DeepPermNet: Visual Permutation Learning

Rodrigo Santa Cruz[1], Basura Fernando[1], Anoop Cherian[1,2], and Stephen Gould[1]
[1]Australian Centre for Robotic Vision, Australian National University, Canberra, Australia
[2]Mitsubishi Electric Research Labs, 201 Broadway, Cambridge, MA

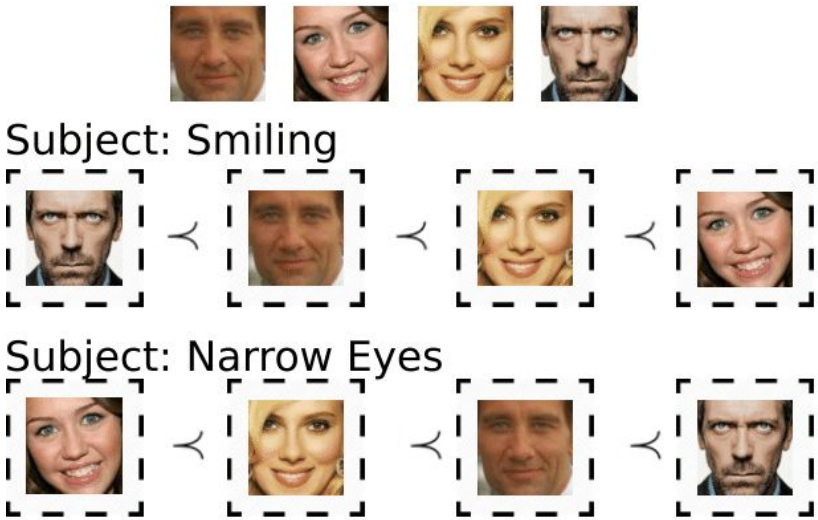
[DeepPermNet: Visual Permutation Learning. Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.]

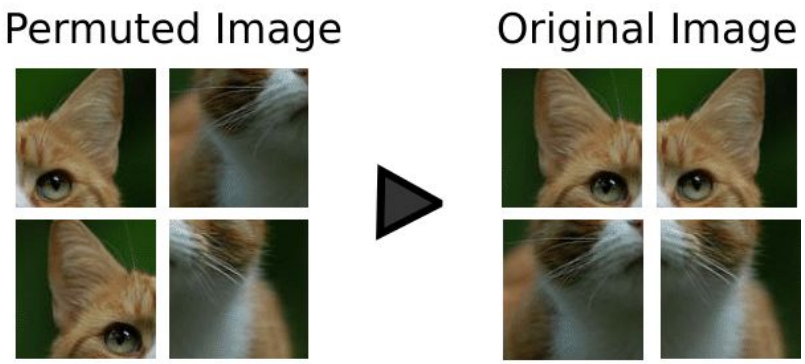
[Visual Permutation Learning. Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould. Pattern Analysis and Machine Intelligence (PAMI)]

Motivation

Consider the following tasks:

- To order a set of images according to a given attribute.
- Given shuffled image patches, can we recover the original image?





