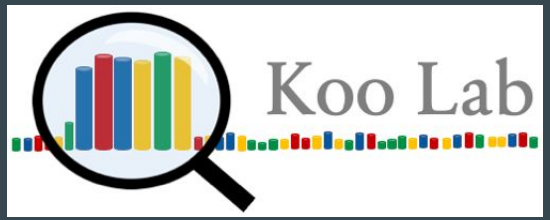
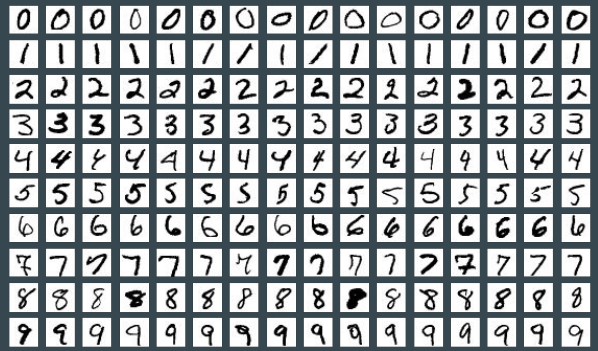
Knowledge Distillation in Genomic Deep Learning

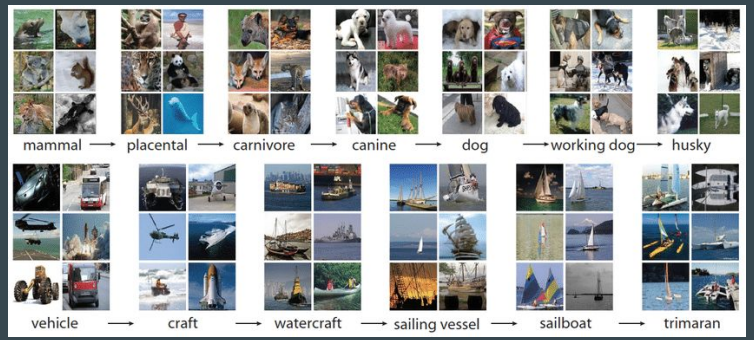
...

Roshan Kenia
Stony Brook University



Deep Neural Networks





Model Compression

TABLE I
SUMMARIZATION OF DIFFERENT APPROACHES FOR MODEL COMPRESSION AND ACCELERATION.

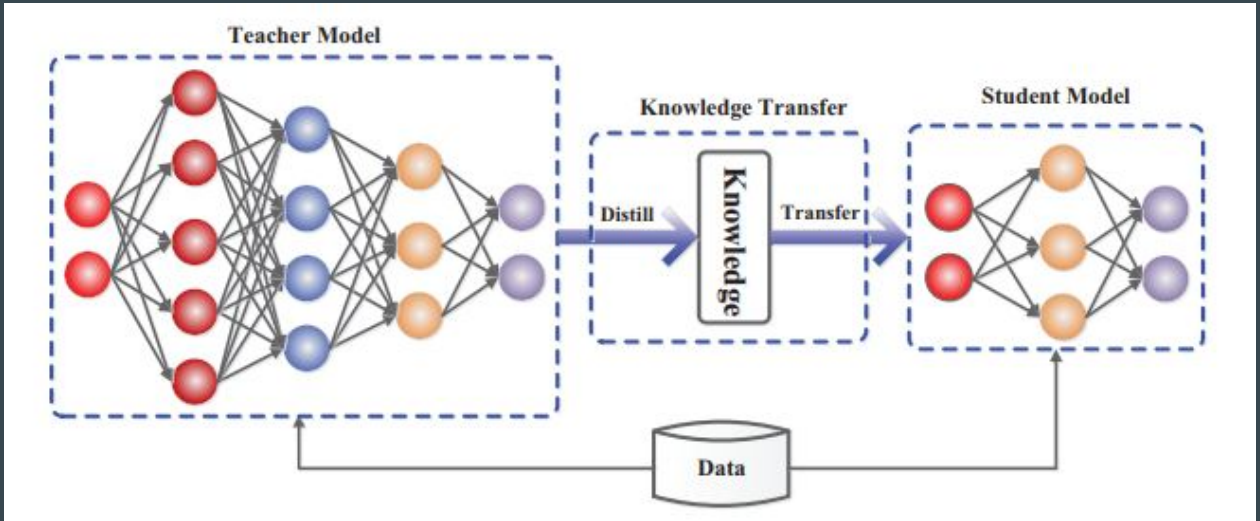
Category Name	Description	Applications	More details
Parameter pruning and quantization	Reducing redundant parameters which are not sensitive to the performance	Convolutional layer and fully connected layer	Robust to various settings, can achieve good performance, can support both train from scratch and pre-trained model
Low-rank factorization	Using matrix/tensor decomposition to estimate the informative parameters	Convolutional layer and fully connected layer	Standardized pipeline, easily to be implemented, can support both train from scratch and pre-trained model
Transferred/compact convolutional filters	Designing special structural convolutional filters to save parameters	Convolutional layer only	Algorithms are dependent on applications, usually achieve good performance, only support train from scratch
Knowledge distillation	Training a compact neural network with distilled knowledge of a large model	Convolutional layer and fully connected layer	Model performances are sensitive to applications and network structure only support train from scratch

