The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Dress Fashionably: Learn Fashion Collocation with Deep Mixed-Category Metric Learning

Long Chen, Yuhang He
School of Data and Computer Science, Sun Yat-sen University
Xiaoguwei Island, Panyu District, Guangzhou 510006, P. R. China

Abstract

In this paper, we seek to enable machine to answer questions like, given a clutch bag, what kind of skirt, heel and even accessory best fashionably collocate with it? This problem, dubbed fashion collocation, has almost been neglected by researchers due to the large uncertainty lies in fashion collocation and professional expertise required to address it. In this paper, we narrow down the well-collocated samples to be fashion images shared on fashion websites, with which we propose an end-to-end trainable deep mixed-category metric learning method to project well-collocated clothing items to lie close but items violating well-collocation far apart in the deep embedding space. Specifically, we simultaneously model the intra-category exclusiveness and cross-category inclusiveness of fashion collocation by feeding a set of well-collocated clothing items and corresponding bad-collocated clothing items to the deep neural network, further a hard-aware online exemplar mining strategy is designed to force the whole neural network to be trainable and learn discriminative features at the early and later training stages respectively. To motivate more research in fashion collocation, we collect a dataset of 0.2 million fashionably well-collocated images consisting of either on-body or off-body clothing items or accessories. Extensive experimental results show the feasibility and superiority of our method.

Introduction

People by nature care about their outfit look, their clothing collocation choice should reflect their fashion taste, social status and fit the social occasion. While it is widely accepted that the elegant fashion collocation expertise is acquired by a limited group of fashion experts, we dedicate to bring this personal styling to the masses: teaching the machine to grasp this capability so that people can consult it whenever they have any fashion collocation question.

Fashion collocation is extremely difficult. On the one hand, the changes of public taste and consciousness towards fashion collocation are ephemeral and resist to qualitative analysis. It is hard to reach a consensus as it is to some extent a matter of personal taste, and even depends on nationality, gender and temporal season. Also, fashion collocation is individual and instance-aware, an outfit itself being fashionable has to align with someone's "look", body characteristic as well as the temporal occasion he or she is involved in in order to be truly de rigueur. Although facing so many uncertainties and obstacles, we still try to tackle it by narrowing down good fashion collocation to be the images shared by various fashion bloggers or professional fashion websites. Leveraging the public taste as a proxy of fashionability enables us to make fashion collocation to be processible by machines. Moreover, from machine learning perspective, we can get large amounts of images online to delve deeper into collocation with powerful deep learning methods (Krizhevsky, Sutskever, and Hinton 2012).

Figure 1: Motivation: given a denim jacket, we retrieve the best-collocated items from the "wardrobe": chinos trouser, white T-shirt, shoe, golden watch and accessory.

We hereby treat fashion collocation as a retrieval problem: given a single item (either top, bottom, shoe or accessory), we retrieve a couple of cross-category clothing items that best collocate with it (see Fig. 1 for motivation illustration). Unlike traditional image retrieval problems which retrieve images with similar visual appearance to the query image, fashion collocation discussed here tries to retrieve fashion images with both intra-category exclusiveness and cross-category inclusiveness requirements (which would be discussed later). To this end, we embrace the powerful deep convolutional neural networks (CNNs) to get feature representation. Furthermore, we leverage deep metric learning to project these feature representations to deep embedding space where a bunch of well-collocated items lie together but irrelevant items lie far apart, even though they share obvious visual similarities. Specifically, our method initially derives from triplet neural network (Schroff, Kalenichenko,

and Philbin 2015) whose basic idea is to embed two similar items (anchor and positive) lie close but dissimilar items far apart with any predefined feature similarity metric. We extend traditional triplet neural network which usually receives three instances in each iteration to accept multiple instances. These multiple instances comprise a set of well-collocated clothing items as well as a set of bad-collocated clothing items. Moreover, a hard-aware online exemplar mining strategy has been designed to supervise the whole training process: making it to be trainable during the early training stage and forcing it to learn discriminative features during the later training stage. Being dubbed deep mixed-category metric learning, our proposed method simultaneously models the intra-category discrepancy as well as cross-category similarity and it turns out to outperform conventional triplet neural network by a large margin, as is shown in experiment section (on our collected 0.2 million images).

In sum, the contribution of this paper lies in: First, we propose an end-to-end trainable deep mixed-category metric learning method to tackle fashion collocation problem. It expands triplet neural network to accept multiple instances per iteration and the idea can be easily applied to other problems. Second, a hard-aware online exemplar mining strategy is further designated to supervise the training process.

