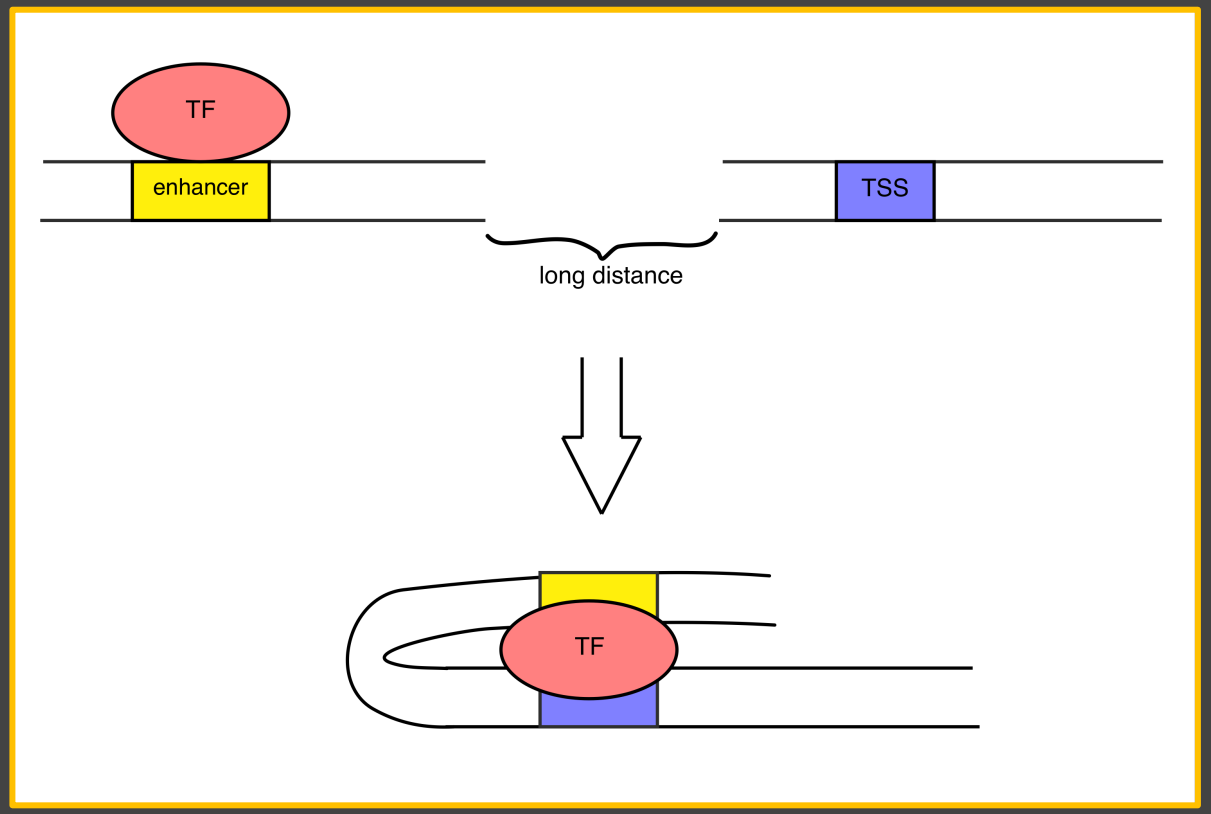
Convolution Neural Networks in enhancer prediction

