# Balanced Activation
## for Long-tailed Visual Recognition

Team Innova
Jiawei Ren*, Cunjun Yu*, Zhongang Cai*, Haiyu Zhao

# Key Observation

- Softmax/Sigmoid regression is biased when the label distribution is different between training and testing (e.g., long-tailed datasets)
- We propose an unbiased extension, *Balanced Activation,* to accommodate the label distribution shift

# Overview

1. **Why is Softmax biased under long-tailed setup?**
2. How to accommodate the label distribution shift?
3. Further discussions on the optimization process

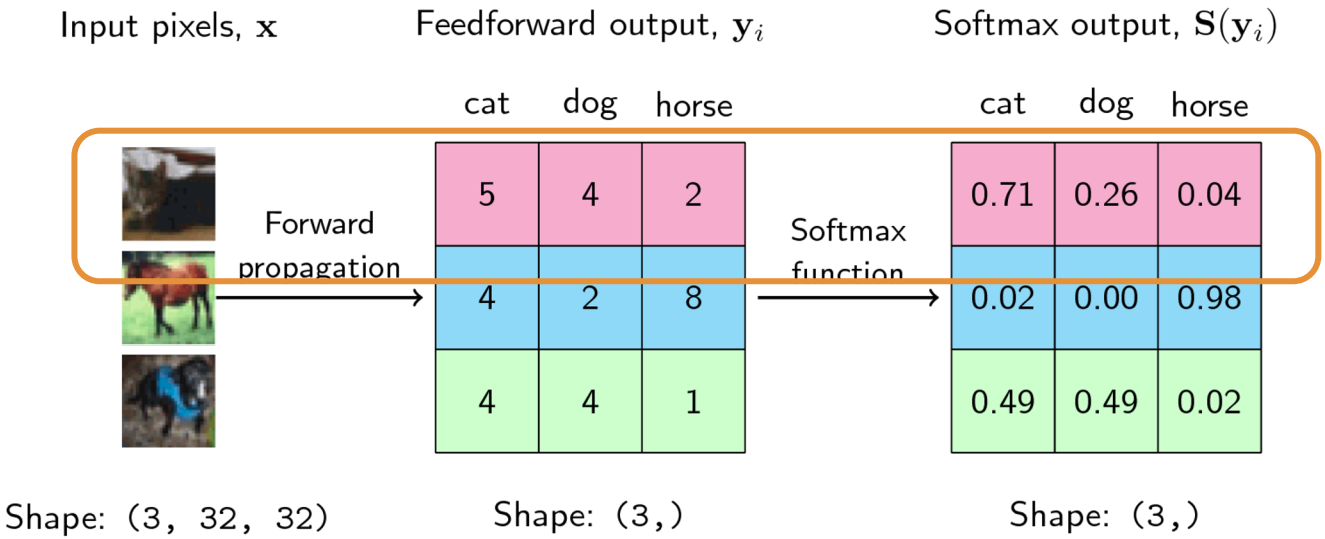
# What are we estimating with Softmax?

Softmax is the default choice for multi-class classification

$$\phi_j = \frac{e^{\eta_j}}{\sum_{i=1}^{k} e^{\eta_i}}$$

$\eta_j$ : Feedforward output

$\phi_j$ : Softmax output



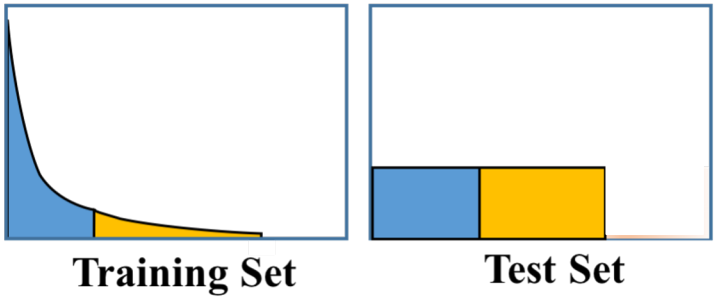

Softmax regression estimates the **p(y|x)** on the training set.

# Long-tail from Bayesian Inference perspective

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- p(x|y): generative distribution
- p(y): label distribution
- p(x): sample distribution

- Under long-tailed setup:
  - p(x|y) are the same on train & test
  - p(y) and p(x) are different on train & test
- **p(y|x)** estimated on the long-tailed training set is biased from **p(y|x)** on the balanced test set!



