Recurrent Neural Networks" (Sundermeyer et al. 2015)



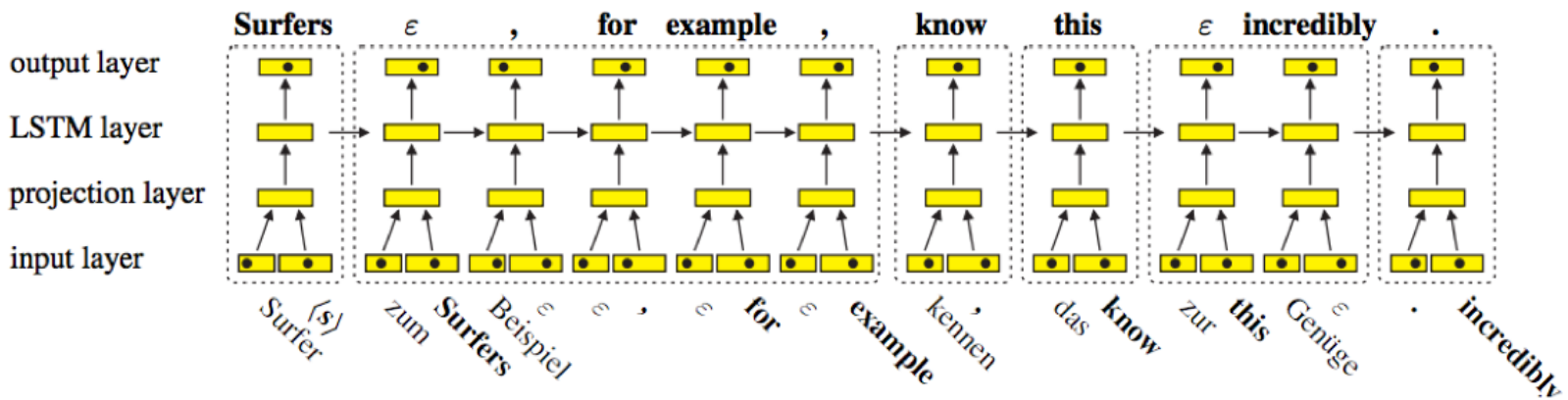
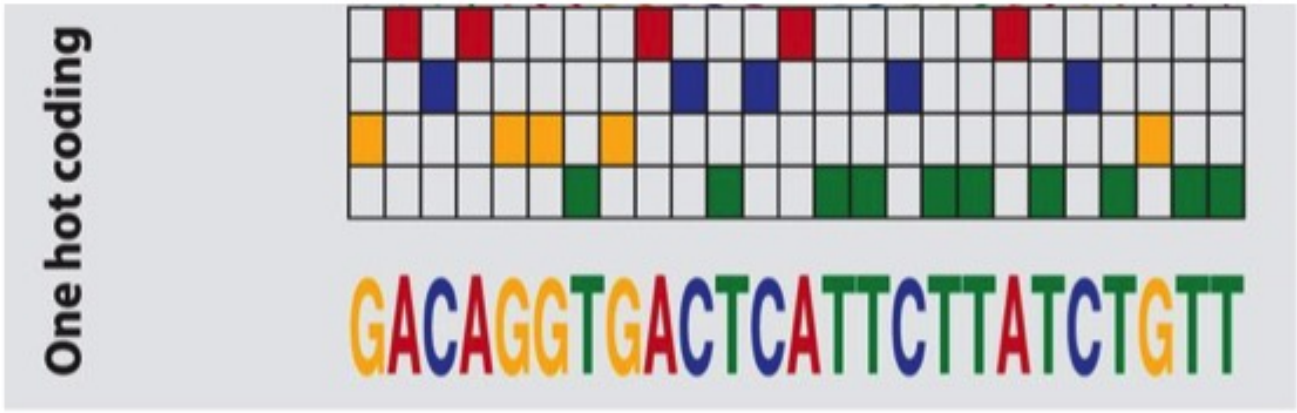


Figure 5: A recurrent phrase-based joint translation model, unfolded over time. Source words are printed in normal face, while target words are printed in bold face. Dashed lines indicate phrases from the example sentence. For brevity, we omit the precise handling of sentence begin and end tokens.

