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* $< \ldots <$ *+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
  - ...

Motivation

We claim not all ordinal classification problems are the same:

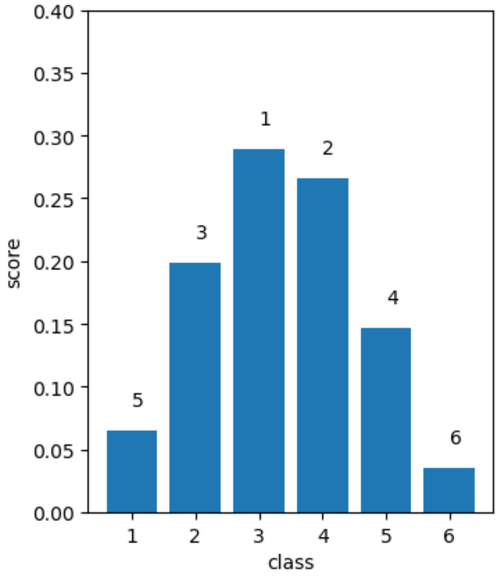
1. Categories are the result of the quantization of a continuous measure (distance, years) $\rightarrow$ Minimize difference between groundtruth and predicted *labels*
2. Distance among categories is unknown $\rightarrow$ Difference between numeric class labels is arbitrary and probably suboptimal as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

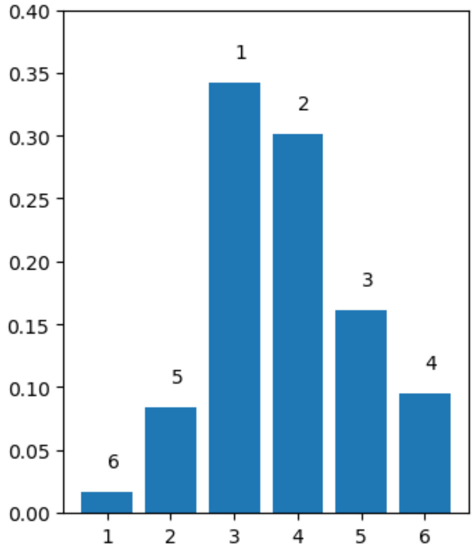
Idea

- 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 grountruth 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.

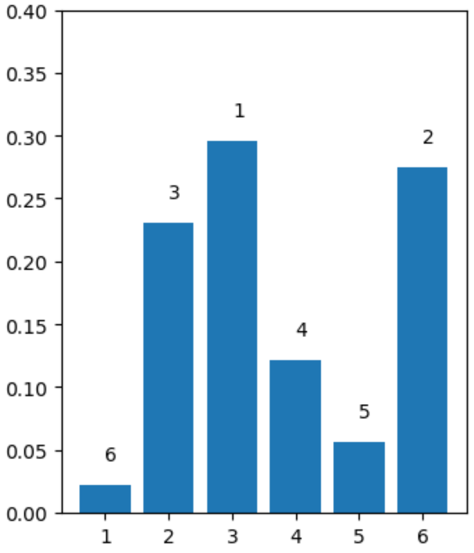
Groundtruth class: 3



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:

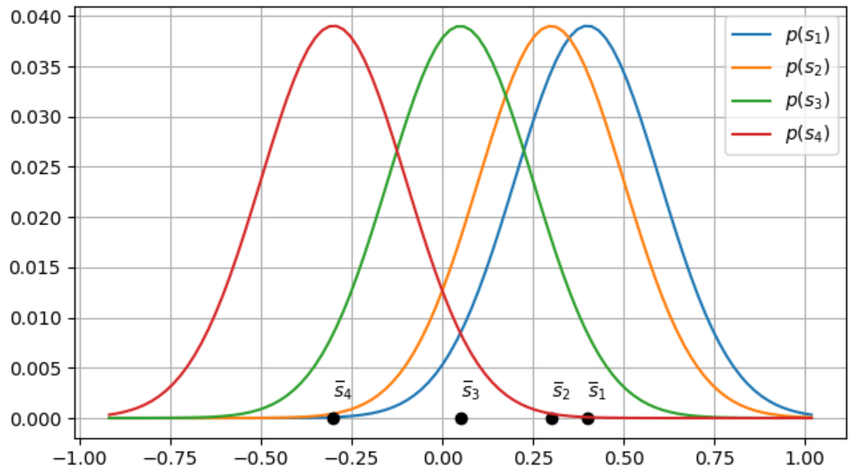
1. Convert the logits/scores output of the network into ranks of classes (first most, second most, . . . , least probable)
2. Define a loss that measures a distance between two rank vectors

