Quasibinary Classifier for Images with Zero and Multiple Labels

Shuai Liao, Efstratios Gavves, ChangYong Oh, Cees Snoek

University of Amsterdam

Which classifier to choose?

Image classification task with 3 classes: { Bird, Cat, Dog }

[✓]Bird [ ]Cat [ ]Dog

One-vs-rest classification

[ ]Bird [✓]Cat [✓]Dog

Multi-labels classification

[ ]Bird [ ]Cat [ ]Dog

Zero-label classification

Softmax classifier

$$p_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$$

Ensemble of binary classifiers

$$p_k = \frac{\exp(z_k)}{1 + \exp(z_k)}$$ (Sigmoid activiation)

Motivation

Softmax classifier:

$$p_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$$

- ✓ Leverage prior knowledge #label=1.
- ✓ Scale-up to large number of classes.
- ✓ Stable gradient, stable training.
- ✗ Unable handle 0/N-label classification.

Ensemble of binary classifiers:

$$p_k = \frac{\exp(z_k)}{1 + \exp(z_k)}$$ (Sigmoid activiation)

- ✓ Flexible to handle 0/N-label.
- ✗ Do not model correlation.
- ✗ Do not scale-up to large number of classes.
- ✗ Unstable to train (saturate easily).

Observation: Softmax and Binary classifier have similar form.
The difference is on the denominator (normalization factor).

Ours: Quasibinary classifier

Goal: Learn a shared normalization factor as a function $\mathrm{C}(X)$
w.r.t. entire dataset X={ $x^{(\mathrm{i})}$ } and all K classes.

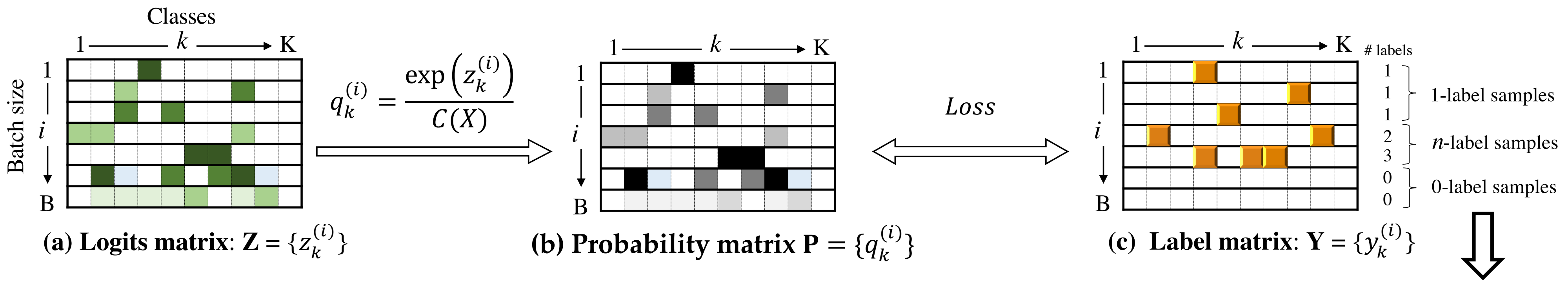
$$p(y = k|x) = q_k = \frac{\exp(z_k)}{\mathrm{C(X)}} \qquad \text{(Quasibinary activation)}$$

Constraints:

I) $\sum_{k=1}^{K} q_k = \#\text{label}$ (see proof in main paper).

II) $q_k^{(i)} \in [0,1]$.

Mini-batch training



(a) Logits matrix: $\mathbf{Z} = \{z_k^{(i)}\}$ **(b) Probability matrix** $\mathbf{P} = \{q_k^{(i)}\}$ **(c) Label matrix**: $\mathbf{Y} = \{y_k^{(i)}\}$

$\sum_i \#\text{label}^{(i)} = N$

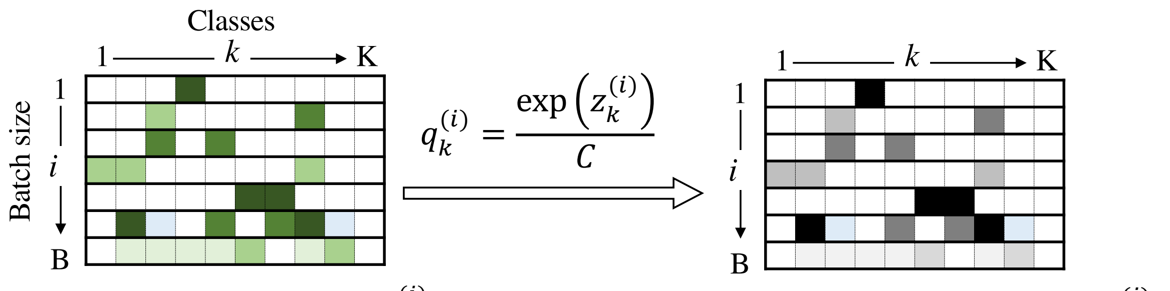
Training: Maximize Likelihood Estimation, *i.e.* $\sum_{i,k} y_k^{(i)} \log(q_k^{(i)})$, with

1) $\sum_i^B \sum_k^K q_k^{(i)} = \sum_i \#\text{label}^{(i)} = N \;\Rightarrow\; C = \frac{1}{N} \sum_{i=1}^{B} \sum_{k=1}^{K} \exp(z_k^{(i)})$

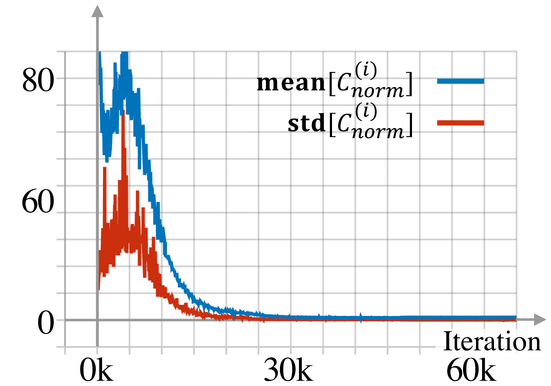
2) Minimize $-\max(\log q_k^{(i)}, 0)$ (Penalize any violation of $q_k^{(i)} > 1$).

