南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Towards Calibrated Multi-label Deep Neural Networks

Jiacheng Cheng Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{jicheng,nvasconcelos}@ucsd.edu

CVPR 2024

Introduction

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Calibration:

A classifier is calibrated if it predicts a posterior class probability of p when the selection of the class is correct p × 100% of the time. The importance of calibration has been noted for many applications. For example, in medical diagnosis, probabilities can be used to determine which examples require human inspection, thus avoiding the cost of manually inspecting all images. However, the process can only be trusted if the DNN provides accurate posterior estimates.

A high-accuracy model may not be well-calibrated

Sample ID	Predicted Probability ($\hat{p}$)	Predicted Label	True Label	Correct Prediction?
1	0.9	+1	+1	Yes
2	0.8	+1	+1	Yes
3	0.7	+1	+1	Yes
4	0.6	+1	-1	No
5	0.4	-1	-1	Yes
6	0.3	-1	-1	Yes
7	0.2	-1	-1	Yes
8	0.1	-1	+1	No
9	0.9	+1	+1	Yes
10	0.85	+1	+1	Yes

Model Accuracy:

- Total samples = 10, correctly predicted samples = 8;
- Accuracy = $\frac{8}{10} = 80\%$.

Calibration Issues:

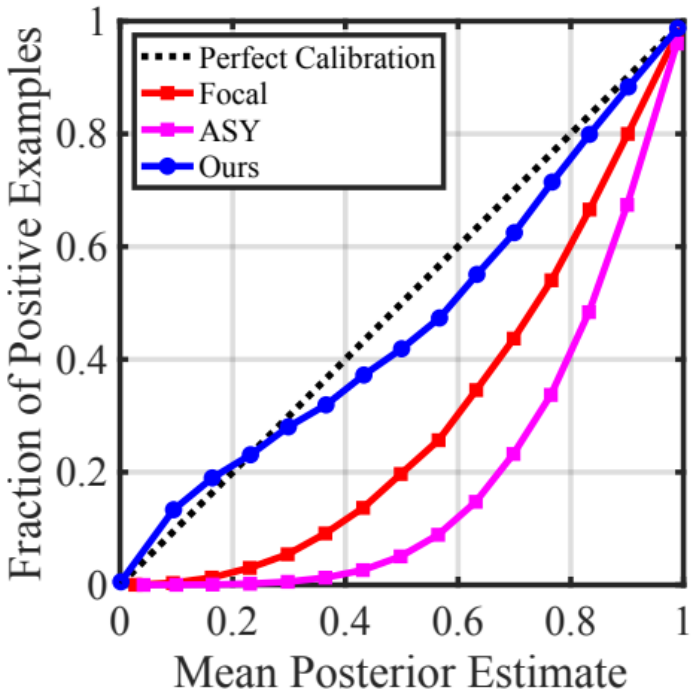
- In the [0.8, 0.9] probability range (Samples 1, 2, 9, 10), the average predicted probability is $\hat{p} = 0.86$, but the actual positive class frequency is 100%.
- In the [0.6, 0.7] probability range (Sample 4), the predicted probability is $\hat{p} = 0.6$, but the actual positive class frequency is 0%.

This indicates that the model, despite being accurate, is poorly calibrated because its predicted probabilities do not align with actual frequencies.

Introduction

Multi-label DNNs can be trained with class probability estimation (CPE) losses that encourage probability calibration, such as the binary cross-entropy (BCE) loss. However, the multi-label setting is highly imbalanced, due to the sparseness of positives, as most tags are absent from any given image. In result, asymmetric losses such as such as the focal loss of or the asymmetric (ASY) loss of tend to produce much higher labeling accuracy than the BCE.

Our preliminary studies reveal that multi-label DNNs trained with existing losses tend to produce poorly-calibrated probabilities. This is illustrated by the calibration curves in Figure 1, where the calibration of the focal and ASY losses are drastically far from perfect. We argue that these popular multi-label losses are poorly suited for class-probability estimation because they are not **strictly proper**. This is a property that denotes the family of **losses uniquely minimized by the true posterior probability.**



Preliminary

Given observation $\mathbf{x} \in \mathcal{X}$, the goal is to estimate the vector $\boldsymbol{\eta}(\mathbf{x}) = \left[\eta^{(1)}(\mathbf{x}), \cdots, \eta^{(T)}(\mathbf{x})\right]^\top$ of class-posterior probabilities

$$\eta^{(t)}(\mathbf{x}) = P\left(y^{(t)} = 1|\mathbf{x}\right), \forall t \in \{1, \cdots, T\}. \quad (1)$$