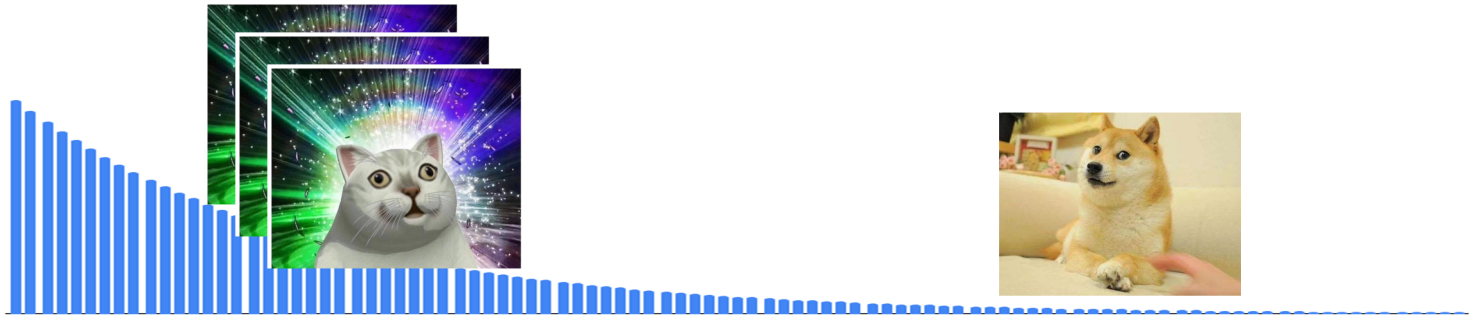
# Example of the Bias between Train & Test

Consider classifying cat/dog by genders: X={M, F}, Y={Cat, Dog}.

Genders are uniformly distributed: p(M|Cat)=p(F|Cat)=p(M|Dog)=p(F|Dog)=0.5

On balanced test set: Cat: {5M, 5F}, Dog: {5M, 5F} => p(Cat|M) =5/10=**0.5**

On long-tailed training set: Cat: {9M, 9F}, Dog: {1M, 1F} => p(Cat|M) =9/10 =**0.9**



# Extension to Class-Balanced Metrics

- For some long-tailed datasets, the test set is NOT balanced (e.g., LVIS)
- However, a class-balanced metric (e.g., mAP) views all classes equally important
- **Class-Balanced Metric** = **Standard Metric** on a balanced test set

$$\frac{1}{n}\sum p(y|x) = \frac{1}{n}\sum \frac{p(y|x)}{p(y)} \cdot p(y)$$

**Standard Metric**

**when p(y) = 1/k**

$$\frac{1}{k}\frac{1}{n}\sum \frac{p(y|x)}{p(y)}$$

**Class-Balanced Metric**

# Overview

1. Why is Softmax biased under long-tailed setup?
2. **How to accommodate the label distribution shift?**
3. Further discussions on the optimization process

# Link between Softmax & Bayesian Inference

Softmax: Exponential Family

$$\eta_j = \log \frac{\phi_j}{\phi_k}$$

**Canonical Link Function**

Bayesian Inference:

$$\frac{\phi_j}{\hat{\phi}_j} = \frac{n}{kn_j} \cdot \frac{\hat{p}(x)}{p(x)}$$

$$\hat{\phi}_j = \frac{n_j e^{\eta_j}}{\sum_{i=1}^{k} n_i e^{\eta_i}}$$

**Balanced Softmax (Ours)**

Please refer to our technical report for the full derivation

# Balanced Activation

$$\hat{\phi}_j = \frac{n_j e^{\eta_j}}{\sum_{i=1}^{k} n_i e^{\eta_i}}$$

Balanced Softmax

$$\hat{\phi}_j = \frac{e^{\eta_j - \log(\frac{n/k}{n_j} \cdot \frac{n-n_j}{n-n/k})}}{1+e^{\eta_j - \log(\frac{n/k}{n_j} \cdot \frac{n-n_j}{n-n/k})}}$$

Balanced Sigmoid

**Balanced Activation**

We derive Balanced Sigmoid similarly. Their combination is named **Balanced Activation**.

