Image ordinal classification with deep multi-view learning

Chao Zhang✉, Xun Xu and Ce Zhu

Image ordinal classification has drawn substantial attention from the research community due to the ordering relation between image categories. Recent advancements towards image ordinal classification lie in applying deep neural networks [convolutional neural network (CNN)]. Nevertheless, the lack of ordinal training data prevents deep models from generalising to testing data. In this work, two multi-view learning approaches are proposed to tackle the insufficient data issue. On one hand, a multi-view ordinal classification with multi-view max pooling (MVMP) approach is proposed, in which each image is randomly blocked with some grids thus creating multiple views of the original data. All views are then used to train multi-view CNN for classification. On the other hand, in order to account for the ordinal relation, the authors propose a double-task learning on MVMP for classification and average pooling for regression. The task of regression benefits that of classification, mainly focusing on improving classification's recognition accuracy. The two proposed approaches are validated on Adience dataset, and show very compelling results. The code and models will be available online.

Introduction: Image ordinal classification aims to predict image's category with ordinal relationship. Examples including age estimation [1] and image quality estimation [2] belong to this scope. Their classification labels are discrete but bear incremental relations between categories. With the proliferation of deep learning, many works have demonstrated that ordinal classification can be improved with state-of-the-art deep neural networks [3–6]. Nevertheless, the limited amount of labelled ordinal data restricts training more complex deep models. To make matters worse, labelling ordinal image requires more extensive exposure to image pairs/triplets which is much more costly than labelling ordinary images, i.e. separating dog and cat.

In this work, we propose an alternative approach towards data augmentation by randomly blocking training images and aggregate different blocked images via multi-view learning. Random blocking is underpinned by the fact that human have no difficulty recognising images with certain patches masked, as shown in Fig. 1.

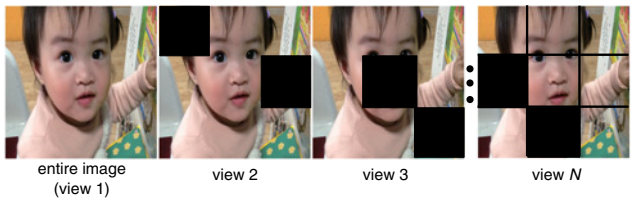

Fig. 1 *Image ordinal classification with multiple views. Human can recognise the age of the girl from different views*

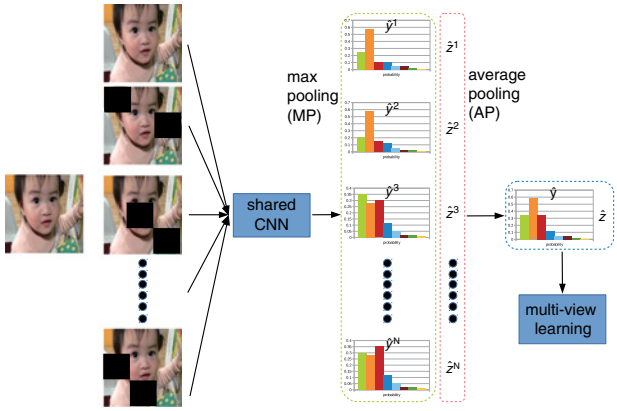

Fig. 2 *Pipline of MVMP and MVAP using max pooling and average pooling (best viewed in magnifier)*

Some previous multi-view learning approaches have been extensively studied [7–9] to exploit data or information from multiple sources. Conventional multi-view learning models are often categorised into co-training [10], multi-kernel/subspace learning [11–13]. In general, they learn multiple classifiers or feature spaces from different domain knowledges, and then jointly aggregate them. While in our model, different blocked images instead of multi-domain knowledges are as multiple views. That is to say, we embed the multi-view aggregation step into deep learning model with only one domain knowledge.