A multi-label DNN typically performs this probability estimation in two steps. First, it maps $\mathbf{x} \in \mathcal{X}$ into a real-valued score vector $\mathbf{v}(\mathbf{x}) = \left[v^{(1)}(\mathbf{x}), \cdots, v^{(T)}(\mathbf{x})\right]^\top \in \mathbb{R}^T$. The embedding $\mathbf{v} : \mathcal{X} \to \mathbb{R}^T$ is composed by a sequence of linear and nonlinear operations. Each $v^{(t)}(\mathbf{x})$ is then mapped into a class-posterior probability estimate with

$$\widehat{\eta}^{(t)}(x) = \widehat{P}\left(y^{(t)} = 1|\mathbf{x}\right) = [\Psi]^{-1}(v^{(t)}(\mathbf{x})), \quad (2)$$

where $[\Psi]^{-1}(\cdot)$ is an inverse link function. This can be any strictly increasing function $[\Psi]^{-1} : \mathbb{R} \to \Delta$, but is usually the logistic inverse link, or sigmoid activation function

$$\sigma(v) = \frac{1}{1 + e^{(-v)}}, \quad (3)$$

Given a CPE loss $\ell : \Delta \times \{-1, +1\} \to \mathbb{R}$ that assigns a cost $\ell(\widehat{\eta}, \pm 1)$ for predicting $\widehat{\eta}$ as the class-posterior probability of positive class when the true label is $y = \pm 1$,[1] the optimal posterior probability estimator minimizes the *risk*:

$$\mathcal{R}(\widehat{\boldsymbol{\eta}}) = \mathbb{E}_{\mathbf{x},\mathbf{y}} \left[\sum_{t=1}^{T} \ell \left(\widehat{\eta}^{(t)}(\mathbf{x}), y^{(t)} \right) \right], \tag{4}$$

期望风险

where $\mathbb{E}$ denotes expectation. To train a multi-label probabilistic DNN, this is approximated by the *empirical risk*

$$\widehat{\mathcal{R}}(\widehat{\boldsymbol{\eta}}; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \ell(\widehat{\eta}^{(t)}(\mathbf{x}_i), y_i^{(t)}) \tag{5}$$

经验风险

on a training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ of i.i.d. samples from $\mathcal{X} \times \mathcal{Y}$ [74].

BCE Loss

$$\begin{cases} \ell_{+1}^{BCE}(\widehat{\eta}) = -\log(\widehat{\eta}), \\ \ell_{-1}^{BCE}(\widehat{\eta}) = -\log(1-\widehat{\eta}). \end{cases} \tag{6}$$

ASY Focal Loss

$$\begin{cases} \ell_{+1}^{AsyFocal}(\widehat{\eta}) = -(1-\widehat{\eta})^{\gamma^+}\log(\widehat{\eta}), \\ \ell_{-1}^{AsyFocal}(\widehat{\eta}) = -(\widehat{\eta})^{\gamma^-}\log(1-(\widehat{\eta})), \end{cases} \tag{7}$$

ASY Focal Loss

$$\begin{cases} \ell_{+1}^{ASY}(\widehat{\eta}) = -(1-\widehat{\eta})^{\gamma^+}\log(\widehat{\eta}), \\ \ell_{-1}^{ASY}(\widehat{\eta}) = -(\widehat{\eta}-m)_+{}^{\gamma^-}\log(1-(\widehat{\eta}-m)_+), \end{cases} \tag{8}$$

Method

$$\mathcal{R}(\widehat{\boldsymbol{\eta}}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\sum_{t=1}^{T}\ell\left(\widehat{\eta}^{(t)}(\mathbf{x}), y^{(t)}\right)\right], \tag{4}$$

Since (4) can be rewritten as

$$\begin{aligned}\mathcal{R}(\widehat{\boldsymbol{\eta}}) =& \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[\sum_{t=1}^{T}\ell\left(\widehat{\eta}^{(t)}(\mathbf{x}), y^{(t)}\right)\middle|\mathbf{x}\right]\right] \qquad (9)\\ =& \mathbb{E}_{\mathbf{x}}\left[\sum_{t=1}^{T}C\left(\eta^{(t)}(\mathbf{x}), \widehat{\eta}^{(t)}(\mathbf{x})\right)\right] \qquad (10)\end{aligned}$$

where C is the (pointwise) *conditional risk*

$$C\left(\eta(\mathbf{x}), \widehat{\eta}(\mathbf{x})\right) = \eta(\mathbf{x})\ell_{+1}(\widehat{\eta}(\mathbf{x})) + (1-\eta(\mathbf{x}))\ell_{-1}(\widehat{\eta}(\mathbf{x})). \tag{11}$$

Definition 1. *(Strict Properness [4, 19, 68]) The pair of partial losses $\{\ell_{-1}, \ell_{+1}\}$ or $\{\ell_{-1,\Psi}, \ell_{+1,\Psi}\}$ is strictly proper if the conditional risk $C(\eta, \widehat{\eta})$ of (11) is uniquely minimized by $\widehat{\eta} = \eta$ for all $\eta \in [0, 1]$.*

