Rank-based ordinal classification

Joan Serrat, Idoia Ruiz

Computer Vision Center and Compt. Science Dept.
Universitat Autònoma de Barcelona, Spain

ICPR 2020
Ordinal classification

- Nominal (mainstream) classification: categories are considered independent, unrelated
- Ordinal classification: categories follow a certain relative order

Applications

- Assessment of aesthetic quality of an image: very bad < flawed < ordinary < professional < exceptional
- Age prediction from face images: 0–2 < 3–6 < ... < +60 years, or one class per year
- Stage of a progressive illness in medical imaging: mild nonproliferative retinopathy < moderate < severe < proliferative
- Building damage assessment from satellite images: no damage < moderate < severe < destruction
- Monocular depth estimation: 0–1.5 < 1.5–5 < 5–10 < +10 meters
- ...
We claim not all ordinal classification problems are the same:

1. Categories are the result of the quantization of a continuous measure (distance, years) → Minimize difference between groundtruth and predicted labels
2. Distance among categories is unknown → Difference between numeric class labels is arbitrary and probably suboptimal as a loss

- How much worse is predicting a building is destroyed when the groundtruth is severe damage?
- What’s the distance between a professional photo and a flawed one?
Idea

- Predict a ranking or order of all the ordinal classes, from most to least probable
- Propose a new ordinal classification loss that does not need to define a distance between classes because it compares groundtruth vs predicted rankings
- Enforce both the accuracy and consistency of prediction: the order of the classes corresponds to some unimodal distribution, which mode is the groundtruth class.

Groundtruth class: 3

Predicted ranking [3,4,2,5,1,6] consistent  [3,4,5,6,2,1] consistent  [3,6,2,4,5,1] not consistent
Method

Goals:

1. Convert the logits/scores output of the network into ranks of classes (first most, second most, ..., least probable)
2. Define a loss that measures a distance between two rank vectors

Difficulties:

- Ranking, like sorting, is a non differentiable operation
- Metrics to compare rankings (Spearman cross-correlation, Kendall’s tau-beta) are not differentiable
Method : from logits to rank probabilities

Let $C = 4$ be the number of classes and $\bar{s}_i, i \in [1, 4]$ their scores, approximated as the mean of a Gaussian distribution $p(s_i) = \mathcal{N}(s_i | \bar{s}_i, \sigma^2)$ for $\sigma = 0.2$

$$P = [p_j(r)], \; j, r \in [1, C]$$ probability for each score to have each possible rank

Same after 50 iterations of Sinkhorn transform
(See algorithm details in our paper)

Method: rank-based losses for ordinal classification

In classification, the groundtruth is the label of the true class \( l \in [1, C] \), where \( C \) is the number of classes.

There are several possible valid groundtruth rankings \( V(l) \) within \( S_C \), the group of permutations of \( 1 \ldots C \)

\[
V(l) = \{ c \in S_C \mid c_1 = l, \\
l \leq c_i < c_j \text{ if } i < j \text{ and } \\
c_i < c_j \leq l \text{ if } i > j, \forall i \neq j \}
\]

For instance, if \( l = 3, C = 4 \), \( V(l) = \{ [3, 4, 2, 1], [3, 2, 4, 1], [3, 2, 1, 4] \} \) corresponding to the ranks of unimodal 4-tuples of scores with maximum at the third position.

We have designed 3 losses to compare groundtruth and predicted rankings.
Method: rank-based losses for ordinal classification

**One-configuration loss** $L_{oc}(P, l)$. The only valid ranking for the true label $l$ is $c_l = [l, l + 1, l - 1, l + 2, l - 2 \ldots]$

$$L_{oc}(P, l) = \sum_{r=1}^{C} \text{NLL}([p_j(r)]_{j=1\ldots C}, \text{OneHot}(c_l[r]))$$

**All configurations loss** $L_{ac}(P, l)$. The groundtruth ranking follows the rank distributions $P$ we have obtained.

$$L_{ac}(P, l) = \min_{v \in V(l)} L_{oc}(P, v)$$

**Valid pairs loss** $L_{vp}(P, l)$. Let $(i, j, a, b) \in [1, C]^4$ be 4-tuples where $i, j$ are indices in a vector of ranks and $a, b$ classes.

$$L_{vp}(P, l) = -\sum_{i<j \leq l, a>b \atop l \leq i<j, a<b} \log p_i(a)p_j(b)$$
### Results: Adience

