Supplementary Material for Label2Label

Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu*

[1] Department of Automation, Tsinghua University, China
[2] Beijing National Research Center for Information Science and Technology, China
{wanhua016,caozx00}@gmail.com, {jfeng,jzhou,lujiwen}@tsinghua.edu.cn

1 Evaluation Metrics

For pedestrian attribute prediction, we adopted five evaluation metrics. We present the details of these metrics. The only label-based metric is the mean accuracy (mA) metric, which is the mean of positive accuracy and negative accuracy for each attribute. Mathematically, the mA is calculated by:

$$mA = \frac{1}{2M}\sum_{j=1}^{M}(\frac{TP_j}{P_j} + \frac{TN_j}{N_j}), \tag{1}$$

where M is the number of attributes, P_j and TP_j represent the numbers of positive samples and correctly predicted positive samples of the j-th attribute respectively, N_j and TN_j are the numbers of negative samples and correctly predicted negative samples of the j-th attribute respectively.

We also consider four example-based metrics: accuracy, precision, recall, and F1 score:

$$\begin{aligned} Acc &= \frac{1}{N}\sum_{i=1}^{N}\frac{|\boldsymbol{Y}_i \cap \boldsymbol{Y}_i'|}{|\boldsymbol{Y}_i \cup \boldsymbol{Y}_i'|}, Prec = \frac{1}{N}\sum_{i=1}^{N}\frac{|\boldsymbol{Y}_i \cap \boldsymbol{Y}_i'|}{|\boldsymbol{Y}_i'|}, \\ Rec &= \frac{1}{N}\sum_{i=1}^{N}\frac{|\boldsymbol{Y}_i \cap \boldsymbol{Y}_i'|}{|\boldsymbol{Y}_i|}, F1 = \frac{2 * Prec * Rec}{Prec + Rec}, \end{aligned} \tag{2}$$

where N denotes the number of samples, $\boldsymbol{Y}_i$ is the positive labels of the i-th sample and $\boldsymbol{Y}_i'$ is the predicted positive values for the i-th sample.

2 Weighting Strategy

For facial attribute recognition and clothing attribute recognition, we follow the common practice which does not utilize the weighting strategy for loss functions. Therefore, we have:

$$\begin{aligned} \mathcal{L}_{mlm}(\boldsymbol{x}) &= \sum_{j=1}^{M} y_j \log(p_j) + (1-y_j)\log(1-p_j), \\ \mathcal{L}_{aqn}(\boldsymbol{x}) &= \sum_{j=1}^{M} y_j \log(l_j) + (1-y_j)\log(1-l_j). \end{aligned} \tag{3}$$