$C = \eta(x)[-\log(\hat{\eta}(x))] + (1-\eta(x))(-\log(1-\hat{\eta}(x)))$

$C' = -\frac{\eta(x)}{\hat{\eta}(x)} + \frac{1-\eta(x)}{1-\hat{\eta}(x)}$

C为凹函数

$\hat{\eta}(x) = \eta(x)$ 时，C最小

Theorem 2. *Denote the conditional risk of (11) defined by the ASY loss ℓ^{ASY} of (8) by $C^{ASY}(\eta, \widehat{\eta})$. Let*

$$\widehat{\eta}^{ASY,*}(\mathbf{x}) = \arg\min_{\widehat{\eta}\in[0,1]} C^{ASY}(\eta(\mathbf{x}), \widehat{\eta}(\mathbf{x})) \tag{12}$$

be the minimizer of this risk for any $\mathbf{x} \in \mathcal{X}$. Then there is a mapping ϕ such that

$$\eta(\mathbf{x}) = \phi\left(\widehat{\eta}^{ASY,*}(\mathbf{x}); \gamma^-, \gamma^+, m\right), \tag{13}$$

where

$$\phi(z; \gamma^-, \gamma^+, m) = \frac{h((z-m)_+; \gamma^-)}{h((z-m)_+; \gamma^-) + h(1-z; \gamma^+)}, \tag{14}$$

$$h(z; \gamma) = \frac{z^\gamma - \gamma z^{\gamma-1}(1-z)\log(1-z)}{1-z}. \tag{15}$$

$\phi : \Delta \to \Delta$ is a bijective map if and only if $m = 0$.

Method

rather than designing a CPE loss $\ell_{\pm 1}(\widehat{\eta})$ explicitly, we consider the separate design of a composite loss $\ell_{\pm 1,\Psi}(v)$ and an inverse link $[\Psi]^{-1}(v)$ as below

$$\begin{cases} \ell_{+1,\Psi}(v) = -\frac{1}{\zeta^+}\log\left(\frac{1}{1+e^{(-k^+(v-b^+))}}\right), \\ \ell_{-1,\Psi}(v) = -\frac{1}{\zeta^-}\log\left(\frac{1}{1+e^{(+k^-(v-b^-))}}\right), \end{cases} \tag{16}$$

$$[\Psi]^{-1}(v) = \frac{\frac{k^-}{\zeta^-}\sigma(-k^-(v-b^-))}{\frac{k^-}{\zeta^-}\sigma(-k^-(v-b^-)) + \frac{k^+}{\zeta^+}\sigma(k^+(v-b^+))}, \tag{17}$$

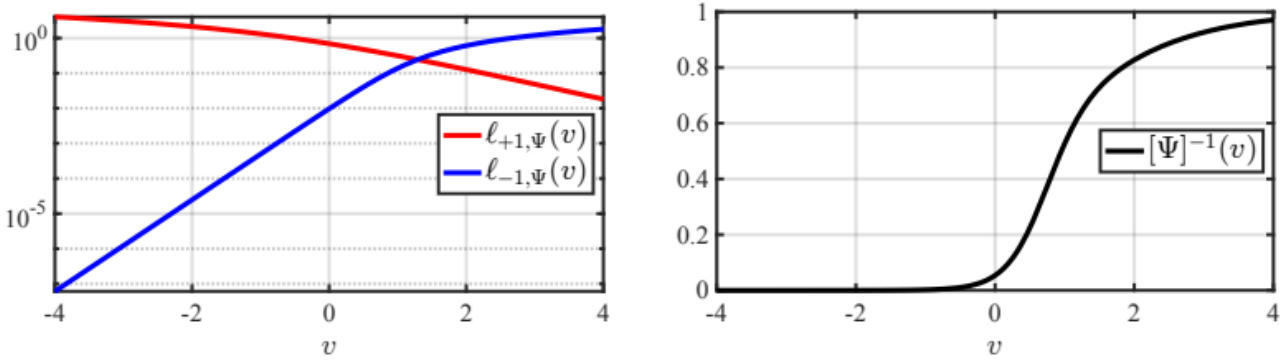


Figure 2. Example of the SPA losses (left) and the associated inverse link (right) where $(\zeta^+, k^+, b^+, \zeta^-, k^-, b^-) = (1, 1, 0, 5, 3, 1)$.

where $\sigma(\cdot)$ is the sigmoid function of (3). According to the Theorem 3 below, the CPE loss composed of (16) and (17) is strictly proper and thus referred to as the *Strictly Proper Asymmetric* (SPA) loss in this work. In practice, following the practice in the literature [61], we reduce the positive partial loss to the BCE loss by setting $(\zeta^+, k^+, b^+) = (1, 1, 0)$ and only tune the hyperparmeters ζ^-, k^-, b^- of the negative partial loss. The motivation for these hyperparmeters is simple and intuitive: i) k^-, b^- define an affine transformation of the logits $v^{(t)}(\mathbf{x})$ that enables control of the rate at which ℓ_{-1} decays to 0 as $v \to 0$. ii) ζ^- is a scale factor that controls the overall weight of negative examples. Note that unlike prior losses (6)-(8), SPA does not directly operate on the probability estimate $\widehat{\eta}(\mathbf{x})$, and that the introduction of these hyperparameters induces the need for the inverse link of (17) to achieve the strict properness, rather than simply using the sigmoid of (3).

