Towards Latent Attribute Discovery From Triplet Similarities

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

Abstract

This paper addresses the task of learning latent attributes from triplet similarity comparisons. Consider, for instance, the three shoes in Fig. 1(a). They can be compared according to color, comfort, size, or shape resulting in different rankings. Most approaches for embedding learning either make a simplifying assumption - that all inputs are comparable under a single criterion, or require expensive attribute supervision. We introduce Latent Similarity Networks (LSNs): a simple and effective technique to discover the underlying latent notions of similarity in data without any explicit attribute supervision. LSNs can be trained with standard triplet supervision and learn several latent embeddings that can be used to compare images under multiple notions of similarity. LSNs achieve state-of-the-art performance on UT-Zappos-50k Shoes and Celeb-A Faces datasets and also demonstrate the ability to uncover meaningful latent attributes.

1. Introduction

Supervised learning has undeniably revolutionized computer vision and machine learning. Such supervision is often provided in terms of high-level semantic concepts, which are straightforward to define for many domains, such as the space of objects [27, 8, 20]. But they are less clear for others - what is the right ontology of concepts for attributes [9, 3, 24] or actions [5, 11]? Indeed, even object labels may be culturally ambiguous to define - for example, overseas annotators struggled to correctly label `hotdogs` in COCO [20]! The difficulty associated with defining and obtaining such labels gives rise to a considerable body of work in unsupervised learning [10, 16, 18, 26, 32].

Our work explores a third avenue for learning through *similarity embeddings*, where human annotators provide labels for similar and dissimilar objects [4, 7, 28, 31]. Such supervision is easier to scale as labels can be extracted from web-scale click data [29] and relies less on cultural and linguistics norms. A large body of work in the psychophysics community on Just-Noticeable-Differences illustrates the universality of similar and dissimilar comparisons [23].

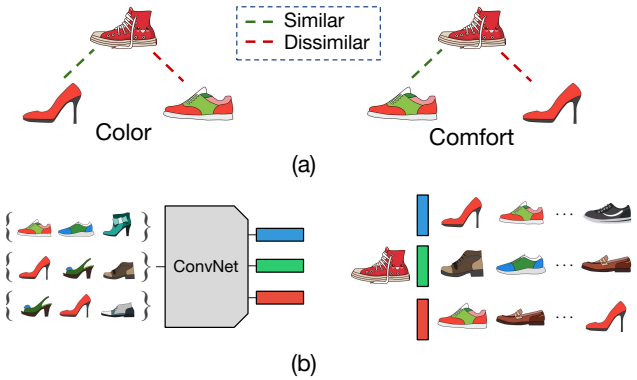

Figure 1: **(a)** The first triplet corresponds to the notion of `color` while the second triplet corresponds to the notion of `comfort`. Relative similarity measurements encode useful contextual information. We explore the problem of learning latent attribute concepts from the context encoded in triplet comparisons. **(b)** (Left) Our proposed method learns latent attribute embeddings from triplet comparison data. (Right) The latent attribute embeddings learned by our system are useful for knowledge representation as well as for applications such as multi-attribute image retrieval.

Specifically, our work demonstrates that one can discover discrete latent concepts from large-scale similarity-based object comparisons. Intuitively, concepts correspond to *latent attributes*. The heart of our approach is based on the observation that objects can be similar (or dissimilar) in different ways. For example, an annotator might label two particular shoes as being similar according to `color` (Fig. 1 (a)), but another annotator might label them as different according to their `comfort`. A single embedding space for shoes cannot capture both notions of similarity.

One simple solution is learning *different* embeddings for each notion of similarity, such as `color`, `comfort`, etc. Given a pair of images and an *observed* attribute label, one can compute the distance between the images in the appropriate attribute-specific embedding. As we argue above, such semantic attribute labels are non-trivial and sometimes even impossible to obtain. Indeed, the task of

uncovering user preferences from search queries is a vibrant topic in information retrieval [1]. Consider a user searching for a shoe online. They may not be able to *verbalize* their precise goals, but can easily compare any given pair of shoes according to their innate preferences, generating a training example for learning. A practical multi-attribute embedding learning approach must be able to discover *hidden* notions of similarity in the data without explicit supervision.

Latent Similarity Networks: In this paper we propose Latent Similarity Networks (**LSNs**) - a method for discovering latent concepts *without* any supervision beyond similarity comparisons. Given a training triplet, we compute the loss under multiple candidate embeddings and generate a gradient update for the optimally matched embedding. This can be seen as a form of hard-assignment expectation maximization, where attributes are treated as latent variables that are marginalized out. As a result, our method implicitly clusters the training triplets into latent attributes. The training is performed in an end-to-end fashion, which results in learning a deep representation which is disentangled with respect to the attributes. We demonstrate that this formulation is a special case of Multiple Choice Learning [12].

