Computational Systems Biology: Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

Jacqueline Liu
Thursday, March 9

Deep Motif (DeMo): Visualizing Genomic Sequence Classifications

Jack Lanchantin, Ritambhara Singh, Zeming Lin, Yanjun Qi

http://mit6874.github.io

Table of Contents

Overview

Key Claim

- Deep Motif (DeMo) achieves state-of-the-art accuracy in motif classification and provides visual representations of positive binding sites

Importance

- DeMo can be used to more quickly screen for TF binding sites

Issues

- The paper does not share accuracy metrics (besides AUC) and may not be applicable for finding new binding sites
- Vague information and perhaps not reproducible

Two Main Goals

Classifying TF binding sites with a neural network

Visualizing motifs via class optimization

Assumptions

Classification

- Neural networks are good for identifying TF binding sites because of their scalability
- Deeper models are better at detecting relevant features and doing binary classification
- Highway MLP is more effective than standard MLP

Visualization

- Need better, more accessible visualization tools

Methods: Classification

Aim: make binary classifications to see if there's a positive TFBS

Input: raw nucleotide characters

- Similar model as in NLP

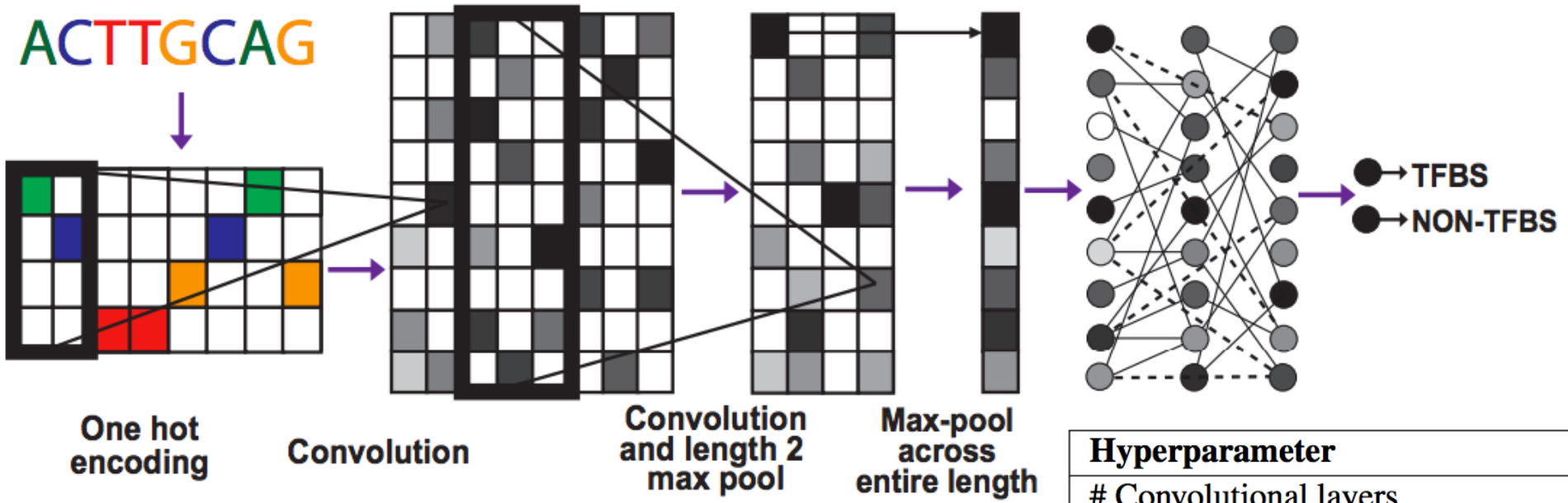
Output: classification

Multiple convolutional layers

Highway multi-layer perceptron (MLP)

