Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction

Edgar Simo-Serra and Hiroshi Ishikawa
Waseda University, Tokyo, Japan

Waseda University
Department of Computer Science & Engineering

JST CREST

Objective

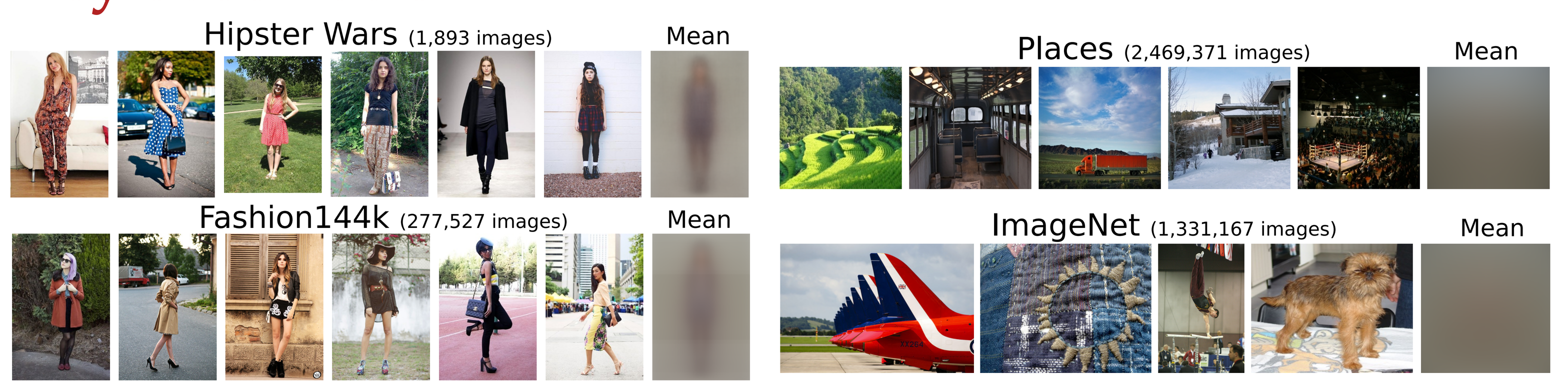
- Learn compact, discriminative representations of images with Convolutional Neural Networks.
- Exploit weak data in the form of incomplete and noisy user-provided tags.
- Optimize for comparisons with L_2 distance.

Main features

- Able to exploit data with **missing and incomplete** tags.
- Obtains compact 128-float representations of **whole images**.
- Adaptable to new datasets **without needing annotating**.
- Outperforms pre-trained features for fashion style prediction.

http://hi.cs.waseda.ac.jp/~esimo/research/stylenet/

Key observation



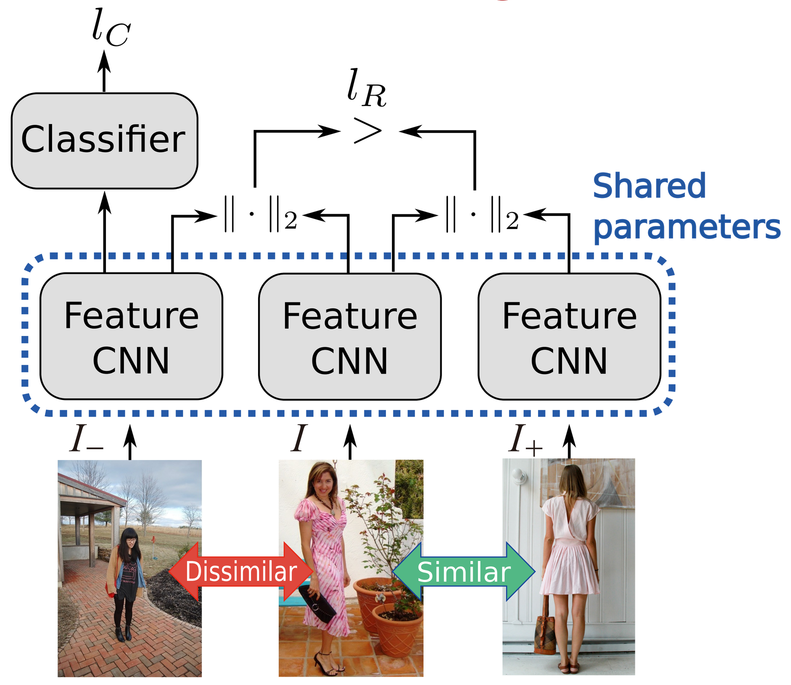
1. Pre-trained imagenet networks limits both architecture and target applications, i.e., images should be similar to imagenet.
2. Lots of data with user-provided tags on the internet. However, these tags are often **incomplete and noisy**. We want to exploit this data to train new networks from scratch.
3. *Problem?* Standard training with classification losses is **not robust to noisy data**.
4. *Solution:* Jointly use a **ranking loss with a classification loss**. The ranking loss allows comparing vectors and is robust to noise, while the classification loss is critical for training.