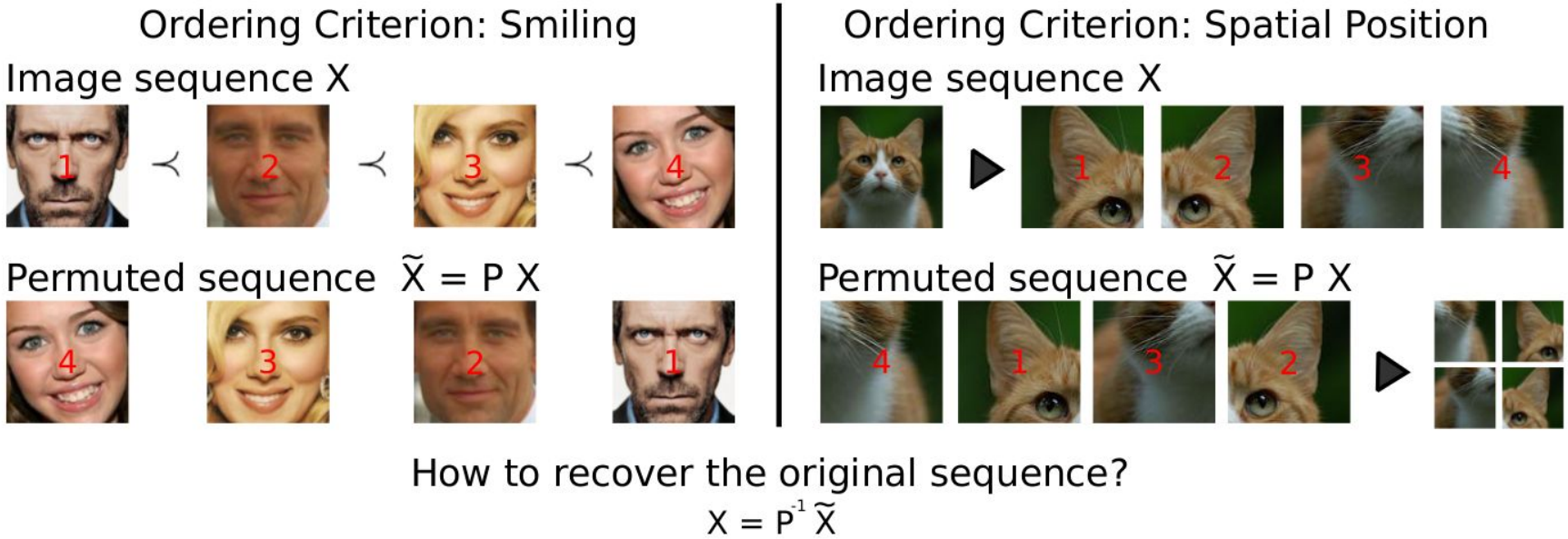
Motivation

- These tasks essentially involve learning how to recover the order, i.e., infer the shuffling permutation.

- Tasks in different fields can be reduced to this problem:
  - Computer graphics: Jigsaw puzzle
  - Biology: DNA and RNA modeling
  - Archeology: Re-assembling relics
  - **Computer Vision: Image Ranking and Self-supervised representation learning.**

- We propose a generic formulation to learn structural concepts in image sequences by predicting the shuffling permutation.

Visual Permutation Learning - Task



We hypothesize that the model trained to solve such task is able to capture high-level semantic concepts, structure and shared patterns in visual data.

Visual Permutation Learning - Learning

- Let us define a training set,

$$\mathcal{D} = \{(X, P) \mid X \in \mathcal{S}^c \text{ and } \forall P \in \mathcal{P}^l\}$$

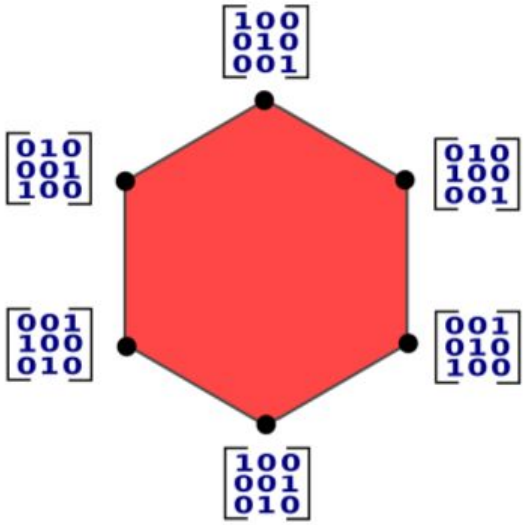
- We propose to learn a function that maps from fixed length image sequence to permutation matrices. Then our permutation learning problem can be described as,

$$\underset{\theta}{\text{minimize}} \quad \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, f_\theta(\tilde{X})\right) + R(\theta)$$

Geometry of Permutation Matrices

- Permutation Matrices form discrete points in the Euclidean space which imposes difficulties for gradient based optimization solvers.

- According to the Birkhoff-von Neumann theorem, the Birkhoff polytope (which is the set of l × l doubly-stochastic matrices), forms a convex hull for the set of l × l permutation matrices.