Related Work

Fashion related research has many open challenges awaiting to be tackled. Various attempts have already been made to unveil fashion code from various perspectives, including fashion style analysis (Kiapour et al. 2014)(Serra et al. 2015)(Yamaguchi, Kiapour, and Berg 2013)(Vittayakorn et al. 2015), fashion feature learning (Serra and Ishikawa 2016)(He and Chen 2016), fashion likelihood prediction(Wang et al. 2015)(He, Lin, and McAuley 2016), clothing fashion annotations and retrieval (Liu et al. 2016b). Although promising results have been achieved and various accompanying datasets have even been made publicly available, fashion collocation still remains as a virgin territory. The existing work closely relating to fashion collocation is the work by A. Veit *et al.* (Veit et al. 2015), in which they assume clothing items being "frequently brought together" by consumers are well collocation templates. We argue, however, that this assumption cannot withstand scrutiny as it cannot accurately represents human being's understanding towards fashion collocation. On the contrary, we narrow down its definition to be the fashion images shared by various fashion websites, which reflect people's preference towards fashion as well as instant fashion trend. This definition has two main benefits: dataset collection or algorithm formulation. From the dataset side, we can easily collect hundreds of thousands of fashion images with high quality. From the algorithm side, we can fully exploit deep convolutional neural network (CNNs) and deep metric learning that have already shown state of the art performance on various vision tasks to solve fashion collocation problem.

The algorithm principle of our deep mixed-category metric learning is simple: given a bunch of training samples, we learn a deep embedding space where a set of well-collocated clothing items are embedded close, whereas items violating well-collocation are embedded far apart even though they share large visual similarity. Our deep mixed-category metric learning initially derives from triplet neural network, which has been successfully applied to various vision tasks, including face recognition (Schroff, Kalenichenko, and Philbin 2015), video representation (Wang and Gupta 2015), cross-domain clothing image retrieval (Huang et al. 2015) and fine-grained vehicle recognition (Liu et al. 2016a), person re-identification (ReID) (Hermans, Beyer, and Leibe). However, triplet neural network is also notorious for the difficulty to train and the cubic size explosion with offline training samples preparation, resulting in the hardship of utilizing all training triples. To mitigate this dilemma, various online hard exemplar mining (OHEM) strategies (Cui et al. 2016)(Simo-Serra et al. 2015)(Wang and Gupta 2015)(Liu et al. 2016a)(Hermans, Beyer, and Leibe) and distance metrics (Huang, Loy, and Tang 2016)(Ustinova and Lempitsky 2016)(Song et al. 2015)(Yuan, Yang, and Zhang 2016) have been proposed. Our deep mixed-category metric learning method tries to avoid these dilemmas from two aspects: First, we modify our neural network to accept multiple instances per iteration and these instances form a complete well-collocation set and other clothing items sharing visual similarities with the well-collocated items. This multiple instances input strategy has successfully avoided the randomness of offline training sample creation. Moreover, we further involve hard-aware online hard exemplar mining strategy to force the deep neural network to be trainable and to learn discriminative features in the early and later training stages respectively. In feature representation module, we leverage deep convolutional neural networks (CNNs). We combine CNNs with deep metric learning together to formulate the whole neural network to be end-to-end trainable.

Deep Mixed-Category Metric Learning

The goal of deep mixed-category metric learning is to learn an embedding space, where a bunch of well-collocated clothing items are embedded together while bad-collocated or irrelevant clothing items are embedded far apart. With this embedding space, a query clothing item can be used to retrieve other clothing items in terms of fashion collocation by indexing its closeness with all clothing items in the database. Mathematically, deep metric learning aims at learning a function $f_\theta(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$ parameterized by θ which maps user defined similar instances from the data manifold in $\mathbb{R}^F$ onto metrically close instances in $\mathbb{R}^D$, but analogously dissimilar instances in $\mathbb{R}^F$ onto metrically distant instances in $\mathbb{R}^D$. Before directly entering into our method, we briefly introduce triplet neural network.

Triplet Neural Network

Triplet neural network has been successfully applied to various vision tasks, including online-offline clothing image retrieval (Huang et al. 2015), face recognition (Schroff, Kalenichenko, and Philbin 2015), video representation learning (Wang and Gupta 2015). The basic idea beneath

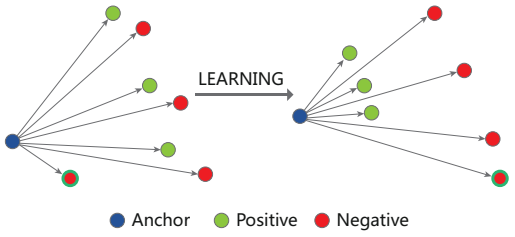


Figure 2: Deep mixed-category metric learning. The three positive and negative instances are initially mutually lying close in the deep embedding space, but far apart from anchor instance. Also, the anchor-negative instance pair lies very close. After training, all the positive instances are pulled closer to the anchor instance, while all negative instances are pushed far from the anchor instance.