* Corresponding author

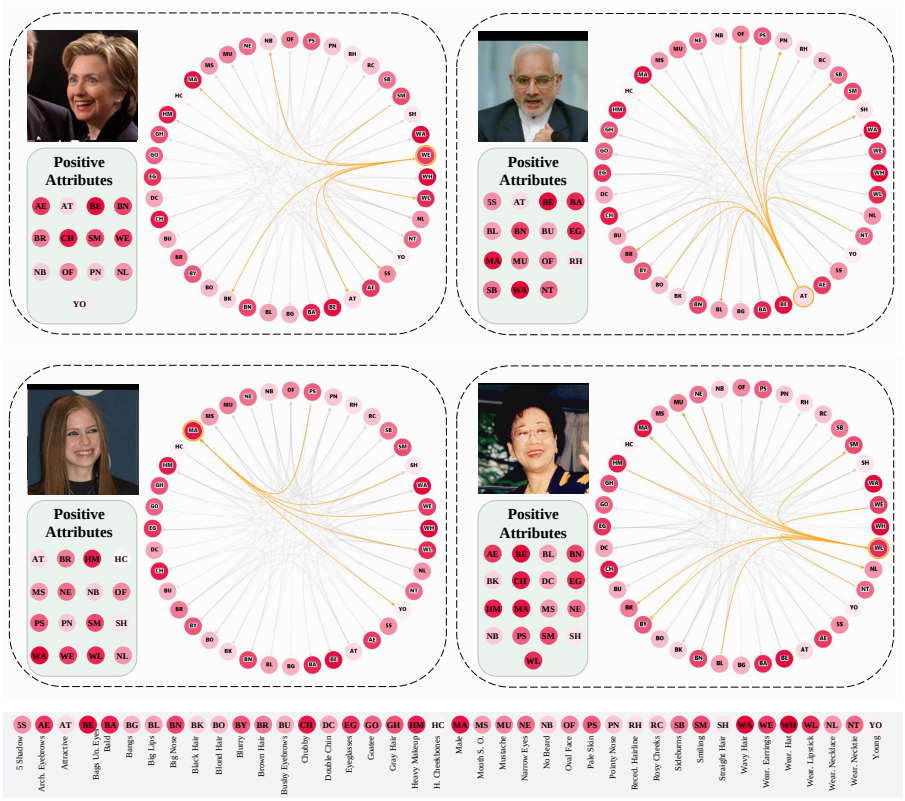

Fig. 1. More visualization results of attention scores in the self-attention layer. We show the attention of the first head at layer 1 with four samples. The positive ground truth attribute labels of each sample are listed in the corresponding bottom-left corner.

For pedestrian attribute recognition, we follow the widely used weighted binary-entropy strategy in [5, 11]. In this way, we have:

$$\begin{aligned} \mathcal{L}_{mlm}(\boldsymbol{x}) &= \sum_{j=1}^{M} w_j (y_j \log(p_j) + (1-y_j)\log(1-p_j)), \\ \mathcal{L}_{aqn}(\boldsymbol{x}) &= \sum_{j=1}^{M} w_j (y_j \log(l_j) + (1-y_j)\log(1-l_j)), \\ w_j &= y_j e^{1-\gamma_j} + (1-y_j)e^{\gamma_j}, \end{aligned} \tag{4}$$

where γ_j is the positive example ratio of the j-th attribute.

Table 1. Ablation experiments on the position embeddings of word representations.

Method	Pos	Error(%)
Label2Label	✗	12.49
	✓	12.51

Table 2. Ablation experiments on the position embeddings of visual features.

Method	Pos	Error(%)
AQN	✗	13.51
	✓	**13.36**
Label2Label	✗	12.98
	✓	**12.49**

3 More Ablation Studies

3.1 Position Embeddings for Word Representations

We conducted ablation experiments on the position embeddings of word representations. Since we are dealing with unordered "sentences", we randomly define three different label sequences and use the corresponding position embeddings respectively. We report the average performance of three different label sequences on the LFWA database in Table 1. We found no additional performance gain from the position embeddings of the word representations. The reason is that our "sentences" are essentially made up of unordered "words".

3.2 Position Embeddings for Visual Features

In our paper, we add 2D-aware position embeddings to visual feature vectors to retain positional information. We conduct experiments to verify their effectiveness and show the results on the LFWA database in Table 2. We observe that introducing position embeddings in visual features is beneficial for performance.

3.3 Comparisons with Transformer-based Multi-label Classification Methods

Many Transformer-based multi-label classification methods [6,7] have been proposed in recent years. To further verify the effectiveness of the proposed method, we conducted experiments on the three datasets used in our paper. Table 3 shows the results. We see our method consistently outperforms C-Tran [6] and Q2L [7], which shows the superiority of our method.

3.4 The Need of Masking

To verify the effectiveness of masking, we construct three pure reconstruction (without masking) baselines. 1) Feature Reconstruction: direct reconstruction of the word features $\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_M$. 2) Score Reconstruction: direct reconstruction of the predicted scores $l_1, l_2, ..., l_M$. 3) Label Reconstruction: direct reconstruction of the labels: $y_1, y_2, ...y_M$. Table 4 shows the results on the LFWA dataset.

Table 3. Comparisons of our method with other Transformer-based methods.

Dataset	LFWA	PA100K			Clothing
Meteic	Error	mA	Accuracy	F1	Accuracy
C-Tran [6]	14.66	81.53	78.97	86.86	90.00
Q2L [7]	13.28	80.72	78.78	86.73	91.81
Ours	**12.49**	**82.37**	**79.03**	**86.96**	**92.87**

Although the Label Reconstruction works competitively, it is still inferior to our method with masking. Just as found in [4], although Autoencoder (reconstruction) works well, the Masked Autoencoder (masking) is the key factor to learning better features. In BERT, the masked word is replaced with the `[mask]` token or a random word. So the MLM has two tasks: mask-recovering and error-correcting. Both increase the training difficulty. In our method, the wrong predictions are like random words in BERT. See Table 4, Ours ($\alpha = 0$) outperforms Label Reconstruction (12.55 vs 12.70). The only difference is that the input of our IC-MLM contains wrong predictions while Label Reconstruction does not, which proves that our proposed IC-MLM also benefits from handling this special "mask".