Jurijs Nazarovs

UW-Madison

December 13, 2017

Biological perspective

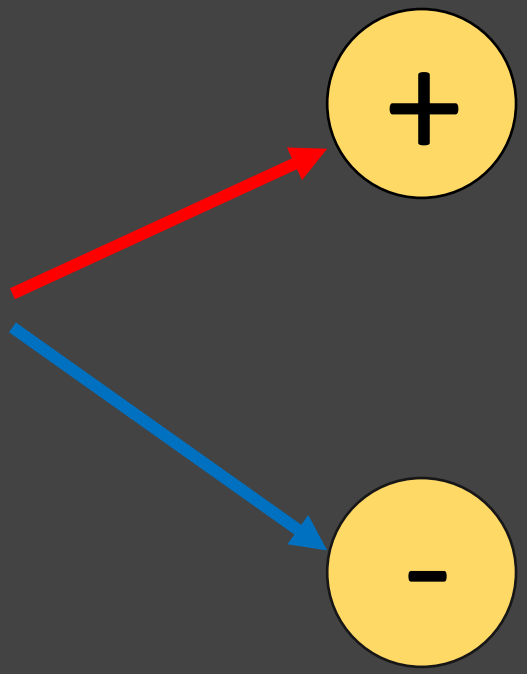


1. Transcription factor binds to an enhancer

2. DNA strand loops and brings transcription factor to the transcriptional start site of specific gene to activate or suppress

Problem statement

Chromosome 1: 1234 - 1298



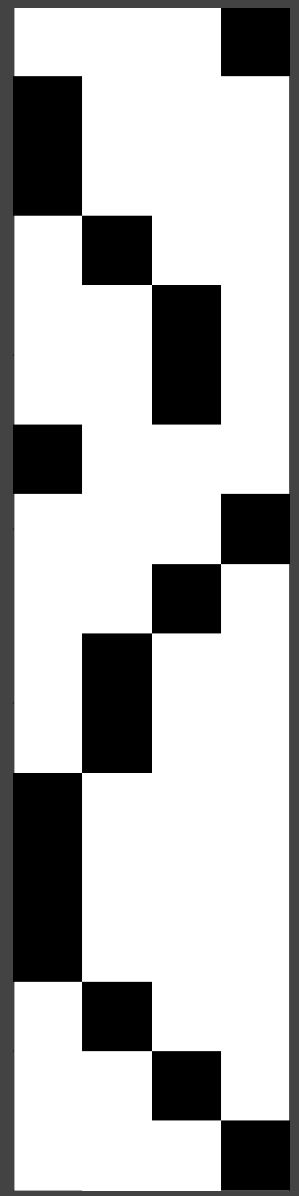
Supervised binary classification problem

Data description and preparation

Original data – positive sample:
chromosome, start, end, …

Nucleotide sequence + truncation:
AAACTCCG….