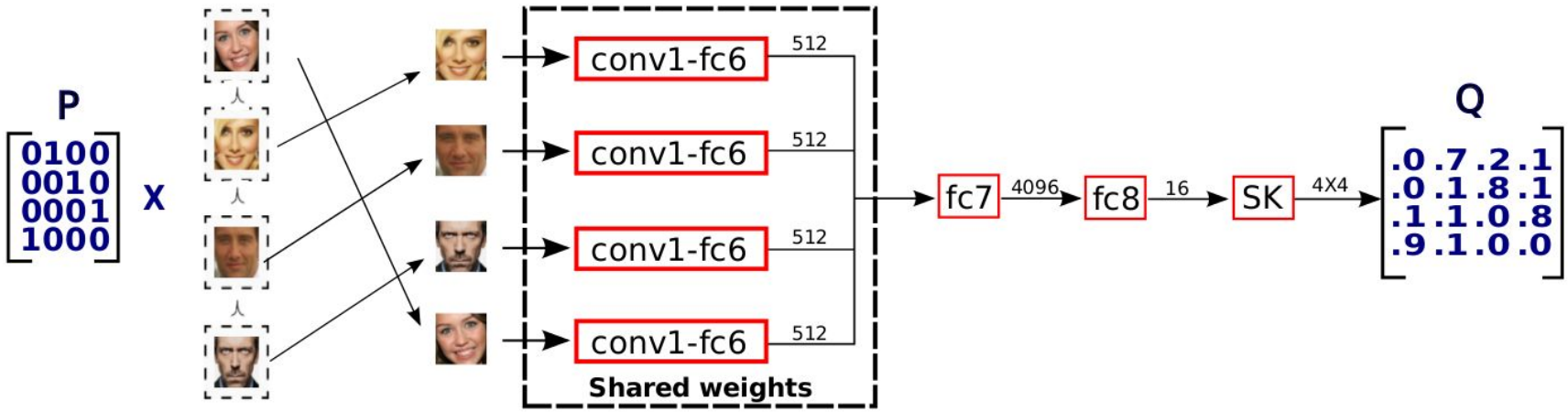


Then, we propose to approximate inference over permutation matrices to inference over their nearest convex-surrogate, the doubly stochastic matrices.

$$f_\theta \quad : \quad \mathcal{S}^c \quad \rightarrow \quad \mathcal{B}^l$$

DeepPermNet - Model

- We also wish to learn the image representation that captures the structure behind our sequences.



- Incorporating the DSM structure in our predictors can avoid the optimizer from searching over imposible solutions.

Sinkhorn Layer

- Sinkhorn's theorem: Any non-negative square matrix can be converted to a DSM by alternating between rescaling its rows and columns to one.
- Function:

$$R_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{i,k}}; \quad C_{i,j}(Q) = \frac{Q_{i,j}}{\sum_{k=1}^{l} Q_{k,j}}$$

$$S^n(Q) = \begin{cases} Q, & \text{if } n = 0 \\ C\left(R\left(S^{n-1}(Q)\right)\right), & \text{otherwise.} \end{cases}$$

- Gradient (Row normalization):

$$\frac{\partial \Delta}{\partial Q_{p,q}} = \sum_{j=1}^{l} \frac{\partial \Delta}{\partial R_{p,j}} \left[\frac{[\![j = q]\!]}{\sum_{k=1}^{l} Q_{p,k}} - \frac{Q_{p,j}}{\left(\sum_{k=1}^{l} Q_{p,k}\right)^2} \right]$$

Bi-level Optimization

Note that this problem can be reformulated as a bi-level optimization problem,

$$
\begin{aligned}
\underset{\theta}{\text{minimize}} \quad & \sum_{(X,P)\in\mathcal{D}} \Delta\left(P, \hat{Q}\right) + R(\theta) \\
\text{subject to} \quad & \hat{Q} \in \underset{\hat{Q} \in \mathbb{R}_+^{n\times n}}{\text{argmin}} \quad \left\| \hat{Q} - f_\theta(\tilde{X}) \right\| \\
& \qquad \text{subject to} \quad \hat{Q}\,\mathbf{1} = \mathbf{1} \\
& \qquad\qquad\qquad\quad \hat{Q}^T\,\mathbf{1} = \mathbf{1}
\end{aligned}
$$

- We refer to "On differentiating parameterized argmin and argmax problems with application to bi-level optimization" by Gould et al. for a detailed explanation about computing gradients of argmin functions.