Architecture

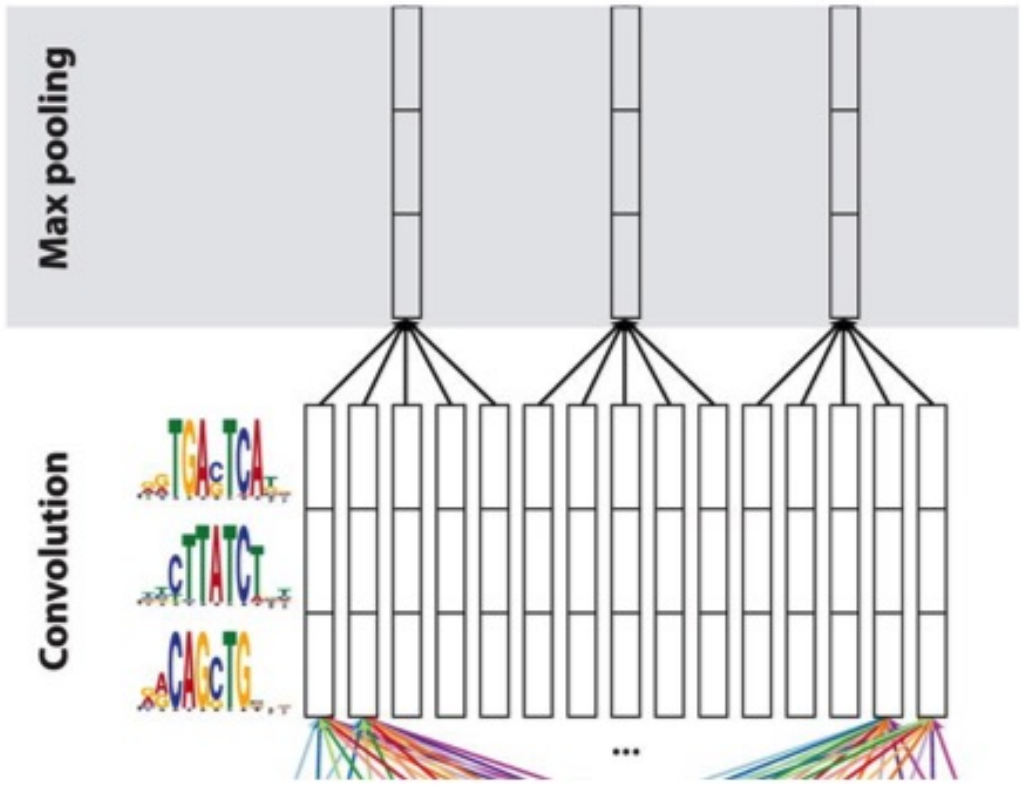
an input sequence is first one hot encoded into a 4-row bit matrix



Daniel Quang, and Xiaohui Xie Nucl.
Acids Res. 2016;44:e107

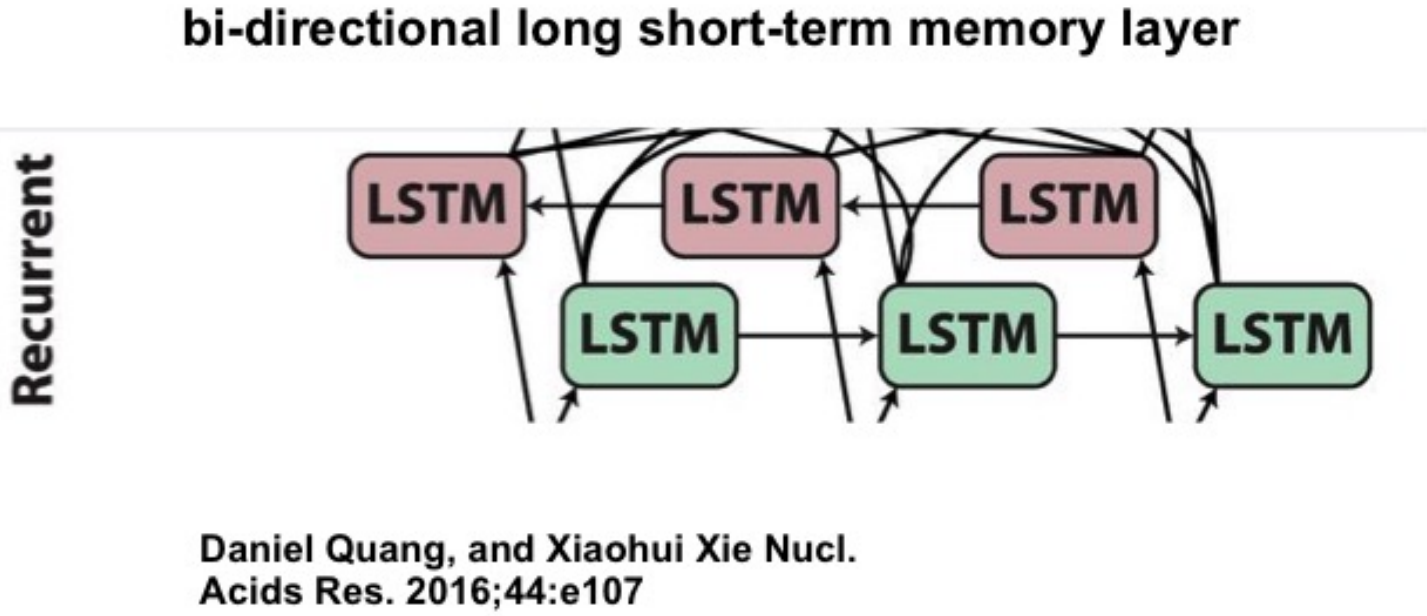
Architecture

a convolution layer acts as a motif scanner
max pooling reduces the size of the output matrix