Difficulties:

- Ranking, like sorting, is a non differentiable operation
- Metrics to compare rankings (Spearman cross-correlation, Kendall's tau-beta) are not differentiable

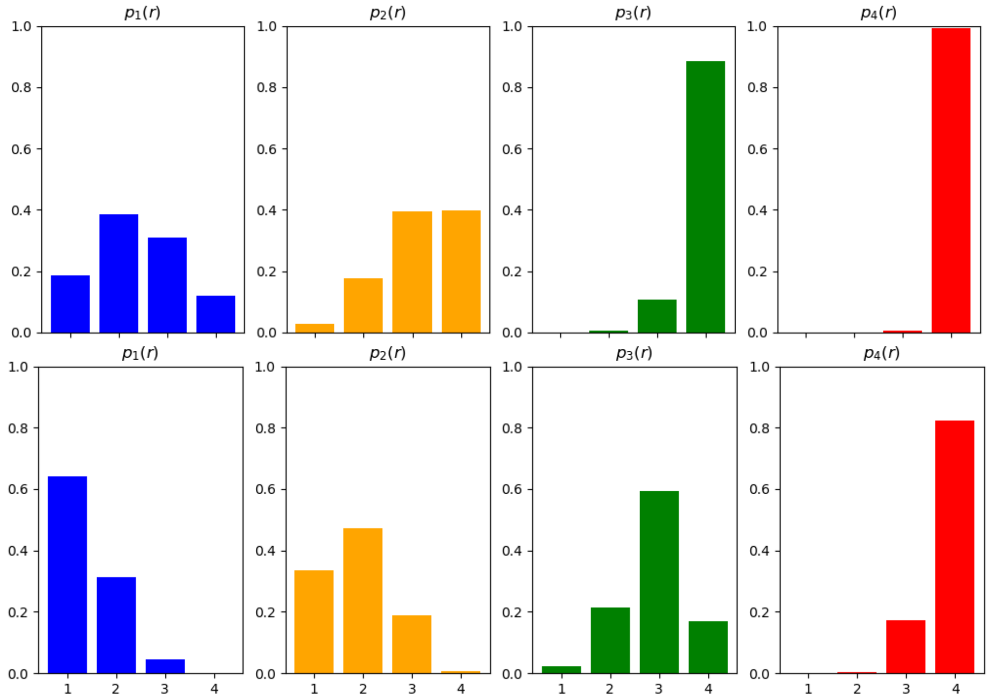
Method : from logits to rank probabilities

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)



[1] M. Taylor, J. Guiver, S. Robertson, and T. Minka. Softrank: Optimising non-smooth rank metrics, in First ACM Int. Conf. on Web Search and Data Mining (WSDM), 2008.

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 $1 \dots C$

$$
\begin{aligned}
V(l) = \{ \quad & c \in S_C \mid c_1 = l, \\
& l \leq c_i < c_j \text{ if } i < j \text{ and} \\
& c_i < c_j \leq l \text{ if } i > j,\ \forall i \neq j \quad \}
\end{aligned}
$$

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

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

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

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

Valid pairs loss $L_{\mathbf{vp}}(P,l)$. 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.

$$L_{\mathsf{vp}}(P,l) = -\sum_{\substack{i<j\leq l, a>b \\ l\leq i<j, a<b}} \log p_i(a)\, p_j(b)$$

0-2	4-6	8-12	15-20	25-32	38-43	48-53	60+
15%	13%	13%	10%	27%	13%	5%	5%

	EMD	SORD	CNN-POR	One config	All configs.
Accuracy	**62.2** † —	59.6 ± 3.6 †	57.4 ± 5.8 †	55.3 ± 4.4	55.2 ± 3.7
	53.0 ± 5.3 ⋆	48.8 ± 6.9 ⋆	—	59.1 ± 5.2	59.0 ± 3.7
MAE	—	**0.49** ± 0.05 †	0.55 ± 0.08 †	0.57 ± 0.05	0.56 ± 0.05
	0.76 ± 0.09 ⋆	1.31 ± 0.21 ⋆	—	**0.49** ± 0.06	**0.49** ± 0.05