Visual Permutation Learning - Inference

Finally, we can recover the correctly ordered sequence from a permuted sequence by,

- Solving a approximation problem (or argmax rows/cols)

$$\begin{aligned} \hat{P} \in \underset{\hat{P}}{\operatorname{argmin}} \quad & \left\| \hat{P} - Q \right\|_F \\ \text{subject to} \quad & \hat{P} \cdot \mathbf{1} = \mathbf{1} \\ & \mathbf{1}^T \cdot \hat{P} = \mathbf{1} \\ & \hat{P} \in \{0,1\}^{l \times l} \end{aligned}$$

- Permuting the shuffled image sequence by the inverse permutation

$$X = \hat{P}^T \tilde{X}$$

Visual Permutation Learning - Recap

- Given a set of ordered images S^c according to *c*, we build a data set *D* as,

$$\mathcal{D} = \{(X, P) \mid X \in \mathcal{S}^c \text{ and } \forall P \in \mathcal{P}^l\}$$

- Using *D,* we learn a function (CNN) which maps shuffled image sequences to its DSM matrix employing the **sinkhorn layer** or **bi-level optimization**.

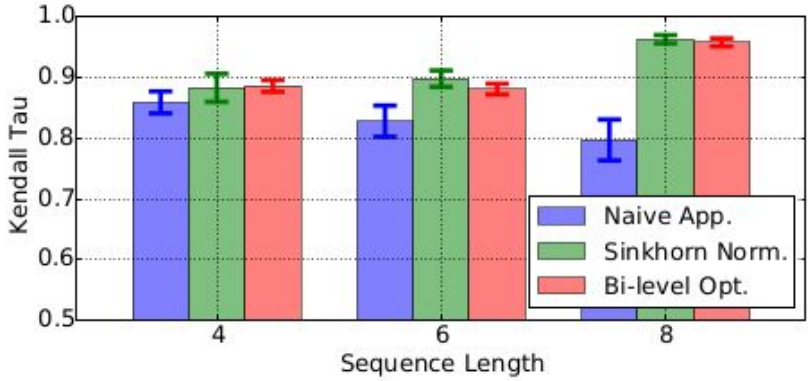
$$f_\theta : \mathcal{S}^c \rightarrow \mathcal{B}^l$$

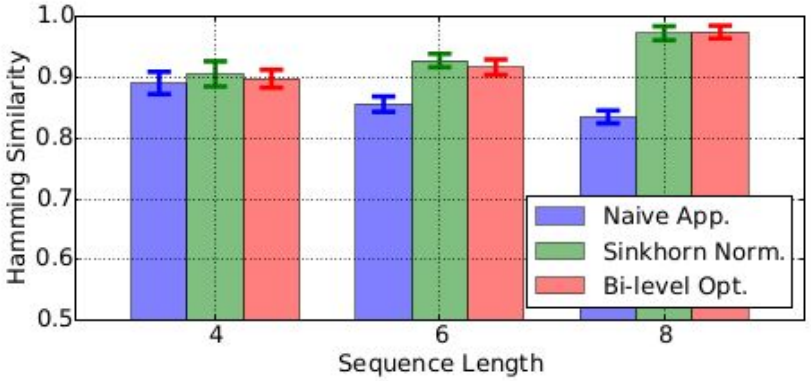
- During test time, we receive a shuffled image sequence and reorder it according to *c* by doing,

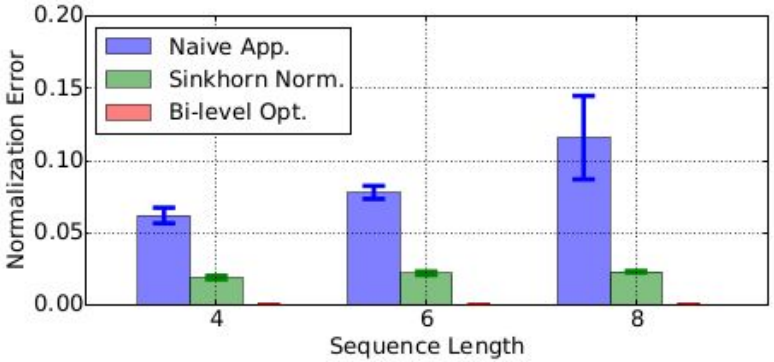
$$\tilde{X} \rightarrow f_\theta(\cdot) \rightarrow \text{Infer P} \rightarrow X = P^T \tilde{X}$$