Daniel Quang, and Xiaohui Xie Nucl. Acids Res. 2016;44:e107

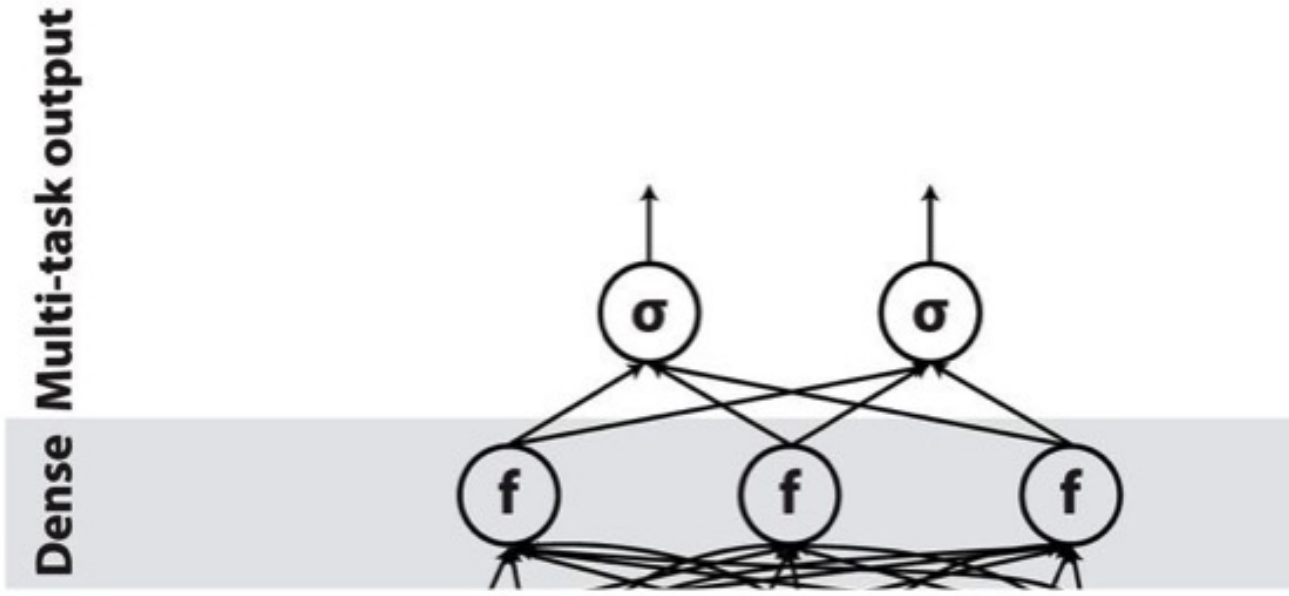
Architecture



Daniel Quang, and Xiaohui Xie Nucl.
Acids Res. 2016;44:e107

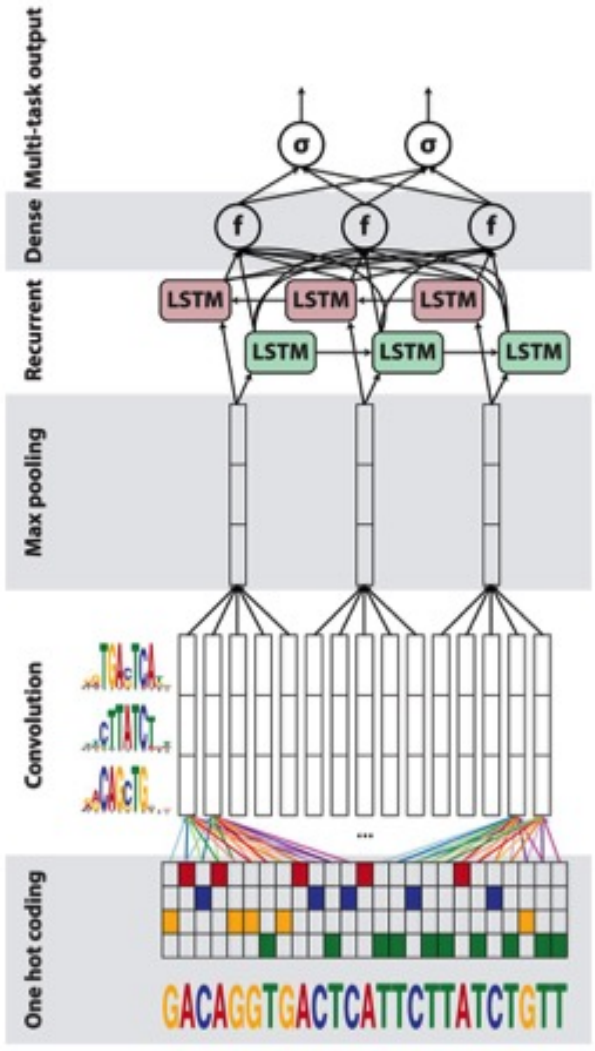
Architecture

fully connected layer of rectified linear units
→ sigmoid non-linear transformation



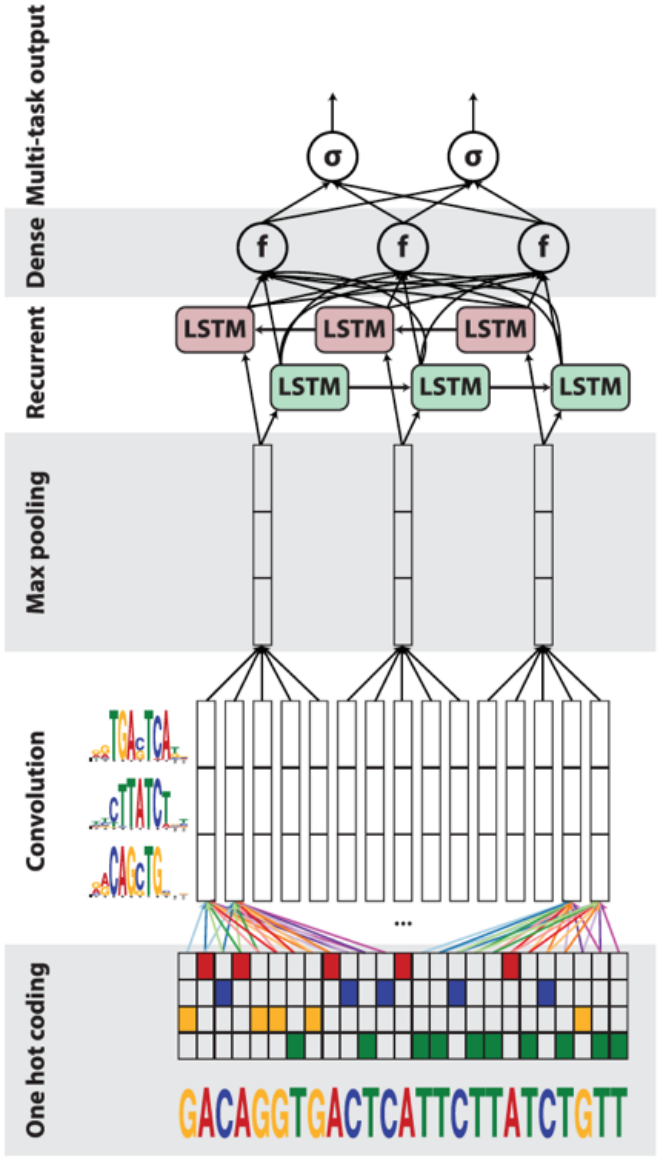
Daniel Quang, and Xiaohui Xie Nucl.
Acids Res. 2016;44:e107