triplet neural network is simple: it tries to learn a metric which automatically projects user-defined similar images to lie close while forcing dissimilar images to be far away in the embedding space. Specifically, triplet neural network usually consists of three identities: an anchor identity associating with a positive identity and a negative identity. The anchor and positive identity are similar but anchor and negative identity are different. The goal of triplet neural network is to minimize the distance between anchor and positive identity, while maximizing the distance between anchor and negative identity,

$$d(x_i^a, x_i^p) + \alpha < d(x_i^a, x_i^n), \forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T} \quad (1)$$

where x_i^a, x_i^p and x_i^n are anchor, positive and negative exemplar, respectively. α is a margin factor that controls the distance difference between anchor-positive and anchor-negative identity pair. The whole network loss thus is,

$$\mathcal{L}_{triplet} = \sum_i^N [d(f(x_i^a), f(x_i^p)) - d(f(x_i^a), f(x_i^n)) + \alpha]_+ \quad (2)$$

where $f(\cdot)$ is the image feature extractor and currently CNNs has often been adopted to learn feature representations. As to the similarity measurement $d(\cdot)$, cosine distance is widely adopted because it ignores the magnitude of the two feature representation (Liu et al. 2016a)(Schroff, Kalenichenko, and Philbin 2015),

$$d(x_1, x_2) = \frac{x_1 \cdot x_2}{\|x_1\| \|x_2\|} \quad (3)$$

Note that, in addition to cosine distance, Euclidean distance and other self-adaptive distances (*i.e.* position dependent deep metric (Huang, Loy, and Tang 2016)) are often applied to metric two features' similarity and we would discuss their performance on fashion collocation in the experiment section. By optimizing Eqn. (2), we learn an embedding space where the similarity is ruled by the anchor-positive-negative triples in the training database. Triplet neural network works well for categorical problems. For example, an image belongs to a single specific category and it is different to any image coming from other categories. Existing applications (Huang et al. 2015)(Schroff, Kalenichenko, and Philbin 2015) unanimously follow this rule. For instance, J. Huang *et al.* (Huang et al. 2015) aim at online-offline clothing retrieval. A query clothing item is categorically either similar or dissimilar to any clothing item in the database. F. Schroff *et al.* (Schroff, Kalenichenko, and Philbin 2015) harness triplet neural network to tell whether two faces belong to the same person. What's more, triplet neural network is of $\mathcal{O}(N^3)$ offline training size explosion, making it nearly impossible to consider all samples. Whereas fashion collocation is fundamentally different with them and there are two main challenges lie in fashion collocation:

Cross-category Inclusiveness the goal of fashion collocation is to retrieve several (usually 3-6) clothing items or accessories regarding any query item, these retrieved items come from different categories and thus hold large visual appearance discrepancy. For example, the denim jacket in Fig. 1 best collocates with chinos trouser, white T-shirt, shoe, golden watch. That is, we have to embed items across categories to the close localization in the deep embedding space.

Intra-category Exclusiveness Fashion collocation retrieves items across categories. On the one hand, given any query image, no item belonging to the same category with the query item should be retrieved. Nobody would be happy if a pair of sandal is retrieved for a canvas shoe. On the other hand, each individual retrieved item must be exclusive too. That is, there is one and only one white T-shirt best collocates denim jacket. The blue or black T-shirt is not a well-collocated candidate, although they share similar style or outlook with the T-shirt.

Deep Mixed-Category Metric Learning

Our deep mixed-category metric learning method simultaneously models the cross-category inclusiveness and intra-category exclusiveness. Generally, instead of inputing an instance triple, as the triplet neural network does, we directly feed multiple positive instances and negative instances to the neural network per time, where the positive instances comprise items forming well-collocation, negatives instances are randomly chosen but they individually share the same category name with positive instances one by one. Note that the positive instances are cross-category but the positive-negative pairs are intra-category, this is why we call our method deep mixed-category metric learning. The advantage of mixed-category training dataset preparation is threefold: First, we successfully avoid $\mathcal{O}(N^3)$ training sample explosion problem inherently exists in triplet neural network because the number of available mixed-category training samples is fixed or at least limited in size (depending on how many instances are involved in either positive or negative set). Second, mixed-category training dataset preparation method dynamically models cross-category inclusiveness within the positive instances and intra-category exclusiveness from positive-negative pairs. Third, it endows us with much flexibility for online hard exemplar mining, enabling us supervise the whole training process according

to our needs. An intuitive illustration of our deep mixed-category triplet neural network is shown in Fig. 2, in which the positive-negative pairs initially lie close to each other in the deep embedding space. After training, however, the positive instances cluster together, while all negative instances are automatically pushed far away in the deep embedding space.