Experiments - Permutation Prediction

Unpermute 20K shuffled sequences:







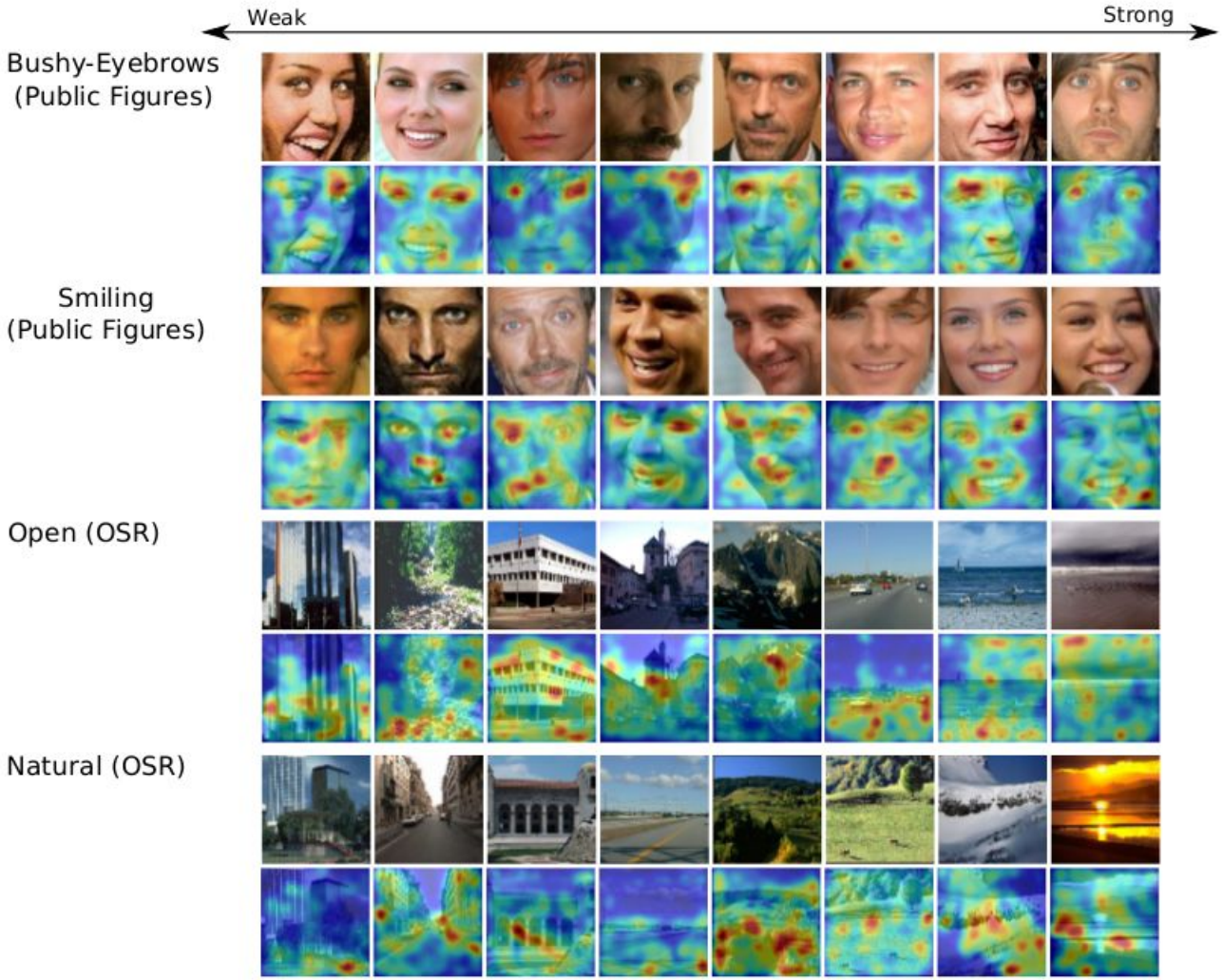
Experiments - Relative Attributes

TABLE 1
Evaluating the proposed model applied to the relative attributes task on the Public Figures Dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Lips	Eyebrows	Chubby	Male	Eyes	Nose	Face	Smiling	Forehead	White	Young	Mean
Parikh and Grauman [59]	79.17	79.87	76.27	81.80	81.67	77.40	82.33	79.90	87.60	76.97	83.20	80.56
Li et al. [46]	81.87	81.84	79.97	85.33	83.15	80.43	86.31	83.36	88.83	82.59	84.41	83.37
Yu and Grauman [82]	90.43	89.83	87.37	91.77	91.40	89.07	86.70	87.00	94.00	87.43	91.87	89.72
Souri et al. [71]	93.62	94.53	92.32	95.50	93.19	94.24	94.76	95.36	97.28	94.60	94.33	94.52
DeepPermNet (Sinkhorn Norm.)	**99.55**	**97.21**	97.66	**99.44**	96.54	**96.21**	99.11	97.88	**99.00**	**97.99**	**99.00**	**98.14**
DeepPermNet (Bi-level Opt.)	99.53	96.65	**98.54**	98.99	**97.21**	94.72	**99.44**	**98.55**	98.77	95.66	98.77	97.89

TABLE 2