Upon the convergence of training, C(X) also converges to a constant C.

Test



(a) Logits matrix: $\mathbf{Z} = \{z_k^{(i)}\}$ **(b) Probability matrix P** $= \{q_k^{(i)}\}$



Convergence of C

Test:

- #label per test sample is unknown.
- Use the learned constant C.
- $q_k^{(i)} = \min\left(\exp(z_k^{(i)})\,/C, 1\right)$

Quasibinary Classifier for Images with Zero and Multiple Labels

QUVA Deep Vision Lab
University of Amsterdam
Shuai Liao, Efstratios Gavves, ChangYong Oh, Cees Snoek

ICPR 2020

Motivation

Consider a classification problem with K=3 classes: { *Bird, Cat, Dog* }.

One-vs-rest problem
(Single-label problem)
[✓]**Bird** [ ]Cat [ ]Dog

Multi-labels problem
[ ]Bird [✓]**Cat** [✓]**Dog**

Zero-label problem
[ ]Bird [ ]Cat [ ]Dog

(I) Softmax classifier
✓ Use prior #label=1.
✓ Scale-up to large number of classes.
✓ Stable to train.
✗ Unable handle 0/N-label classification.

Softmax activation:
$$p_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}$$

(II) Ensemble of binary classifiers
✓ Flexible to handle 0/N-label classification.
✗ Lack of correlation: independently trained.
✗ Do not scale-up to large number of classes.
✗ Unstable to train.

Sigmoid activation:
$$p_k = \frac{\exp(z_k)}{1+\exp(z_k)}$$

Observation: Binary classifier and softmax have similar form. The difference is on the denominator (normalization factor).

Method

Goal: Learn a shared normalization factor as a function $C(X)$ w.r.t. entire dataset X={ $x^{(i)}$ } and all K classes.

$$q_k^{(i)} = q(y_k^{(i)} | x^{(i)}, X_{\backslash(i)}) = \frac{\exp(z_k^{(i)})}{C(x^{(i)}, X_{\backslash(i)})} = \frac{\exp(z_k^{(i)})}{C(X)}$$ (Probability of i-th image in being k-th class.)

- Leverage $\sum_{k=1}^{K} q_k$ = #lable for each image (Generalize to 0/1/N labels. *See proof in main paper.*)
- Ensure a valid probability $q_k^{(i)} \in [0,1]$.

Training: Maximize likelihood function, i.e. $\sum_{i,k} y_k^{(i)} \log(q_k^{(i)})$, with

1) $\sum_i^B \sum_k^K \exp(z_k^{(i)})/C = \sum_i \#\text{label}^{(i)} \Rightarrow C = \frac{1}{N} \sum_{i=1}^{B} \sum_{k=1}^{K} \exp(z_k^{(i)})$
(Compute C(X) over a mini-batch with size B.)

2) Minimize $-\max(\log q_k^{(i)}, 0)$ (Penalize any violation of $q_k^{(i)} > 1$).

Upon the convergence of training, C(X) also converges to a constant C.

Test:
- Unknown #label per test sample.
- Use the learned constant C.
- $q_k^{(i)} = \min\left(\exp(z_k^{(i)})/\text{C}, 1\right)$
- Number of output labels is a hyperparameter.

Experiment

Multi-label image classification

Setup: Follow Li *etal. CVPR2017.*

	MS-COCO		NUS-WIDE	
	$F_1\uparrow$	ECE(%)↓	$F_1\uparrow$	ECE(%)↓
Binary classifier [16], [18], [30]	51.2	26.8	40.7	23.6
Softmax [31]	**54.7**	32.2	43.2	25.8
Softmax w/ temperature [32]	**54.7**	31.4	43.2	24.6
Quasibinary classifier	**54.7**	**2.8**	**43.5**	**3.3**

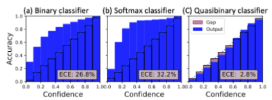


Fig. 3: **Reliability diagrams for multi-label classification**

Conclusion: Quasibinary classifier is both accurate and credible.

One-vs.-rest image classification (1-label)

Setup: Resnet18 with 32x32 and 224x224 input.

	CIFAR10	CIFAR100	Tiny-ImageNet	ImageNet
Binary classifier	**4.8**	35.4	×	×
Quasibinary classifier (*Ours*)	4.9	**21.9**	**42.9**	25.4
Softmax classifier	5.2	22.2	43.3	**23.9**

Conclusion: Quasibinary classifier is better than binary classifier, and comparable with softmax classifier.

Zero-label image classification

Setup: CIFAR60+40 dataset, with images of 40 classes from original CIFAR100 being treated as 0-label.

	IN Accuracy ↑	OUT MMC ↓	BOTH AU-ROC↑
Binary classifiers [16], [18], [30]	77.8 %	14.7 %	0.901
Softmax + $\mathcal{L}_{\text{Uniform}}$ [6]	**80.7** %	59.8 %	0.800
Softmax + $\mathcal{L}_{\text{MaxConf}}$ [11]	45.2 %	7.4 %	0.764
Quasibinary classifier	80.6 %	**6.9** %	**0.913**

Conclusion: Quasibinary classifier achieves good performance on all measures.

More information & time for questions at our poster.

Thank you!