Theorem 3. *For any $\zeta^+, \zeta^-, k^+, k^- \in \mathbb{R}_{++}$ and $b^-, b^+ \in \mathbb{R}$, the CPE loss composed of composite loss (16) and inverse link function (17) is strictly proper.*

Method

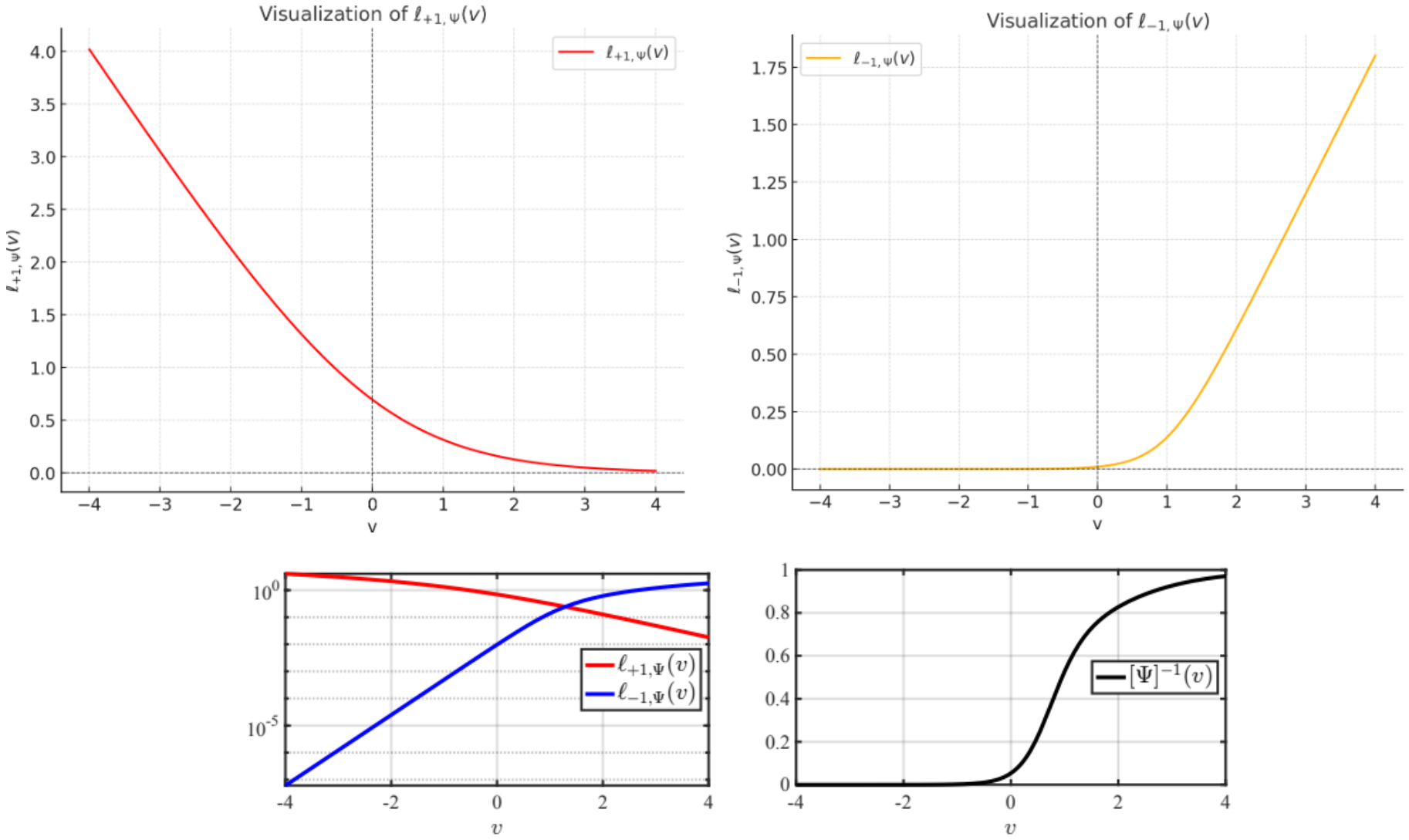


Figure 2. Example of the SPA losses (left) and the associated inverse link (right) where $(\zeta^+, k^+, b^+, \zeta^-, k^-, b^-) = (1, 1, 0, 5, 3, 1)$.