Mathematically, the mini-batch is a mixed-category set $\{(x^p_{i,k}, x^n_{i,k}), i = 1, 2, 3, k = 1, 2, \cdots, K\}$, where $x^p_{i,k}$ and $x^n_{i,k}$ come from the same category, K is the mini-batch size. The mini-batch is feed to the same CNNs to learn feature representation $f(x)$. After feature representation, we select a set of triples $\sum < f(x^a_i), f(x^p_i), f(x^n_i) >$ to calculate the final loss according to the exemplar mining strategy that will be discussed in the next section.

$$\mathcal{L}_{mixed} = \sum^{M} [d(f(x^a_{i1}), f(x^p_{i2})) - d(f(x^a_{i1}), f(x^n_{i3})) + \alpha]_+ \tag{4}$$

where the anchor instance x^a_{i1} is directly sampled from the positive instances pool as they are mutually equivalent, M is the number of triples we sampled within the mini-batch. $i1, i2, i3$ here indicate the three instances that may come from any instance in the mini-batch. Instead of generating these triples offline, we choose to form triples from input samples online, it assigns us with much freedom to create whatever desirable triples needed for model training.

Hard-aware Online Exemplar Mining

Triplet neural network is notorious for the difficulty to train although various tricks have been added to ease the training (Cui et al. 2016)(Simo-Serra et al. 2015)
(Wang and Gupta 2015)(Liu et al. 2016a). Since our goal is to make sure all instances in the training database fulfil the requirement in Eqn. (1), it is theoretically desirable to select instances that violate Eqn. (1) to reinforce the model to learn meaningful things, as many existing work does (Cui et al. 2016)(Simo-Serra et al. 2015)(Wang and Gupta 2015)(Liu et al. 2016a). However, in real practice, choosing either the hardest positive instance or the negative instance for an anchor instance easily leads to the model collapsing into local optima at an early training stage or to be unstable, partially due the erroneous offline data preparation or intra-category exclusiveness property in fashion collocation. Furthermore, we find in practice that the training loss, since we fine-tune all model pre-trained on ImageNet dataset (Russakovsky et al. 2015), hardly reduce even without any hard exemplar mining. Under this circumstance, we propose a hard-aware online exemplar mining strategy to assist the whole training process. Specifically, we adopt a simple/semi-hard/hard exemplar mining strategy within a mini-batch for the early, intermidiate and later training stage respectively, to enable the model to be trainable during the first several epochs training but also reinforce the model to learn meaningful things at a later stage.

In simple exemplar mining, we first find two positive instances in the multiple positive pool sharing the smallest Euclidean distance. Then for each one of them, we choose the one which has the largest Euclidean distance to it as the its corresponding negative instance from the multiple negative pool.

$$\begin{aligned} &\arg\min_{x^p_1, x^p_2} \|f(x^p_1) - f(x^p_2)\|^2_2 \quad \arg\max_{x^n_1} \|f(x^p_1) - f(x^n_1)\|^2_2 \\ &\arg\max_{x^n_2} \|f(x^p_2) - f(x^n_2)\|^2_2 \\ &s.t. \quad x^n_1, x^n_2 \in \sum \big\{\{x^p_i\}, \{x^n_i\}\big\} \end{aligned} \tag{5}$$

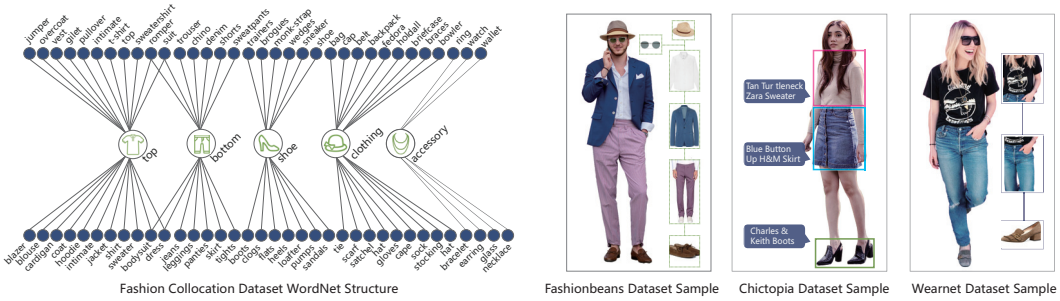
Note that in the simple exemplar mining strategy, the selection of negative instance for each selected positive instance is within the whole mini-batch pool, ignoring its micro and macro category label (see the dataset introduction section for the dataset organisation manner). The main reason is that, during the early course of training, we eagerly anticipate our model to learn useful fashion representation since the model is pre-trained on ImageNet (Russakovsky et al. 2015). Thus the pre-constructed similarity degree among clothing items is not that much important and we can directly ignore it. On the contrary, in the semi-hard case mining during the later training stage, we select the two positive instances and two negative instances by going the opposite way presented in Eqn. 5, but with a soft constraint,

$$\begin{aligned} &\arg\max_{x^p_1, x^p_2 \in \{x^p_i\}} \|f(x^p_1) - f(x^p_2)\|^2_2 \\ &\arg\min_{x^n_1 \in \{x^n_i\}} \|f(x^p_1) - f(x^n_1)\|^2_2 \quad \arg\min_{x^n_2 \in \{x^n_i\}} \|f(x^p_2) - f(x^n_2)\|^2_2 \\ &s.t. \ \|f(x^a_2) - f(x^p_2)\|^2_2 < \|f(x^a_2) - f(x^n_2)\|^2_2 \end{aligned} \tag{6}$$

In hard exemplar mining, all the four instances are rigidly selected within a single exemplar which contains a pair of multiple positive instances and negative instances with a specified similarity degree. Note that the reason why we select a quadruplet instead of triplet in hard case mining is that we want to fully exploit the available instances and according to C. Huang *et al.* (Huang, Loy, and Tang 2016), quadruplet selection best represents feature distribution.