Evaluating the proposed model applied to the relative attributes task on the OSR dataset. We report the pairwise accuracy as well as its mean across the attributes.

Method	Depth-Close	Diagonal-Plane	Natural	Open	Perspective	Size-Large	Mean
Parikh and Grauman [59]	87.53	86.5	95.03	90.77	86.73	86.23	88.80
Li et al. [46]	89.54	89.34	95.24	92.39	87.58	88.34	90.41
Yu and Grauman [82]	90.47	92.43	95.7	94.1	90.43	91.1	92.37
Singh and Lee [69]	96.1	97.64	98.89	97.2	96.31	95.98	97.02
Souri et al. [71]	97.65	98.43	**99.4**	97.44	96.88	96.79	97.77
DeepPermNet (Sinkhorn Norm.)	96.09	94.53	97.21	96.65	96.46	98.77	96.62
DeepPermNet (Bi-level Opt.)	97.99	98.21	97.76	97.10	97.21	96.65	97.49
DeepPermNet (Sinkhorn Norm. + VGG16)	96.87	97.99	96.87	**99.79**	**99.82**	**99.55**	**98.48**
DeepPermNet (Bi-level Opt. + VGG16)	**98.12**	**99.92**	98.13	97.78	98.72	97.87	98.42


Weak
Strong
Bushy-Eyebrows (Public Figures)
Smiling (Public Figures)
Open (OSR)
Natural (OSR)

Experiments - Learning to rank

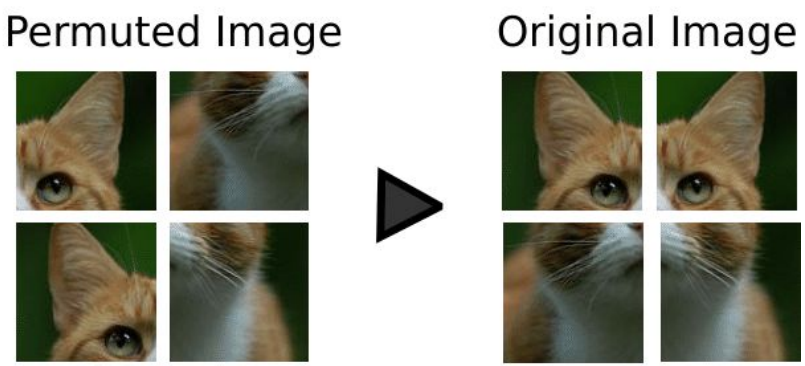
Permutation prediction + Sorting Algorithm