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Label Pair Regularizer

A strictly proper CPE loss only guarantees perfect label-probability estimates for asymptotically large datasets. In practice, for finite datasets, empirical risk minimization does not guarantee the recovery of true risk minimizer. In this case, probability calibration can usually be improved by adding regularization terms to the loss function. In this work, we propose a regularizer specifically designed for multi-label learning.

Under the independence Assumption 1, the probability estimates of each label in $\{1, \cdots, T\}$ are supervised independently, *i.e.* there is no explicit supervision for the joint prediction of multiple classes. This is consistent with the decomposition of the risk of (4) into a sum of t label-specific risks. However, an example $\mathbf{x}$ still provides joint constraints on the probability estimates of different labels. To see this, consider any $\mathbf{x} \in \mathcal{X}$ and pair (t, t') of labels with different values. The probability of $y^{(t)} = +1$ is then

Assumption 1. *For any* $\mathbf{x} \in \mathcal{X}$ *and* $i \neq j \in \{1, \cdots, T\}$, $y^{(i)}$ *and* $y^{(j)}$ *are independent given* $\mathbf{x}$, i.e. $y^{(i)} \perp\!\!\!\perp y^{(j)} | \mathbf{x}$.

$$
\begin{aligned}
\beta^{tt'}(\mathbf{x}) &= P\left(y^{(t)} = +1 | y^{(t)} \neq y^{(t')}, \mathbf{x}\right) \\
&= \frac{P(y^{(t)} = +1, y^{(t')} = -1 | \mathbf{x})}{P\left(y^{(t)} = +1, y^{(t')} = -1 | \mathbf{x}\right) + P\left(y^{(t)} = -1, y^{(t')} = +1 | \mathbf{x}\right)} \\
&= \frac{\eta^{(t)}(\mathbf{x})(1 - \eta^{(t')}(\mathbf{x}))}{\eta^{(t)}(\mathbf{x})(1 - \eta^{(t')}(\mathbf{x})) + (1 - \eta^{(t)}(\mathbf{x}))\eta^{(t')}(\mathbf{x})}
\end{aligned}
\tag{18}
$$

where the last equality follows from the Assumption 1. Let

$$
\widehat{\beta}^{tt'} = \frac{\widehat{\eta}^{(t)}(1 - \widehat{\eta}^{(t')})}{\widehat{\eta}^{(t)}(1 - \widehat{\eta}^{(t')}) + \widehat{\eta}^{(t')}(1 - \widehat{\eta}^{(t)})} \tag{19}
$$

be the plugin estimator of $\beta^{tt'}$. The following result shows that the accurate estimation of $\widehat{\beta}^{tt'}$ is a necessary condition for the accurate estimation of $\widehat{\eta}^{(t)}$ and $\widehat{\eta}^{(t')}$.

Label Pair Regularizer

Lemma 1. *For any $t \neq t'$, $\widehat{\beta}^{(tt')} = \beta^{(tt')}$ is a necessary condition for $\widehat{\eta}^{(t)} = \eta^{(t)}$ and $\widehat{\eta}^{(t')} = \eta^{(t')}$.*

Since the example $\mathbf{x}$ can be seen as a calibration constraint for the estimation of $\widehat{\beta}^{tt'}$ in addition to those that it already provides for the individual calibration of $\widehat{\eta}^{(t)}$ and $\widehat{\eta}^{(t')}$, this suggests that introducing calibration supervision on $\widehat{\beta}^{tt'}$ can help improve the calibration of $\eta^{(t)}, \eta^{(t')}$. We leverage this observation by introducing a new *Label Pair Regularizer* (LPR) o calibrate the estimate $\widehat{\beta}^{ij}$, implemented with the BCE loss

$$\mathcal{L}^{\text{LPR}}(\mathbf{x}, \mathbf{y}) = \frac{2}{T(T-1)} \sum_{t \neq t'} \mathbb{1}_{\substack{y^{(t)}=+1 \\ y^{(t')}=-1}} \left[-\log \widehat{\beta}^{tt'} \right]. \tag{20}$$

Finally, in our proposed training approach, a multi-label DNN is trained by a joint optimization of SPA and LPR:

$$\mathcal{L}^{\text{overall}}(\mathbf{x}, \mathbf{y}) = \frac{1}{T} \sum_{t=1}^{T} \ell_{y^{(t)},\Psi}(v^{(t)}(\mathbf{x})) + \lambda \mathcal{L}^{\text{LPR}}(\mathbf{x}, \mathbf{y}) \tag{21}$$

where λ is the multiplier balancing the two terms.

