National Center of Excellence for Computational Drug Abuse Research

Learning Deconfounded Representations through Neural Networks, with Applications in Genetic Data

Haohan Wang

School of Computer Science

Carnegie Mellon University

haohanw@cs.cmu.edu

@HaohanWang

Carnegie Mellon University
School of Computer Science

Outline

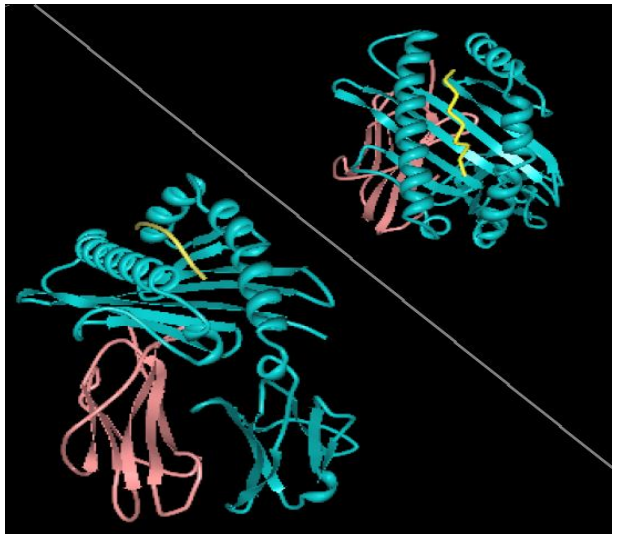
- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

Outline

- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

Confounding Factors in GWAS

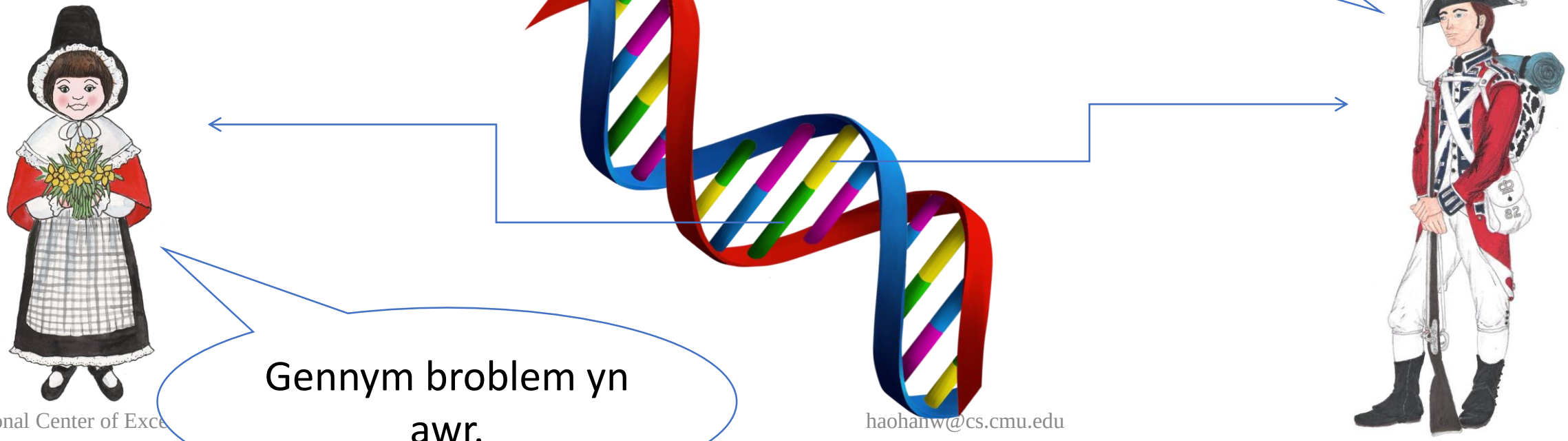
- Chopstick usage behavior prediction from genetics



(Vilhjálmsson and Nordborg, 2013)

Confounding Factors in GWAS

- (Weale *et al* 2002)
- In a large data collection drawn



haohanw@cs.cmu.edu

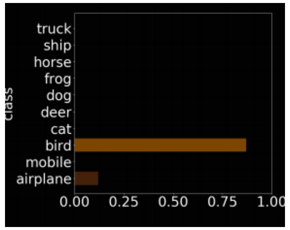
Deep Learning Era and Prediction Tasks

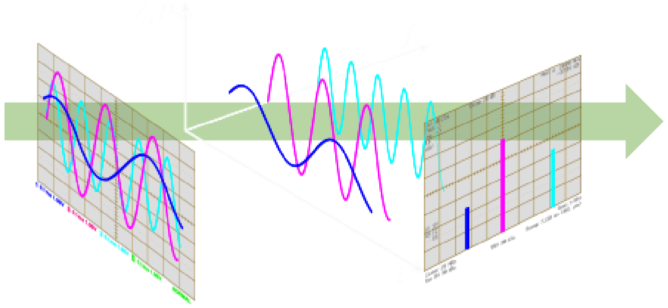
- “Universal Approximation” can consider anything as a predictive signal
- Confounding variables can degrade generalization performance of radiological deep learning models
  - (Zech et al, 2018)
- Removing confounding factors associated weights in deep neural networks improves the prediction accuracy for healthcare applications
  - (Wang et al, 2019)

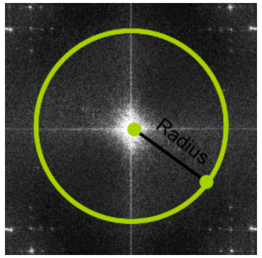
High Frequency Component Helps Explain the Generalization of Convolutional Neural Networks

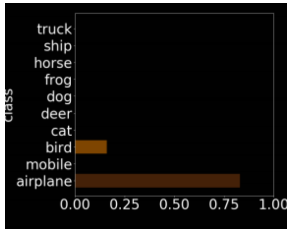
- **Haohan Wang**, Xindi Wu, Zeyi Huang, and Eric P. Xing
- https://arxiv.org/abs/1905.13545