Evaluation: Quantitative evaluation of latent attribute discovery is challenging, since it requires ground truth annotations of the very concepts that are to be discovered. A common approach to evaluate such a scenario is to *simulate* the discovery process by processing data with known ground-truth concepts. In the domain of attribute discovery, this corresponds to constructing triplets with respect to a known set of latent attributes [15]. We follow this path using the UT Zappos-50k Shoes dataset [35] and the Celeb-A Faces dataset [21] for quantitative evaluation.

Contributions: Our work makes the following contributions: **(1)** We propose the first deep-learning based method for unsupervised attribute discovery; **(2)** Our method achieves state-of-the art results on the UT-Zappos-50k and Celeb-A datasets in an unsupervised learning scenario; **(3)** We provide qualitative analysis of the discovered latent attributes.

2. Related Work

Embedding learning has a long history in computer vision [4, 7, 33, 30, 25, 28]. In this section we focus on the most relevant topics: supervised and unsupervised multi-attribute embedding learning, and learning to predict multiple outputs.

Supervised multi-attribute embedding learning makes use of multiple explicitly-labeled measures of similarity. Whittle Search [17] uses multi-attribute feedback for interactive image search, allowing users to reason about images based on relative attributes [24], by interactively *whitteling* away the search space to retrieve an image of interest. Yu and Grauman [35] introduced UT-Zappos50k, a multi-attribute shoe dataset, which was used to train a Conditional Similarity Networks (CSN) [34] for encoding multiple attribute-specific embeddings. Importantly, CSNs require training examples with ground-truth attribute labels, while our method discovers latent notions of similarity (making it applicable for scenarios where ground-truth attributes are unavailable or difficult to verbalize).

Unsupervised multi-attribute embedding learning has received relatively little attention, with the important exception of Amidi and Ukkonen [2]. Their work represents attribute spaces as soft linear combinations of a fixed set of handcrafted features, while we jointly learn both the feature embedding and discrete latent attribute concepts in a end-to-end fashion. Moreover, they use latent attributes as hidden variables in a single embedding, and so do not compare images under different notions of similarity (which is our focus). Sec. 5 extensively compares our approach to theirs, demonstrating that end-to-end learning is essential for performant latent attribute discovery.

Multiple Choice Learning (MCL) [12, 19] is a method for learning multiple hypothesis predictors with an "oracle" loss that, given a training example, evaluates all hypotheses, but only updates the minimal-loss hypothesis. This was used to train systems that generate multiple image classifications, segmentations, and captions. We demonstrate that one can repurpose this loss function for latent attribute discovery (rather than multiple hypothesis prediction).

3. Preliminaries

We begin by introducing a few concepts and the associated notation used in the paper.

3.1. Triplet loss

We wish to learn an image embedding, where $f_\theta(x_i)$ is a nonlinear embedding of image x_i parameterized by θ. This is usually done by minimizing the Euclidean distance between similar images and maximizing it between dissimilar ones:

$$D_{ij} = ||f(x_i) - f(x_j)||_2^2, \tag{1}$$

Most contemporary approaches rely on triplet supervision [25, 28], where $T = \{(x_q, x_p, x_n)_i\}_{i=1}^K$ is a set of triplets composed of a query image x_q, a positive image x_p, and a negative image x_n. The goal is to learn an embedding function f where positives are more similar to the query than negatives:

$$D_{qp} < D_{qn} \text{ for all } (x_q, x_p, x_n) \in T \tag{2}$$

This is commonly achieved with the Triplet Loss:

$$\mathcal{L}_{triplet}(x_p, x_q, x_n) = [D_{qp} - D_{qn} + M]_+ \tag{3}$$

where M is the margin and $[\cdot]_+$ is the hinge function. The loss explicitly encourages positive images to be closer to query images than to negative images by a fixed margin M.

3.2. Multi-attribute networks

Classical embedding learning approaches are based on an assumption that all the triplets in T are defined with a *single* notion of similarity. This assumption, however, may not hold in many practical scenarios. Coming back to Fig. 1, shoes can be compared along different attributes of comfort, color, style, etc. In the extreme, a particular triple of shoes (x_i, x_j, x_k) may swap positives and negatives depending on the particular attribute. One naive solution is learning a separate embedding $f_{\theta_A}(x)$ for each attribute A:

$$D_{ij}(A) = ||f_{\theta_A}(x_i) - f_{\theta_A}(x_j)||_2^2, \quad \forall A \in \mathcal{A}. \quad (4)$$

Rather than learning separate attribute-specific networks, Veit et al. [34] learn separate linear projection masks m_A over a single embedding:

$$D_{ij}(A) = ||f(x_i) \odot m_A - f(x_j) \odot m_A||_2^2, \quad (5)$$

where $\odot$ denotes the elementwise product. The above can be learned with triplets augmented with supervised attribute labels $T = \{(x_q, x_p, x_n, A)_i\}_{i=1}^K$. The associated Supervised Loss can be written as:

$$\mathcal{L}_{SUP}(x_p, x_q, x_n, A) = [D_{qp}(A) - D_{qn}(A) + M]_+ \quad (6)$$

In practice, such ground-truth attribute supervision is hard and sometimes even impossible to obtain. In the next section we propose a method that is able to learn multi-attribute embeddings without such supervision.