Multi-label Image Retrieval

A natural score for image ranking is then the posterior probability

$$s(\mathbf{x}) = P\left(\{y^{(j)}=+1\}_{j\in\mathcal{P}}, \{y^{(k)}=-1\}_{k\in\mathcal{N}} \middle| \mathbf{x}\right)$$
$$= \prod_{j\in\mathcal{P}} \eta^{(j)}(\mathbf{x}) \prod_{k\in\mathcal{N}} (1-\eta^{(k)}(\mathbf{x})) \quad (22)$$

where the equality follows from the Assumption 1. This can be estimated by a probabilistic multi-label networks as

$$\widehat{s}(\mathbf{x}) = \prod_{j\in\mathcal{P}} \widehat{\eta}^{(j)}(\mathbf{x}) \prod_{k\in\mathcal{N}} (1-\widehat{\eta}^{(k)}(\mathbf{x})). \quad (23)$$

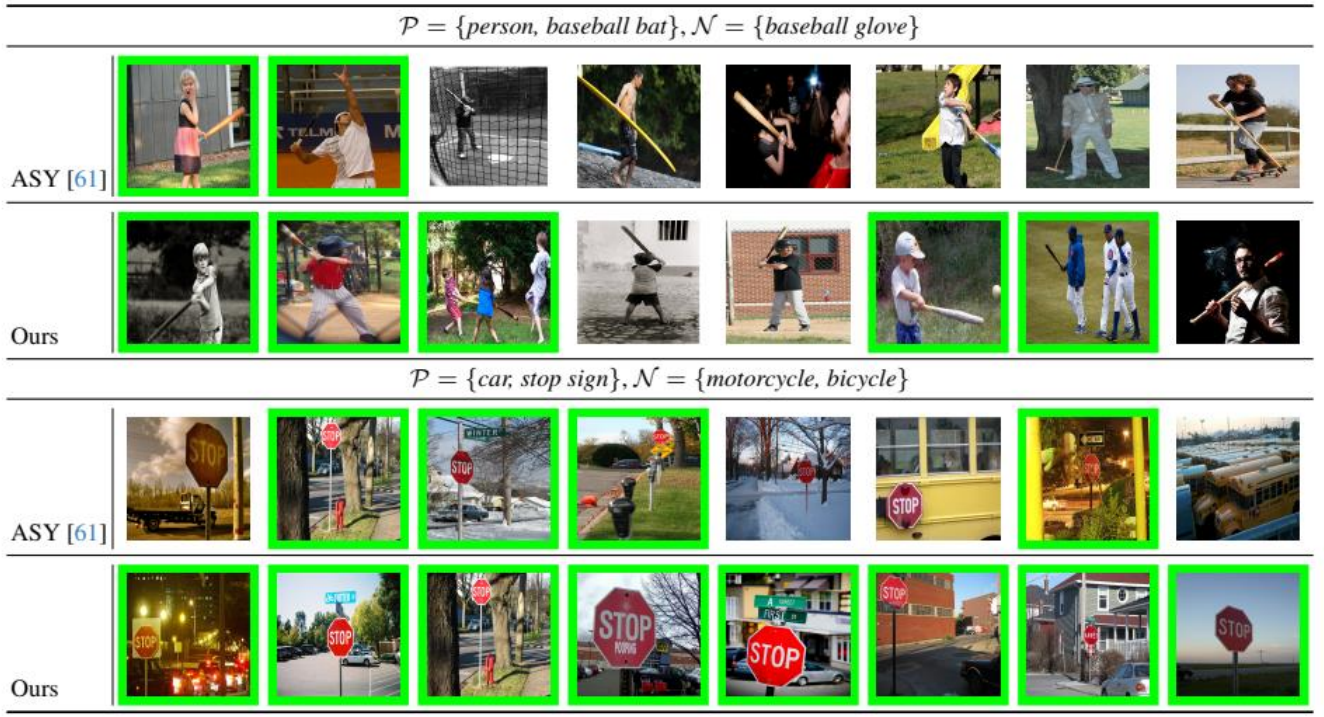


Figure 4. Qualitative results of multi-label image retrieval. Correct retrieval results are highlighted in green.