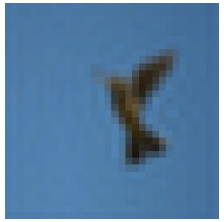
CNN’s tendency in high-frequency data





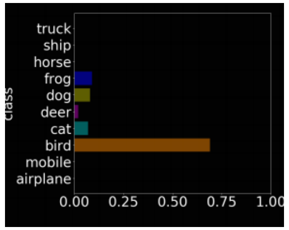


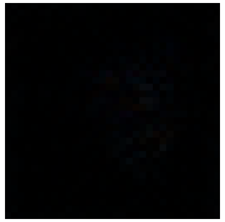




Reconstruction

Low frequency components

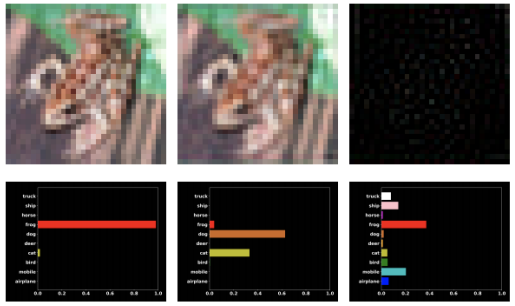




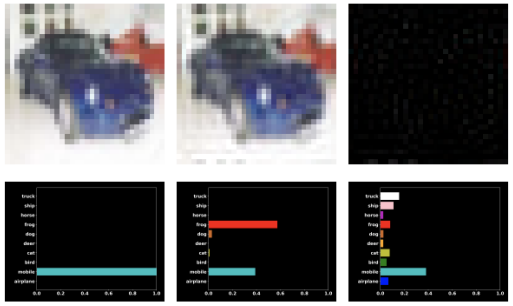
Reconstruction

High frequency components

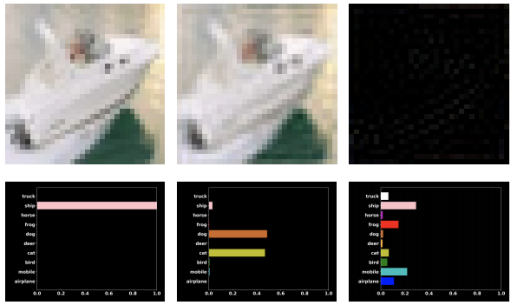
CNN's tendency in high-frequency data



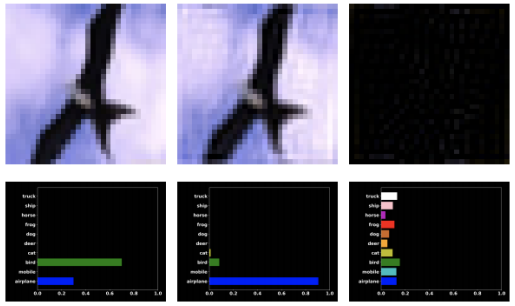
(a) A sample of frog



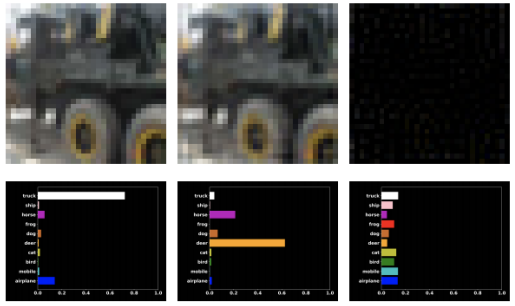
(b) A sample of mobile



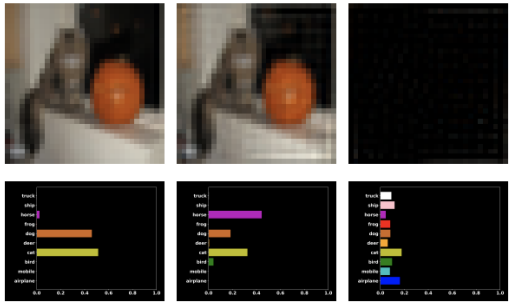
(c) A sample of ship



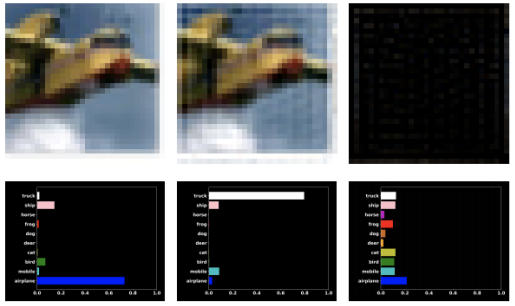
(d) A sample of bird



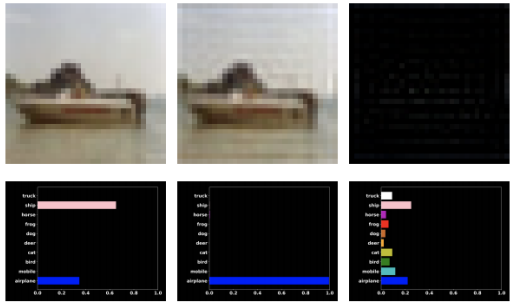
(e) A sample of truck



(f) A sample of cat



(g) A sample of airplane