Architecture



Daniel Quang, and Xiaohui Xie Nucl. Acids Res. 2016;44:e107

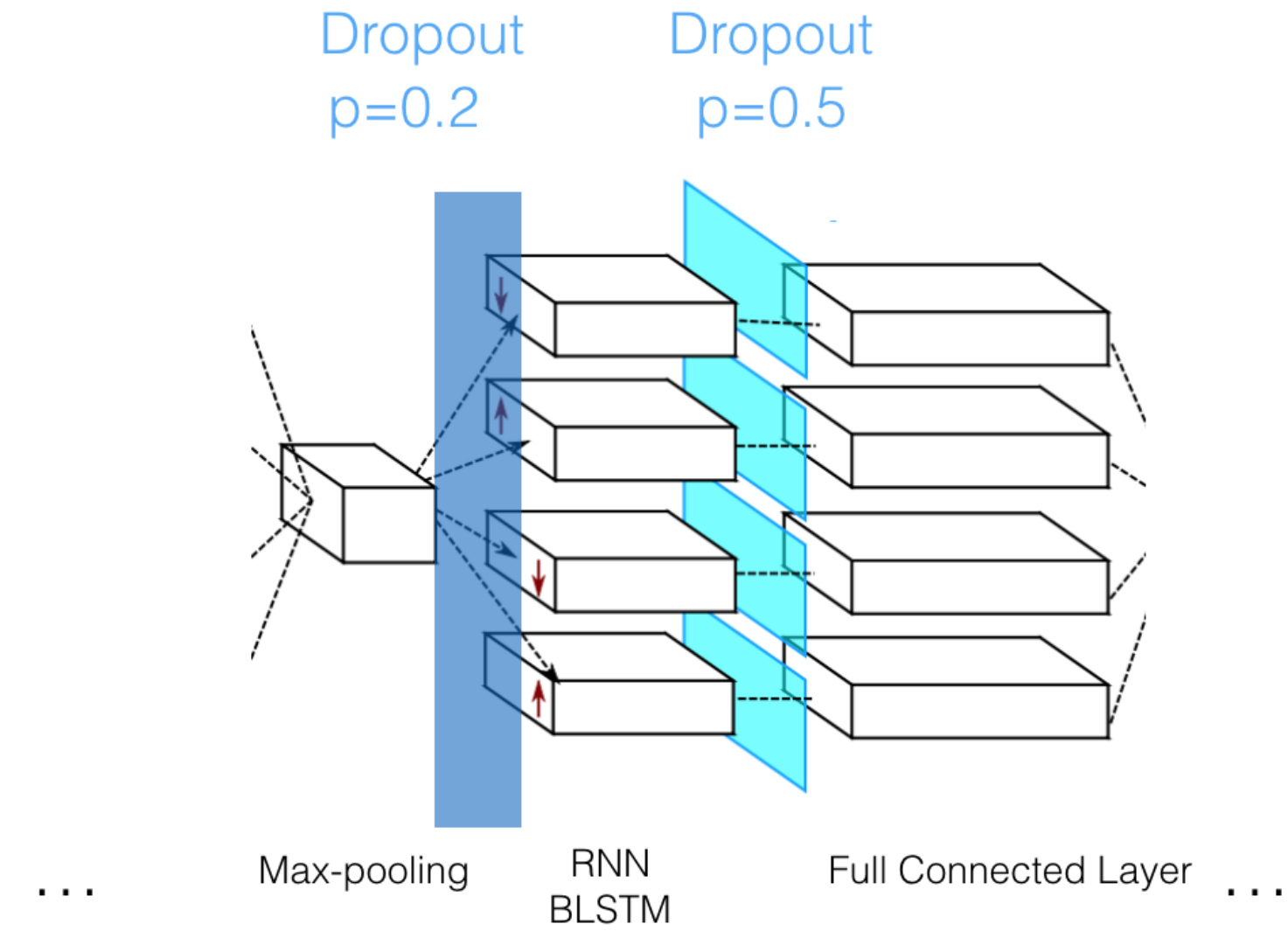
Training - Dropout



dropout rate p=0.5

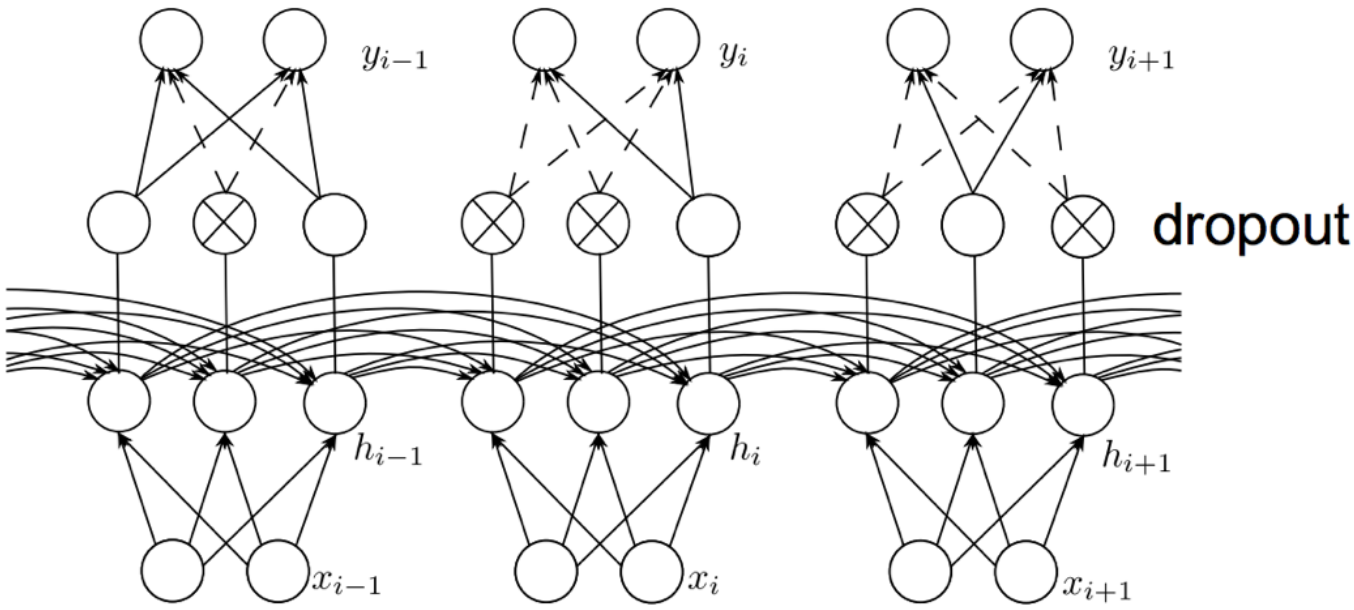
dropout rate p=0.2

Training - Dropout



Training - Dropout

Dropout is only applied to feed-forward connections in RNNs. The recurrent connections are kept untouched.



Pham, Vu, et al. "Dropout improves recurrent neural networks for handwriting recognition." Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014.

Training - Initialization

DanQ	All 320 convolution kernels are initialized randomly from unif (− 0.05,0.05).Biases set to 0.