Fashion Collocation Dataset

We introduce a new dataset as there is no publicly available dataset for fashion collocation research. First, we narrow down "well-collocation" to be fashionably dressing so that outfits wearing by fashion model or being recommended by fashion experts or trendsetters can be safely treated as fashionable well-collocation. Under this assumption, we have crawled about 0.2 million images from three famous fashion websites: chictopia, .wearnet.com and fashionbeans, each image is associated with three or more independent items indicating fashion experts' preference in creating the well-collocated outfit (so we gathered more than 700,000 images in total). Fashionbeans (30,000 images) is designed for male fashion collocation only, each recommended single item is independent and ready for wearing (we call off-body module). The clothing items in Wearnet (70,000 images) are either independent (off-body module) or dependent (on-body module). Chictopia (120,000) does not directly involve corresponding clothing items for each recommended outfit, but



Fashion Collocation Dataset WordNet Structure | Fashionbeans Dataset Sample | Chictopia Dataset Sample | Wearnet Dataset Sample

Figure 3: Left: WordNet hierarchy structure of fashion collocation dataset. Right: sample images crawled by the three websites.

we label each single clothing item with a tight bounding box. We organize the dataset in hierarchical WordNet architecture (see Fig. 3), in which each node is depicted by a variety of relevant clothing items. Besides, the WordNet tree is also sorted in coarse-to-fine manner from the top to bottom. We first roughly classify all items into five categories: top, bottom, shoe, clothing accessory and accessory (we call macro category), each one of which corresponds to multiple fine-grained clothing categories (we call micro category). Clothing accessory here specially indicates cloth-related items such as hat, gloves and satchel.

Experiment

We evaluate our proposed framework both quantitatively and qualitatively. In detail, we test both on-body and off-body module, which is achieved by evaluating on Fashionbeans (off-body), Wearnet (both off-body and on-body) and Chictopia (on-body) dataset separately. For experiment setup, we adopt GoogleNet (Szegedy et al. 2014) pre-trained on ImageNet ILSVRC-2012 (Russakovsky et al. 2015). Caffe deep learning framework (Jia et al. 2014) is used to train our model, the initial learning rate is set 0.005 but further divided by 10 per 50,000 iterations. The maximum iteration we set is 300,000 and the batch size is 30. The GPU servers are 4 TITAN X Pascals. Within our method, we want to figure out the performance influence of various training samples preparation methods on fashion collocation. Also, we would like to test the effect of deep embedding space size on fashion collocation. Outside our method, we test various distance metrics to see their impact on fashion collocation. For evaluation metric, we record Recall@K results and mean average precision (mAP). Recall@K is the average recall scores over all the query images in test set (Yuan, Yang, and Zhang 2016). Specifically, the recall score will be 1 if at least one positive image is retrieved in the top-K images and 0 otherwise. Note that we just record Recall@20 value because our retrieval is cross-category retrieving, we usually assume the retrieval result is good if the positive fashion images appears in top-20 results.

Moreover, we evaluate on both item-level and category-level so as to test our method's performance at different granularities, where item-level requires to retrieve the specific item regarding any query item so as to be the correct retrieval result, while category-level mitigates this harsh condition and merely requires to retrieve an item coming from different category with the query item to be a successful test. In fact, item-level measures our model's generalization ability to retrieve unseen items for any unseen query item, while category-level focuses on our model's capability on modelling cross-category inclusiveness and intra-category exclusiveness.

Sampling Method and Embedding Size Test We consider three sampling methods: siamese neural network(McAuley et al. 2015). We sampled 70,000 pairs from Chictopia, Wearnet and Fashionbeans dataset, respectively. Conventional triplet neural network, which samples a triplet per time. We also sampled 70,000 triplets for these three datasets respectively. For embedding size test, we set "loss1/loss1", "loss2/loss2" and "loss3/loss3" of GoogleNet to be 128 size, so we can form different embedding size by concatenating part of them together. In sum, we test 128 ("loss3/loss3"), 256 (concatenate "loss2/loss2" and "loss3/loss3") and 384 (concatenate "loss1/loss1", "loss2/loss2" and "loss3/loss3"). Moreover, we want to know the generalization capability of model leaned with fashion classification on fashion collocation. To this end, we trained another GoogleNet classification (category-level) model. All of these models are trained with cosine metric.