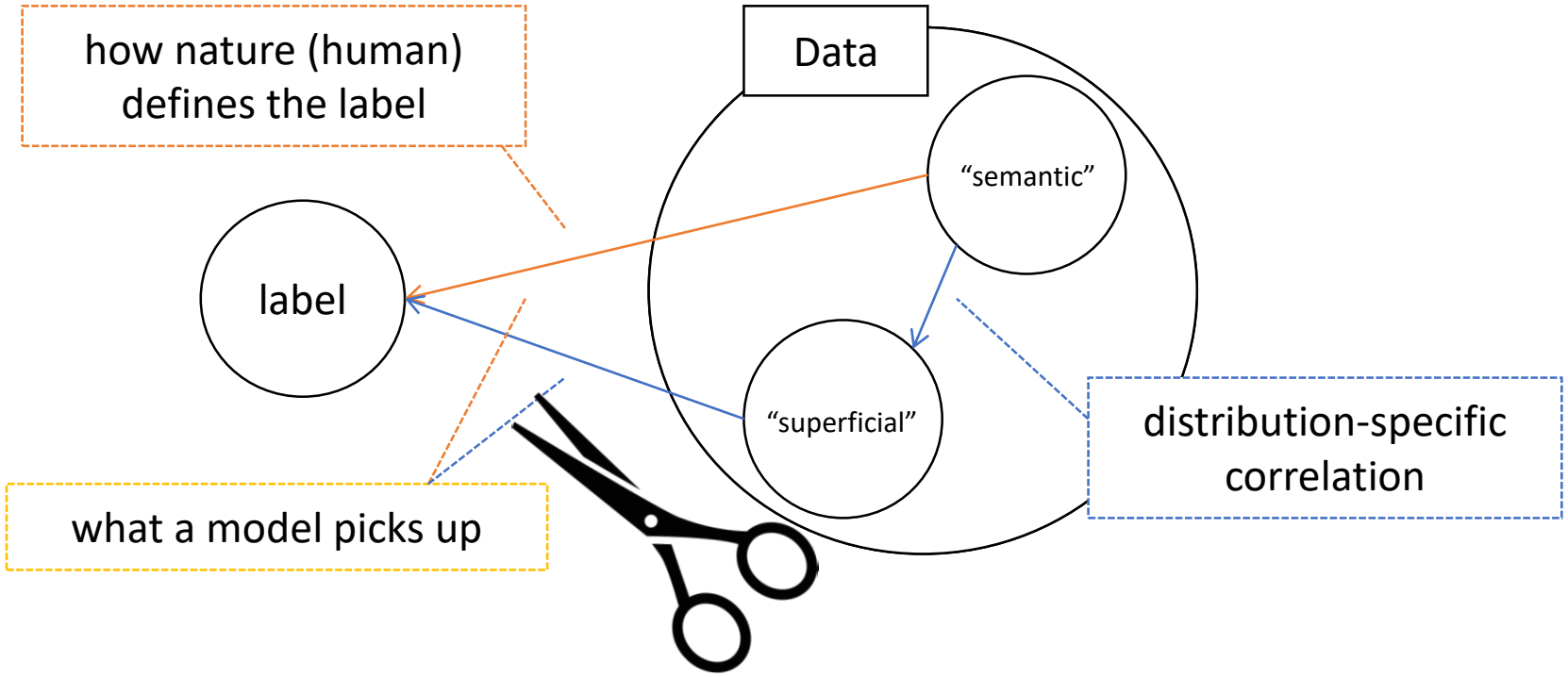
(h) A sample of ship

High-frequency Component Helps Explain the Generalization of Convolutional Neural Networks

- Take home messages:
  - CNN sees data differently form human
  - Accuracy should not be the only thing to aim at
    - Trade-off between accuracy and robustness
  - New explanations to previously elusive facts:
    - Rethinking data before rethinking generalization
    - The effectiveness of Batch Normalization
    - The underlying cause of adversarial vulnerability
  - https://arxiv.org/pdf/1905.13545.pdf

Confounding Factors in Data

- Where the problem comes from



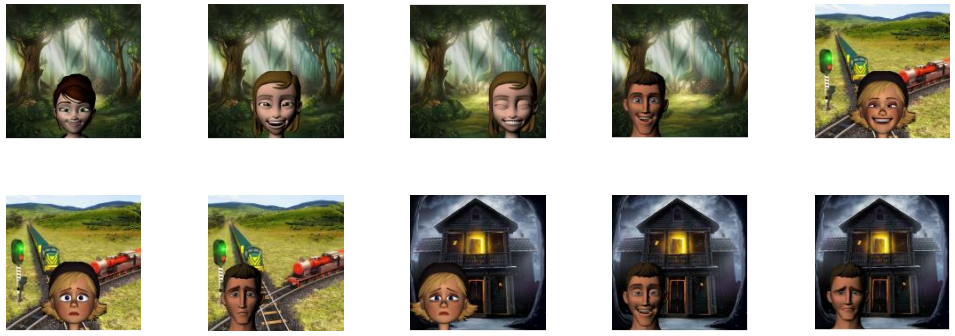
Outline

- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

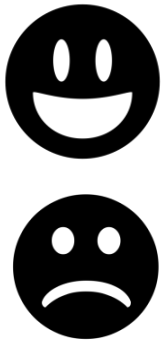
Problem Setup: a Sentiment Classification