DanQ-JASPAR	Initialize 1024 convolutional kernels from known motifs in JASPAR: About half of the subsections of each kernel are set to the values of the position frequency matrix of known motif minus 0.25. Bias set to unif (− 1.0,0.0).

Training - Optimization

RMSprop - with minibatch of size 100

rmsprop: Keep a moving average of the squared gradient for each weight

$$MeanSquare(w,t) = 0.9\ MeanSquare(w,\ t-1) + 0.1\left(\partial E / \partial w(t)\right)^2$$

Dividing the gradient by $\sqrt{MeanSquare(w,\ t)}$ makes the learning work much better (Tijmen Tieleman, unpublished).

Loss function - Average multi-task binary cross entropy

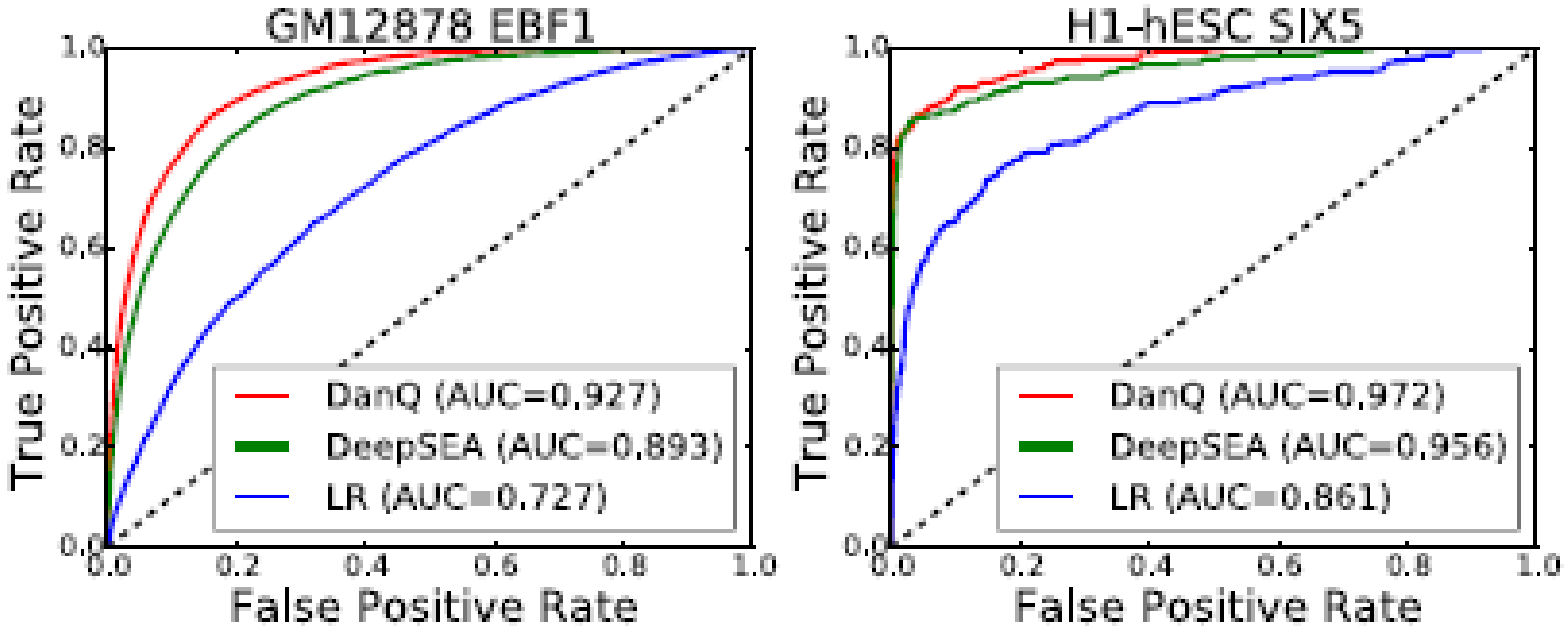
$$L(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)\right],$$