[1]

Sources:

1. Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2017). A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282.

Knowledge Distillation



[1]

Sources:

1. Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2020). Knowledge distillation: A survey. arXiv preprint arXiv:2006.05525.

Distillation

$$q_i = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

[1]

Sources:

1. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Distillation

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

[1]

Sources:

1. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Distillation

- A typical loss function would look something like this:

$$\text{loss} = \text{alpha} * \text{student\_loss} + (1 - \text{alpha}) * \text{distillation\_loss}$$

- The alpha value decides how much weight you would like to put on the student loss and on the distillation loss.

MNIST data

Network	Teacher: Neural net with two hidden layers of 1200 rectified linear hidden units	Scratch Student: Smaller net with two hidden layers of 800 rectified linear hidden units	Distilled Student: Smaller net with two hidden layers of 800 rectified linear hidden units
Test Errors	67	146	74

[1]

Sources:

1. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

MNIST data

Class:	Teacher	Distilled Student	Scratch Student
Accuracy:	0.9875	0.9765	0.9799

- We used an alpha value of 0.1 and a temperature of 10.

Extending to Biological Data

- With the MNIST example we only showed that we can use this altered loss function in training and also obtain similar accuracy.
- We hypothesize that the use of a distilled student network trained on a deeper teacher network will allow for smoother features in our function, meaning better interpretability.

Saliency Maps

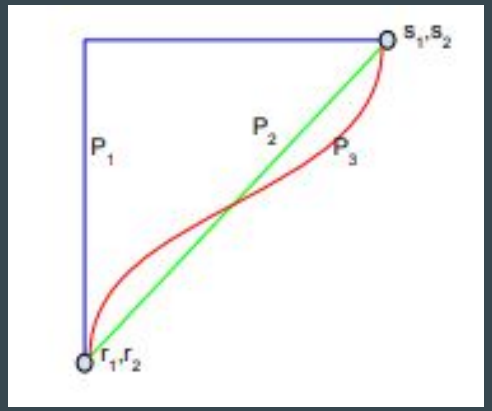
[1]

$$\frac{\Delta \, Predicted \; Class}{\Delta \, Pixel \; Value}$$

Sources:

1. Mundhenk, T. N., Chen, B. Y., & Friedland, G. (2019). Efficient Saliency Maps for Explainable AI. arXiv preprint arXiv:1911.11293.

Integrated Gradients

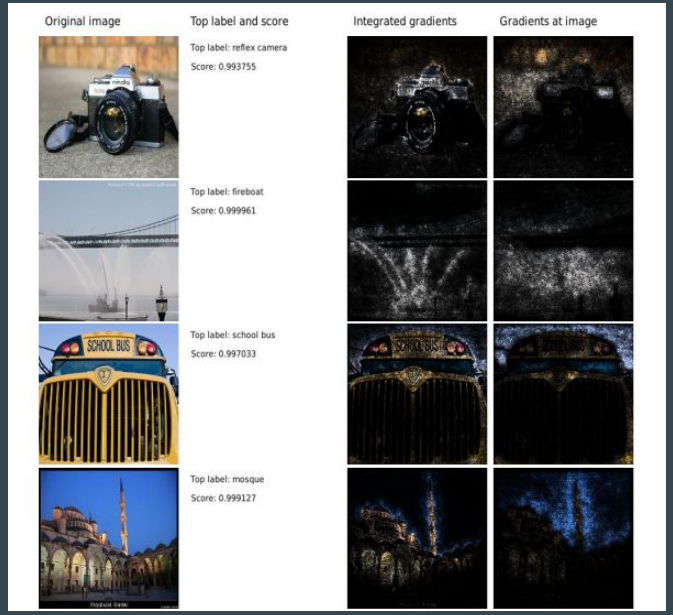


[1]

Sources:

1. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR.

