# Improving Image Classifiers with IQ Loss functions: Non-adversarial f-Divergence Minimization

**Avi Gupta, Div Garg, Roy Yuan**

## Introduction

- Image classifiers exhibit calibration issues and proper calibration is essential to reliability for downstream tasks
- Deep neural networks applied to computer vision tasks have exhibited overconfidence in predictions due to overfitting to loss function
- New loss functions have been proposed to address miscalibration, energy-based modeling has proved particularly promising

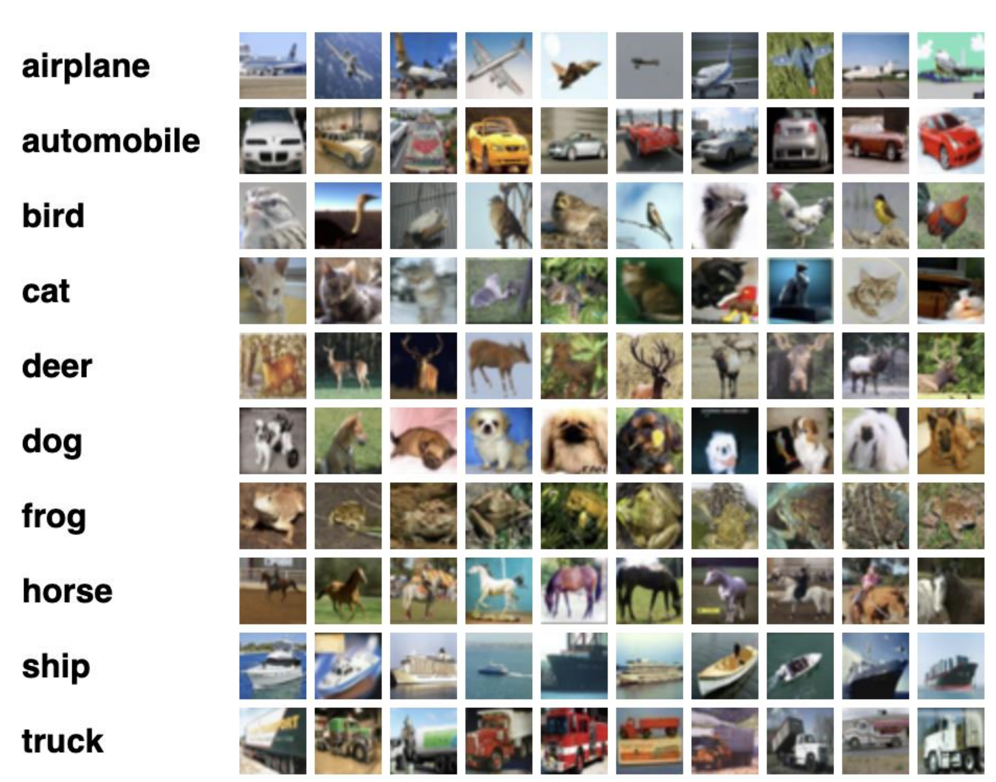
## Problem Statement

- Overfitting to loss functions produces miscalibration in deep neural network classifiers

- Improved loss functions can produce more robust and better-calibrated models

## Datasets

- CIFAR-10 and CIFAR-100
- OOD testing using SVHN
- Improved loss functions can produce more robust and better-calibrated models




CIFAR10 Dataset

## Method

- Idea: Interpret classifier as energy-based model
- Apply novel loss functions that minimize different f-divergences via non-adversarial training using regularized energy-based models (REMs)
- We present ***IQ Loss***: non-adversarial loss function that can minimize different f-divergences
- IQ Loss achieves superior calibration and robustness to distributional shifts compared to existing loss functions