TABLE 3
Evaluation on supervised learning to rank

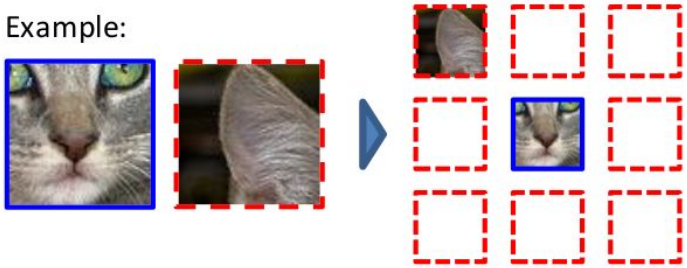
Method	Scene interestingness			Car chronology		
	NDCG	KT	Pair. Acc.	NDCG	KT	Pair. Acc.
Joachims [29]	0.870	0.317	65.8	0.928	0.482	74.1
Xu and Li [64]	0.745	-0.077	46.1	0.827	0.118	55.9
Wu et al. [62]	0.860	0.315	64.3	0.935	0.409	70.6
Cao et al. [8]	0.821	0.118	55.9	0.872	0.291	64.5
Xia et al. [63]	0.862	0.282	64.1	0.854	0.278	63.9
Fernando et al. [19]	0.887	0.347	67.4	.949	0.553	76.9
DeepPermNet (Sinkhorn Norm.)	0.922	0.360	68.0	**0.968**	**0.724**	**86.2**
DeepPermNet (Bi-level Opt.)	**0.923**	**0.363**	**68.2**	0.964	0.700	84.9

Experiments - Self-Supervised Learning

- The main idea is to exploit supervisory signals, intrinsically in the data, to guide the learning process.
- In practice, we define a supervised proxy task, where labels are obtained with almost zero cost, to train the model before finetune for the target task.



[Zhang et al., *ECCV16]*



[Doersch et al., *ICCV 2015]*

Experiments - Self-Supervised Learning

Pre-training Method	Cls.	Det.	Seg.
ImageNet	78.2	56.8	48.0
Random Gaussian	53.3	43.4	19.8
Agrawal et al. [3]	52.9	41.8	-
Doersch et al. [15]	55.3	46.6	-
Wang and Gupta [76]	58.4	44.0	-
Pathak et al. [60]	56.5	44.5	29.7
Donahue et al. [17]	58.9	45.7	34.9
Zhang et al. [83]	65.6	47.9	35.6
Noroozi and Favaro [55]*	67.6	53.2	37.6
Owens et al. [58]	61.3	44.0	-
Bojanowski and Joulin [6]	65.3	49.4	-
Noroozi et al. [56]	67.7	51.4	36.6
Lee et al. [44]	63.8	46.9	-
Pathak et al. [61]	61.0	52.2	-
Zhang et al. [84]	67.1	46.7	36.0
Larsson et al. [43]	65.9	-	38.0
Jenni and Favaro [35]	69.8	52.5	38.1
Gidaris et al. [25]	**72.97**	54.4	39.1
Kim et al. [37]	69.2	52.4	39.3
Nathan Mundhenk et al. [53]	69.6	**55.8**	**41.2**
Ren and Jae Lee [62]	68.0	52.6	-
DeepPermNet (Sinkhorn Norm.)*	69.4	49.5	37.9
DeepPermNet (Bi-level Opt.)*	65.5	45.7	36.4

Conclusion

- We tackle the problem of learning visual concepts from image sequences and introduce a formulation based on permutation matrix prediction.
- We propose novel CNN layers which explores the structure of permutation matrices.
- We show applications on relative attributes, learning-to-rank and Self-supervised representation learning.

DeepPermNet: Visual Permutation Learning

Rodrigo Santa Cruz[1], Basura Fernando[1], Anoop Cherian[1,2], and Stephen Gould[1]
[1]Australian Centre for Robotic Vision, Australian National University, Canberra, Australia
[2]Mitsubishi Electric Research Labs, 201 Broadway, Cambridge, MA

[DeepPermNet: Visual Permutation Learning. Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.]

[Visual Permutation Learning. Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, Stephen Gould. Pattern Analysis and Machine Intelligence (PAMI).]