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

very relevant
35.5%

relevant
42%

irrelevant
22.5%

Samples for the query "Beach"

Query	CNN-POR		One config.		All configs.		Valid pairs	
	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE
Baby	50.00	0.636	51.26	0.590	**51.51**	0.592	51.35	**0.578**
Cat	52.89	0.598	54.07	0.534	**54.82**	0.536	54.09	**0.530**
Beach	51.11	0.596	**55.30**	0.496	54.85	0.503	55.27	**0.489**
Fish	66.33	0.355	67.48	0.337	66.63	0.337	**68.80**	**0.324**

LeNet and mean of 3 runs like in CNN-POR

Results: Schiffanella's ImageAesthetics

unacceptable 0.3%

flawed 4.3%

ordinary 72.4%

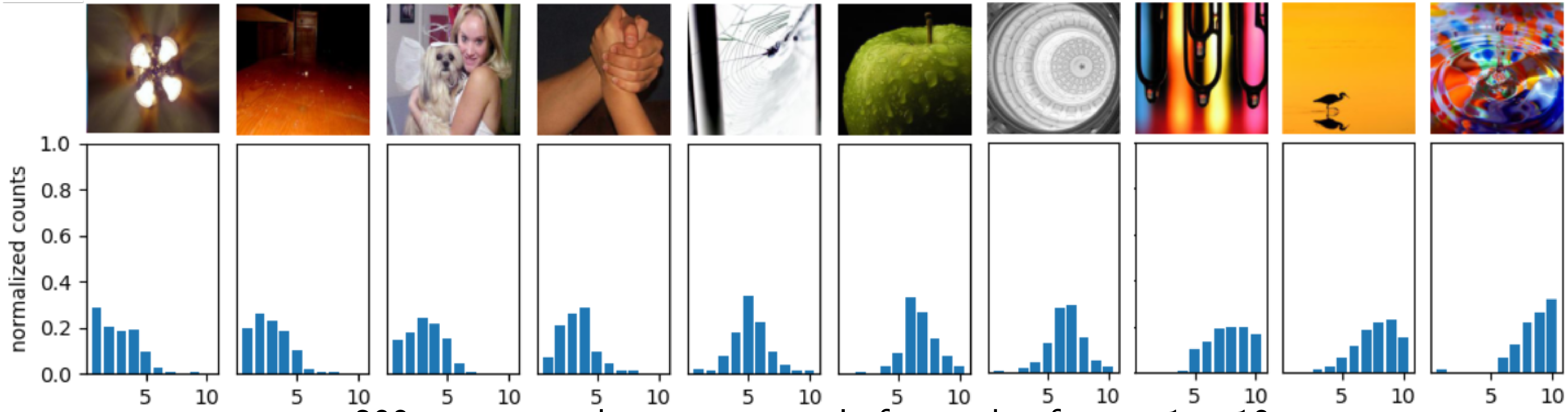
professional 22.0%

exceptional 1.0%

Category	EMD		SORD		Valid pairs	
	Acc.	MAE	Acc.	MAE	Acc.	MAE
Nature	71.96*	0.342*	73.59†	0.271†	**74.77**	**0.261**
	72.06*	0.317*	71.04*	0.381*	**74.95**	**0.260**
Animals	66.98*	0.408*	**70.29**†	**0.308**†	69.32	0.318
	67.17*	0.405*	64.76*	0.555*	**70.07**	**0.310**
Urban	70.89*	0.342*	**73.25**†	**0.276**†	72.98	0.281
	70.64*	0.303*	67.75*	0.498*	**73.41**	**0.276**
People	67.97*	0.429*	70.59†	**0.309**†	**70.73**	**0.309**
	67.04*	0.421*	65.50*	0.571*	**70.79**	**0.307**

Top rows VGG16, bottom rows ResNet18. $\star$ as computed by our implementation. $\dagger$ as reported in papers.
CNN-POR not included because SORD is better in all categories.