Multi-view max pooling (MVMP) for classification: We consider the problem of learning a mapping f from image feature space $\mathcal{X}$ to label space $\mathcal{Y} \in \mathbb{R}^C$, i.e. $f : \mathcal{X} \to \mathcal{Y}$, where C is the number of the classes. Suppose there are M images in the training set $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_M, \boldsymbol{y}_M)\} \in (\mathcal{X} \times \mathcal{Y})^M$. For each original image $\boldsymbol{x}_m$, we generate N blocked images $\mathcal{X}_m = \{\boldsymbol{x}_m^1, \boldsymbol{x}_m^2, \ldots, \boldsymbol{x}_m^N\}$ by randomly masking, as shown in Figs. 1 and 2. The output of the mapping f for each blocked image is denoted as $\hat{\boldsymbol{y}}_m^i = f(\boldsymbol{x}_m^i), \hat{\boldsymbol{y}}_m^i \in \mathbb{R}^C$.

Next we consider the aggregation of multiple outputs $\hat{\boldsymbol{y}}_m^i$. A naive method is to make a vote for the output set $\{\hat{\boldsymbol{y}}_m^i\}_{i=1\ldots N}$. Instead of counting the vote directly, for each category, we use max pooling on N views. Here class probability $\hat{\boldsymbol{y}}_{mc}^i$ of each blocked image $\boldsymbol{x}_m^i$ is introduced for multi-view aggregation, and which satisfies $\sum_{c=1}^C \hat{\boldsymbol{y}}_{mc}^i = 1$ for all i and m. The aggregated probability $\hat{\boldsymbol{y}}_{mc}$ for each input image $\boldsymbol{x}_m^i$ is achieved by max pooling all N views, as shown in Fig. 2. The aggregated class probability is written as

$$\hat{\boldsymbol{y}}_{mc} = \max_i \{\hat{\boldsymbol{y}}_{mc}^i\}_{i=1}^N \tag{1}$$

After the aggregation, the sum of the aggregated class probability is unequal to 1, i.e. $\sum_{c=1}^C \hat{\boldsymbol{y}}_{mc} \neq 1$ for any m. Here we use $\mathtt{softmax}$ operation to normalise the probability $\hat{\boldsymbol{y}}_{mc}$. Finally, we simply adopt the cross-entropy loss as the training objective. In contrast to conventional convolutional neural network (CNN) training pipeline, we propose multiple views via random blocking and define a novel MVMP training loss by taking all these views' class probability into account. The loss is written as

$$L_{\text{MVMP}} = -\sum_{m=1}^{M}\sum_{c=1}^{C} \boldsymbol{y}_{mc} \log\left(\text{softmax}(\hat{\boldsymbol{y}}_{mc})\right) \tag{2}$$

Multi-view average pooling (MVAP) for regression: Image ordinal classification often refers to two important hints: class category and ordinal score. In above section, MVMP only considers categorical information but without ordinal scores. In this section, we propose MVAP approach in which regression benefits classification for multi-task learning.

The ordinal score $z_m \in \mathbb{R}$ is predicted via a continuous regression mapping, i.e. written as $h : \boldsymbol{x}_m \to z_m$. For N randomly blocked images $\{\boldsymbol{x}_m^i\}_{i=\{1,2,\ldots,N\}}$, N estimated scores $\{\hat{z}_m^i\}_{i=\{1,2,\ldots,N\}}$ are generated by the mapping h. Considering that $\hat{z}_m^i$ is scalar value, we use average pooling to aggregate N outputs for each raw image

$$\hat{z}_m = \frac{1}{N}\sum_{i=1\ldots N} \hat{z}_m^i \tag{3}$$

For the task of score regression, the L2 loss is adopted as

$$L_{\text{MVAP}} = \sum_{m=1}^{M} \|z_m - \hat{z}_m\|_2^2 = \sum_{m=1}^{M} \|z_m - \sum_{i=1\ldots N} \hat{z}_m^i\|_2^2 \tag{4}$$

Max pooling and average pooling are applied to MVMP and MVAP, respectively. However, the former runs max pooling on multiple categorical distributions $\hat{\boldsymbol{y}}_m^i \in \mathbb{R}^C$, and the latter one runs average pooling on multiple scalar values $\hat{z}_m^i \in \mathbb{R}$, as shown in Fig. 2. They both use the idea of multi-view learning, but with different operations.

