Towards Latent Attribute Discovery from Triplet Similarities
Supplementary Material

Ishan Nigam Pavel Tokmakov Deva Ramanan
Robotics Institute, Carnegie Mellon University
{inigam,ptokmako,deva}@cs.cmu.edu

A. Qualitative analysis for UT-Zappos Shoes

Fig. 6 in the main manuscript visualizes the qualitative trends in PCA embeddings of the learnt LSN embeddings for Zappos-Human shoes. We now adopt a different modus operandi for investigating the learnt latent embeddings. We project each 16-D latent embedding in the learnt LSN model to two dimensions and visually inspect these embeddings in an effort to uncover the low-level features that may have been learnt in the subspaces. Fig. 3 shows the results of this qualitative experiment.

B. Experimental setup details

Below, we describe the datasets used in our study and the implementation details for training the networks.

B.1. UT-Zappos-50k shoes dataset

Yu and Grauman [9] introduced the UT Zappos-50k Shoes Dataset, consisting of 50,025 shoe images along with pairwise human preferences - perceived `comfort`, visual `open`-ness, visual `pointy`-ness, and perceived `sporty`-ness. We refer to this triplet comparison data as Zappos-Human. Fig. 4a in the main manuscript illustrates the general nature of the attributes. Additionally, UT-Zappos Shoes also consists of meta-data labels which have been treated as attribute labels in the study conducted by Veit et al. [8]. The attributes are `type`, `gender`, `heel-height`, and `closing-mechanism`. We refer to this triplet similarity comparison data as Zappos-Meta.

Zappos-Human triplets: UT Zappos-50k provides pairwise annotations of similar and dissimilar shoes for four fine-grained attributes. These pairwise human annotated labels only provide annotations for a few hundred shoes each. Whittle Search [4] uses these similarity labels to learn linear rank-SVMs [2] to generate per-attributed ranked lists. We follow this strategy to obtain scores for the per-attribute linear SVMs to mine triplets for each attribute.

Zappos-Meta triplets: We simply use the triplets released by Veit et al. [8] at link.

Figure 1: Samples from attributed ranked lists generated using rank-SVMs, as described in [4]. We observe that the ranked lists are noisy and may not strictly correspond to the attributes in the dataset.

We note that the attributes are used solely for generating triplets and are not available to our learning method during training. We perform all analysis on human-labeled attributes since this data closely follows the real-world scenario of obtaining weak supervision from the web. The noisy nature of the rank-SVM generated similarity scores parallels the noise observed in webly obtained data (see Figure 1).

B.2. Celeb-A faces dataset

The Celeb-A dataset [6] contains 202,599 face images labeled with 40 binary visual attributes. Fig. 2b illustrates the general nature of a few of these attributes. We select eight visual attributes for ablative analysis - `Eyeglasses`, `Male`, `Smiling`, `Young`, `Attractive`, `Wearing_Lipstick`, `5_o_Clock_Shadow`, and `Bags_Under_Eyes`.

Triplets: The Celeb-A dataset, unlike the UT-Zappos-50k dataset, is exhaustively labeled with attributes. The dense labels allow us to mine triplets based on the presence or absence of attributes in the images.

B.3. Methods

We now describe in detail how we implement the various methods in our quantitative analysis.

Singular Similarity Networks (SSN): SSN is a Resnet-18 network pre-trained on ImageNet [7] with a single embedding for satisfying all notions of similarity. The UT-Zappos

Figure 2: Mean face images for all 40 attributes in Celeb-A Faces. We observe that a number of the attributes are correlated. For example, `Goatee` and `Mustache` are correlated. Surprisingly, `Male` and `Big-Nose` are also visually correlated!

Shoe images are small (resized to 112×112) and (following Veit et al. [8]) we chop off the Resnet-18 network after the last max-pooling layer to accommodate the small shoe images. The Celeb-A Faces images are much larger (resized to 224×224) and all experiments are performed using the standard Resnet-18 network. For all experiments, the smallest possible embedding which does not suffer from overfitting is used to report Supervised-Eval performance. All experiment, except two, utilize a 16-D embedding. The Celeb-A Faces experiment with 8 attributes (Sec. 5.3.2) utilizes a 32-D embedding, and the Celeb-A Faces experiment with all 40 attributes (Sec. 5.4.2) utilizes a 128-D embedding.