Experiments

Dataset	Method	ECA-ResNet50-T Accuracy mAP@y ↑	ECA-ResNet50-T Accuracy mAP@x ↑	ECA-ResNet50-T Calibration ACE ↓	ECA-ResNet50-T Calibration MCE ↓	ViT-B/32 Accuracy mAP@y ↑	ViT-B/32 Accuracy mAP@x ↑	ViT-B/32 Calibration ACE ↓	ViT-B/32 Calibration MCE ↓
COCO	BCE	72.2	82.4	8.2	17.0	70.1	81.9	6.0	15.5
	TWL	77.2	87.9	17.2	33.6	76.2	88.0	14.2	31.3
	Focal	74.8	85.9	20.1	34.9	72.0	83.8	23.6	36.2
	Focal+ϕ	-	-	11.1	20.3	-	-	8.5	17.9
	ASY	77.5	88.2	30.6	46.0	76.4	87.7	29.5	48.6
	ASY + ϕ	77.0	87.8	15.0	26.0	76.2	87.3	16.2	26.2
	SPA	**77.8**	88.4	5.3	12.0	**76.8**	87.9	4.8	10.7
	SPA + LPR	77.7	**88.6**	**4.2**	**9.3**	76.6	**88.1**	**2.1**	**5.3**
VOC	BCE	85.1	92.4	6.9	13.8	86.9	92.7	7.8	16.7
	TWL	89.1	93.7	10.8	22.0	90.1	94.5	14.5	28.3
	Focal	87.4	93.3	19.0	35.8	88.4	93.6	16.2	31.7
	Focal + ϕ	-	-	7.0	16.5	-	-	9.3	20.7
	ASY	89.6	**94.6**	26.4	47.7	**90.4**	94.7	31.7	52.8
	ASY + ϕ	89.1	94.3	12.9	24.7	89.8	94.2	15.5	33.0
	SPA	89.5	94.1	5.4	14.9	90.0	93.9	6.1	14.4
	SPA + LPR	**89.9**	94.3	**4.9**	**11.0**	**90.4**	**94.8**	**5.5**	**12.7**
WIDER-A	BCE	74.9	82.8	11.5	25.2	76.7	83.2	8.6	13.3
	TWL	79.9	85.5	14.3	29.0	81.6	87.6	16.7	32.3
	Focal	78.5	84.1	20.1	34.2	80.8	85.2	18.8	31.2
	Focal + ϕ	-	-	9.3	21.0	-	-	7.4	16.5
	ASY	**80.6**	**86.0**	24.2	35.2	82.2	87.8	22.8	36.3
	ASY + ϕ	79.9	85.4	14.3	27.5	81.8	87.2	13.3	20.7
	SPA	80.1	85.8	5.3	11.2	82.0	87.6	4.3	10.5
	SPA + LPR	80.3	85.9	**3.5**	**8.0**	**82.7**	**87.9**	**2.8**	**6.2**
VISPR	BCE	46.3	78.8	9.7	16.6	48.4	80.5	8.2	14.8
	TWL	52.4	84.3	17.9	26.0	**53.6**	85.4	12.4	26.8
	Focal	48.0	81.0	24.0	40.5	50.1	83.2	24.6	35.7
	Focal + ϕ	-	-	8.6	18.2	-	-	9.6	14.1
	ASY	51.6	84.0	28.8	44.2	53.0	85.0	27.7	45.2
	ASY + ϕ	51.4	83.7	14.3	23.7	43.9	52.8	16.4	29.0
	SPA	52.4	84.5	5.8	12.1	53.2	85.3	5.9	10.2
	SPA + LPR	**52.7**	**84.9**	**3.0**	**8.1**	53.4	**85.6**	**2.5**	**7.4**