Table 4. Comparisons with three pure reconstruction baselines.

Method	Reconstruction			Ours (Masking)	
	Feature	Score	Label	$\alpha = 0$	$\alpha = 0.1$
Error(%)	13.45	13.63	12.70	12.55	**12.49**

4 Network Structure Configuration

We show the default hyper-parameters for the Transformer decoder layer of our method in Table 5.

Table 5. Hyperparameters for the Transformer decoder layer of our method.

Component	Hyperparameters
Activation	GELU
Hidden dim	2048
FFN hidden size	2048
Attention heads	4
Attention head szie	512

5 More Visualization Results

We provide more visualization results of the attention scores in Figure 1. We conducted the experiments on the LFWA database. We read out the attention from the self-attention layer of our label decoder. The DODRIO [12] is used for visualization. We show the attention scores of the first head at layer 1 with four examples.

For the first example, the attribute "Wearing Earrings" is strongly related to the existence of "Wearing Lipstick", "No Beard", and "Female" and the absence of "5 o'clock shadow". For the second sample, the attributes "Oval Face", "Pointy Nose", "Sideburns", "Wearing Necktie" and "Male" imply the existence of "Attractive". For the third sample, the attributes "Wearing Earrings" and "Wearing Lipstick" indicate the gender "Female". For the last example, the attribute "Wearing Lipstick" assigns more attention to the existence of "Wearing Earrings", "Wearing Necklace", "Heavy Makeup" and the absence of "Mustache", "Male". We see our method can learn the instance-level attribute relations even if a sample has some wrong labels.

6 Detailed Results

For facial attribute recognition, some methods [2, 10] report the pre-class recognition accuracy. We report the pre-attribute classification error on the LFWA database in Table 6 for a comprehensive comparison.

We observe our method attains very competitive results with a simple framework compared to highly tailored domain-specific methods, which demonstrates the effectiveness of our method.

References

1. Cao, J., Li, Y., Zhang, Z.: Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In: CVPR. pp. 4290–4299 (2018)
2. Han, H., Jain, A.K., Wang, F., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. TPAMI **40**(11), 2597–2609 (2017)
3. Hand, E.M., Chellappa, R.: Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: AAAI (2017)
4. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
5. Jia, J., Chen, X., Huang, K.: Spatial and semantic consistency regularizations for pedestrian attribute recognition. In: ICCV. pp. 962–971 (2021)
6. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: CVPR. pp. 16478–16488 (2021)
7. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)
8. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. pp. 3730–3738 (2015)

Table 6. The classification error (%) obtained by all the competing methods on the LFWA datasets. The accuracy for each attribute obtained by the proposed method is highlighted in bold.

	5 o'clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby
PANDA [13]	16.00	21.00	19.00	20.00	16.00	16.00	27.00	21.00	13.00	6.00	26.00	26.00	21.00	31.00
LNets+ANet [8]	16.00	18.00	17.00	17.00	12.00	12.00	25.00	19.00	10.00	3.00	26.00	23.00	18.00	27.00
NSA [9]	22.41	18.28	19.84	17.38	8.12	9.29	21.03	16.87	7.51	2.53	13.58	19.07	15.74	23.94
MCNN-AUX [3]	22.94	18.22	19.69	16.52	8.06	9.92	20.76	15.02	7.37	2.59	14.77	19.15	15.03	23.14
MCFA [14]	25.00	21.00	23.00	21.00	9.00	11.00	25.00	19.00	9.00	3.00	14.00	23.00	24.00	26.00
PS-MCNN-LC [1]	21.83	16.47	18.16	13.26	7.40	8.55	17.30	13.52	7.04	1.49	12.80	18.13	14.28	21.89
DMTL [2]	20.00	14.00	18.00	16.00	8.00	7.00	23.00	17.00	8.00	3.00	11.00	19.00	20.00	25.00
DMM-CNN [10]	20.82	17.30	18.90	17.30	8.04	8.70	20.18	16.33	8.45	2.83	12.42	18.44	14.67	22.34
Label2Label	**20.76**	**16.67**	**18.28**	**16.10**	**6.93**	**8.06**	**19.40**	**15.00**	**6.96**	**2.25**	**13.00**	**16.88**	**12.89**	**21.61**