Multi-View Triplet Embeddings (MVTE): Amidi and Ukkonen [1] implement their learning algorithm by learning linear classifiers over fixed representations based on Fourier descriptors and color histograms. We provide the MVTE algorithm with better feature representations via Resnet-18 features trained on ImageNet [7]. Thus, our MVTE implementation benefits from better feature representations, while remaining faithful to the original proposed learning algorithm.

Latent Similarity Networks (LSN): The proposed LSNs follow the same network architecture as SSN, apart from learning multiple linear subspace projections on the singular embedding which forms the ultimate SSN layer. LSNs are end-to-end trained using an adapted form of stochastic MCL [5]. We note that the MVTE and LSN methods are identical in terms of network architecture and differ in two major ways: **(1)** LSNs are end-to-end trained while MVTE learns linear classifiers over fixed Resnet-18 features learnt on ImageNet [7], and **(2)** the LSN learning algorithm relies on hard label assignment for each triplet sample, while MVTE relies on soft label assignment.

Conditional Similarity Networks (CSN): The fully supervised CSNs [8] follow the same network architecture as MVTE and LSN. CSNs benefit from the added supervision of learning from each triplet sample conditioned on knowing the underlying latent attribute. We use the implementation provided by Veit et al. [8] for reporting the quantitative performance for CSNs.

B.4. Implementation details

The proposed Latent Similarity Network architecture consists of a Resnet-18 [3] encoder pre-trained on Imagenet [7]. Following [8], we resize UT Zappos-50k images to 112×112 and remove the final max-pool layer in the encoder to accommodate the smaller image size. Celeb-A images are resized to 224×224 to be loaded into the encoder. A final fully-connected layer is added to the encoder, which