4. Latent attribute discovery

We now describe a method for learning multi-attribute embeddings without any attribute supervision. Importantly, our method can *discover* latent notions of discrete attributes encoded in a training set.

4.1. Our approach

Our key insight is to treat attributes as latent variables that are minimized over in the loss function in Equation 6. We call the resulting function Latent Loss:

$$\mathcal{L}_{LAT}(x_p, x_q, x_n) = \min_{A \in \mathcal{A}} [D_{qp}(A) - D_{qn}(A) + M]_+ \quad (7)$$

The details of our method, which we denote as Latent Similarity Networks (LSN), are shown in Fig. 3. LSN relies on the Multiple Choice Learning (MCL) algorithm [12], shown in Fig. 2, to learn from triplets. A triplet, in this case consisting of a red sport shoe, a red female high-heel shoe, and a blue-green sneaker, is passed through an embedding network. The image embedding is then projected in the subspaces corresponding to the latent attributes by applying the masks m_A. The attribute embedding with the smallest triplet loss (`color`, in this case) is selected and used to backpropagate the loss for this triplet. Our latent loss function essentially

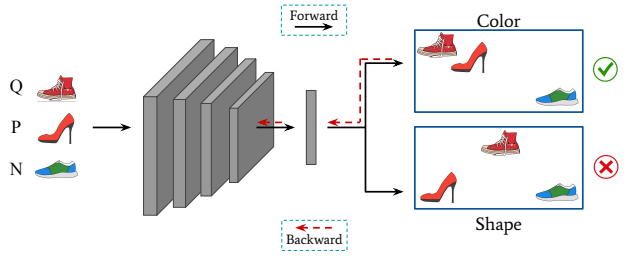

Figure 2: Multiple Choice Learning: A triplet represented by an Query (Q), a Positive (P), and a Negative (N) is passed through a multi-task network. A valid triplet (Q, P, N) requires that the distance between Q and P is less than the distance between Q and N. These separations (D_{QP} and D_{QN}) are computed in each task-specific space (i). The triplet is assigned to the task-space where it is most well separated ($\max_i(D^i_{QN} - D^i_{QP})$). In the figure above, the triplet is most well separated in the Color embedding-space. Hence, it is assigned to the Color embedding-space and we update the parameters associated with this embedding-space, along with the parameters of the shared embedding network.

clusters training triplets into distinct latent attributes, while simultaneously learning an embedding for each attribute.

Our approach can be seen as an instance of MCL for the task of predicting multiple hypotheses for the same input example. This loss function is also commonly referred to as an "oracle" loss, because it predicts the right answer for a multiple choice task. Instead, we use MCL to learn latent notions of similarity for the task of multi-attribute image embedding learning. We note that in addition to providing a model for comparing images according to several latent criterions, our approach implicitly learns a disentangled image representation [6, 22] in the common embedding space using triplet supervision.

MCL [12] optimizes the multiple-choice loss by alternating between assigning examples to their min-loss predictors and training models to convergence using the examples assigned to them. This approach is, however not feasible for deep networks, which can take days to train. Instead, stochastic MCL [19] interleaves the assignment step with batch updates in stochastic gradient descent. We adopt stochastic MCL for our optimization, allowing us to jointly learn the nonlinear parameters of f_θ while estimating the latent variables A in every mini-batch.

Overall, our proposed method takes the form of the following objective function for training:

$$\mathcal{L}(x_q, x_p, x_n) = \mathcal{L}_{LAT}(x_q, x_p, x_n) + \lambda_1 \mathcal{L}_\theta(\mathbf{x}) + \lambda_2 \mathcal{L}(\mathbf{m}), \quad (8)$$

where $\mathcal{L}_\theta = ||f_\theta(\mathbf{x}, \theta)||_2^2$ is the embedding regularizer, $\mathcal{L}(\mathbf{m}) = ||\mathbf{m}||_1$ enforces the sparsity of the collective set of masks $\mathbf{m} = \{m_A\}$, and λ_1, λ_2 are hyper-parameters that balance the relative contribution of the three terms.