Integrated Gradients



[1]

Sources:

1. Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International Conference on Machine Learning (pp. 3319-3328). PMLR.

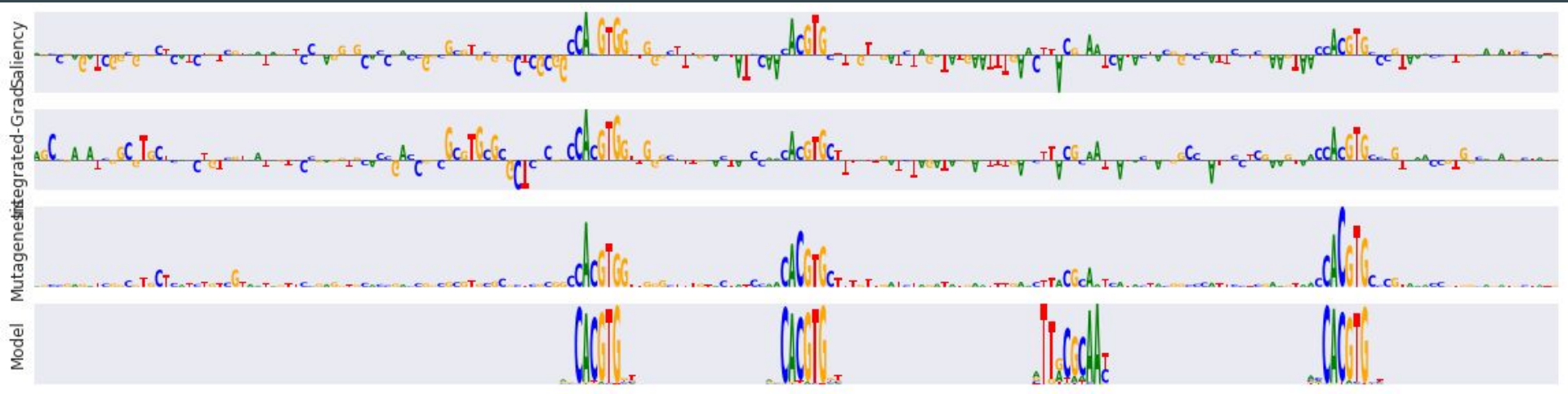
In Silico Mutagenesis

- In silico mutagenesis involves testing every mutation of a nucleotide in order to predict the functional activity of arbitrary sequences.
- Computing the predicted accessibility of all possible mutations to a sequence offers a powerful approach to understand and apply the regulatory grammar that it has learned. [1]

Sources:

1. Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome research, 26(7), 990-999.

Attribution Methods



- The nucleotides that appear the largest have the greatest importance in deciding a class score by the model.