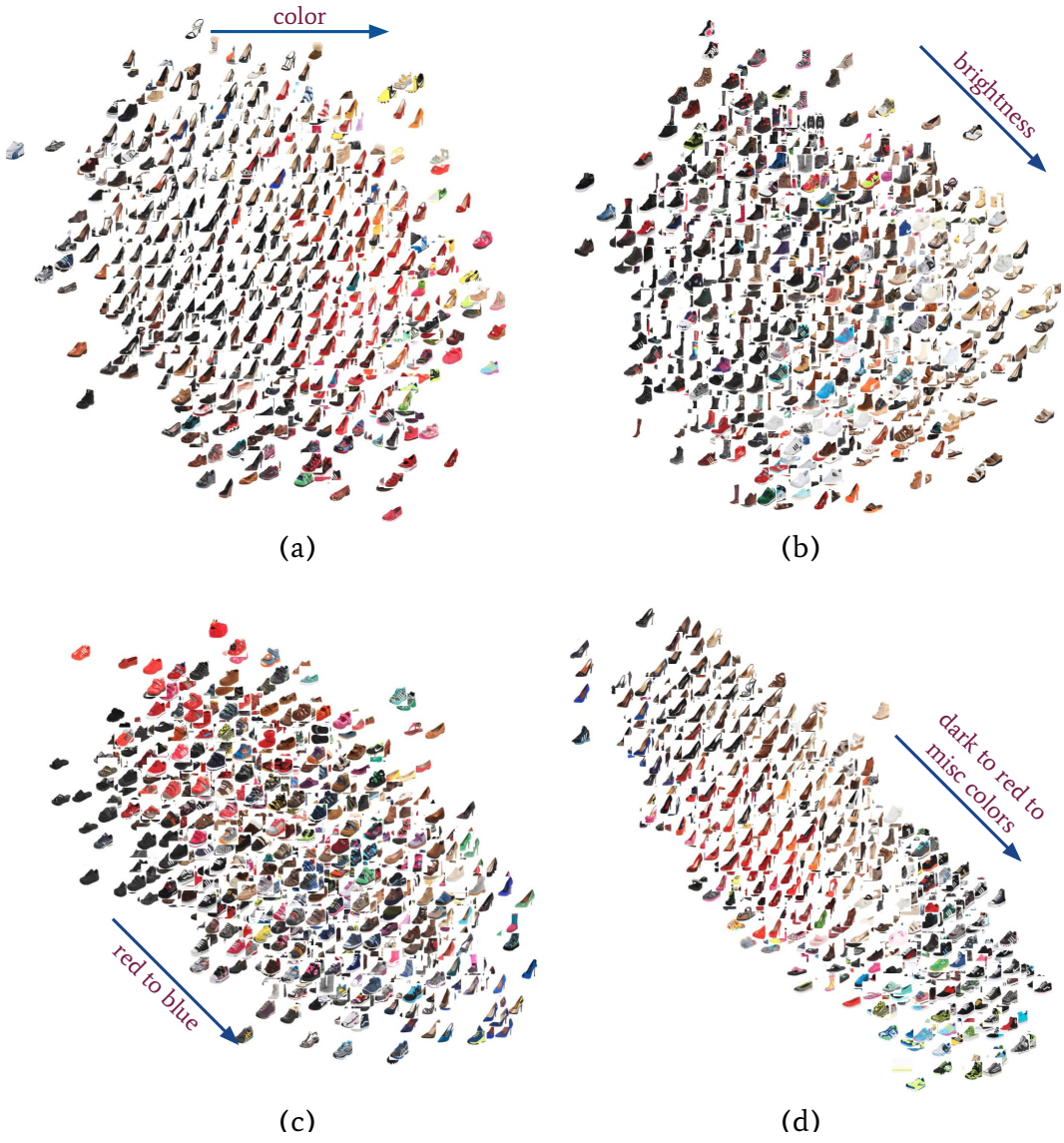

Figure 3: Qualitative Analysis of 2D projections of learnt embeddings for the UT Zappos-50k Shoes dataset: **(a)** A subset of the `comfort` attribute subspace learns to distinguish between colors: it is possible that humans perceive colorful shoes to be more comfortable, **(b)** A subset of the `open` attributes learn to reason about brightness of shoes: it is possible that humans find bright shoes to be more open, **(c)** A set of dimensions in the `pointy` attribute subspace embedding reasons about red versus blue colors, probably due to the fact that a number of pointy shoes in the dataset are red stilettos, and **(d)** A subset of the `sporty` attribute embeddings learn to reasob about the shade of the color of the shoe.

serves as the universal embedding for the networks used in the study. All experiments are performed using a universal embedding dimension of 16. LSNs also include a linear subspaces which are learnt on the universal embedding. We learn the linear subspaces as 16-dimensional projections of the universal embedding, as experiments with 32 or 48 or 64-dimensional projections provided similar performance. The subspaces are initialized from a normal distribution and learned in an end-to-end fashion. The models are trained using Stochastic Gradient Descent with an initial learning rate of 5^{-6}. The loss hyperparameters penalizing the magnitudes of the universal embedding and the linear subspace embeddings are $\lambda_1 = 5^{-3}$ and $\lambda_2 = 5^{-4}$, respectively. Each minibatch is uniformly sampled from the list of triplets. We train each model for 40 epochs and perform early stopping on the validation set. We implemented Multi-view Triplet Embeddings (MVTE) [1] as a competitive baseline for our proposed Latent Similarity Networks (LSN) by learning a