**IQ Loss:** $$\max_{\epsilon} F(\epsilon) = \max_{\epsilon} \mathbb{E}_{\rho_E}[\phi(\epsilon)] - \alpha \log Z$$

with $Z = \int_{x \in X} e^{\epsilon/\alpha}$

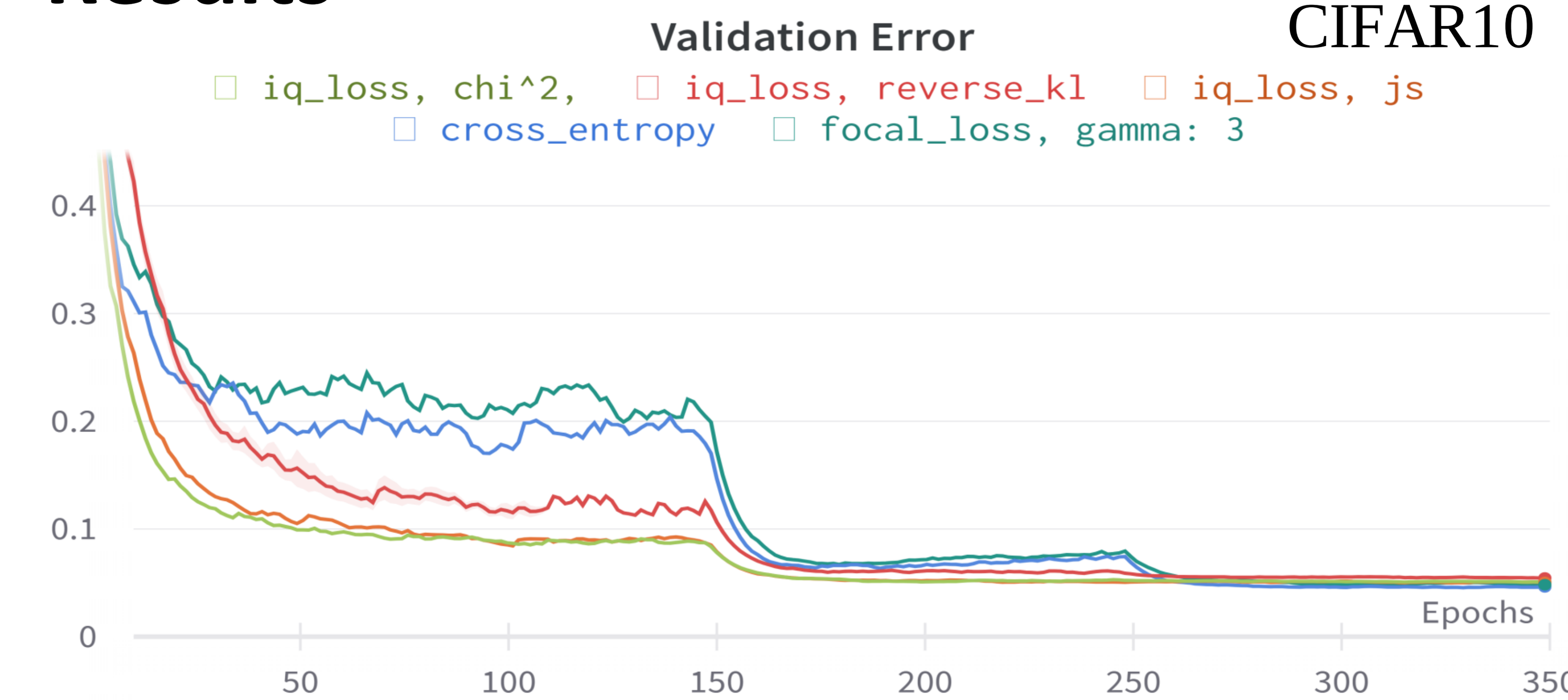
Table 1. List of divergence functions, $\phi$, and optimal energy estimators

| Divergence | $f(t)$ | $\phi(x)$ | $\epsilon$ |
|---|---|---|---|
| Forward KL | $-\log t$ | $1 + \log x$ | $\frac{\rho_E}{\rho}$ |
| Reverse KL | $t \log t - t + 1$ | $-e^{-x}$ | $\log \frac{\rho_E}{\rho}$ |
| Squared Hellinger | $(\sqrt{t} - 1)^2$ | $\frac{x}{1+x}$ | $\sqrt{\frac{\rho_E}{\rho}} - 1$ |
| Pearson $\chi^2$ | $(t-1)^2$ | $x - \frac{x^2}{4}$ | $2(1 - \frac{\rho}{\rho_E})$ |
| Total variation | $\frac{1}{2}\lvert t-1 \rvert$ | $x$ | $\frac{1}{2}\text{sign}(1 - \frac{\rho}{\rho_E})$ |
| Jensen-Shannon | $-(t+1)\log(\frac{t+1}{2}) + t \log t$ | $\log(2 - e^{-x})$ | $\log \frac{1}{2}(1 + \frac{\rho_E}{\rho})$ |