Training Data

data



labels



Testing Data

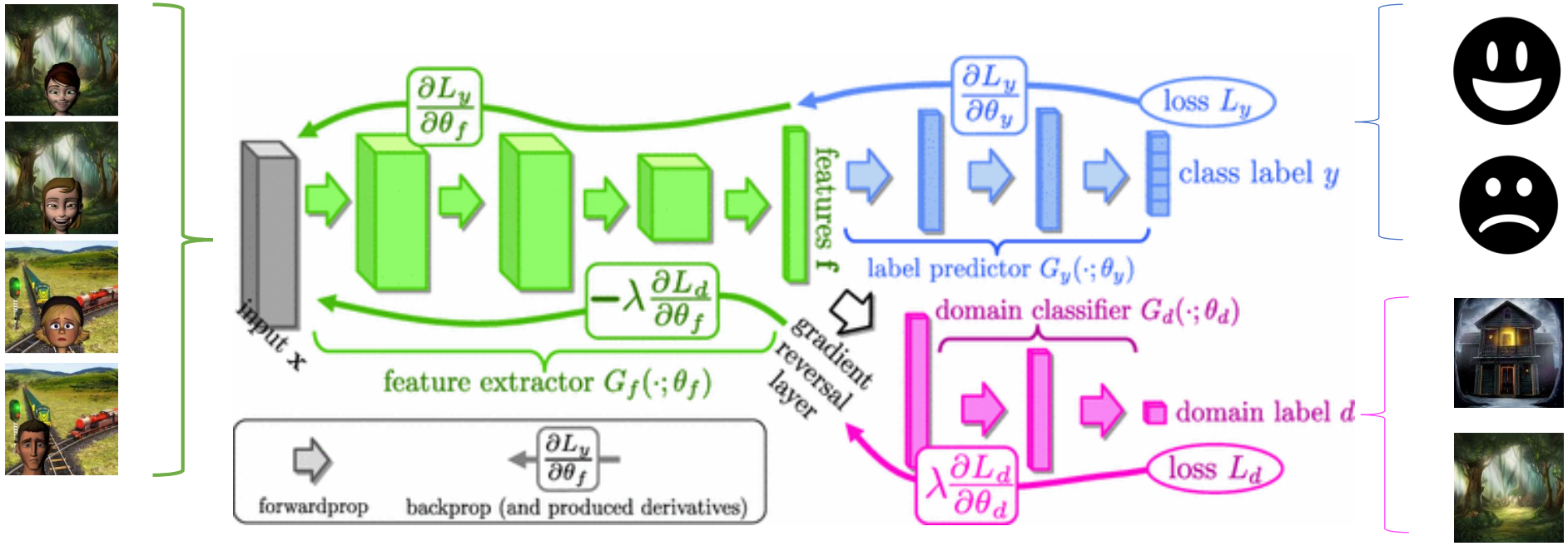
With Labels of Confounding Factors

Solution: forcing invariance towards the labels of confounding factors

- Solutions:
  - Domain Adversarial Neural Network
    - https://arxiv.org/abs/1505.07818
  - Select-Additive Learning
    - https://arxiv.org/abs/1609.05244
  - Confounder Filtering Method
    - https://www.ncbi.nlm.nih.gov/pubmed/30864310

Domain Adversarial Neural Network

- Forcing invariance through negative gradient



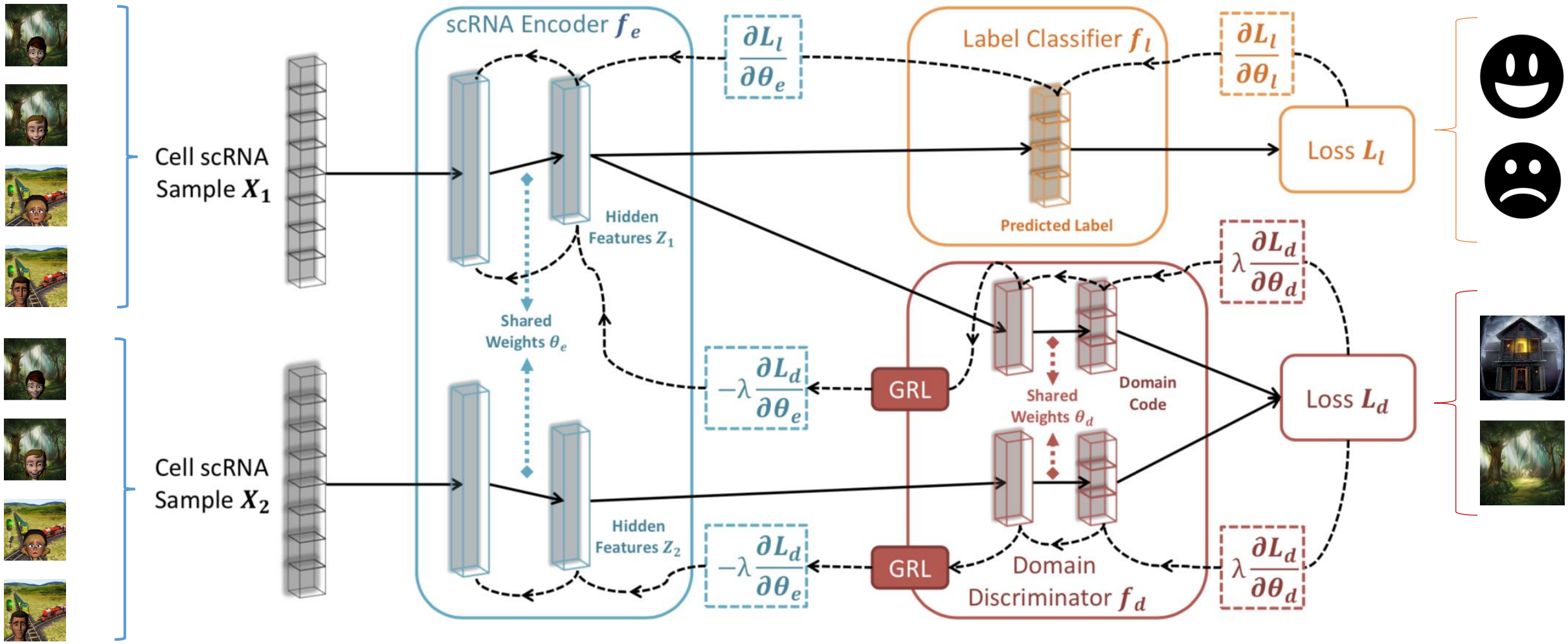
Supervised Adversarial Alignment of Single-Cell RNA-seq Data

- Songwei Ge, **Haohan Wang**, Amir Alavi, Eric P. Xing, and Ziv Bar-Joseph
- RECOMB 2020
- https://www.biorxiv.org/content/10.1101/2020.01.06.896621v1

Background

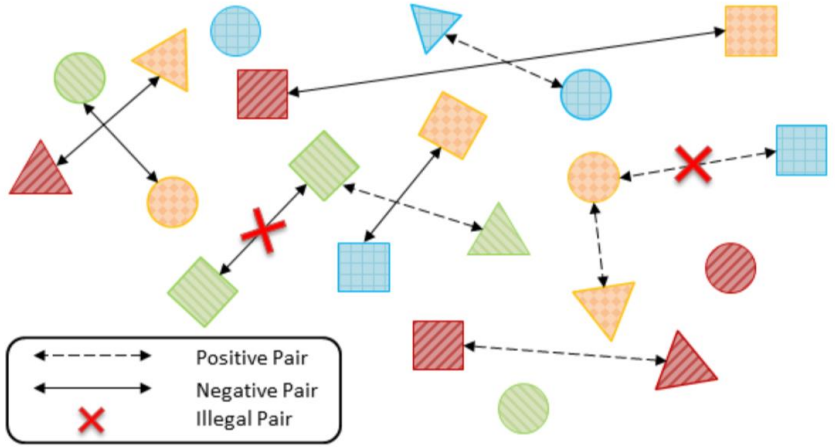
- Challenges of scRNA analysis:
  - How to integrate and compare results from multiple scRNA-seq studies
    - Batch effects as confounding factors
- Available Data:
  - scRNA data, cell types, batch ids
  - We build a model to classify cell types, invariant to batch information
    - So that the representation is more about the cell's nature, less about the batch effects