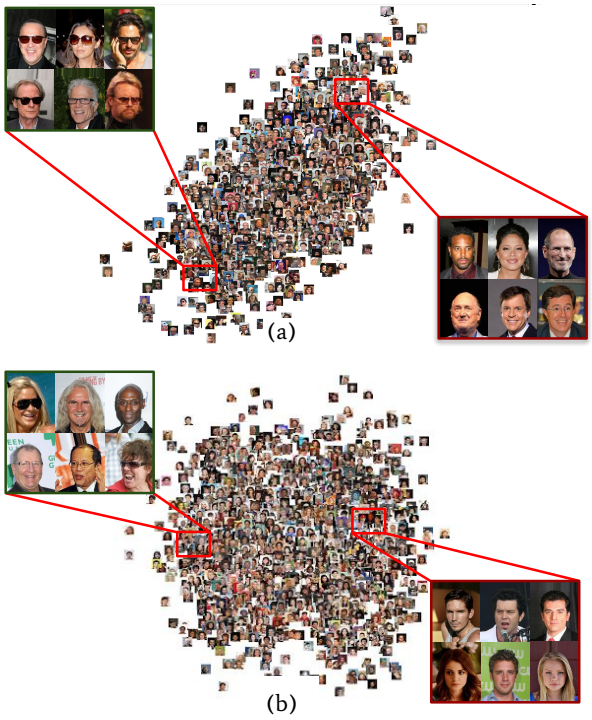

Figure 4: Embedding visualization for discovered latent attributes in the Celeb-A Faces dataset: **(a)** The discovered attribute corresponds to the `eyeglasses` attribute. Our method succeeds in recognizing eyeglasses across age, race, gender. **(b)** The discovered `smile` attribute. Our method learns to discover smiles across age, pose, gender.

linear classifier over a fixed Resnet-18 encoder pre-trained on Imagenet [7].

Fig. 3a shows how the latent embedding space corresponding to `comfort`: it is possible that humans perceive colorful shoes to be more comfortable. Fig. 3b illustrates how the latent embedding space corresponding to `open` learns to reason about brightness of shoes. Fig. 3c shows a projection of the latent embedding space corresponding to `pointy` which shows a continuous progression from red to blue colored shoes; we note that a number of pointy shoes in the dataset are red stilettos. Fig. 3d visualizes a lower-dimensional projection of the `sporty` latent attribute embedding space that learns to reason about the shade of the color of shoes.

C. Qualitative analysis for UT Zappos-50k

Fig. 6 in the main manuscript visualizes the qualitative trends in PCA embeddings of the learnt LSN embeddings for Zappos-Human shoes. We now adopt a different modus operandi for investigating the learnt subspace embeddings. We randomly choose two dimensions for each 16-D embedding in the LSN model, and visually inspect these embeddings in an effort to uncover the low-level features that may have been learnt in the subspaces. Fig. 3 shows the results of this qualitative experiment.

D. Qualitative analysis for Celeb-A Faces

Fig. 5 in the main manuscript visualizes the qualitative trends in PCA embeddings of the learnt LSN embeddings for two attributes in Celeb-A Faces. Fig. 4 presents additional qualitative analysis of the PCA embeddings of the learnt latent spaces for other attributes in the Celeb-A Faces dataset.

Fig. 4a illustrates a 2D PCA visualization of the learnt latent embedding space corresponding to the `eyeglasses` attribute. Our method succeeds in recognizing eyeglasses across age, race, gender. Fig. 4b shows a PCA visualization of the latent embedding space for the `smile` attribute. Our method learns to discover smiles across age, pose, gender.

Fig. 5 in the main manuscript visualizes the qualitative trends in PCA embeddings of the learnt LSN embeddings for two attributes Celeb-A Faces. Fig. 4 now presents qualitative analysis of the clustering of several other attributes in Celeb-A Faces.

References

[1] Ehsan Amid and Antti Ukkonen. Multiview Triplet Embedding: Learning Attributes in Multiple Maps. In *ICML*, 2015. 2, 3

[2] André Elisseeff and Jason Weston. A Kernel Method for Multi-Labelled Classification. In *NeurIPS*, 2002. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks. In *ECCV*, 2016. 2

[4] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In *IEEE CVPR*, 2012. 1

[5] Stefan Lee, Senthil Purushwalkam, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra. Stochastic Multiple Choice Learning for Training Diverse Deep Ensembles. In *NeurIPS*, 2016. 2

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE ICCV*, 2015. 1

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 2, 4

[8] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional Similarity Networks. In *IEEE CVPR*, 2017. 1, 2

[9] Aron Yu and Kristen Grauman. Fine-Grained Visual Comparisons with Local Learning. In *IEEE CVPR*, 2014. 1