Problem formulation

1. Consider set of possible tags T
2. Assume dataset of images with noisy labels $\boldsymbol{l} = (l^t)_{t \in T}$ with $l^t \in 0, 1$
3. Define similarity between two images as $r(\boldsymbol{a}, \boldsymbol{b}) = \frac{|\boldsymbol{a} \wedge \boldsymbol{b}|}{|\boldsymbol{a} \vee \boldsymbol{b}|}$

Black-Pants Blue-Shirt Gray-Scarf White-Sweater Black-Bag Yellow-Shoes	Black-Jeans Brick-Red-Sweater Brown-Bag Dark-Brown-Vest Red-Wedges	Aquamarine-Bag Camel-Heels Ivory-Shorts Ivory-Sunglasses Light-Yellow-Hat Turquoise-Blue-Vest	Black-Bag Black-Boots Black-Sweater White-Hat White-Shirt	Black-Boots Gray-Bag Light-Blue-Jeans Light-Brown-Coat White-Sweater Gray-Sweater	Black-Coat Black-Leggings Bronze-Scarf Dark-Gray-Blouse Neutral-Shoes Tawny-Bag

Model & Training



type	kernel size	output size	params
convolution	3×3	384x256x64	1,792
convolution	3×3	384x256x64	36,928
dropout (25%)		384x256x64	
max pooling	4×4	96x64x64	
batch normalization		96x64x64	128
convolution	3×3	96x64x128	73,856
convolution	3×3	96x64x128	147,584
dropout (25%)		96x64x128	
max pooling	4×4	24x16x128	
batch normalization		24x16x128	256
convolution	3×3	24x16x256	295,168
convolution	3×3	24x16x256	590,080
dropout (25%)		24x16x256	
max pooling	4×4	6x4x256	
batch normalization		6x4x256	512
convolution	3×3	6x4x128	32,896
fully-connected		128	393,344
TOTAL		128	1,572,544

Top: Overview of the proposed model.
Right: Architecture of the feature CNN.

Model is trained from scratch using a classification and ranking loss.

Ranking loss: Defined on triplets of images where one is an anchor I, one is similar to the anchor I_+ with $r(I, I_+) > \tau_s$, and one is dissimilar to the anchor I_- with $r(I, I_-) < \tau_d$.

Loss encourages the distance between output of the anchor and similar image $d_+ = d(I, I_+)$ to be smaller than the distance between the output of the anchor and the dissimilar image $d_- = d(I_-, I)$ [5]:

$$l_R(d_-, d_+) = \left(\frac{\exp(d_-)}{\exp(d_-) + \exp(d_+)} \right)^2$$

Classification loss: Auxiliary network used to predict image labels of the dissimilar image X_- with multi-label cross-entropy loss:

$$l_C(X_-, \boldsymbol{y}_-) = \frac{1}{|T|} \sum_{t \in T} l_\times(X_-^t)$$

with $l_\times(x, y) = -x_y + \log(\exp(x_0) + \exp(x_1))$

Joint loss: $L(d_-, d_+, X_-, \boldsymbol{y}_-) = l_R(d_-, d_+) + \alpha l_C(X_-, \boldsymbol{y}_-)$

Implementation

1. Pre-training with classification loss only.
2. Batches formed by selecting anchor images and randomly sampling until similar/dissimilar criterion is met.
3. Optimization with ADADELTA [6].
4. Fine-tuned VGG to remove poor quality images from training can improve performance.

Experimental results

Trained on **Fashion144k dataset** [3] using 80,554 training and 8,948 testing images with $|T| = 3,303$ tags. Evaluation on **Hipsters Wars dataset** [2] with 1,893 images and 5 class labels.

Table 1: Linear classifier evaluated on 100 random 9:1 train-test splits.

feature	params	dim.	acc.	pre.	rec.	iou
Ours Joint	**1.6M**	**128**	**75.9**	**75.4**	**76.5**	**61.5**
Ours Ranking	**1.6M**	**128**	74.5	74.2	74.5	59.6
Ours Siamese	**1.6M**	**128**	73.3	72.9	74.0	58.2
Ours Class.	**1.6M**	**128**	73.5	71.7	74.1	57.3
Ours Joint Dirty	**1.6M**	**128**	72.9	72.1	73.1	57.0
Kiapour et al. [2]†	‡	39,168	70.6	70.6	70.4	54.6
VGG M [1]	99M	4096	71.9	72.9	70.9	56.2
VGG 16 [4]	134M	4096	70.1	70.5	69.7	54.8
VGG M 1024 [1]	86M	1024	70.4	71.1	69.5	54.2
VGG M 128 [1]	82M	**128**	63.5	62.8	63.5	46.3
VGG 16 Places [8]	134M	4096	57.4	57.6	59.4	41.5

Table 2: Similarity search (no train).

