MCFL: Multi-label Contrastive Focal Loss for Pedestrian Attribute Recognition

Xiaoqiang Zheng, Zhenxia Yu, Lin Chen, Fan Zhu , Shilog Wang*

Chengdu University of Information Technology, China

Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science , China

ICPR 2020 — 25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION — Milan, Italy 10 | 15 January 2021

Abstract

- New loss function called Multi-label Contrastive Focal Loss (MCFL)
- Emphasizing the hard and minority attributes by using a separated re-weighting mechanism imbalance
- Enlarge the gaps between the intra-class of multi-label attributes, to extract more subtle discriminative features
- MCFL with the ResNet-50 backbone is able to outperform other state-of-the-art approaches in term of mean accuracy

Introduction

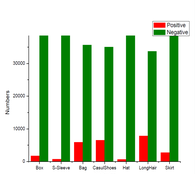
The main task of Pedestrian Attribute Recognition (PAR) is to give a series of semantic pedestrian attributes, such as gender, age, clothing style, or other appearance attributes, to help to locate specific target

Main Challenge:

- Low quality, various in-camera viewing angle, illumination, background, or human poses
- PAR is a multi-label learning as it needs to describe dozens of attributes for a person simultaneously
- Fine-grained attributes (such as muffler and glass) are hard to recognize at a far distance.
- The extremely imbalanced distribution and lack of sufficient training data are also limited to the PAR performance

Pose Variation

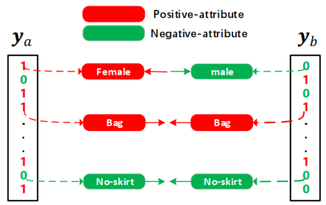
Occlusion



Data Imbalance

Methods

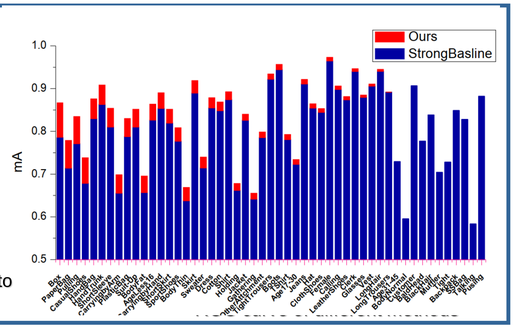
- Multi-label Contrastive Focal Loss(MCFL), integrating multi-label focal loss and contrastive loss simultaneously.
- Focusing on the difficult and error-prone attributes.
- Separating the losses of positive and negative attributes by re-weighting mechanism to decrease the impact of the class imbalance.

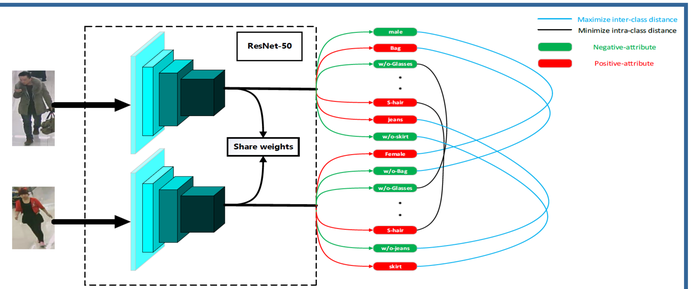


The optimization process of the multi-label contrastive loss function

Results

- MCFL delivers a better performance for the imbalance and error prone attributes including carrying-boxes, body-shape and shoe-types.
- Also, age range seems more difficult to classify correctly due to its ambiguous definition





The workflow of the MCFL model

Methods

➢ Multi-label Focal Loss

$$L_i^p = -\mathbf{sw}_i(1-Pt_i)\mathbf{y_i}log(Pt_i)$$

$$L_i^n = -\mathbf{sw}_i Pt_i(1-\mathbf{y_i})log(1-Pt_i)$$

$$\mathbf{sw}_i = exp(\mathbf{y}_i(1-\frac{\sum_{i=1}^{N}\mathbf{y}_i}{N}) + (1-\mathbf{y}_i)\frac{\sum_{i=1}^{N}\mathbf{y}_i}{N})$$

➢ Multi-label Contrastive Loss

$$S_p = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^p}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^p)^2\sum_{j=1}^{M}(Pt_{bj}w_j^p)^2}}$$

$$S_n = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^n}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^n)^2\sum_{j=1}^{M}(Pt_{bj}w_j^n)^2}}$$

$$S_d = \frac{\sum_{j=1}^{M} Pt_{aj}Pt_{bj}w_j^d}{\sqrt{\sum_{j=1}^{M}(Pt_{aj}w_j^d)^2\sum_{j=1}^{M}(Pt_{bj}w_j^d)^2}}$$