Model


Cell scRNA Sample X_1
Cell scRNA Sample X_2
scRNA Encoder f_e
Hidden Features Z_1
Hidden Features Z_2
Shared Weights θ_e
Label Classifier f_l
Predicted Label
Loss L_l
$\frac{\partial L_l}{\partial \theta_e}$
$\frac{\partial L_l}{\partial \theta_l}$
$\lambda \frac{\partial L_d}{\partial \theta_d}$
$-\lambda \frac{\partial L_d}{\partial \theta_e}$
GRL
GRL
Shared Weights θ_d
Domain Code
Domain Discriminator f_d
Loss L_d

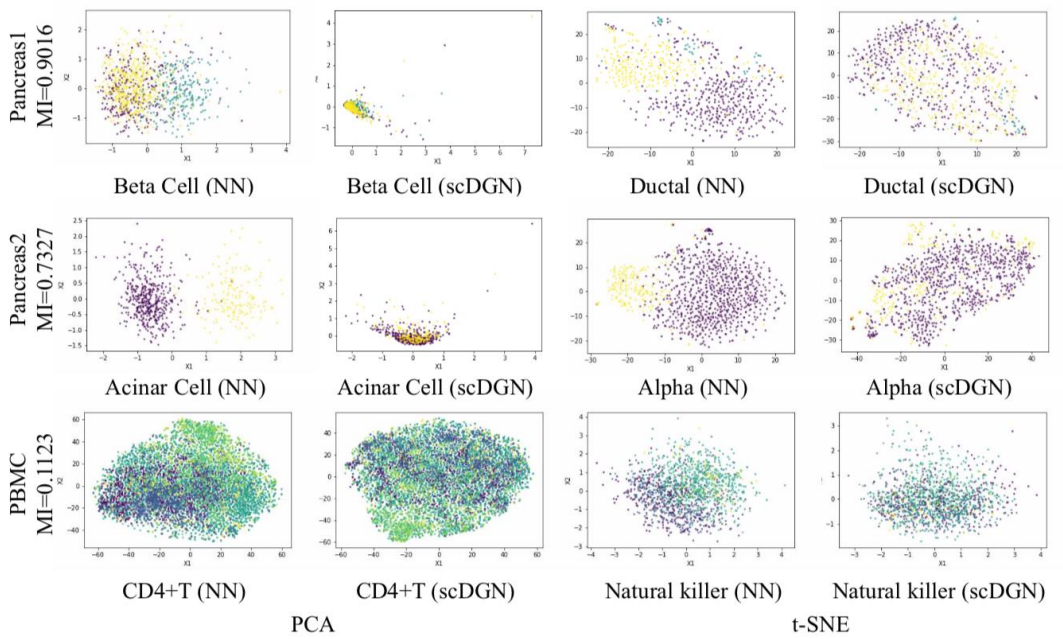
Algorithm: Conditional Domain Generalization

- Only two types of sample pairs are considered
  - Samples from the same domain, with different cell types
  - Samples from different domains, with the same cell type



Results

- Numerical results: cell type classification accuracy
- Visualization



- Key Gene Analysis

Outline

- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

With Knowledge of Confounding Factors

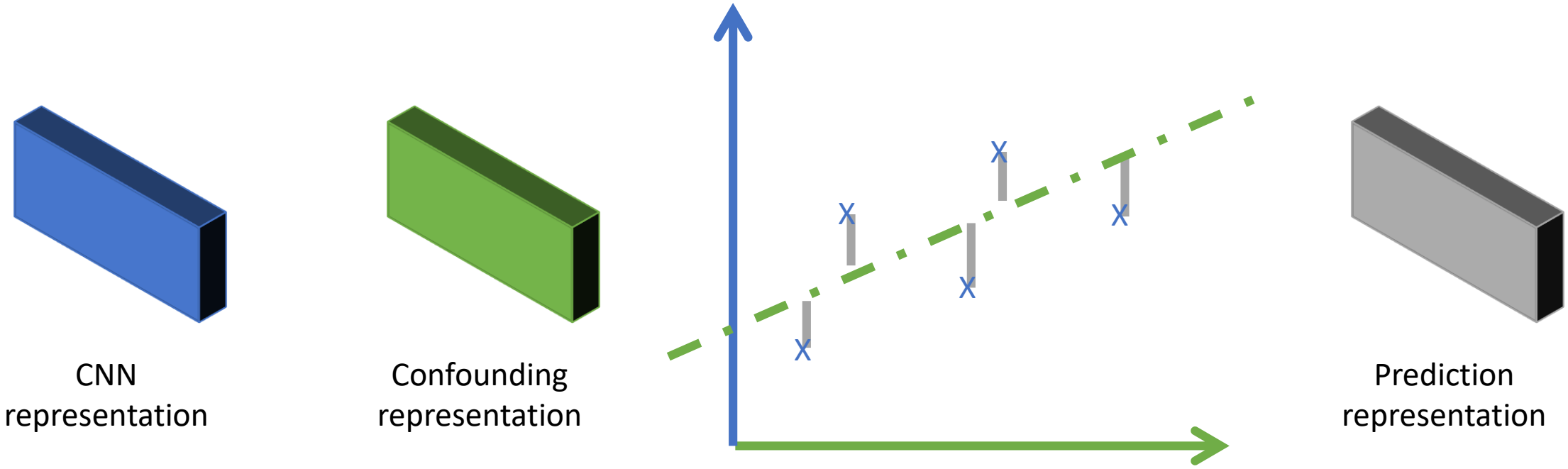
Extra information: we also have labels of the background

Solution: a two-step strategy

- First model the confounding factors **only**
  - Neural-GLCM
    - texture of image
    - https://arxiv.org/abs/1903.06256
  - Patchwise Adversarial Regularization
    - local predictive pattern of images
    - https://arxiv.org/abs/1905.13549
- Then throw it away
  - Through regression
  - Through adversarial regularization

Regression Technique (HEX)