where $\hat{y}_n \equiv g(\mathbf{w}\cdot\mathbf{x}_n)$, with $g(z)$ the logistic function as before.

Training - Optimization time

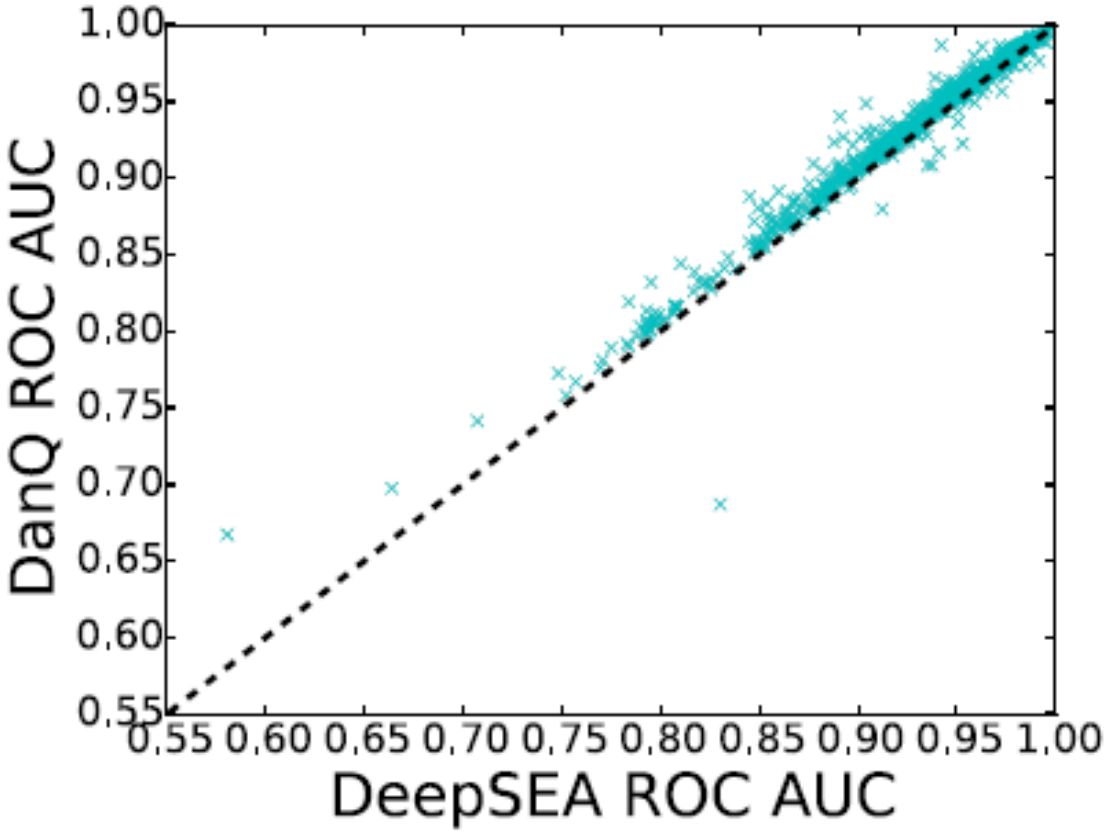
- DanQ Model took 60 epochs to converge (6 hr/epoch)
- DanQ-JASPAR took 30 epochs to converge (12 hr/epoch)
- Hardware: NVIDIA Titan Z GPU

Results - ROC Example



- ROC curves are computed for 919 binary targets
- Notice the "decent" performance of Logistic regression baseline