Hyperparameters were data set dependent

Methods: Classification



Hyperparameter	Values
# Convolutional layers	{3,4}
# Convolutional hidden units	{128}
Max-pooling at each convolutional layer	{2,1}
# Highway MLP layers	{5,7}
# MLP hidden units	{32}

Methods: Visualization

Aim: deconvolute output to get visual representation of the probable motif

Input: softmax output of trained model

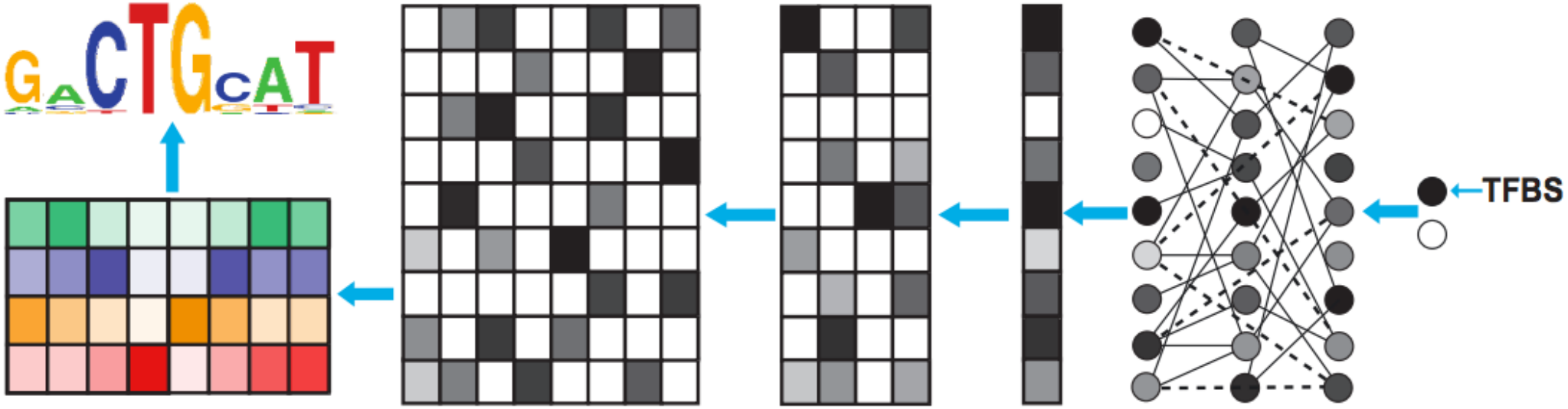
Output: PWM and visual representation of PWM

Find optimal (highest probability) input sequence through backpropagation

Convert optimal input sequence into PWM with Laplace smoothing

Visualize PWM

Methods: Visualization



Data Used

Evaluating classification accuracy

- 108 leukemia cell TF data sets
- Average of 30,819 training sequences each (101 bp per sequence)
- Used by DeepBind (previous state-of-the-art)

Evaluating validity of generated motifs

- JASPAR motifs
- “Gold standard” for positive TF binding sites
- Not guaranteed to be accurate representations of positive TF binding sites
- Can compare 57 of 108 data sets (from classification results) to JASPAR
- Check how identified sequences score comparatively