One – hot representation:
A = 1 0 0 0
C = 0 1 0 0
T = 0 0 1 0
G = 0 0 0 1



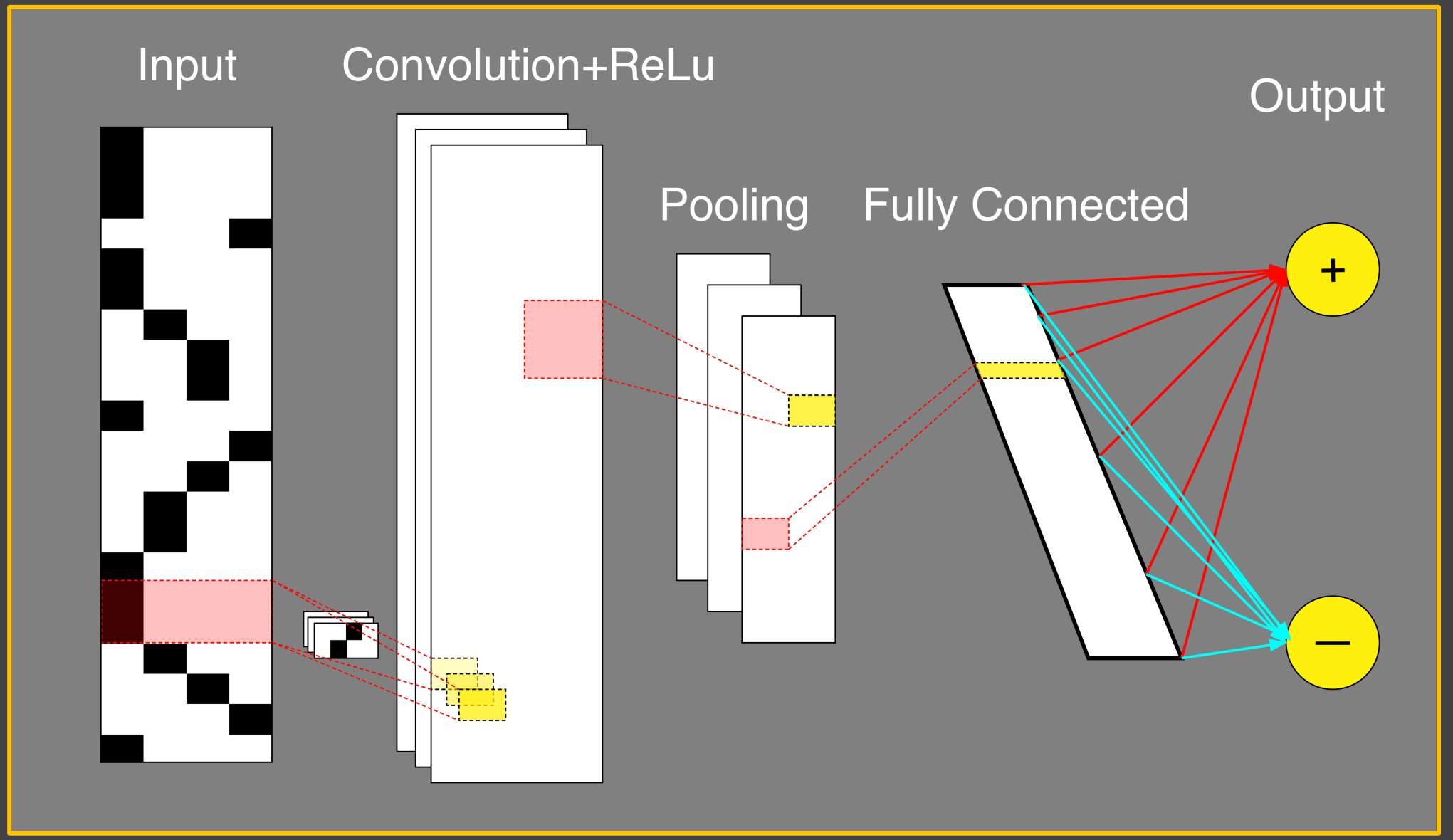
Negative samples

Random complement of positive samples

- Take complement of input region
- Sample random interval of adequate length