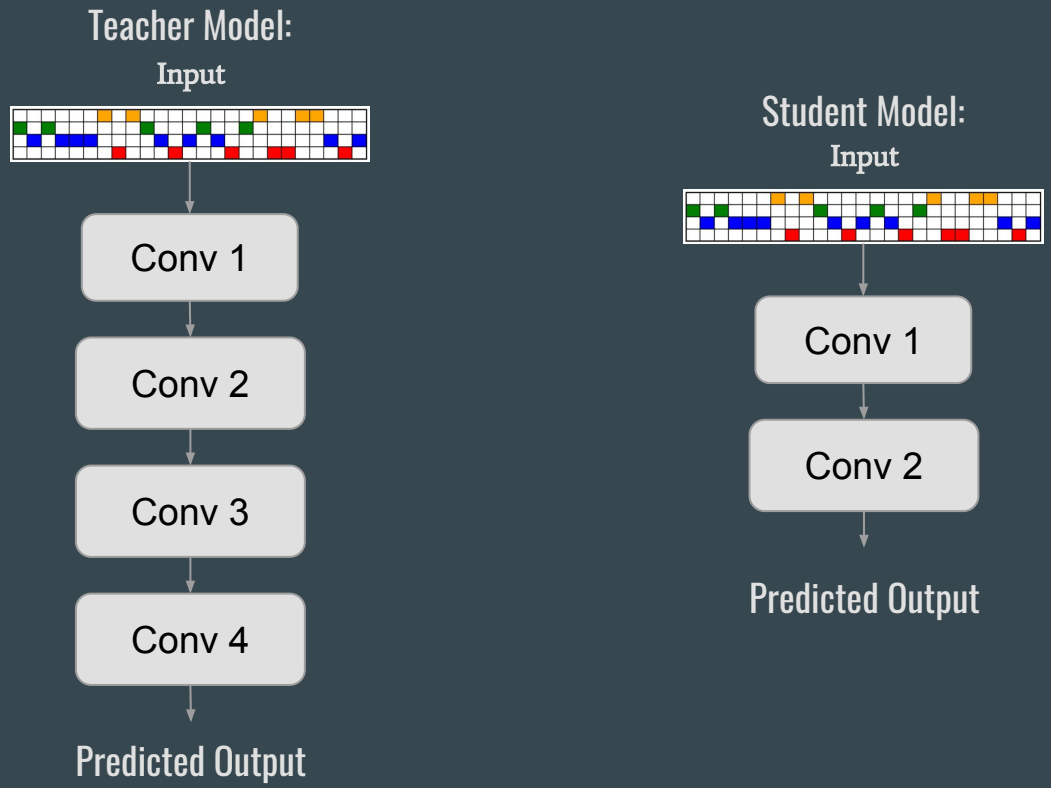
Experiment on Synthetic Data

- The dataset consists of synthetic DNA sequences implanted with known transcription factor binding motifs. [1]
- Each sequence is 200 nucleotides long and is embedded with 1 to 5 TF motifs.

Sources:

1. Koo, P. K., & Ploenzke, M. (2019). Improving convolutional network interpretability with exponential activations. bioRxiv, 650804.

Visualization of Teacher and Student Architecture



Early Results

Class:	Teacher	Distilled Student	Scratch Student
AUPR:	0.974377	0.955244	0.943862
AUROC:	0.975047	0.957035	0.944482

- Conducted with a temperature of 15 and an alpha of 0.5

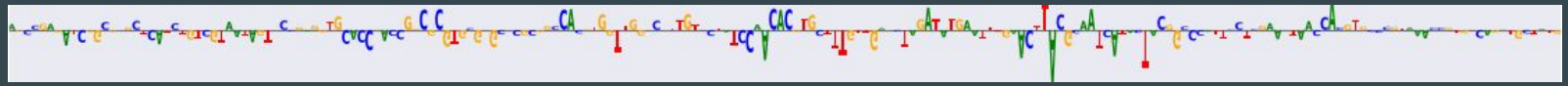
Saliency Map

Ground Truth:

Teacher:

Distilled Student:

Scratch Student:



Integrated Gradient

Ground Truth:

Teacher:

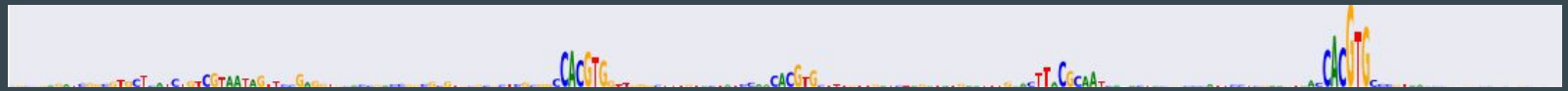
Distilled Student:

Scratch Student:

In Silico Mutagenesis

Ground Truth:

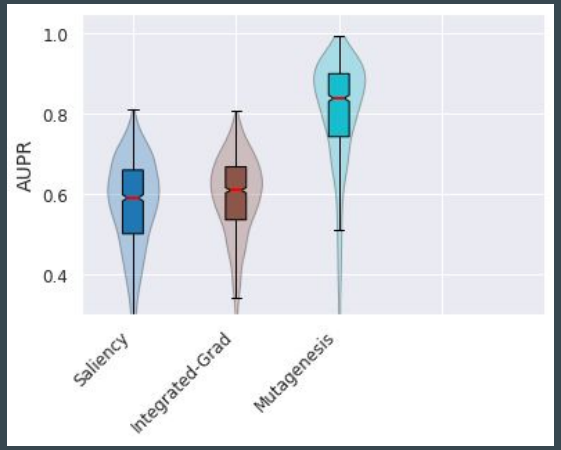
Teacher:



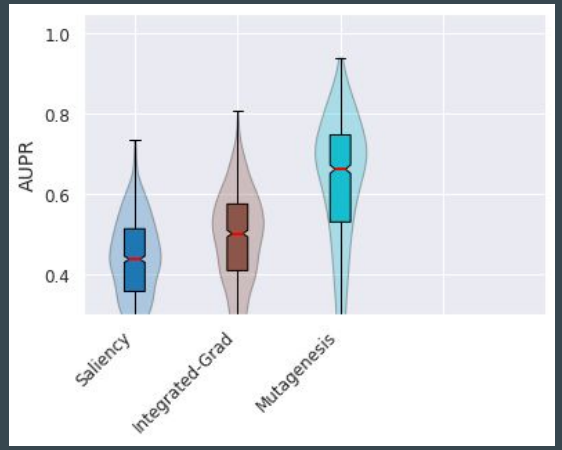
Distilled Student:

Scratch Student:

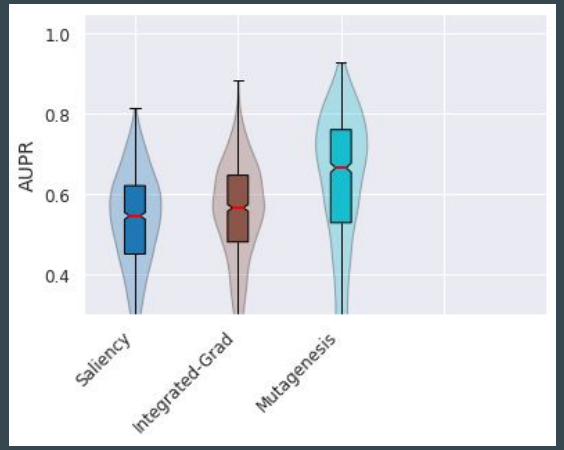
Early Results



Teacher

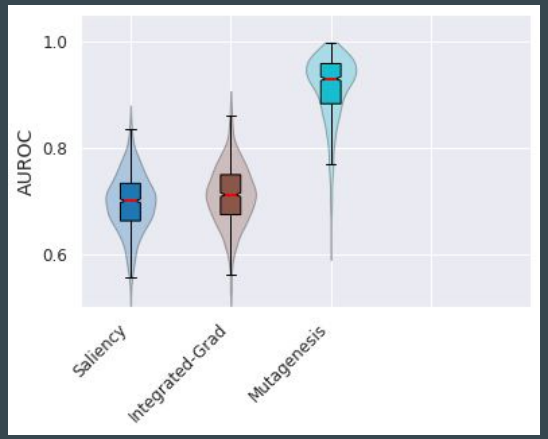


Scratch Student

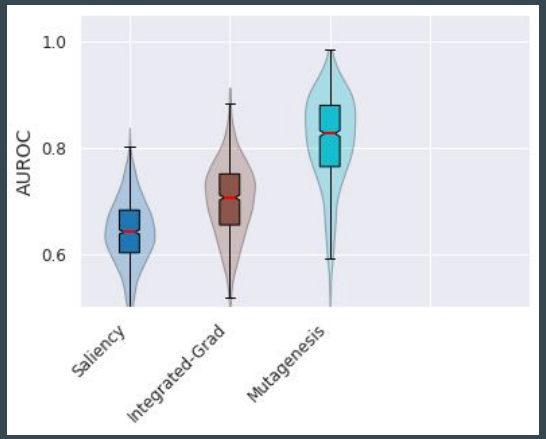


Distilled Student

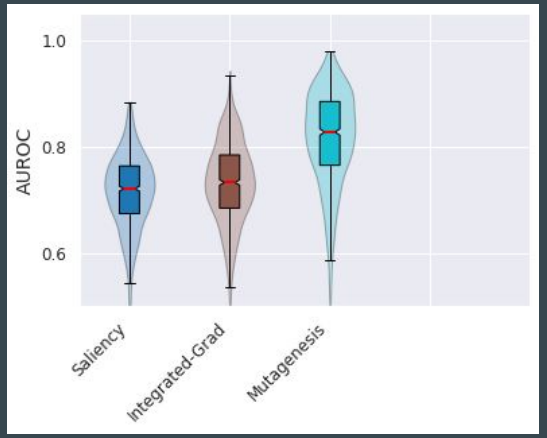
Early Results



Teacher



Scratch Student



Distilled Student

Possible Fixes

- The reason the student model may not be as interpretable as the teacher model for right now could possibly be due to the teacher not being as complex of a model as it should be.
- With a more complex model, the distilled student may be able to learn more interpretable results as the teacher builds a deeper model.

Future Tests

- Once we see that there is merit in using student-teacher networks in improving interpretability for genomic sequences, we can further test this method on other models.
- One such model that may be valuable to look into would be a multi-head attention model which uses self-attention to learn how motifs may interact with each other.

Questions to be answered

- Would there be a better way to perform these experiments in order to maximize the potential “dark knowledge” that can be learned by the student network? [1]

Sources:

1. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.