Results: Classification

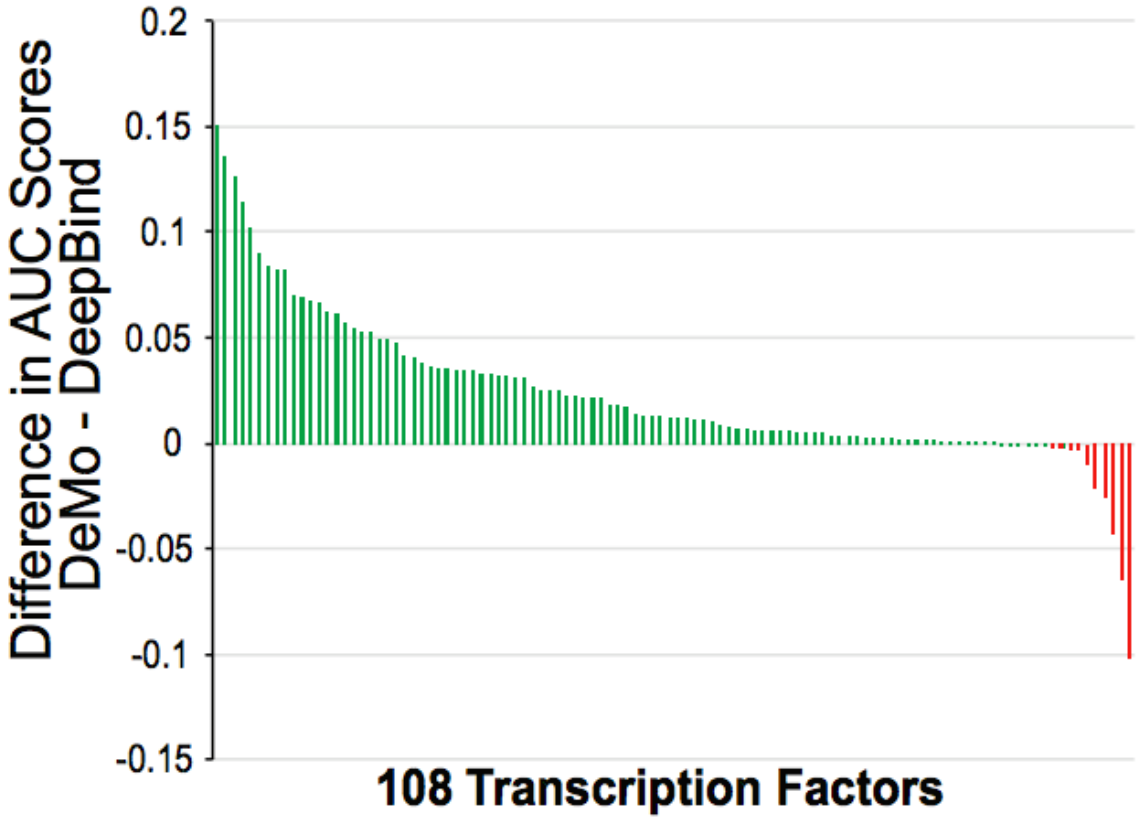
Compared to DeepBind

Using AUC metric

Higher AUC for 92/108 TF data sets

Higher median AUC (0.951 vs. 0.931)