Since training of regression benefits the ordinal classification, we adopt a multi-task fashion by combining both losses together. The final training objective L_{MVMPAP} is written as

$$L_{\text{MVMPAP}} = L_{\text{MVMP}} + L_{\text{MVAP}} \tag{5}$$

Experimental settings: In this section, we conduct ordinal classification on general CNN, MVMP, and MVMPAP. Experiments are carried out on the challenging Adience dataset [3]. This dataset partitions age interval into eight levels $\{y = n\}_{n=0\ldots 7}$. Clearly classification label y_m is same as regression label z_m in (2) and (4). Roughly 26,000 images taken from 2284 persons are included in Adience dataset. We follow the standard protocol [3] to perform five-fold cross-validation, which are denoted as Cross0, Cross1, Cross2, Cross3, Cross4.

Many previous works [1, 3–5] are evaluted on this dataset for ordinal classification, while few of them are under the same condition and with the same network. The work [3] defines their own specific network structure for ordinal group classification. To make our results reproducible and to be fair, we use VGG-net [4, 5, 14] as the base CNN model. In the ablation study, different methods are all based on the same condition.

In the training stage, we initialise the learning rate with 0.001. We apply an exponential decay function to control the learning rate, i.e. decaying every 5000 steps with a base of 0.5. For all models, we freeze some beginning layers from `Conv1-1` to `Conv2-2`. Each image is divided into 5×5 grids with the equal size. The proportion of the dropped grids is 25%, which diversifies Adience dataset $C_{5\times5}^{[5\times5\times0.25]}$ times. We set the number of multi-view instances as 8, i.e. $N = 8$. The is to say, MVMP and MVAP take eight randomly blocked images as input for each original training image. For general CNN, the batch size is 64, and total epoch number is 150. For MVMP and MVAP, the batch size is 8, and the total epoch number is 45. The VGG-Net inherits pre-trained parameters from ImageNet.

In the testing stage, blocking operation is clearly not necessary. In order to use multi-view learning, we also use eight views: four corners and four edge-corners of the testing image. This approach often occurs in image pre-processing for data augmentation.

Experimental results and discussion: To demonstrate the effectiveness of the proposed methods, two comparisons are given in this section. On one hand, two groups of ablation study are carried out under the same setting. On the other hand, we compare the proposed methods with the state-of-the-arts. In fact, our methods are versatile, which also can be embedded into these state-of-the-art works. The results on ablation study and state-of-the-arts are shown in Table 1 and 2, respectively.

Table 1: Results of general CNN, grid-CNN, MVMP, and MVMPAP

Models, %	Cross0	Cross1	Cross2	Cross3	Cross4	Mean
CNN	61.17	41.66	57.83	49.68	51.45	52.36
grid-CNN	62.44	44.31	57.80	49.08	53.51	53.43
MVMP	66.52	52.24	62.44	54.08	**58.35**	58.73
MVMPAP	**68.16**	**54.66**	**63.88**	**57.24**	57.78	**60.34**

Bold values indicate best results for different methods on different datasets.

Table 2: Comparison with existing methods

Models, %	Mean
LBP+FPLBP [1]	45.10
CNN [3]	50.70
cumulative attribute [4] + CNN	52.34
L-w/o-hyper [5] (pretrained on MS-Celeb-1M)	49.46
L-w/o-KL [5] (pretrained on MS-Celeb-1M)	54.52
full model [5] (pretrained on MS-Celeb-1M)	56.01
MVMP	58.73
MVMPAP	**60.34**

Bold values indicate best results for different methods on different datasets.

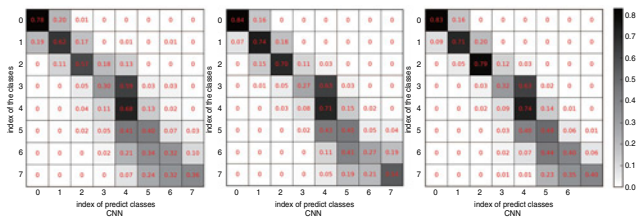

Fig. 3 *Confusion matrix for image ordinal classification (best viewed in magnifier)*

For all cross-validation splits in Table 1, we observe that both MVMP and MVMPAP outperform the general CNN with a large margin (5–13%). This suggests that random masking combined with multi-view learning is a robust approach towards ordinal image classification. Moreover, the task of classification and regression are complementary to each other and benefit the learning when optimised jointly. In order to further exploit the result, the confusion matrix of three methods on Cross0 are given in Fig. 3. It can be observed that by taking multi-view learning into consideration, the performance can be improved accordingly. This greatly shows our proposed methods' robustness.