Experiments

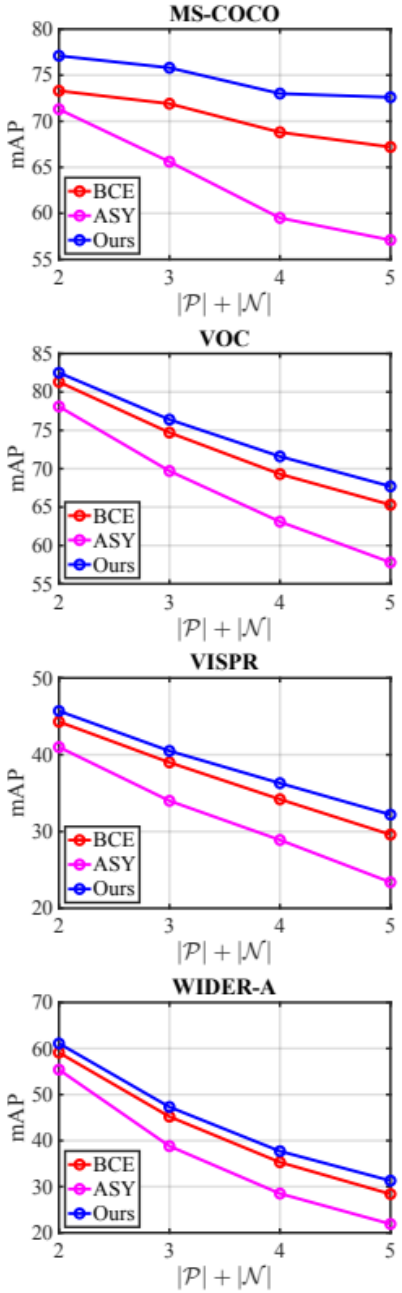
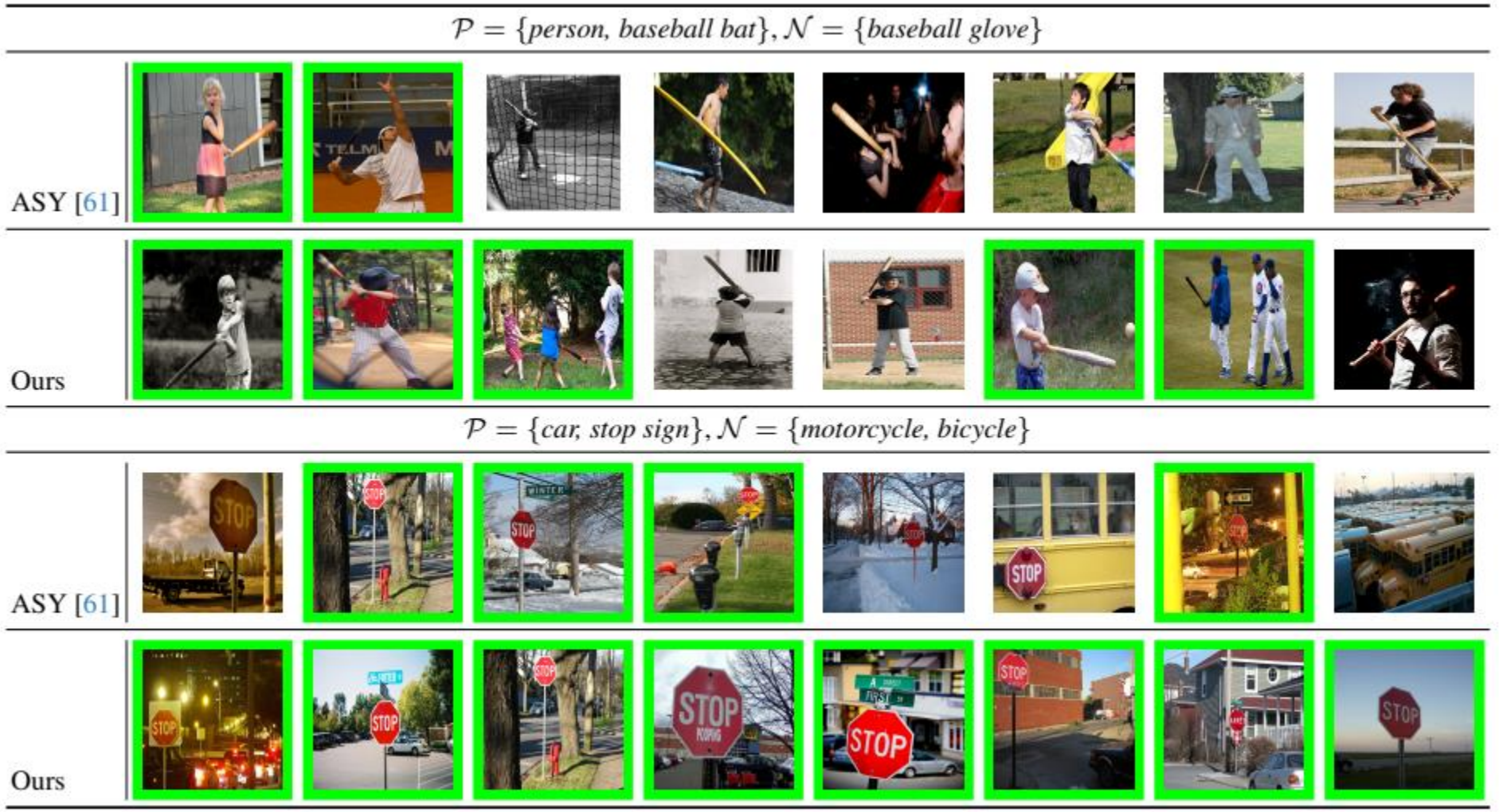


Figure 3. Multi-label image retrieval mAP versus the number of search conditions $|\mathcal{P}| + |\mathcal{N}|$.

Experiments



Experiments

	mAP@**x**↑	mAP@**y**↑	ACE↓	MCE↓
BCE	72.2	82.4	8.2	17.0
BCE + LPR	**72.4**	**82.9**	**6.5**	**14.8**
Focal	74.8	85.9	20.1	**34.9**
Focal + LPR	**75.1**	**86.3**	**19.7**	36.4
ASY	**77.5**	**88.2**	30.6	46.0
ASY + LPR	77.2	88.1	**29.4**	**44.7**

Table 4. Effect of LPR on other CPE losses (COCO, ECA-ResNet50-T).

Thanks