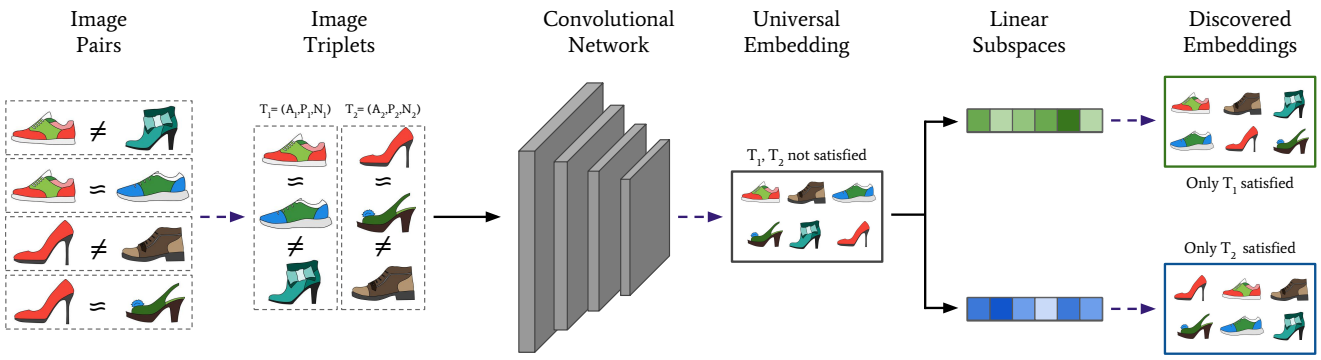

Figure 3: Details of the proposed method: Triplets of images are constructed based on implicit supervision mined from pairs of similar and dissimilar objects. A convolutional network embeds the images in a (shared) universal embedding. Multiple linear subspaces, each corresponding to a distinct latent attribute, are learnt over the universal embedding. The Multiple Choice Learning algorithm (Fig. 2) assigns the triplets to the (latent attribute) subspace. This allows our proposed method to *discover* multiple notions of similarities.

4.2. Evaluating latent attribute discovery

Evaluating latent attribute discovery is challenging. In the most realistic scenario, triplets are collected from user interactions in the wild [17] and the proposed approach may be used to analyze the underlying user preferences. The lack of ground truth attribute annotations in this scenario, however, makes quantitative evaluation challenging. An alternative path is proposed by CSN [34], where the authors utilize datasets with ground truth attribute annotations [35] to mine triplets according to these attributes.

We present evaluation using simulated triplets to validate the ability of our method to discover real user preferences as well as to quantitatively compare our method to prior work. We draw the attention of the reader to and re-emphasize a critical detail - although we utilize ground truth attributes to construct simulated triplets, these attributes are *not* provided to our learning algorithm, but instead only used for evaluation. More details on how the triplets are constructed are provided in the supplementary material. Next, we describe the measure used for evaluation in the paper.

In the supervised setting, train and test triplets include ground-truth attributes, (x_p, x_q, x_n, A). The Supervised-Eval metric measures the test error as the fraction of triplets that are not satisfied under the corresponding embedding:

$$\frac{1}{|S|} \sum_{(x_q,x_p,x_n,A)\in S} I\Big(D_{qp}(A) > D_{qn}(A)\Big) \qquad (9)$$

where S is the test set of triplets and A is the ground truth attribute according to which the images are compared.

The Supervised-Eval metric requires that the learned latent embeddings are mapped to the underlying ground-truth attributes. To this end we propose to utilize a small held-out set (5% of the data) of annotated triplets to determine the mapping of discovered latent embeddings to the ground-truth attributes. This allows us to compare unsupervised approaches to fully-supervised methods. We now describe two strategies to perform this mapping.

One-to-one mapping: One-to-one mapping finds the optimal mapping between learned latent embeddings and the underlying ground-truth attributes. The objective of this mapping strategy is to measure the ability of an unsupervised method to exactly recover all underlying ground-truth attributes. Let E represent the number of learned latent embeddings (which is equal to the number of ground-truth attributes for one-to-one mapping). We consider all $E!$ combinations of the latent embeddings and the ground-truth attributes to report the Supervised-Eval performance.

One-to-many mapping: In practice, latent learning may discover factors of variation that map to multiple correlated attributes (e.g., `male`-ish shoes may tend to be `comfortable` and `sporty`). To allow for such one-to-many mappings, we greedily assign each ground-truth attribute to the latent embedding that produces the best validation error on a held-out test set. This allows us to compute the optimal one-to-many mapping in $\mathcal{O}(EK)$.

5. Experiments

We now demonstrate the effectiveness of LSNs in qualitatively and quantitatively discovering latent attributes. We begin by describing the datasets used in our analysis.

5.1. Datasets

We briefly discuss the data used in our evaluation below. The data and the triplet construction strategies are described in detail in the supplementary material. Our qualitative and quantitative analysis focuses on the UT-Zappos Shoes Dataset [35] and the Celeb-A Faces Dataset [21].

UT-Zappos-50k Shoes: Yu and Grauman [35] 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 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. [34]. The attributes are `type`, `gender`, `heel-height`, and `closing-mechanism`. We refer to this triplet similarity comparison data as Zappos-Meta.

Celeb-A Faces: The Celeb-A dataset [21] contains 202,599 face images labeled with 40 binary visual attributes. We select eight visual attributes for ablative analysis - `Eyeglasses`, `Male`, `Smiling`, `Young`, `Attractive`, `Wearing_Lipstick`, `5_o_Clock_Shadow`, and `Bags_Under_Eyes`. Fig. 4b illustrates the general nature of a few of these attributes. We also study the performance of our method on the entire 40 attribute dataset.

We note that the underlying attributes for both datasets are used solely for generating triplets and are not available to the unsupervised learning methods during training.

5.2. Methods

We study the effectiveness of the proposed Latent Similarity Networks along with three other methods.