Results - ROC AUC



- DanQ performs marginally better than DeepSEA (1-4%) on 94% of the targets

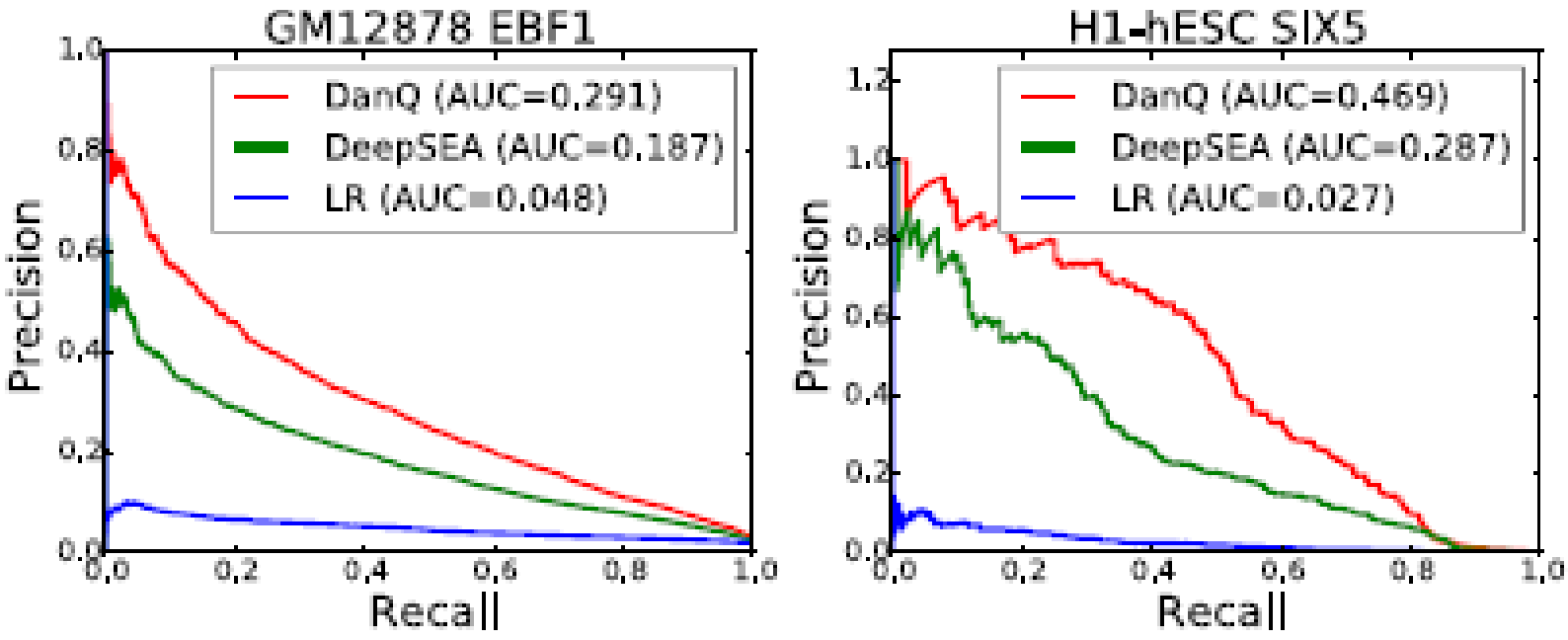
Results - Caveats

- The data is highly unbalanced (There are much more negatives)
- ROC is very optimistic in this scenario
- It makes sense to consider a measure that more robust

Results - Caveats

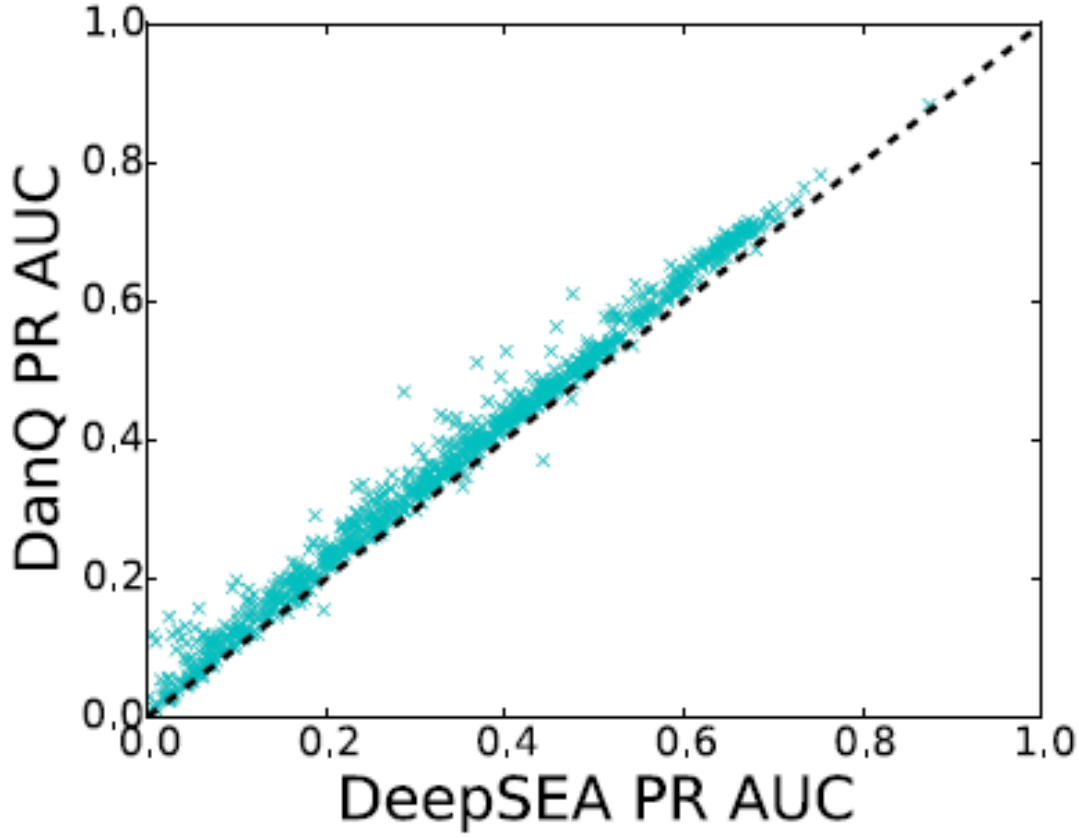
- Precision $\equiv \frac{TP}{TP+FP}$, Recall $\equiv \frac{TP}{TP+FN} \equiv TPR$
- Since there is no direct dependence on TN, these measure are more robust
- In particular, Precision focuses entirely on the quality of the positive predictions