DeMo performs better than DeepBind



Results: Visualization

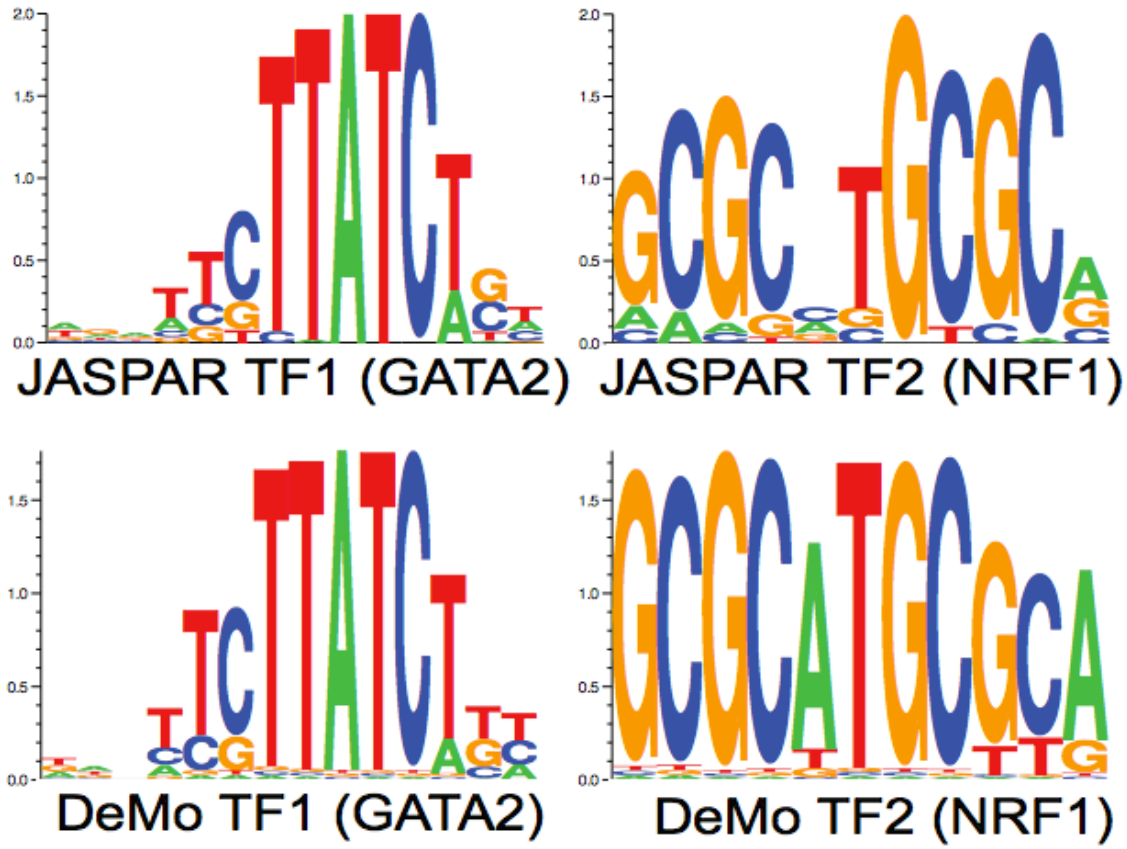
Compared to JASPAR motifs

Method 1

- Use Tomtom to measure similarity between reconstructed motifs and JASPAR motifs
- Found 36/57 significant motifs

Method 2

- Use Average Motif Affinity (AMA) tool to compare how motifs score
- 29/57 motifs outscore JASPAR



Key Claims

DeMo performs better than DeepMind

DeMo can generate accurate motifs for important TF binding sites

- Finding what general positive TFBS classes look like is more important than specific examples

DeMo can be applied to other sequence classification tasks

Analysis

Strength of Claims

Reproducibility

Failings

Other methods/results

Analysis: Strength of Claims

Needs more support and results

Good comparison to DeepMind, but only uses one metric (AUC)

Don’t compare motif generation to previous tools

Analysis: Reproducibility

Possibly…

Vague references to data sources

Don’t specify which hyperparameters for which data

Written in Lua

Trained models are on Github!

Analysis: Other Methods

Comparisons to DeepBind but only in one dataset with one metric

Motif generation and visualization

Mention other tools that try motif generation

- Subset frequency counts (Stormo, 2000)
- Generative frequency based searching (Setty & Leslie, 2015)
- SVMs (Ghandi et al., 2014)
- Blind-deconvolutional approach (Gomes et al., 2014)

No comparison to the other tools

Analysis: Failings

Not enough results and evaluation

Neural networks have been tailored to fit to mentioned data sets

Spelling and grammar mistakes

Would benefit from more detail

Impact

Application to other sequence classification tasks that need visual representation

Can help locate potential positive TF binding sites as a preliminary screening, but not reliable enough to be the only tool

Good for motif visualization

Summary

Key Claim

- Deep Motif (DeMo) achieves state-of-the-art accuracy in motif classification and provides visual representations of positive binding sites

Importance

- DeMo can be used to more quickly screen for TF binding sites

Issues

- The paper does not share accuracy metrics (besides AUC) and may not be applicable for finding new binding sites
- Vague information and perhaps not reproducible

FIN - Thank You