Singular Similarity Networks (SSN): A Resnet-based model with a single embedding space used to perform all triplet similarity comparisons.

Multi-View Triplet Embeddings (MVTE) [2]: A multi-attribute model which learns projections over a fixed embedding space by predicting soft label assignments.

Latent Similarity Networks (LSN): A multi-attribute end-to-end trainable model which *discovers* multiple latent attributes while learning a disentangled embedding space by predicting hard label assignments.

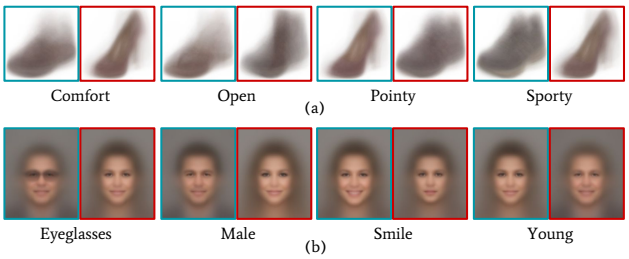

Figure 4: Mean object images illustrating attribute presence or absence: (a) Human-labeled attributes from UT-Zappos-50k, (b) Four attributes from Celeb-A Faces. (The supplementary material illustrates all Celeb-A Faces attributes).

Conditional Similarity Networks (CSN) [34]: A multi-attribute model which conditionally learns multiple attributes in a fully supervised manner.

SSN provides a lower-bound for comparative analysis with our proposed method. On the other hand, CSN is trained in a fully-supervised manner and provides an upper-bound for the unsupervised LSN method. We discuss specific architectural differences in the supplementary material.

Implementation Details: The proposed Latent Similarity Network architecture consists of a Resnet-18 [13] encoder pre-trained on Imagenet [27]. Following [34], 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. The Celeb-A images are resized to 224×224 and provided to the Resnet-18 model. A fully-connected layer is added to the encoder, which serves as the universal embedding for the network. The experiments are performed using a universal embedding dimension of 16, which is the smallest embedding dimension that does not lead to overfitting. The linear subspaces learnt on the universal embedding are initialized as 16-dimensional normally distributed projections on the universal embedding. 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 implement Multi-view Triplet Embeddings (MVTE) [2] as a baseline for our proposed Latent Similarity Networks (LSN) by learning a linear classifier over a fixed Resnet-18 encoder pre-trained on Imagenet [27].

5.3. Recovering latent attributes

In this section we evaluate the ability of proposed approach to exactly recover underlying from triplet similarities. We begin by presenting experiments on a small-scale UT-Zappos dataset in Sec. 5.3.1. Next, we use a much larger Celeb-A dataset to confirm that our method can scale with the number of underlying attributes in Sec. 5.3.2. In these experiments we set the number of learnt latent embeddings to be equal to the number of attributes and use one-to-one mapping (Sec. 4.2) during evaluation.

5.3.1 Latent attribute recovery on UT-Zappos Shoes

We study the latent attribute recovery problem on two attribute label sets in the UT-Zappos Shoes dataset: (1) Zappos-Human consists of 4 human-labeled binary attributes constructed by Yu and Grauman [35], and (2) Zappos-Meta consists of 4 multi-class attributes constructed from the metadata by Veit et al. [34].

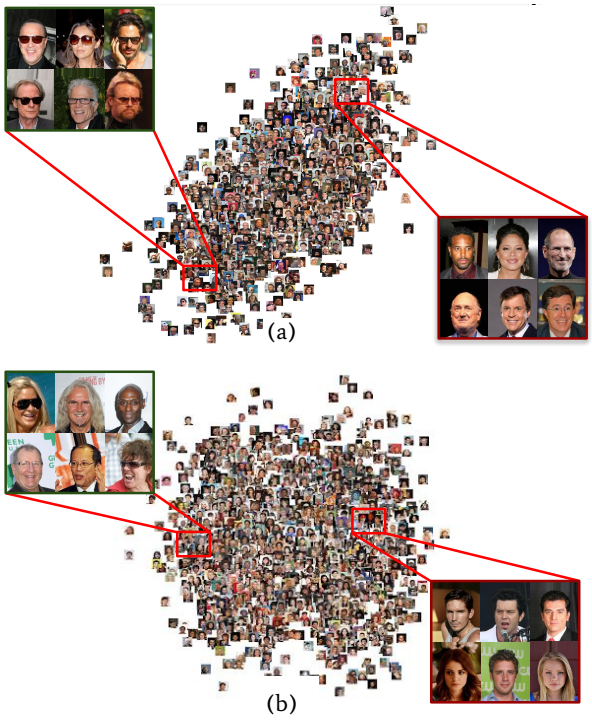

Figure 5: Embedding visualization for discovered latent attributes in the Celeb-A Faces dataset: (a) The discovered attribute corresponds to the `eyeglasses` attribute. Our method succeeds in recognizing eyeglasses across age, race, gender. (b) The discovered `smile` attribute. Our method learns to discover smiles across age, pose, gender.