The result is shown in Table 1. We can clearly see that the fashion classification guided model produces the worst result on all of the three datasets. It is not difficult to understand it because classification oriented model retrieves clothing items of similar visual appearance with the query item, which obviously violates the intra-category exclusiveness requirement of fashion collocation. In addition, the pairwise sampling method (Siamese) works slightly inferior to triplet sampling method (Triplet_conv), attesting that simultaneously modelling the similarity and dissimilarity between items works better than modelling them separately. On the contrary, our mixed-category sampling method outperforms the other two sampling methods by a large margin, which shows involving a complete set of well-collocated instances as well as corresponding items sharing the same category name one by one guarantees the neural network to learn discriminative representations for fashion collocation, benefiting from the fact that our hard-aware online exemplar mining strategy assists the whole training process from coarse to fine, guaranteeing the neural network to be trainable and also

Table 1: Sampling method and embedding size test result in terms of Recall@20 and mean average (mAP).

Dataset	Fashionbeans				Wearnet				Chictopia			
Level	item_level		cate_level		item_level		cate_level		item_level		cate_level	
Metric	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP
GoogleNet_cls	0.07	0.05	0.13	0.07	0.06	0.05	0.07	0.06	0.06	0.04	0.06	0.05
Siamese (2015)	0.16	0.11	0.22	0.18	0.16	0.11	0.23	0.19	0.14	0.09	0.19	0.09
Triplet_conv	0.20	0.15	0.25	0.22	0.19	0.13	0.24	0.20	0.16	0.09	0.20	0.15
Ours_128	0.24	0.19	0.27	0.23	0.22	0.15	0.26	0.23	0.18	0.12	0.23	0.19
Ours_256	0.26	0.23	0.28	0.24	0.24	0.20	0.28	0.25	0.20	0.15	0.26	0.23
Ours_384	**0.34**	**0.25**	**0.37**	**0.29**	**0.32**	**0.22**	**0.34**	**0.27**	**0.32**	**0.20**	**0.33**	**0.29**

to learn meaningful things at different training stages. Moreover, as the embedding size increases, the fashion collocation retrieval performance boosts significantly, which shows concatenating features arising from different CNNs layers truly boost performance.

Loss Function Test The aforementioned experiment leverages cosine distance and hinge loss. Many other loss functions have been designed to boost various deep metric learning methods (Huang, Loy, and Tang 2016)(Ustinova and Lempitsky 2016)(Yuan, Yang, and Zhang 2016)(Lin et al. 2016). In this section, we want to figure out these new loss functions' performance on fashion collocation. We test 6 kinds of loss functions: **HingeLoss** (Schroff, Kalenichenko, and Philbin 2015) with cosine distance. **PDDMLoss** (Huang, Loy, and Tang 2016) proposed a position dependent deep metric unit to measure two instances' similarity. **LiftedStructLoss** (Song et al. 2015) formulated a structured prediction objective function with lifted dense pairwise distance matrix. **HistogramLoss** (Ustinova and Lempitsky 2016): which penalizes the overlap between distance distribution of positive pairs and negative pairs. **HDCLoss** (Yuan, Yang, and Zhang 2016) hierarchically differentiates instances according to their hard level. **AffineLoss** (Lin et al. 2016): which explicitly models the distance between an instance pair with affine transformation.

The quantitative result is shown in Table 2, from which we get several important results. First, AffineLoss performs the worst. This shows explicit affine transform cannot fully grasp the collocation relationship among clothing items. Except for AffineLoss, the other four loss functions overwhelmingly outperform the basic hinge loss by a large margin, attesting the necessity of designing more complex and powerful loss function to tackle fashion collocation problem. Among the three datasets, Chictopia benefits most from these loss functions (20% recall@20 improvement on average), the reason is that those carefully designed loss functions excel at handling occlusion, overlapping and intervention problems that commonly exist for on-body module fashion collocation. Moreover, HDCLoss performs the best due to the hierarchical ranking strategy it exploits to dig the in-depth property of fashion collocation. It also echoes with the embedding size test result shown in previous section where the embedding size 384 deriving from three layers performs the best. In the end, we find the newly designated loss functions specially improve our method's capability in retrieving accessories (we do not give qualitative evaluation on accessory retrieval but it can be directly seen by the huge improvement on Fashionbeans dataset because it contains many accessories), this shows that small object based or fine-grained fashion collocation needs specially designed loss functions, it awaits more deep research to resolve it.