# Empirical Results on CIFAR-10/100-LT

| Dataset | CIFAR-10-LT | | | CIFAR-100-LT | | |
|---|---|---|---|---|---|---|
| Imbalance Factor | 200 | 100 | 10 | 200 | 100 | 10 |
| Softmax | 71.2 ± 0.3 | 77.4 ± 0.8 | 90.0 ± 0.2 | 41.0 ± 0.3 | 45.3 ± 0.3 | 61.9 ± 0.1 |
| CBW | 72.5 ± 0.2 | 78.6 ± 0.6 | 90.1 ± 0.2 | 36.7 ± 0.2 | 42.3 ± 0.8 | 61.4 ± 0.3 |
| CBS | 68.3 ± 0.3 | 77.8 ± 2.2 | 90.2 ± 0.2 | 37.8 ±0.1 | 42.6 ± 0.4 | 61.2 ± 0.3 |
| Focal Loss [7] | 71.8 ± 2.1 | 77.1 ± 0.2 | 90.3 ± 0.2 | 40.2 ± 0.5 | 43.8 ±0.1 | 60.0 ± 0.6 |
| Class Balance Loss [3] | 72.6 ± 1.8 | 78.2 ± 1.1 | 89.9 ± 0.3 | 39.9 ± 0.1 | 44.6 ± 0.4 | 59.8 ± 1.1 |
| LDAM [2] | 71.2 ± 0.3 | 77.2 ± 0.2 | 90.2 ± 0.3 | 41.0 ± 0.3 | 45.4 ± 0.1 | 62.0 ± 0.3 |
| Equalization Loss [8] | 72.8 ± 0.2 | 76.7 ± 0.1 | 89.9 ± 0.3 | 43.3 ± 0.1 | 47.3 ± 0.1 | 59.7 ± 0.3 |
| Balanced Softmax | **79.0** ± 0.8 | **83.1** ± 0.4 | **90.9** ± 0.4 | **45.9**± 0.3 | **50.3** ± 0.3 | **63.1** ± 0.2 |

**Table 1.** Top 1 accuracy for CIFAR-10/100-LT. Softmax denotes the standard cross-entropy loss with Softmax, CBW denotes class-balanced weighting, and CBS denotes class-balanced sampling. Balanced Activation generally outperforms SOTA methods, especially when the imbalance factor is high. Note that for other methods, we reproduce a higher accuracy than the reported value in the original papers.

# Marginal Likelihood of Labels in Prediction

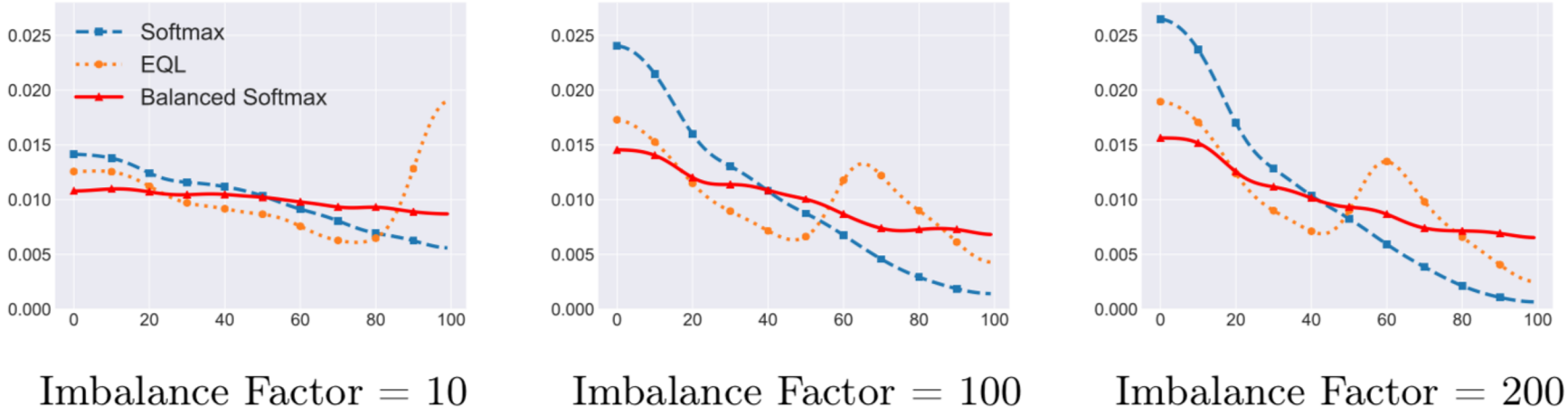



**Fig. 1.** Experiment on CIFAR-100-LT. the x-axis is the class labels with decreasing training samples and the y-axis is the marginal likelihood $p(y)$. Balanced Softmax is more stable under a high imbalance factor compared to the Softmax baseline and SOTA method, Equalization Loss (EQL).