<table>
<thead>
<tr>
<th></th>
<th>EMD</th>
<th>SORD</th>
<th>CNN-POR</th>
<th>One config</th>
<th>All configs.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-2 15%</td>
<td>62.2 †</td>
<td>—</td>
<td>59.6 ± 3.6 †</td>
<td>55.3 ± 4.4</td>
<td>55.2 ± 3.7</td>
</tr>
<tr>
<td></td>
<td>53.0 ± 5.3 *</td>
<td>48.8 ± 6.9 *</td>
<td>57.4 ± 5.8 †</td>
<td>59.1 ± 5.2</td>
<td>59.0 ± 3.7</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>0.49 ± 0.05 †</td>
<td>0.55 ± 0.08 †</td>
<td>0.57 ± 0.05</td>
<td>0.56 ± 0.05</td>
</tr>
<tr>
<td></td>
<td>0.76 ± 0.09 *</td>
<td>1.31 ± 0.21 *</td>
<td>—</td>
<td>0.49 ± 0.06</td>
<td>0.49 ± 0.05</td>
</tr>
</tbody>
</table>

† as reported in these papers, single run of the experiment.
— not reported or implemented.
Odd rows VGG16, even rows ResNet18.
Results: MSRA-MM

very relevant 35.5%
relevant 42%
irrelevant 22.5%

Samples for the query “Beach”

<table>
<thead>
<tr>
<th>Query</th>
<th>CNN-POR</th>
<th>One config.</th>
<th>All configs.</th>
<th>Valid pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>MAE</td>
<td>Acc.</td>
<td>MAE</td>
</tr>
<tr>
<td>Baby</td>
<td>50.00</td>
<td>0.636</td>
<td>51.26</td>
<td>0.590</td>
</tr>
<tr>
<td>Cat</td>
<td>52.89</td>
<td>0.598</td>
<td>54.07</td>
<td>0.534</td>
</tr>
<tr>
<td>Beach</td>
<td>51.11</td>
<td>0.596</td>
<td>55.30</td>
<td>0.496</td>
</tr>
<tr>
<td>Fish</td>
<td>66.33</td>
<td>0.355</td>
<td>67.48</td>
<td>0.337</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LeNet and mean of 3 runs like in CNN-POR
Results: Schifanella’s ImageAesthetics

![Image Aesthetics](image)

unacceptable 0.3% flawed 4.3% ordinary 72.4% professional 22.0% exceptional 1.0%

<table>
<thead>
<tr>
<th>Category</th>
<th>EMD</th>
<th>SORD</th>
<th>Valid pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc.</td>
<td>MAE</td>
<td>Acc.</td>
</tr>
<tr>
<td>Nature</td>
<td>71.96*</td>
<td>0.342*</td>
<td>73.59†</td>
</tr>
<tr>
<td></td>
<td>72.06*</td>
<td>0.317*</td>
<td>71.04*</td>
</tr>
<tr>
<td>Animals</td>
<td>66.98*</td>
<td>0.408*</td>
<td>70.29†</td>
</tr>
<tr>
<td></td>
<td>67.17*</td>
<td>0.405*</td>
<td>64.76*</td>
</tr>
<tr>
<td>Urban</td>
<td>70.89*</td>
<td>0.342*</td>
<td>73.25†</td>
</tr>
<tr>
<td></td>
<td>70.64*</td>
<td>0.303*</td>
<td>67.75*</td>
</tr>
<tr>
<td>People</td>
<td>67.97*</td>
<td>0.429*</td>
<td>70.59†</td>
</tr>
<tr>
<td></td>
<td>67.04*</td>
<td>0.421*</td>
<td>65.50*</td>
</tr>
</tbody>
</table>

Top rows VGG16, bottom rows ResNet18. * as computed by our implementation. † as reported in papers. CNN-POR not included because SORD is better in all categories.
Results: AVA

≈ 200 votes per photo, one sample for mode of votes 1...10

New task analogous to mean score regression: predict the most voted score.

<table>
<thead>
<tr>
<th></th>
<th>SOTA</th>
<th>One config.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean score prediction</td>
<td>( \rho_s )</td>
<td>0.64</td>
</tr>
<tr>
<td>Most voted score prediction</td>
<td>Acc. %</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>MAE</td>
<td>—</td>
</tr>
</tbody>
</table>
Results: Building damage assessment

- Pre-event: No visible damage, Moderate damage, 14K, Severe damage, 12K, Destroyed, 3.5K
- Post-event: No visible damage, Moderate damage, 14K, Severe damage, 12K, Destroyed, 3.5K

Accuracy, MAE, Mathew’s corr. coef.
Summary

- New method for ordinal classification that does not depend on the difference/distance between class labels
- Three loss functions that compare groundtruth and predicted rankings, also enforcing consistency in the prediction
- We compare our method with SOTA on three different datasets, achieving similar or better results in all of them
- We tackle a new task on image aesthetics assessment, namely, the prediction of the most voted class
- We present results on a last application, building damage assessment from remote sensing images