Nucleotides shuffle in positive samples

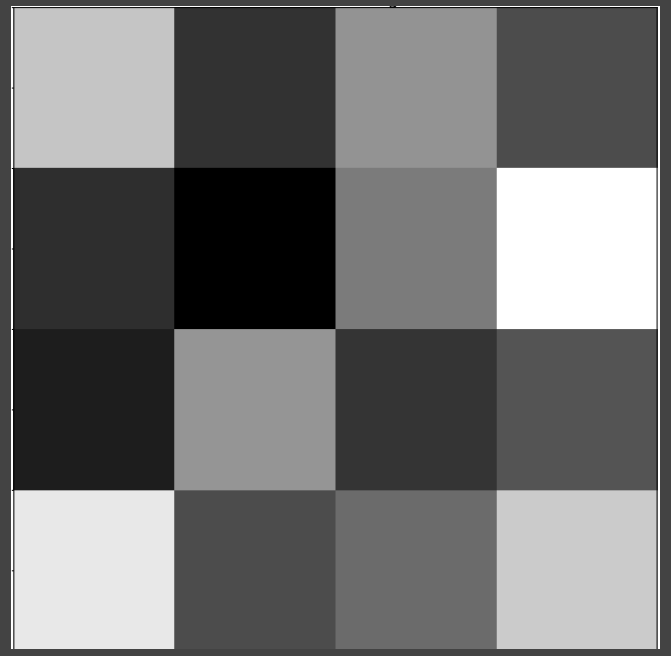
Convolution Network Description

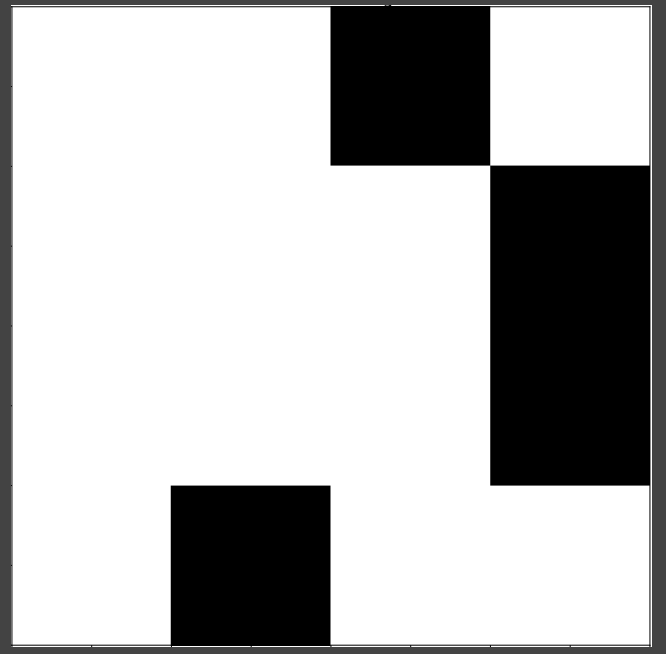


Convolution features

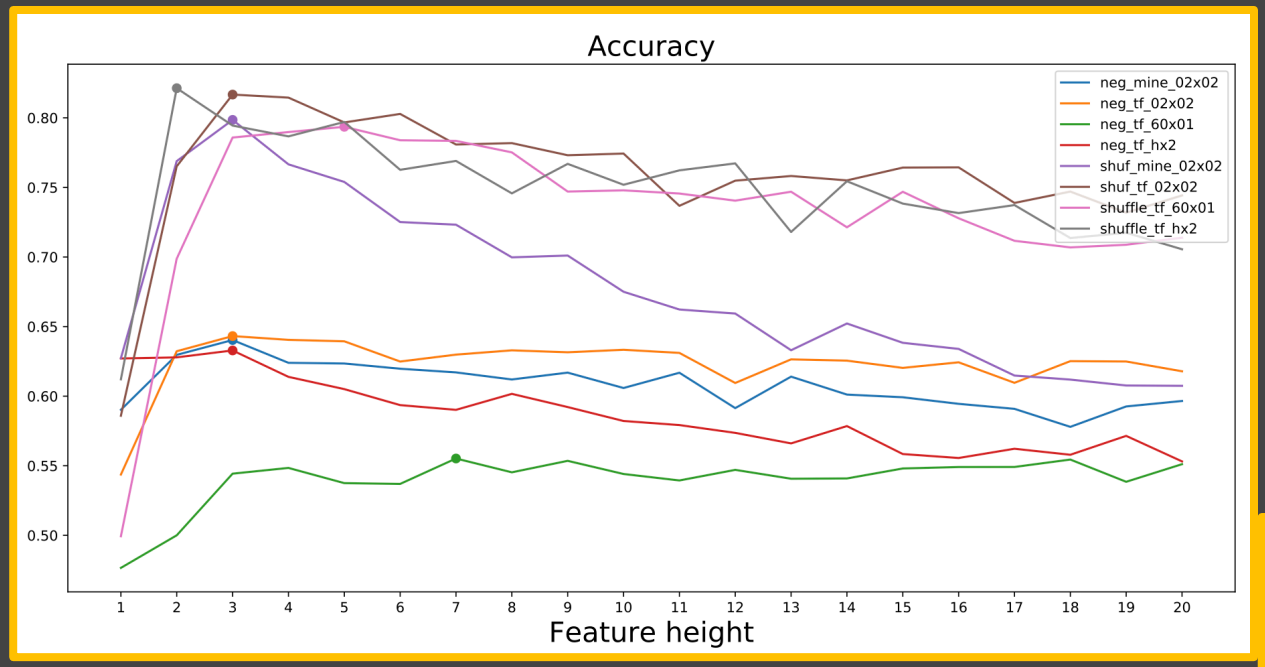
Tensorflow adaptive features with continuous values