- Prediction with the regression residual



Conveniently done with one line of code

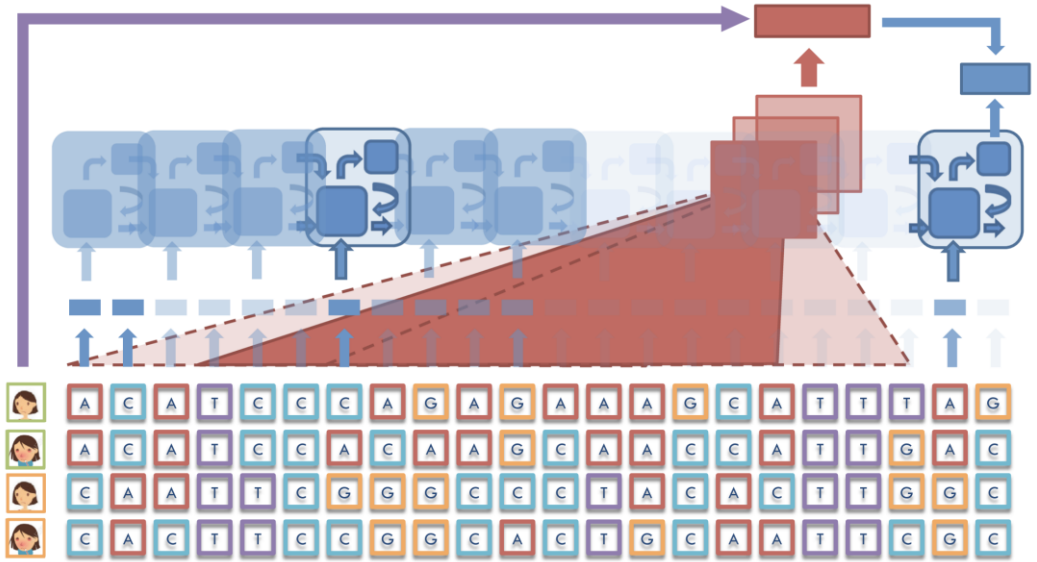
Deep Mixed Model for Marginal Epistasis Detection and Population Stratification Correction in Genome-wide Association Studies

- **Haohan Wang,** Tianwei Yue, Jingkang Yang, Wei Wu, and Eric P. Xing
- BMC Bioinformatics 2019
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3300-9

Background

- Deep learning offers an effective way in modeling epistasis
  - But GWAS naturally has the confounding issues of population stratification, family structure, and cryptic relatedness
  - How do we deal with it within the neural network regime

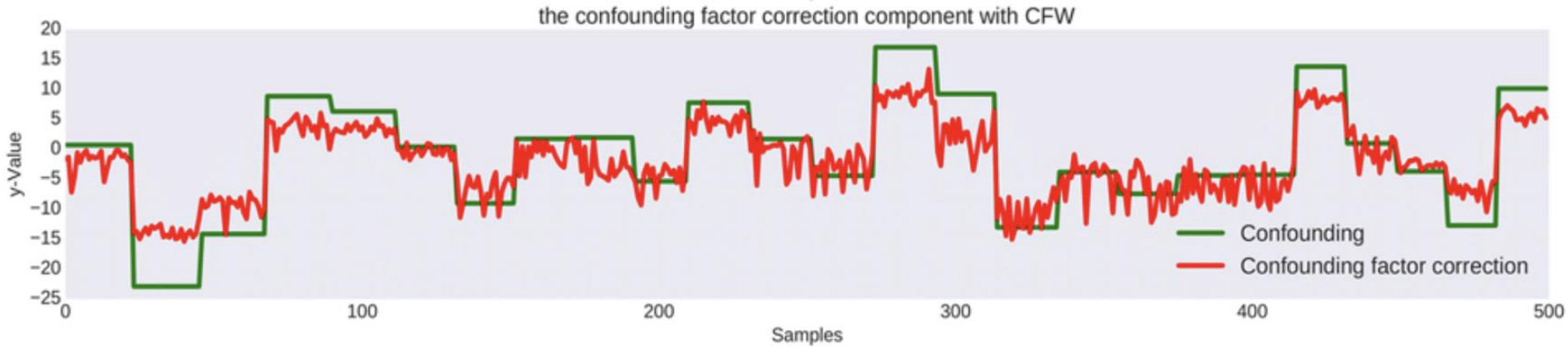
Design Rationale



- Traits are associated with SNPs
  - Only a couple of SNPs matter
  - Locus-specific effect sizes
    - **Lasso (across the whole genome) + LSTM**
- Population structure is associated with SNPs
  - A large number SNPs work together
  - Locus-independent effect sizes
    - **Convolution with a large kernel**

Results

- Simulation performance superior to baselines
- Investigation at internal working mechanism



- Real data study for Alzheimer's disease

Poly(A)-DG: a Neural-network-based Domain Generalization Method to Identify Cross-species Ploy(A) Signal without Prior Knowledge

- Yumin Zheng, **Haohan Wang,** Yang Zhang, Eric P. Xing, and Min Xu
- In preparation

Background

- Poly(A) Signal
  - defining feature of eukaryotic protein-coding genes
  - an essential process during mRNA maturation
    - promote downstream transcriptional termination
    - gene expression can be drastically affected
  - a central motif and other flanking, auxiliary elements
- Poly(A) Signal Identification
  - An identification of MOTIFs
  - Can we identify poly(A) signals across species?
    - To reveal the connections between the underlying mechanisms of different mammals