Tab. 1 summarizes the performance of all methods on the attribute recovery task. LSNs significantly outperform SSNs, demonstrating that multi-attribute latent learning is beneficial. LSN also outperforms MVTE; our method's superior performance may be attributed to the more robust nature of our hard label assignment strategy. The fully-supervised CSN benefits from the additional attribute supervision, outperforming all unsupervised methods.

Method	**Zappos-Human**	**Zappos-Meta**
SSN	80.52	76.24
MVTE [2]	77.94	77.53
LSN **(Proposed)**	88.91	83.09
CSN [34]	97.36	89.27

Table 1: Latent attribute recovery on UT-Zappos Shoes with four human-labeled attributes (Zappos-Human) and four metadata attributes (Zappos-Meta). Our method outperforms all unsupervised methods on both datasets. Our unsupervised learning algorithm is only surpassed in performance by the fully-supervised CSN algorithm.

Finally, we observe that MVTE, as originally proposed by Amid and Ukkonen [2], learns projections over a fixed embedding space. However, our method (LSN) learns embeddings in an end-to-end manner. For a more direct comparison, we train MVTE end-to-end as well. This baseline performs 3.14% worse on Zappos-Human, further validating the superiority of our approach.

5.3.2 Latent attribute recovery on Celeb-A Faces

We now move towards latent attribute recovery on the significantly larger Celeb-A Faces dataset. The presence of 40 attributes allows us to study the effect of changing the number of underlying attributes. To this end, we construct triplets based on 2, 4, 6, 8 visual attributes and learn separate models for each set of attributes. The attributes used in the experiments are described in the supplementary material.

Tab. 2 summarizes the performance of all methods. An increased number of underlying attributes results in a more complex distribution, resulting in decreased performance for all methods as the number of attributes increases. LSNs outperform both unsupervised methods. The performance of fully-supervised CSNs exceeds all unsupervised methods.

We also compare to the end-to-end trained variant of MVTE on Celeb-A. This baseline performs 4.72% worse than our method on this dataset. We attribute the higher margin, compared to Zappos, to the overlapping nature of the Celeb-A ground-truth attributes, which makes the task harder for MVTE.

5.3.3 Qualitative analysis

We now qualitatively analyze the latent embeddings discovered by our method. We use PCA [14] to visualize the embeddings learned on UT-Zappos Shoes and Celeb-A Faces. Fig. 5 visualizes two discovered latent attributes in Celeb-A Faces. In the first embedding, the discovered attribute corresponds to `eyeglasses`. Our method succeeds in recognizing eyeglasses across age, races, genders but fails to recognize frameless eyeglasses. In the second embedding, our method discovers smiles across age, poses, genders.

Fig. 6 demonstrates four discovered latent attributes for UT-Zappos Shoes. Fig. 6(a) shows that the discovered at-

Method	**2**	**4**	**6**	**8**
SSN	86.38	81.12	75.54	71.40
MVTE [2]	88.24	83.38	81.18	75.78
LSN **(Proposed)**	92.33	90.36	87.71	83.53
CSN [34]	99.47	98.23	95.05	90.43

Table 2: Latent attribute recovery on Celeb-A Faces. Our method consistently outperforms all unsupervised methods while learning to recover the underlying attributes.