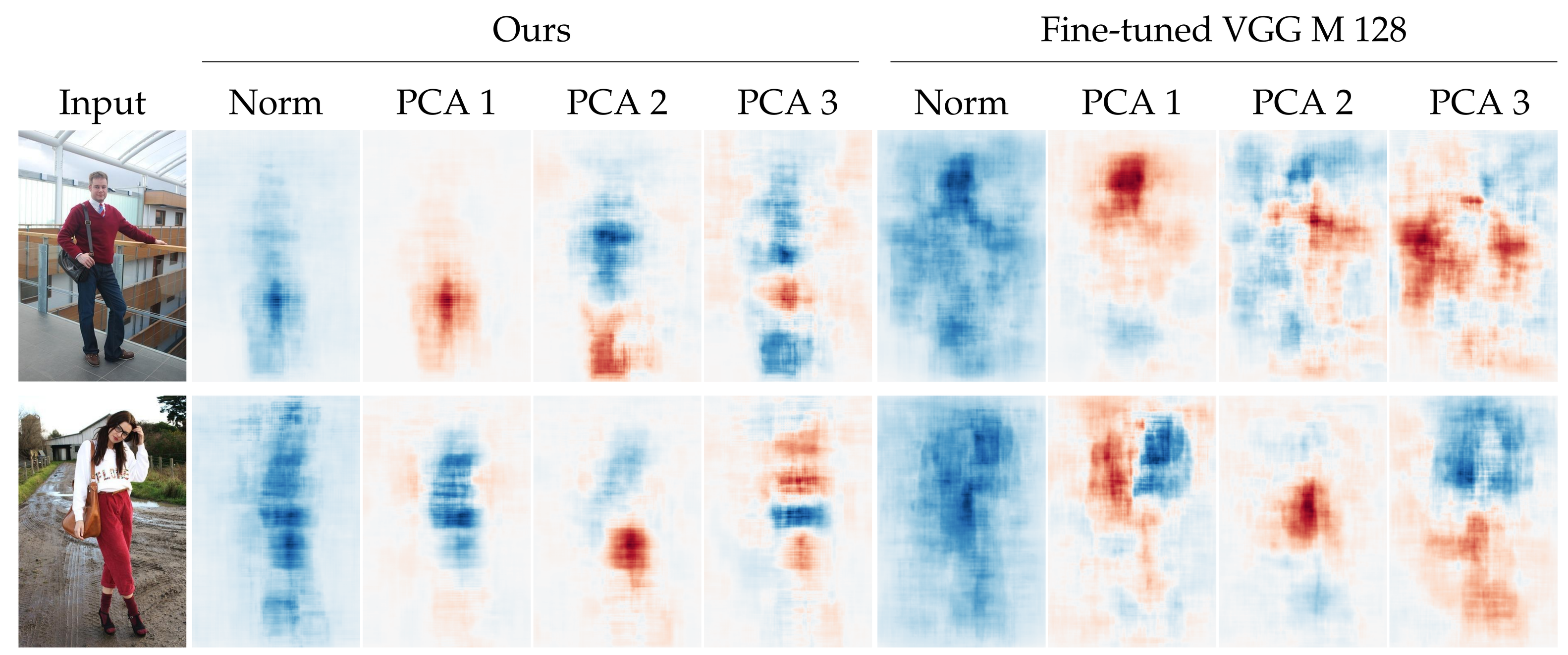
feature	dim.	top-1	top-2	top-3
Ours Joint	**128**	**63.5**	**79.9**	**86.3**
VGG M [1]	4096	53.2	71.7	81.3
VGG 16 [4]	4096	53.2	71.5	80.4
VGG M_128 [1]	**128**	44.6	64.0	76.2
VGG 16 Places [8]	4096	40.1	61.0	72.0

Table 3: Fine-tuning on 1:1 split.

feature	dim.	acc.	pre.	rec.	iou
Ours Joint	**128**	68.4	66.1	67.9	51.0
VGG M	4096	68.4/64.6	67.3/64.2	68.8/63.0	51.8/46.8
VGG 16	4096	63.8/63.3	62.6/62.6	63.5/61.9	46.5/45.4
VGG M 128	**128**	62.6/57.2	60.4/55.1	62.1/56.9	44.5/39.0

Visualizing descriptors

Occlude parts of the image to visualize change [7]. For single descriptors show descriptor norm and PCA basis.



Visually compare the similarity according to the feature CNN $f(\cdot)$ of two images I_1 and I_2 by occluding with a bounding box $B(u, v)$:

$$I_M(u, v) = (f(I_1 \circ B(u, v)) - f(I_1)) \cdot (f(I_2) - f(I_1))$$

References

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.
[2] M. Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. Hipster wars: Discovering elements of fashion styles. In *ECCV*, 2014.
[3] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. In *CVPR*, 2015.
[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
[5] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.
[6] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.
[7] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.
[8] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.