Results: AVA



$\approx$ 200 votes per photo, one sample for mode of votes $1 \ldots 10$

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

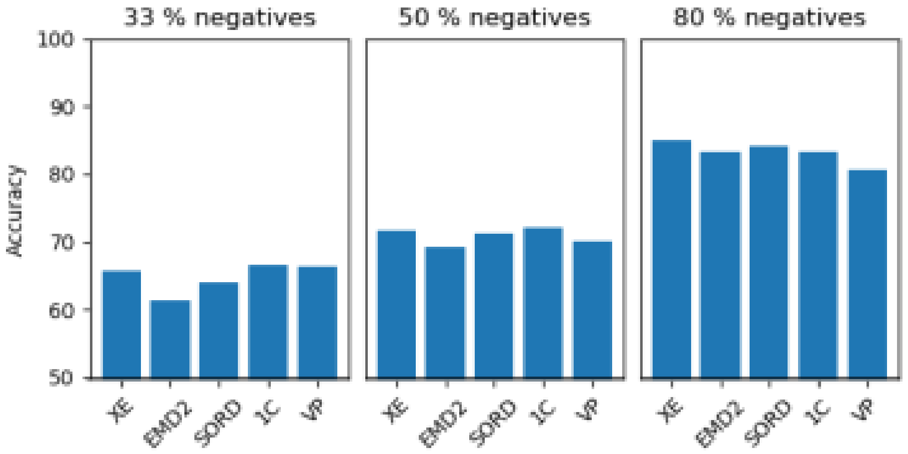
		SOTA	One config.
Mean score prediction	ρ_s	0.64	0.58
Most voted score prediction	Acc. %	—	63.15
	MAE	—	0.41

Results: Building damage assessment

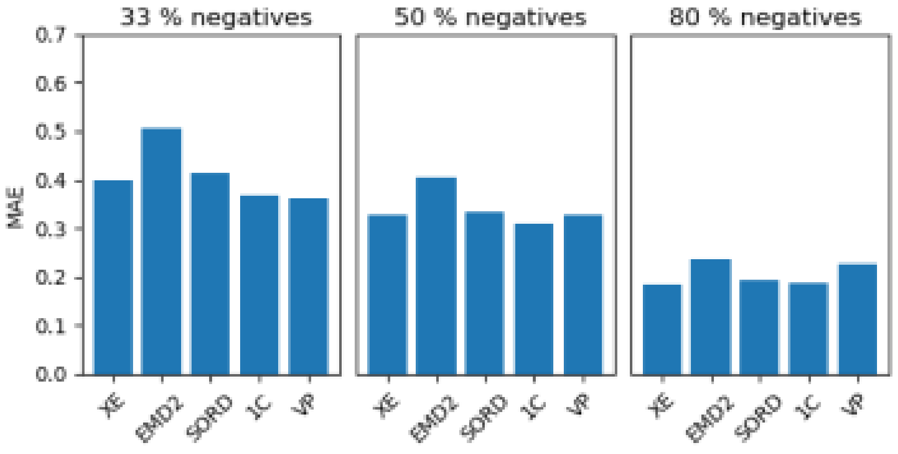
pre-event

post-event

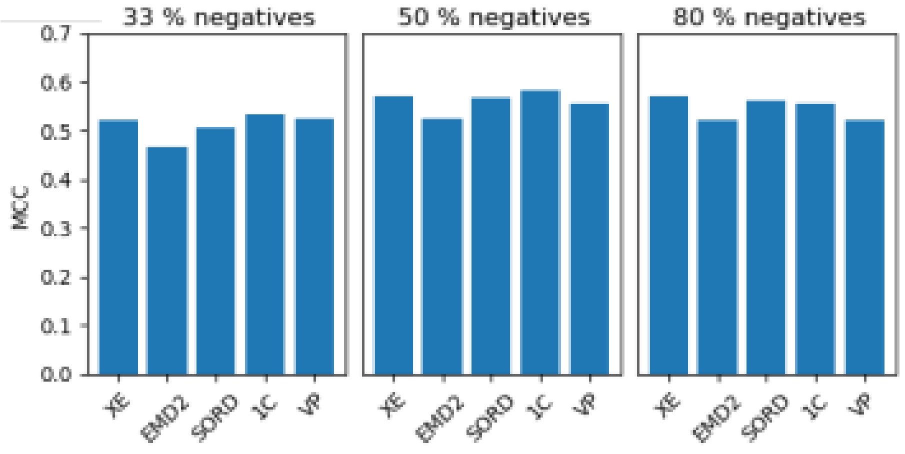
No visible damage | Moderate damage, 14K | Severe damage, 12K | Destroyed, 3.5K



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