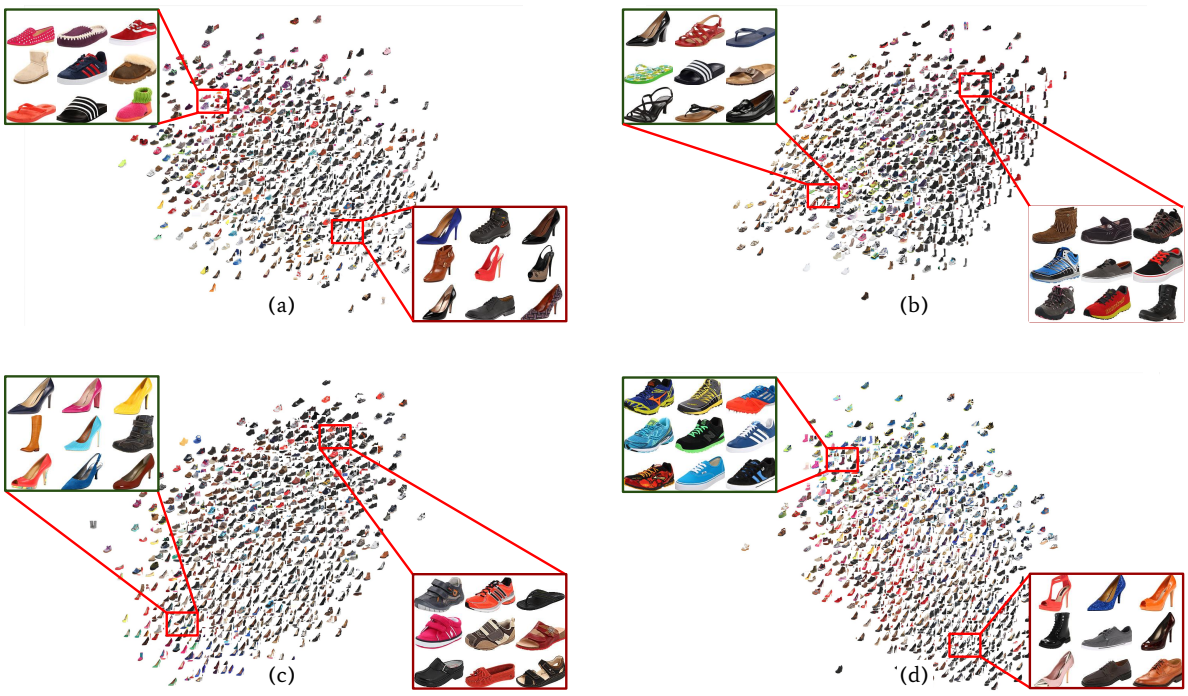

Figure 6: Embedding visualization for discovered latent attributes in the UT Zappos-50k dataset: (a) The discovered attribute corresponds to the `comfort` attribute. Our method learns to capture the notion of comfort across a wide range of visual appearance differences. (b) Our method learns to capture `open`-ness. (c) Our method learns to capture the `pointy`-ness attribute fairly accurately by across a wide range of visual appearance differences. (d) Our method learns the `sporty`-ness attribute and learns to reason that formal shoes (two examples in bottom row of negatives) is the inverse of sporty-ness.

tribute corresponds to `comfort`. Fig. 6(b) show that our method learns to capture `open`-ness. Fig. 6(c) indicates that the embedding captures `pointy`-ness across a wide range of visual appearance differences. Fig. 6(d) shows that our method learns the `sporty`-ness attribute.

So far we have assumed that the underlying attributes are uncorrelated and we recover all underlying latent attributes. We now move towards a more realistic scenario - several ground-truth attributes may correspond to a single underlying notion of similarity - and evaluate whether our model can discover these latent attributes in an unsupervised way.

5.4. Discovering latent attributes

We now attempt to *discover* the underlying latent concepts in the data by learning a small number of latent embeddings, and examining which ground truth attributes end up being grouped together. We use many-to-one mapping strategy for evaluation (Sec. 4.2). In addition, we quantitatively examine the utility of the discovered latent attributes. To this end, we compare our method to a baseline learnt by randomly assigning the ground-truth attributes to a small number of latent embeddings.

5.4.1 Latent attribute discovery with UT-Zappos Shoes

We study the latent attribute discovery problem on both label sets of the UT-Zappos dataset. Fig. 7(a) shows that our method trained with two latent embeddings (LSN-2) is suffi-

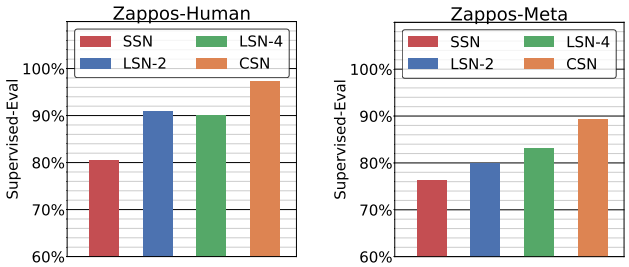

Figure 7: Latent attribute discovery on UT-Zappos Shoes. (a) LSNs trained with 2 latent embeddings (LSN-2) outperform LSNs trained with 4 latent embeddings (LSN-4) on Zappos-Human, indicating the ground-truth attributes are correlated and can be modeled with 2 latent concepts. (b) On Zappos-Meta performance is maximized with 4 latent attributes, indicating that the ground-truth attributes are uncorrelated.

cient for modeling the four ground-truth attributes in Zappos-Human and achieves nearly identical Supervised-Eval performance compared to our method trained with four latent embeddings (LSN-4). The correlated attributes `comfort`, `pointy`, and `sporty` map to the first latent embedding, while `open` maps to the second embedding.

Fig. 7(b) suggests that our method trained with two latent embeddings (LSN-2) on the four ground-truth attributes in Zappos-Meta does not achieve the same Supervised-Eval performance as our method trained with four latent embeddings (LSN-4), suggesting that `closure`, `gender`, `heel`, `type` are not correlated.

To further validate the utility of the discovered latent attributes we compare them to a random assignment baseline. In particular, we randomly assign Zappos-Human ground-truth attributes to latent embeddings and train the network with such randomly obtained supervision. The Supervised-Eval performance over 3 random assignments of 4 ground-truth attributes to 2 embedding spaces is $6.69 \pm 0.53\%$ lower than our method. This baseline is closer to SSN than to our proposed method, demonstrating that our method learns a meaningful clustering of the ground-truth attributes.