Collocation Result Visualization We provide visualization in Fig. 4. The results are generated by HDCLoss with deep mixed-category metric learning. Several interesting phenomenons can be observed. First, our proposed deep mixed-category metric learning method achieves cross-category inclusiveness requirement of fashion collocation and can retrieve clothing items w.r.t. the query item across category in most cases. Second, our method is still color-biased. It prioritizes the item with the same color with the query item while first meeting the cross-category inclusiveness requirement. For example, in the Chictopia test in the second row, the floral top is first retrieved from the database regarding the query short skirt with the same floral pattern. Our fashion collocation model emphasizes color similarity when intra-category exclusiveness and cross-category inclusiveness are met. In the end, large ambiguity still exists in both our model and fashion collocation dataset itself. For instance, in the Fashionbeans test (third row), after the dark blue top, bottom as well as the khaki trench coat being successfully retrieved, our model meets an ambiguity that whether it should retrieve the portfolio or the leather shoe first? In fact both of them are fashionably well-collocated with the dark blue top. Moreover, to figure out the gap between our fashion collocation dataset and human being's understanding towards fashion collocation, we asked several human beings to collocate with a query item, and further compare the collocation results with recommended collocation by our fashion collocation dataset. Although most of them (about 70%) are compatible, ambiguity still exists. As is shown in the last row in Fig. 4, the human collocated items enjoy more color variety than the items recommended by our dataset. Moreover, we find our algorithm is likely to retrieve items sharing similar color information, showing our algorithm's preference to color while collocating. We also want to figure out the performance of our framework on traditional retrieval tasks. To this end, we conducted experiment on the In-Shop fashion retrieval task (Liu et al. 2016b) and got R@1:74.12% R@10: 92.43%, achieving the state of the art performance among reported results.

More Discussion As the hard-aware online exemplar mining strategy proposed by us is of vital importance to supervise the whole training process, we naturally want to figure out the impact of various online exemplar mining strategies on the deep neural network training. The loss curves of various exemplar mining strategy are shown in Fig. 5,

Table 2: Loss function test result in terms of Recall@20 and mean average accuracy (mAP).

Dataset	Fashionbeans				Wearnet				Chictopia			
Level	item_level		cate_level		item_level		cate_level		item_level		cate_level	
Metric	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP	Recall@20	mAP
HingeLoss (2015)	0.34	0.25	0.37	0.29	0.32	0.22	0.34	0.27	0.32	0.20	0.33	0.29
PDDMLoss (2016)	0.44	0.34	0.48	0.35	0.42	0.28	0.44	0.36	0.40	0.30	0.42	0.36
LiftedStructLoss (2015)	0.45	0.33	0.50	0.35	0.43	0.29	0.44	0.37	0.41	0.31	0.42	0.37
HistogramLoss (2016)	0.43	0.33	0.52	0.44	0.42	0.28	0.43	0.36	0.40	0.30	0.42	0.32
AffineLoss (2016)	0.32	0.24	0.36	0.28	0.30	0.21	0.32	0.27	0.29	0.20	0.31	0.28
HDCLoss (2016)	**0.49**	**0.40**	**0.57**	**0.42**	**0.44**	**0.37**	**0.46**	**0.38**	**0.43**	**0.36**	**0.45**	**0.36**

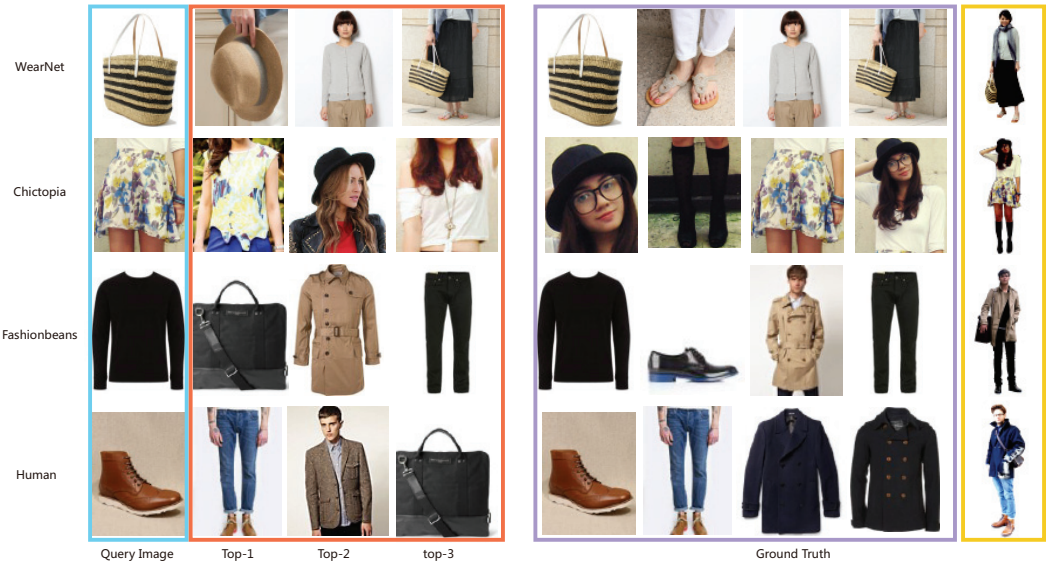


Figure 4: Sample images of clothing fashion collocation result by deep mixed-category metric learning. The leftmost image is the query image and next three adjacent images are samples retrieved from top-20 result. The four ground truth images as well as the overall collocation result are in the right side.

