Image ordinal classification with deep multi-view learning

Chao Zhang\textsuperscript{st}, 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 insufficiency of 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.

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 $X$ to label space $Y \subseteq \mathbb{R}^c$, i.e. $f: X \rightarrow Y$, where $c$ is the number of the classes. Suppose there are $M$ images in the training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\} \subseteq (X \times Y)^M$. For each original image $x_m$, we generate $N$ blocked images $x_m' = \{(x_1', \ldots, x_N')\}$ by randomly masking, as shown in Figs. 1 and 2. The output of the mapping $f$ for each blocked image is denoted as $y'_m = f(x'_m) \in \mathbb{R}^N$.

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

$$\hat{y}_m = \max_i (\hat{y}'_m)_i$$

After the aggregation, the sum of the aggregated class probability is unequal to 1, i.e. $\sum_{m=1}^{M}\hat{y}_m \neq 1$ for any $m$. Here we use softmax operation to normalise the probability $\hat{y}_m$. 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 loss by taking all these views’ class probability into account. The loss is written as

$$L_{MVMP} = -\sum_{m=1}^{M} \sum_{c=1}^{C} y_{cm} \log (\text{softmax}(\hat{y}_m))$$

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 $h: X \rightarrow \mathbb{R}$. For $N$ randomly blocked images $\{x'_m, m = 1, \ldots, N\}$, $N$ estimated scores $\{z'_m, m = 1, \ldots, N\}$ are generated by the mapping $h$. Considering that $z'_m$ is scalar value, we use average pooling to aggregate $N$ outputs for each raw image

$$z'_m = \frac{1}{N} \sum_{n=1}^{N} z'_{nm}$$

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

$$L_{MVAP} = \sum_{n=1}^{M} \|z'_m - z_m\|^2 = \sum_{n=1}^{M} \|z'_m - z'_m\|^2$$

Max pooling and average pooling are applied to MVMP and MVAP, respectively. However, the former runs max pooling on multiple categorical distributions $p'_m \in \mathbb{R}^N$, and the latter one runs average pooling on multiple scalar values $z'_m \in \mathbb{R}$, 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 $L_{MVMPA}$ is written as

$$L_{MVMPA} = L_{MVMP} + L_{MVAP}$$

**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 $[y_1 = y_{100}, \ldots, y_8 = y_{1000}]$. 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 evaluated on this dataset for ordinal classification, while few of them are under the same condition and with the same network. The work [3] defines its 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_{5×5,0.25} N times. We set the number of multi-view instances as 8, i.e. N = 8. The is to say, MVMP and MVMPAP 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 can also 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.

**Conclusion**: This Letter proposes a multi-view learning approach that randomly drops out 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.

© The Institution of Engineering and Technology 2018

One or more of the Figures in this Letter are available in colour online.

Chao Zhang and Ce Zha (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**