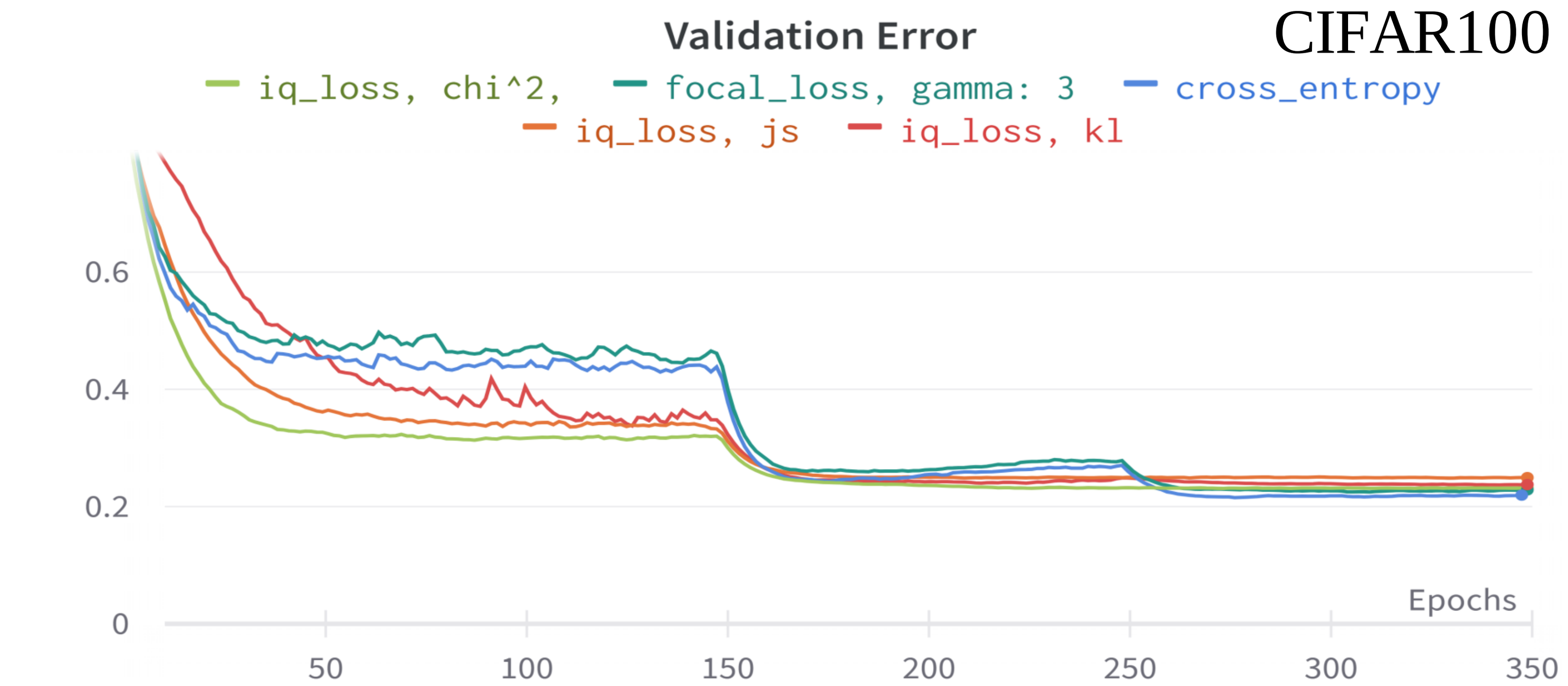
## Experiments

- Trained ResNet50 models on CIFAR-10 and CIFAR-100 datasets
- Compared models trained with IQ Loss (with different divergences), focal loss, and cross-entropy
- Standardized optimization and learning rate scheduling
- Performed out-of-distribution analysis using SVHN Dataset to determine robustness to distributional shifts