Results - PR Curve Example



- Notice that the logistic regression performance falls

Results - PR AUC



- Observe that the results are much more pessimistic!
- This figure also shows that there is a lot of room for improvement

Results: Kernel Interpretation

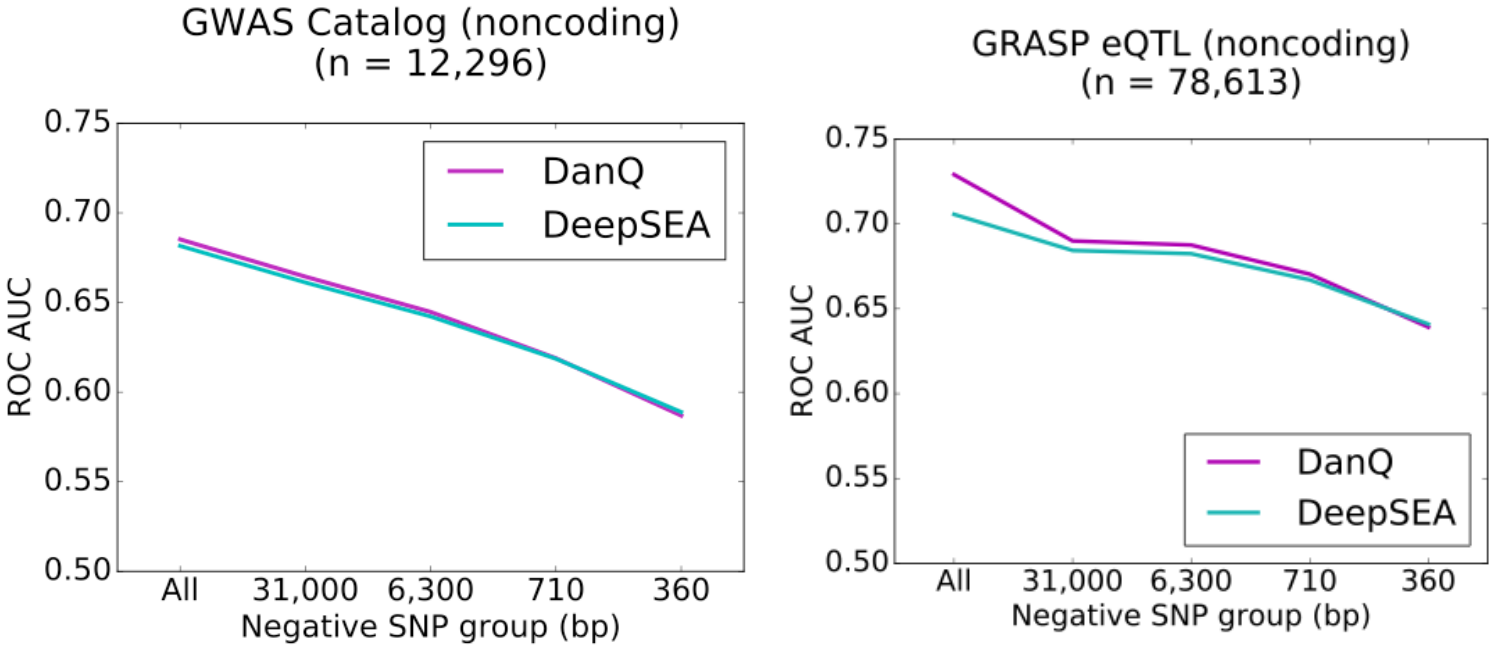
- The authors use the same methodology as DeepBind paper to construct motifs that correspond to the convolutional kernels
- 166 of the convolutional kernels that the network learns significantly match known motifs

Limitations / Future Work

- Model has fixed input length (choice of 1000 bp inputs for context)
  - Making the model fully recurrent could allow processing of arbitrary length (e.g. whole chromosome sequences)
- Model uses $unif[-0.05, 0.05]$ for initialization
  - Consider using Glorot (Gaussian) initialization (On the importance of initialization and momentum in deep learning, by Sutskever et al 2013.)
- Model uses dropout after max pooling + after LSTMs
  - FC layers have many more parameters, often need more regularization
  - (Gal, Yarin, and Zoubin Ghahramani. "Bayesian convolutional neural networks with Bernoulli approximate variational inference." arXiv:1506.02158 (2015))

Limitations / Future Work

- In training/testing data for functionality predictor, SNPs may not be *causal variants*,
  - This is a problem for training data coming from GWAS studies.



- "The DanQ framework outperforms the DeepSEA framework across most of the testing sets, with the performance difference being 0.5–2% in terms of the ROC AUC metric"

Limitations / Future Work

- In training/testing data for functionality predictor, SNPs may not be *causal variants*,
  - This is a problem for training data coming from GWAS studies.

- Data used:
  - DeepSea: all locations with at least one peak for the TF-binding sites ($\approx 17\%$ of all sites)
  - DanQ: all locations with at least one peak for any of the data provided ($\approx 90\%$ of all sites)

Questions?

Thank you for listening!