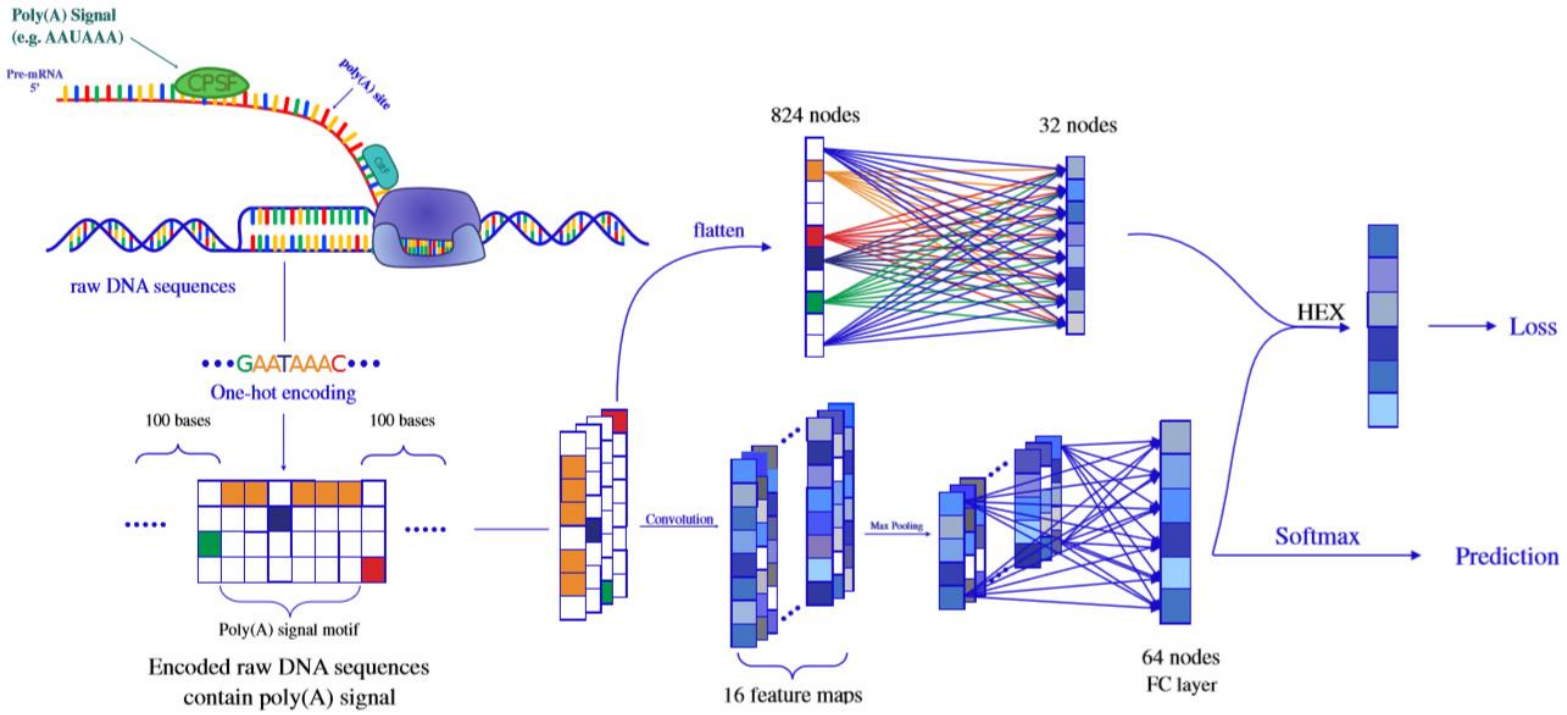
Model Confounding Factors

- A function learns specie distributional information without learning motifs
  - A simple MLP over data
    - But with shuffled sequences as Input

	Original Sequence	Shuffled Sequence
Species Signals	0.379±0.006	0.353±0.002
Poly(A) Signal	0.753±0.053	0.534±0.001

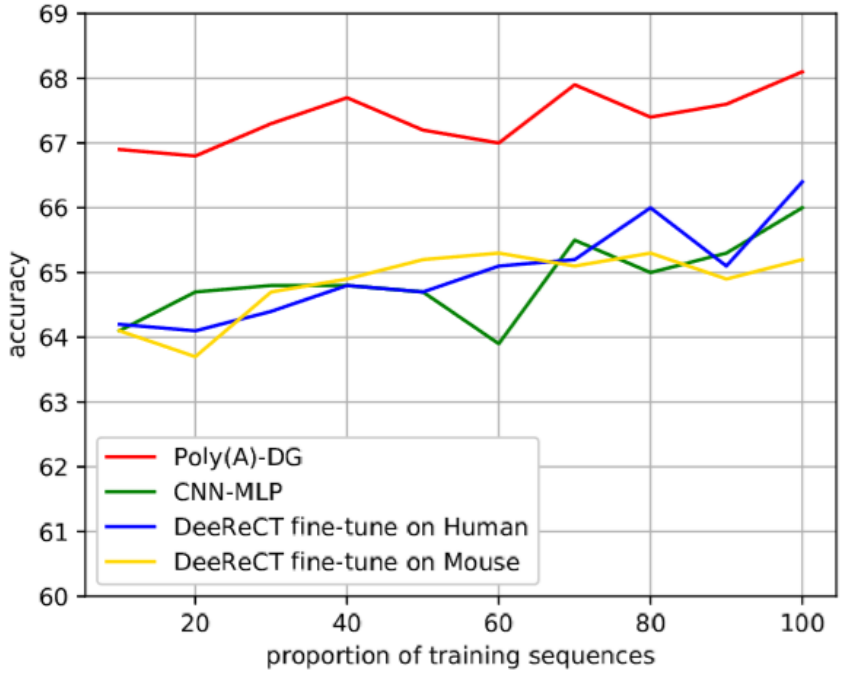
Model

- Model Architecture



Results

- Across-species Prediction
  - Train the model over two species
  - Predict over a 3rd specie