Generated discrete values corresponding to one-hot representation of nucleotides



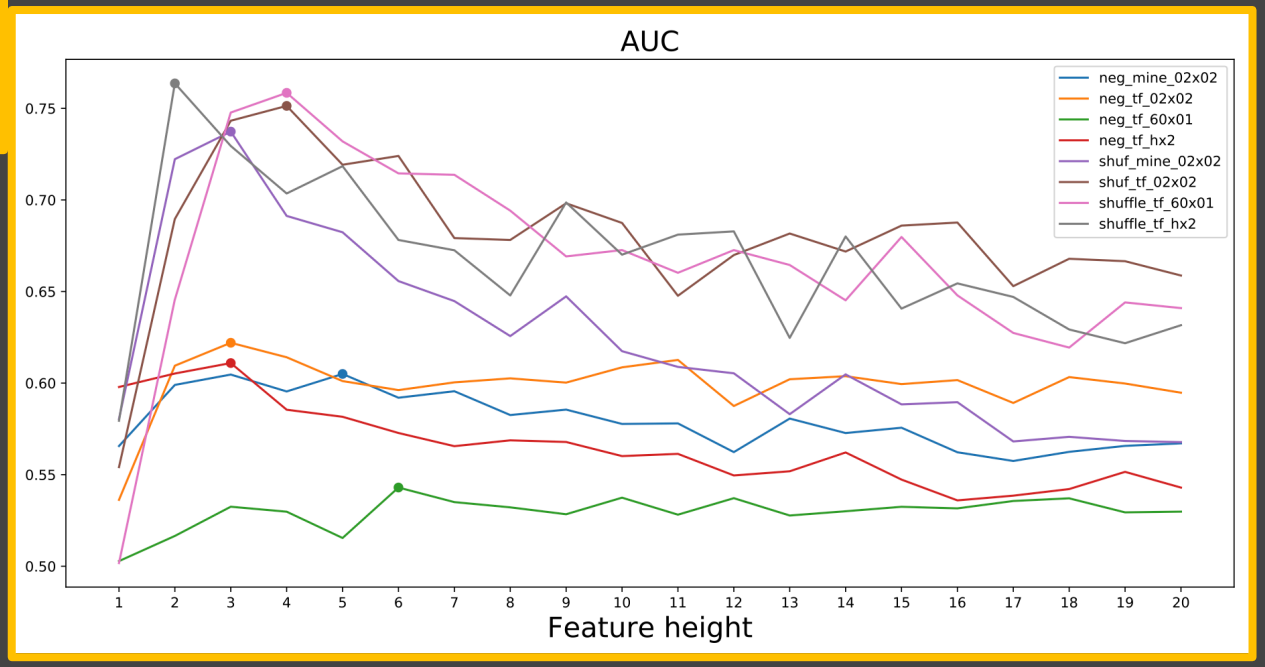


Different CNN comparison



Accuracy best: shuffled nucleotides as negative sample with feature height = 2, 3, 4 and pooling matrix = 2x2

AUC best is the same



Final CNN structure and results

Structure:

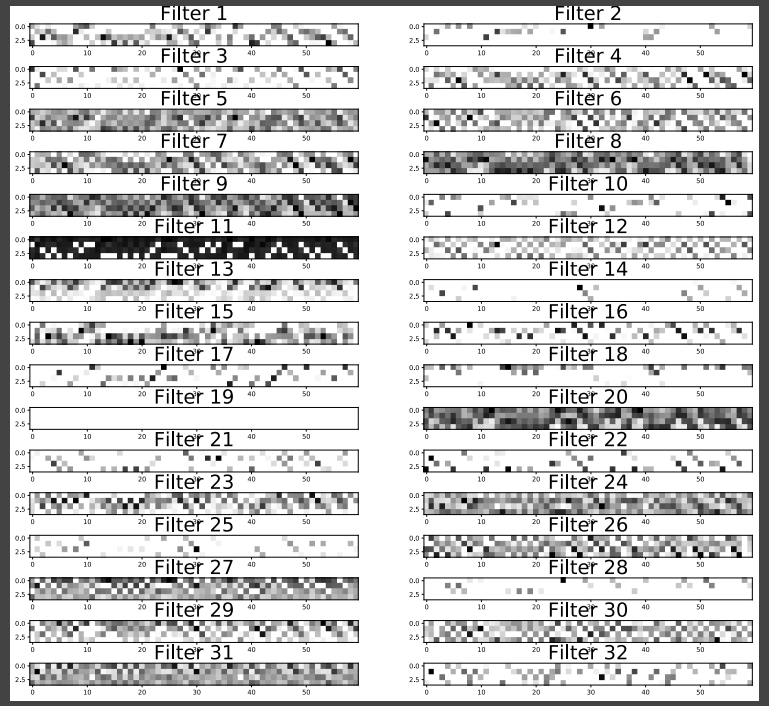
- 1 Layer CNN
- Input image size = 60x4
- Tensorflow adaptive features
- Feature size = 4x4
- 32 Filters in convolution layer
- Pooling size = 2x2
- Fully connected: 200
- Number of epochs: 200

Accuracy: 83%

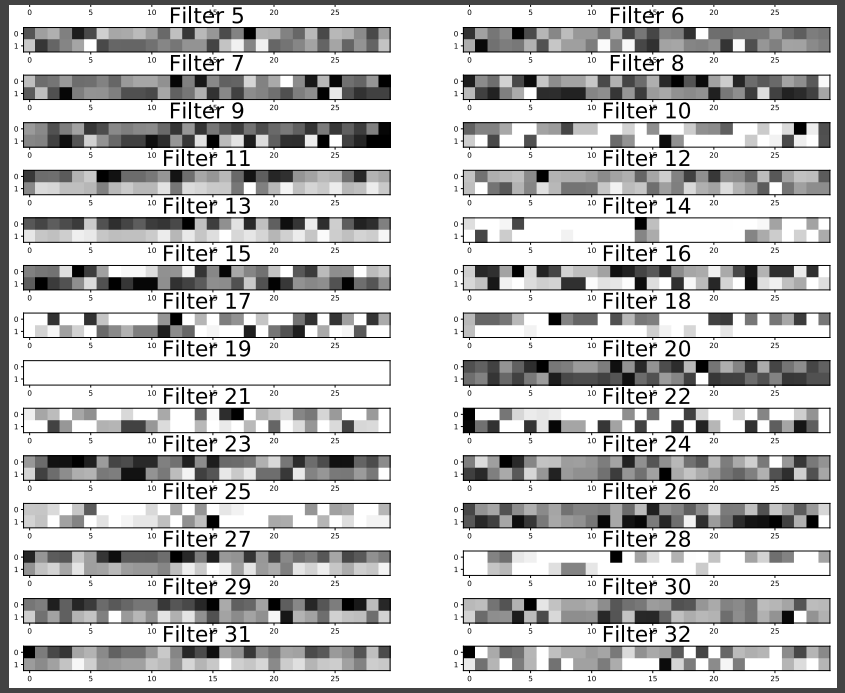
AUC: 85%

Final CNN output of layers

Convolution



Maxpool



Fully connected

Comparison

DeepEnchancer - AUC is about 91%

- More Layers performed
- Negative sample is constructed in a more complicated considering other biological information