➢ Multi-label Contrastive Focal Loss

$$L_{intra\_p} = \mathbf{w}^p L_a^p(\frac{1}{S_p+eps}+\alpha)$$

$$L_{intra\_n} = \mathbf{w}^n L_b^n(\frac{1}{S_n+eps}+\beta)$$

$$L_{inter\_p} = \mathbf{w}^d(\sum_{i=a}^{b} L_i^p)(S_d+\alpha)$$

$$L_{inter\_n} = \mathbf{w}^d(\sum_{i=a}^{b} L_i^n)(S_d+\beta)$$

$$L_{total} = L_{intra\_p} + L_{intra\_n} + L_{inter\_p} + L_{inter\_n}$$

Results

The experiment results on the RAP dataset

Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	76.12	65.39	77.33	78.79	78.05
LGNet(BMVC18) [18]	InceptionV2	78.68	68.00	80.36	79.82	80.09
PGDM(ICME18) [11]	CaffeNet	74.31	64.57	78.86	75.90	77.35
JLPLS-PAA(TIP19) [29]	–	81.25	67.91	78.56	81.45	79.98
RA(AAAI19) [35]	InceptionV3	81.16	–	79.45	79.23	79.34
AAP(AAAI19) [5]	ResNet50	81.42	68.37	**81.04**	80.27	80.65
ALM(ICCV19) [30]	BNInception	81.87	68.17	74.71	**86.48**	80.16
StrongBaseline(2020) [8]	ResNet50	80.52	68.44	79.91	80.64	79.89
MCFL(ours)	ResNet50	**82.06**	**69.01**	77.47	84.91	**81.02**

The experiment results on the PETA dataset

Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	74.21	72.19	82.97	82.09	82.53
LGNet(BMVC18) [18]	InceptionV2	76.96	75.55	86.99	83.17	85.04
PGDM(ICME18) [11]	CaffeNet	74.95	73.08	84.36	82.24	83.29
AAP(AAAI19) [5]	ResNet50	80.56	78.30	**89.49**	84.36	86.85
ALM(ICCV19) [30]	BNInception	80.68	77.08	84.21	**88.84**	86.46
MT-CAS(ICME20) [33]	ResNet50	77.20	78.09	88.46	84.86	86.62
StrongBaseline(2020) [8]	ResNet50	80.50	78.84	87.24	87.12	86.78
MCFL(ours)	ResNet50	**81.11**	**79.01**	86.67	88.15	**87.41**

Ablation study on RAP dataset

Methods	mA	Acc	Prec	Rec	F1
ResNet50+BCE	80.11	67.92	79.43	80.42	79.55
ResNet50+MFL	80.82	**69.33**	**80.30**	80.79	80.54
ResNet50+MCFL(ours)	**82.06**	69.01	77.47	**84.91**	**81.02**

The experiment results on the PA100K dataset

Methods	BackBone	mA	Acc	Prec	Rec	F1
HPNet(ICCV17) [19]	InceptionNet	81.77	76.13	84.92	83.24	84.07
PGDM(ICME18) [11]	CaffeNet	82.97	78.08	86.86	84.68	85.76
RA(AAAI19) [35]	InceptionV3	86.11	–	84.69	88.51	86.56
MsVAA(ECCV18) [22]	ResNet101	84.59	78.56	86.79	86.12	86.46
JLPLS-PAA(TIP19) [29]	–	84.88	**79.46**	87.42	86.33	**86.87**
MT-CAS(ICME20) [33]	ResNet50	83.17	78.78	**87.49**	85.35	86.41
StrongBaseline(2020) [8]	ResNet50	85.19	79.14	87.11	86.18	86.36
MCFL(ours)	ResNet50	**86.84**	78.78	83.68	**89.97**	86.71

- The proposed MCFL obtains the new state-of-the-art performance on RAP and PA100K datasets in terms of both mA and F1 metrics and performs best on the PETA dataset under mA metric with comparable F1 values.
- MCFL significantly outperforms the StrongBaseline by 1.51% and 1.13% on the RAP dataset, 0.61%, and 0.63% on PA100K dataset, 1.65%, and 0.35% on PETA in terms of mA and F1, respectively.

Conclusion

- In this paper, we introduced a method with novel multi-label contrastive focal loss (MCFL) function to improve PAR performance.
- MCFL separates the losses of positive and negative attributes with different weights in order to emphasize the hard samples from minority class.
- the multi-label contrastive loss is proposed to force backbone CNNs to extract more discriminative features.
- Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods in comparison and can achieve better prediction results on test datasets.