Robust and Trustworthy AI

Gap between innovation and deployment

- Stems from accidents in safety critical domains
- 2 Key notions
  - Bias
  - Unmitigated uncertainty

Bias

- Can be quantified when talking about ML
- Most intuitive bias comes from data
  - Selection and sampling bias
- Can be propogated towards models
  - Lack of uncertainty and benchmark metrics
- Also appears in deployment
  - Distribution shifts and feedback loops
- Evaluation and interpretation
  - Bulk metrics don't account for subsystems

Industry Example: Facial Detection

- Selection bias: Data does not reflect the real world
- Evaluation bias: Models were not evaluated by subgroups

Class imbalance

What happens when some classes are more represented than others

- In the most example, everything is one class

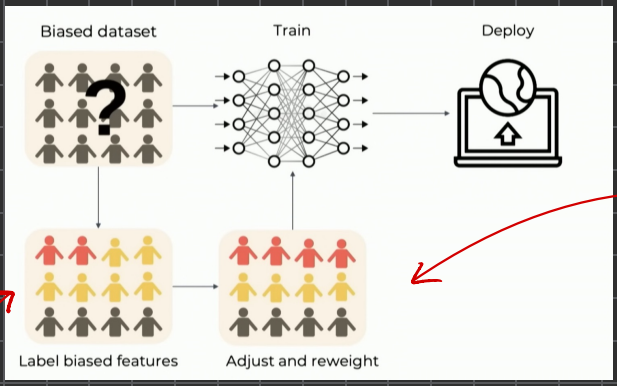
To mitigate

- Sample reweighting
  - Sample more data from underrepresented classes
- Loss reweighting
  - Some mistakes are worse than others!
- Batch Selection
  - Randomly batch for even data spread

What about latent features?

Variations within the same class are important to capture while debiasing

Why is debiasing latent features difficult?



Annotating data with features is data intensive

How do we know what the biased feature is?

Still problems!

VAE Recap:

- Probabalistic twist on autoencoders

How to de-bias through learned latent structure

1.) Learn latent structure

2.) Estimate distribution

- Sample different data sources based on probability distribution

3.) Adaptively Guide Learning

4.) Learn from fair distributions

Approximate the distribution of the latent space with a joint histogram over the latent variables

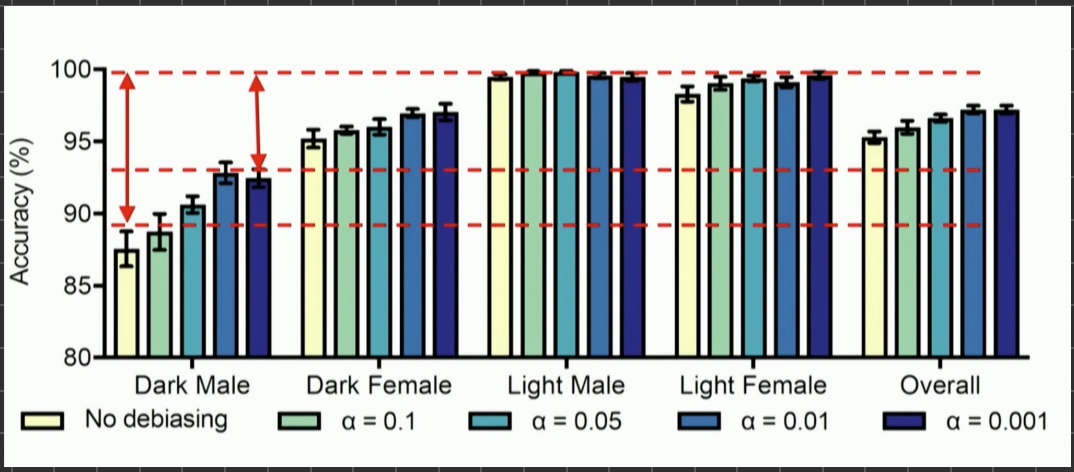
$$\hat{Q}(z|X) \propto \prod_i \hat{Q}_i(z_i|X) \longrightarrow W(z(x)|X) \propto \prod_i \frac{1}{\hat{Q}_i(z_i(x)|X)+\alpha}$$

Independence to approximate

Histogram for every latent variable z_i

Debiasing parameter

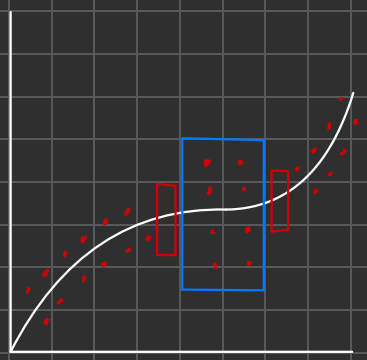
Evaluate: Decreased Categorical Bias



What is uncertainty?

- It is not likelihood
- Uncertainty estimation gives a confidence rating

Types of uncertainty in neural networks



Data uncertainty: 🔵

Very similar inputs have drastically different outputs

Model uncertainty: 🔴

Points here are out of distribution

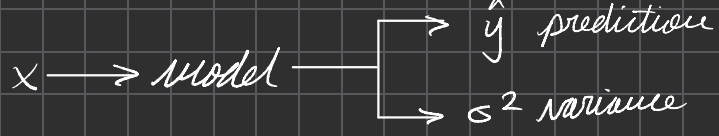
What if you added new data?

- Reduces model uncertainty, does not help data uncertainty.

High data uncertainty = Aleatoric uncertainty

High model uncertainty = Epistemic uncertainty

Estimating Aleatoric Uncertainty: Regression



$$f_\theta(x) \longrightarrow \hat{y}, \sigma^2 \quad \text{(variance is not constant)}$$

Negative Log Likelihood is a generalization of MSE

$$L = \frac{1}{N} \times \sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{2\sigma_i^2} + \ln \sigma_i^2$$

Estimating Epistemic Uncertainty

- What if we train the same model twice and compare results

```
num_ensembles = 5
for i in range(num_ensembles):
    model = create_model(...)
    model.fit(...)

raw_predictions = [models[i].predict(x)
    for i in range(num_ensembles)]
mu = np.mean(raw_predictions)
uncertainty = np.var(raw_predictions)
```

Introducing Stochasticity

Dropout layers!

Compute forward passes multiple times

You can also use reconstruction errors in VAEs

Evidential Deep Learning

Learn variance directly