## Results

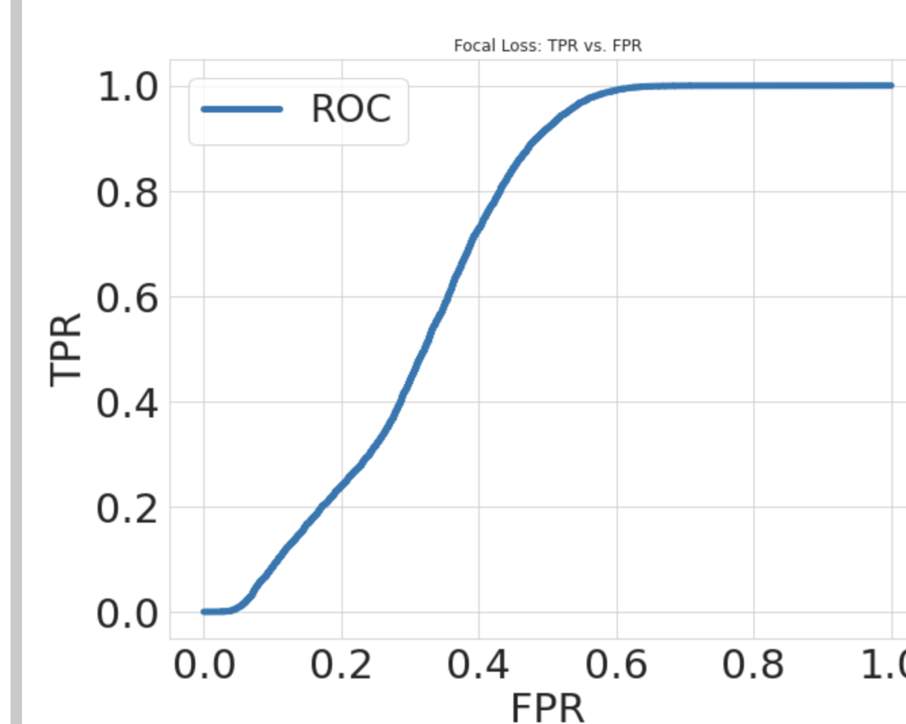



CIFAR10

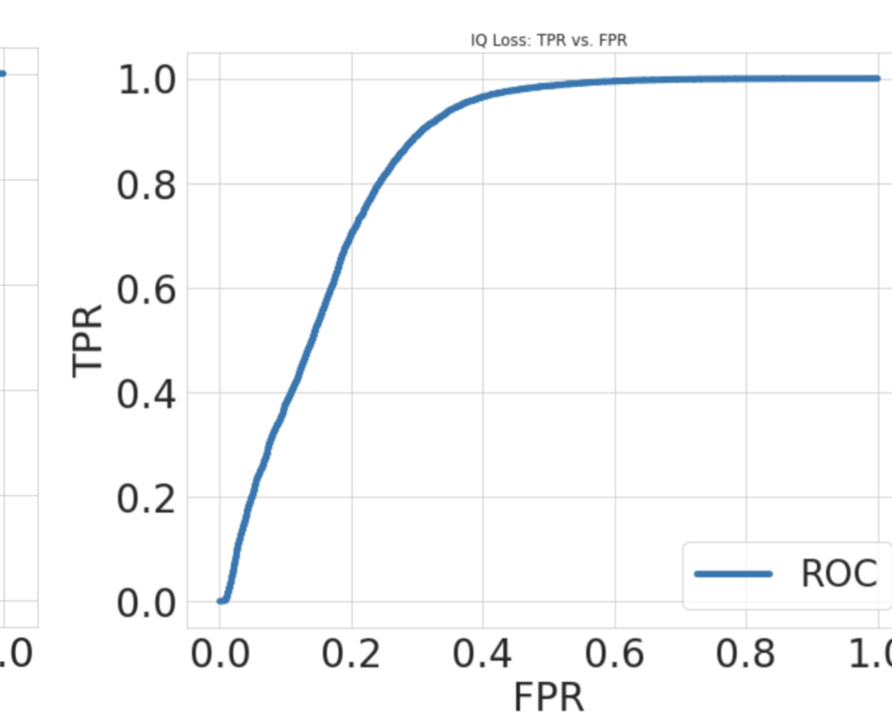



CIFAR100

### OOD Testing: CIFAR-10 vs SVHN




Focal Loss




IQ Loss

| Loss Function | AUROC |
|---|---|
| Focal Loss | 0.689 |
| IQ Loss | 0.843 |

Comparison of Area Under the curve (AUROC) for IQ Loss and Focal Loss during OOD testing on CIFAR10