	Double Chin	Eyeglasses	Goatee	GrayHair	Heavy Makeup	High Cheekbones	Male	MouthOpen	Mustache	NarrowEyes	NoBeard	OvalFace	PaleSkin	PointyNose
PANDA [13]	25.00	11.00	25.00	19.00	7.00	14.00	8.00	22.00	13.00	27.00	25.00	28.00	16.00	24.00
LNets+ANet [8]	22.00	5.00	22.00	16.00	5.00	12.00	6.00	18.00	8.00	19.00	21.00	26.00	16.00	20.00
NSA [9]	19.51	8.50	16.99	11.54	4.61	11.66	7.40	17.50	7.03	17.25	19.23	23.20	9.03	15.80
MCNN-AUX [3]	18.48	8.70	17.03	11.07	4.15	11.62	5.98	6.49	6.57	17.14	17.85	22.61	6.68	15.86
MCFA [14]	23.00	9.00	20.00	12.00	6.00	15.00	7.00	22.00	9.00	22.00	21.00	26.00	18.00	20.00
PS-MCNN-LC [1]	13.30	7.22	15.89	8.96	3.40	11.23	4.82	15.40	5.53	16.49	17.99	22.10	5.03	12.48
DMTL [2]	22.00	8.00	14.00	12.00	5.00	11.00	7.00	14.00	5.00	18.00	19.00	25.00	9.00	16.00
DMM-CNN [10]	19.02	7.17	17.18	10.62	4.32	11.87	5.86	15.55	5.54	16.33	17.52	23.06	8.14	15.49
Label2Label	**16.24**	**7.38**	**15.34**	**10.13**	**3.88**	**10.36**	**5.78**	**16.38**	**5.89**	**15.52**	**16.45**	**20.26**	**8.47**	**15.09**

	RecedingHairline	RosyCheeks	Sideburns	Smiling	Straight Hair	WavyHair	WearingEarrings	WearingHat	WearingLipstick	WearingNecklace	WearingNecktie	Young		Average
PANDA [13]	16.00	27.00	24.00	11.00	27.00	25.00	8.00	18.00	7.00	14.00	21.00	18.00		18.97
LNets+ANet [8]	15.00	22.00	23.00	9.00	24.00	24.00	6.00	12.00	5.00	12.00	21.00	14.00		16.15
NSA [9]	15.10	12.92	18.24	9.20	21.09	21.72	5.25	9.77	5.93	10.41	18.60	14.32		14.18
MCNN-AUX [3]	13.75	12.08	16.87	8.17	21.47	18.39	5.05	9.93	4.96	10.06	19.34	14.16		13.69
MCFA [14]	15.00	15.00	22.00	12.00	23.00	21.00	7.00	9.00	6.00	11.00	18.00	13.00		16.37
PS-MCNN-LC [1]	12.50	11.19	15.58	7.30	20.35	16.65	4.46	8.79	4.30	9.08	17.82	13.12		12.64
DMTL [2]	15.00	14.00	20.00	8.00	21.00	20.00	6.00	8.00	7.00	9.00	19.00	13.00		13.85
DMM-CNN [10]	13.70	13.56	17.01	7.76	20.80	20.13	5.86	9.16	4.89	10.53	18.72	11.06		13.44
Label2Label	**12.46**	**10.65**	**14.90**	**7.85**	**16.83**	**17.17**	**5.12**	**8.02**	**4.96**	**9.74**	**16.03**	**13.82**		**12.49**

9. Mahbub, U., Sarkar, S., Chellappa, R.: Segment-based methods for facial attribute detection from partial faces. TAC **11**(4), 601–613 (2018)
10. Mao, L., Yan, Y., Xue, J.H., Wang, H.: Deep multi-task multi-label cnn for effective facial attribute classification. TAC (2020)
11. Tang, C., Sheng, L., Zhang, Z., Hu, X.: Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In: ICCV. pp. 4997–5006 (2019)
12. Wang, Z.J., Turko, R., Chau, D.H.: Dodrio: Exploring transformer models with interactive visualization. In: ACL (2021)
13. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: CVPR. pp. 1637–1644 (2014)
14. Zhuang, N., Yan, Y., Chen, S., Wang, H.: Multi-task learning of cascaded cnn for facial attribute classification. In: ICPR. pp. 2069–2074 (2018)