# Empirical Results on LVIS-1.0

| Method | Backbone | $AP_{mask}$ | $AP_f$ | $AP_c$ | $AP_r$ | $AP_{box}$ | Split |
|---|---|---|---|---|---|---|---|
| Repeat Factor Sampling | ResNet-50 | 22.1 | 27.8 | 21.1 | 11.6 | 23.6 | Val |
| cRT | ResNet-50 | 23.2 | 27.8 | 22.5 | 14.7 | 24.8 | Val |
| LWS | ResNet-50 | 23.3 | 27.8 | 22.5 | 14.9 | 24.8 | Val |
| Equalization Loss | ResNet-50 | 24.2 | 27.5 | **24.2** | 16.5 | 25.8 | Val |
| Balanced Sigmoid | ResNet-50 | **25.1** | **29.3** | 24.0 | **18.3** | **26.7** | Val |
| Repeat Factor Sampling | ResNet-101 | 23.2 | 28.8 | 22.0 | 13.5 | 24.7 | Val |
| cRT | ResNet-101 | 23.9 | 28.7 | 22.7 | 15.8 | 25.6 | Val |
| LWS | ResNet-101 | 24.2 | 28.8 | 23.1 | 16.5 | 24.7 | Val |
| Equalization Loss | ResNet-101 | 25.4 | 28.8 | 25.3 | 18.5 | 27.3 | Val |
| Balanced Sigmoid | ResNet-101 | **26.4** | **30.2** | **25.4** | **20.1** | **28.4** | Val |
| Balanced Sigmoid | ResNet-101 | 26.1 | 30.4 | 25.0 | 19.5 | - | test-dev |

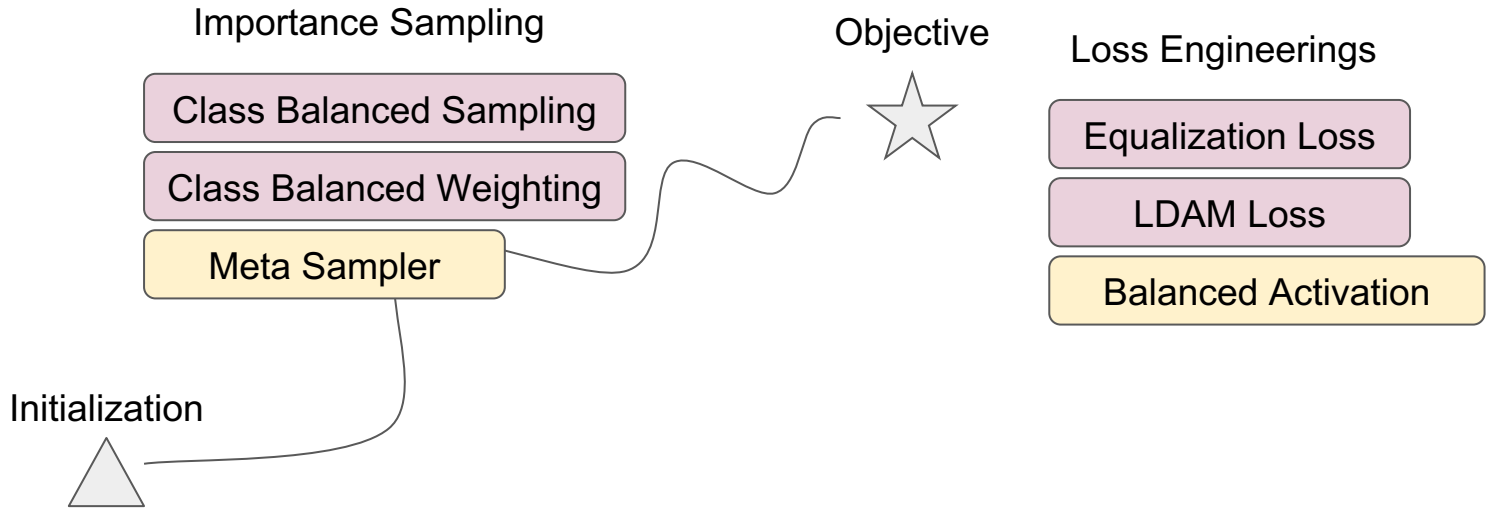
**Table 2.** Comparison of various methods using ResNet-50 and ResNet-101 on LVIS-1.0. Balanced Activation generally outperforms SOTA methods by a healthy margin.

# Overview

1. Why is Softmax biased under long-tailed setup?
2. How to accommodate the label distribution shift?
3. **Further discussions on the optimization process**

# Improve Optimization with Meta Sampler

Standard importance sampling leads to an 'over-balance' issue with Balanced Activation. We proposed Meta Sampler to help the optimization process. Checkout our arxiv paper: **Balanced Meta-Softmax for Long-Tailed Visual Recognition**



Thank you