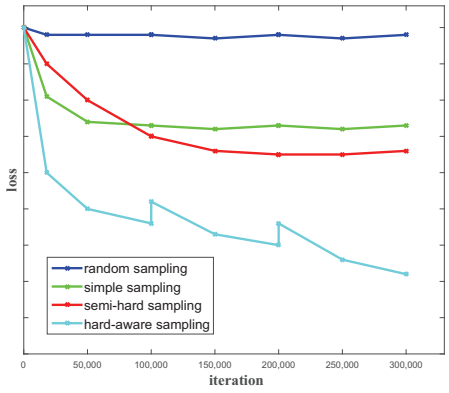


Figure 5: Loss curves variation trend.

from which we can clearly see that the loss curve hardly reduces with random sampling strategy, which means fashion collocation needs special instance triples to learn meaningful things. Simple exemplar mining and semi-hard exemplar mining alone lead to the loss curve to reduce during the several early epochs but soon level off at a high loss. However, our hard-aware online example mining strategy enables to loss curve to gradually go down to small loss value, although two inflexion points are encountered during simple/semi-hard, semi-hard/hard mining strategy transformation. This shows that applying simple exemplar mining, semi-hard exemplar mining and hard exemplar sequentially assigns the whole neural network with different goal at different training stages.

Conclusion

We have proposed a deep mixed-category metric learning framework to address fashion collocation by combing deep convolutional neural network and deep metric learning. A hard-aware online exemplar mining strategy has been proposed to supervise the whole training process. Experimental results show the feasibility and superiority of the method.

Acknowledgement This work is supported in part by the National Natural Science Foundation of China Under Grant 61773414, 41401525.

References

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. 2016. Fine-grained categorization and dataset boostrapping using deep

metric learning with humans in the loop. In *Proc. CVPR*.

He, Y., and Chen, L. 2016. Fast fashion guided clothing image retrieval: Delving deeper into what feature makes fashion. In *Proc. ACCV*.

He, R.; Lin, C.; and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proc. WWW*.

Hermans, A.; Beyer, L.; and Leibe, B. In defense of the triplet loss for person re-identification. In *arXiv preprint arXiv: 1703.07737*.

Huang, J.; Feris, R.; Chen, Q.; and Yan, S. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proc. ICCV*.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Local similarity-aware deep feature embedding. In *Proc. NIPS*.

Jia, Y.; Evan, S.; Jeff, D.; Sergey, K.; Jonathan, L.; Ross, G.; Sergio, G.; and Trevor, D. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM MM*.

Kiapour, M.; Yamaguchi, K.; Berg, A.; and Berg, T. 2014. Hispster wars: Discovering elements of fashion styles. In *Proc. ECCV*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*.

Lin, L.; Wang, G.; Zuo, W.; Feng, X.; and Zhang, L. 2016. Cross-domain visual matching via generalized similarity measure and feature learning. *Transaction of pattern recogniton and machine intelligence (PAMI)*.

Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; and Huang, T. 2016a. Deep relative distance learning: Tell the difference between similar vehicles. In *Proc. CVPR*.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016b. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. CVPR*.

McAuley, J.; Targett, C.; Shi, Q.; and Hengel, A. 2015. Image-based recommendations on style and substitutes. In *Proc. ACM SIGIR*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*.

Serra, E. S., and Ishikawa, H. 2016. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *Proc. CVPR*.

Serra, E.; Fidler, S.; Noguer, F.; and Urtasun, R. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proc. CVPR*.

Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; and Moreno-Noguer, F. 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proc. ICCV*.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2015. Deep metric learning via lifted structured feature embedding. In *Proc. NIPS*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2014. Going deeper with convolutions. In *arXiv preprint arXiv: 1409.4842*.

Ustinova, E., and Lempitsky, V. 2016. Learning deep embeddings with histogram loss. In *Proc. NIPS*.

Veit, A.; Kovacs, B.; Bell, S.; McAuley, J.; Bala, K.; and Belongie, S. 2015. Learning visual clothing style with heterogeneous dyads. In *Proc. ICCV*.

Vittayakorn, S.; Yamaguchi, K.; Berg, A. C.; and Berg, T. L. 2015. Runway to realway: Visual analysis of fashion. In *Proc. WACV*.

Wang, X., and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *Proc. ICCV*.

Wang, J.; Nabi, A.; Wang, G.; Wan, C.; and Ng, T. 2015. Towards predicting the likeability of fashion images. *arXiv preprint arXiv: 1511.05296*.

Yamaguchi, K.; Kiapour, M.; and Berg, T. 2013. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proc. ICCV*.

Yuan, Y.; Yang, K.; and Zhang, C. 2016. Hard-aware deeply cascaded embedding. In *arXiv preprint arXiv: 1611.05720*.