- Other comparisons
  - With limited data
  - With imbalanced Data

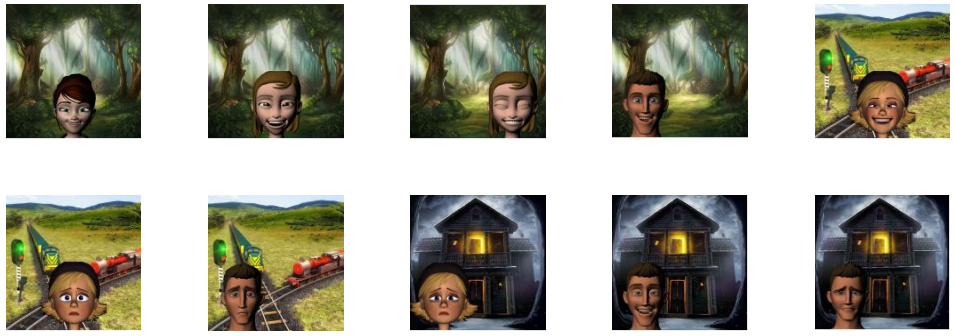
Outline

- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

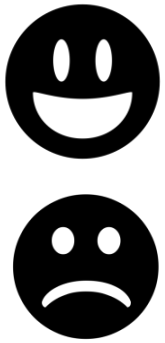
What if we have nothing else

Training Data

data

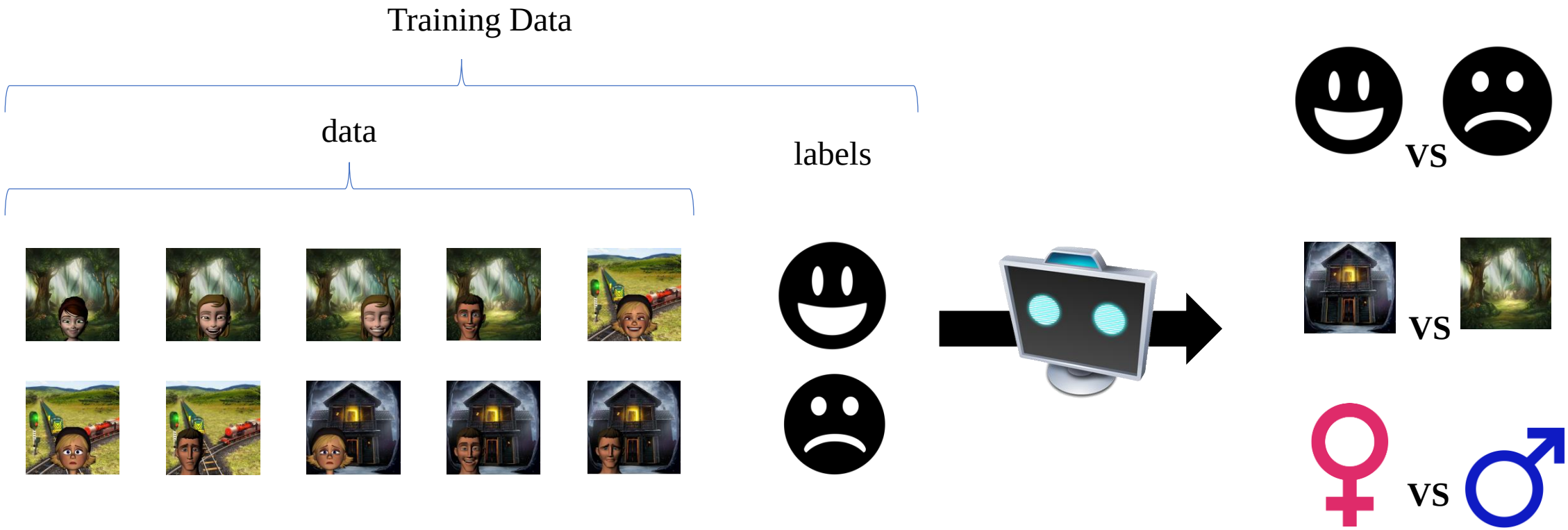


labels

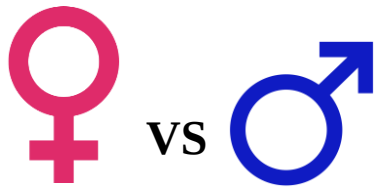


Testing Data

Let’s just do it



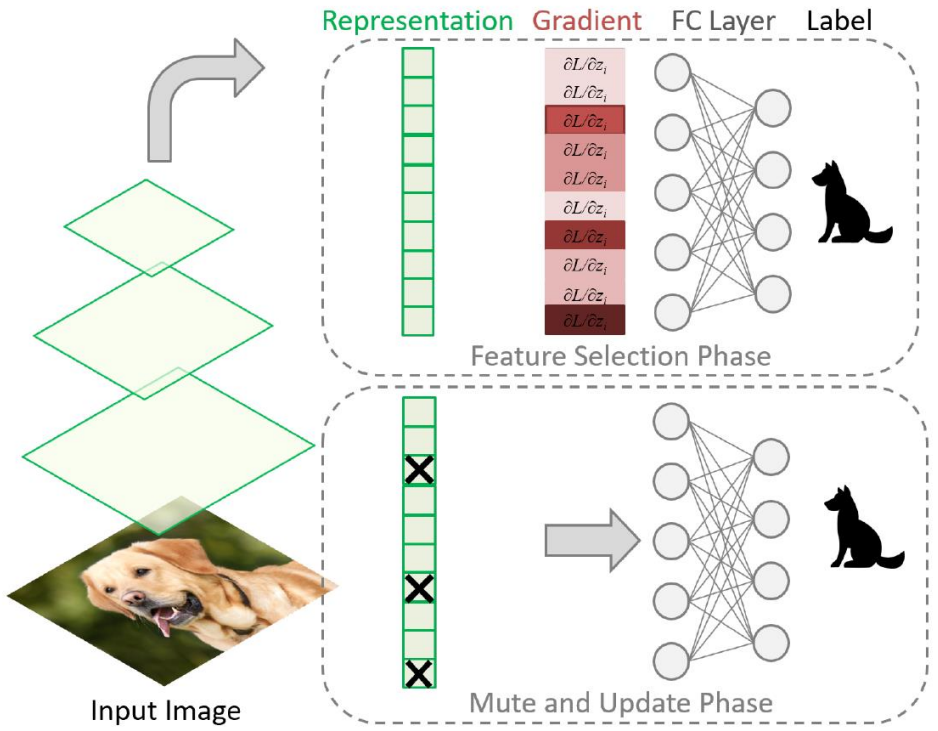
Multiple Classifiers



- At least, one of them is correct
- The correct one has the least training error
- So, if we force everyone to be the same, and if we force everyone to have the smallest training error possible…

Self-Challenging

- We force the model to challenge itself
  - Whatever features are most helpful
    - Don’t use them



Results

- Results over Standard ImageNet

ImageNet	backbone	Top-1 Acc ↑	Top-5 Acc ↑	#Param. ↓
Baseline	ResNet50	76.13	92.86	25.6M
RSC(ours)	ResNet50	77.18	93.53	25.6M
Baseline	ResNet101	77.37	93.55	44.5M
RSC(ours)	ResNet101	78.23	94.16	44.5M
Baseline	ResNet152	78.31	94.05	60.2M
RSC(ours)	ResNet152	78.89	94.43	60.2M

Outline

- Background: confounding factors in the data
- Solutions:
  - With labels of confounding factors
  - With knowledge of confounding factors
  - At the least informed situation

Thanks

- Questions?