5.4.2 Latent attribute discovery with Celeb-A Faces

We now study latent attribute discovery on the much larger Celeb-A Faces dataset using the entire attribute label set. Fig. 9 summarizes the performance of all methods learnt with a varying number of latent embeddings. The performance of our method (shown in green) increases from 2 to 8 latent embeddings, but drops slightly afterwards. This result suggests that the 40 Celeb-A attributes can be captured with 8 to 16 disentangled latent concepts. Fig. 8 provides a qualitative visualization of the discovered attribute clusters.

We also compare our method to the random baseline described in Sec. 5.4.1 on Celeb-A Faces. The average

Figure 8: LSNs trained with 8 latent embeddings learn to optimally model the 40 Celeb-A Faces attributes with the highest Supervised-Eval performance. The mean face images for the 40 attributes, as shown in the supplementary material, suggest that the optimal clustering illustrated above does indeed appear to be visually similar.

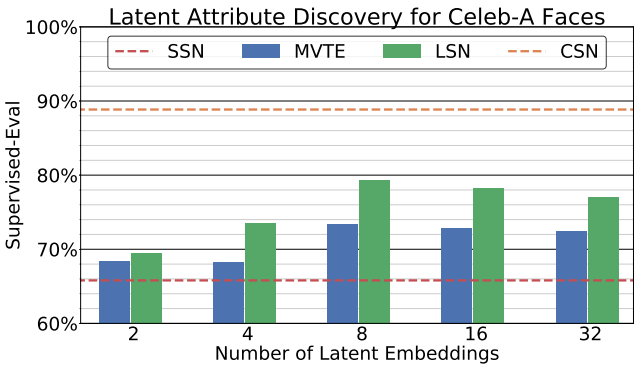

Figure 9: Latent attribute discovery on Celeb-A Faces. Our proposed LSNs trained on 40 ground-truth attributes with 8 latent-embeddings achieve optimal Supervised-Eval performance. The results suggest that there are between 8 and 16 underlying latent concepts in Celeb-A Faces.

performance over 5 random assignments of the 40 attributes to 8 embedding spaces is $3.21 \pm 0.68\%$ lower than that of our discovered attribute clustering. We attribute the smaller margin between the random baseline and our method on this dataset to the overlapping nature of attributes in it.

6. Conclusion

We introduced Latent Similarity Networks (LSNs) - an approach for discovering latent concepts from triplet similarity comparisons. Our model demonstrated state-of-art performance on UT Zappos-50k Shoes and Celeb-A Faces datasets without making use of ground-truth attributes. Further, we performed qualitative experiments to demonstrate that the subspaces learnt by LSNs are semantically interpretable. The design and successful experimental validation of LSNs suggests that practical image retrieval systems may benefit from modeling contradicting user preferences. We hope our proposed method spurs the community to further investigate latent concept learning from similarity data.

Acknowledgements: This research is based upon work supported in part by NSF Grant 1618903, the Intel Science and Technology Center for Visual Cloud Systems (ISTC-VCS), Google, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *ACM SIGIR*, 2006.

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

[3] Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, 2010.

[4] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature Verification Using a Siamese Time Delay Neural Network. In *NeurIPS*, 1994.

[5] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE CVPR*, 2017.

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *NeurIPS*, 2016.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a Similarity Metric Discriminatively with Application to Face Verification. In *IEEE CVPR*, 2005.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 2010.

[9] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing Objects by Their Attributes. In *IEEE CVPR*, 2009.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NeurIPS*. 2014.

[11] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions.

[12] Abner Guzmán-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple Choice Learning: Learning to Produce Multiple Structured Outputs. In *NeurIPS*, 2012.

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

[14] Ian T Jolliffe. Principal Component Analysis and Factor Analysis. In *Principal Component Analysis*, pages 115–128. Springer, 1986.

[15] Kun Ho Kim, Oisin Mac Aodha, and Pietro Perona. Context Embedding Networks. In *IEEE CVPR*, 2018.

[16] Diedrik Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

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

[18] Quoc V. Le, Marc'Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen, Greg S. Corrado, Jeff Dean, and Andrew Y. Ng. Building High-level Features Using Large Scale Unsupervised Learning. In *ICML*, 2012.

[19] 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.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

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

[22] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling Factors of Variation in Deep Representation Using Adversarial Training. In *NeurIPS*, 2016.

[23] Stephen E Palmer. *Vision Science: Photons to Phenomenology*. MIT press, 1999.

[24] Devi Parikh and Kristen Grauman. Relative Attributes. In *IEEE CVPR*, 2011.

[25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep Face Recognition. In *BMVC*, 2015.

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*, 2016.

[27] 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.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A Unified Embedding for Face Recognition and Clustering. In *IEEE CVPR*, 2015.

[29] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In *WWW*, 2014.

[30] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively Learning the Crowd Kernel. *ICML*, 2011.

[31] Evgeniya Ustinova and Victor Lempitsky. Learning Deep Embeddings with Histogram Loss. In *NeurIPS*, 2016.

[32] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *ICML*, 2016.

[33] Laurens Van Der Maaten and Kilian Weinberger. Stochastic Triplet Embedding. In *IEEE MLSP Workshop*, 2012.

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

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