In order to further show the effectiveness of multi-view learning, another ablation study is also given. The grid-CNN randomly blocks some grids for each training image in the same way. For each training image in each iteration, one blocked image instead of multiple view images, as the input, is used to train the deep model. For each raw image it does not combine different views. As shown in the Table 1, both MVMP and MVMPAP have better performance than grid-CNN. To be sure, multi-view learning really plays a large role in aggregating different views.

Besides implementing ablation study, we also compare multi-view learning approaches with some state-of-the-art methods. As shown in Table 2, MVMPAP achieves the best of all performance. More importantly, our model pre-trained on general ImageNet dataset is able to outperform the work [5] that is pre-trained on the facial dataset (MS-Celeb-1M).

Conclusion: This Letter proposes a multi-view learning approach that randomly dropouts some grids in the training image, and then aggregate these blocked images. The prediction of each training image is jointly determined by multi-view learning on multiple blocked images. In experiments, we implement ablation study and give a comparison with state-of-the-art methods, showing very competitive results.

Submitted: *26 April 2018* E-first: *20 September 2018*
doi: 10.1049/el.2018.5101
One or more of the Figures in this Letter are available in colour online.

Chao Zhang and Ce Zhu (*School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, People's Republic of China*)

✉ E-mail: galoiszhang@gmail.com

Xun Xu (*Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, Singapore*)

Chao Zhang: Also with Sichuan Police College, Luzhou, People's Republic of China

References

1 Eran, E., Roee, E., and Tal, H.: 'Age and gender estimation of unfiltered faces', *IEEE Trans. Inf. Forensics Sec.*, 2014, **9**, (12), pp. 2170–2179
2 Ding, Y., Deng, R., and Shang, X.: 'Image quality assessment employing joint structure-colour histograms as quality-aware features', *Electron. Lett.*, 2017, **53**, (25), pp. 1644–1645
3 Gil, L., and Tal, H.: 'Age and gender classification using convolutional neural networks'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, Boston, MA, USA, June 2015
4 Ke, C., Shaogang, G., Tao, X., *et al.*: 'Cumulative attribute space for age and crowd density estimation'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, June 2013
5 Yunxuan, Z., Li, L., Cheng, L., *et al.*: 'Quantifying facial age by posterior of age comparisons'. ArXiv preprint arXiv:1708.09687v2, 2017
6 Zhang, C., Zhu, C., Xiao, J., *et al.*: 'Image ordinal classification and understanding: grid dropout with masking label'. Int. Conf. on Multimedia and Expo (ICME), San Diego, CA, USA, July 2018
7 Zhao, J., Xie, X., Xu, X., *et al.*: 'Multi-view learning overview: recent progress and new challenges', *Inf. Fusion*, 2017, **38**, pp. 43–54
8 Sun, S.: 'A survey of multi-view machine learning', *Neural Comput. Appl.*, 2013, **23**, (7-8), pp. 2031–2038
9 Xu, C., Tao, D., and Xu, C.: 'A survey on multi-view learning'. Int. Conf. Machine Learning (ICML), Montreal, QC, Canada, June 2009
10 Blum, A., and Mitchell, T.: 'Combining labeled and unlabeled data with co-training'. COLT, Madison, WI, USA, July 1998
11 Chaudhuri, K., Kakade, S.M., Livescu, K., *et al.*: 'Multi-view clustering via canonical correlation analysis'. Int. Conf. Machine Learning (ICML), Montreal, QC, Canada, June 2009
12 Diethe, T., Hardoon, D.R., and Shawe, T.J.: 'Multiview fisher discriminant analysis'. Neural Information Processing Systems (NIPS) Workshops, Vancouver, BC, Canada, September 2008
13 Xu, X., Cheong, L.F., and Li, Z.: 'Exploiting multiple geometric models for motion segmentation'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, June 2018
14 Karen, S., and Andrew, Z.: 'Very deep convolutional networks for large-scale image recognition'. ArXiv preprint arXiv:1409.1556, 2014