Metric Learning for Semantic-Based Clothes Retrieval

YANG Bo[1], GUO Caili[1,2], LI Zheng[1]

(1. Beijing Laboratory of Advanced Information Networks, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTECOM.202201010

https://kns.cnki.net/kcms/detail/34.1294.TN.20220307.1444.002.html, published online March 9, 2022

Manuscript received: 2021-11-15

Abstract: Existing clothes retrieval methods mostly adopt binary supervision in metric learning. For each iteration, only the clothes belonging to the same instance are positive samples, and all other clothes are "indistinguishable" negative samples, which causes the following problem. The relevance between the query and candidates is only treated as relevant or irrelevant, which makes the model difficult to learn the continuous semantic similarities between clothes. Clothes that do not belong to the same instance are completely considered irrelevant and are uniformly pushed away from the query by an equal margin in the embedding space, which is not consistent with the ideal retrieval results. Motivated by this, we propose a novel method called semantic-based clothes retrieval (SCR). In SCR, we measure the semantic similarities between clothes and design a new adaptive loss based on these similarities. The margin in the proposed adaptive loss can vary with different semantic similarities between the anchor and negative samples. In this way, more coherent embedding space can be learned, where candidates with higher semantic similarities are mapped closer to the query than those with lower ones. We use Recall@K and normalized Discounted Cumulative Gain (nDCG) as evaluation metrics to conduct experiments on the DeepFashion dataset and have achieved better performance.

Keywords: clothes retrieval; metric learning; semantic-based retrieval

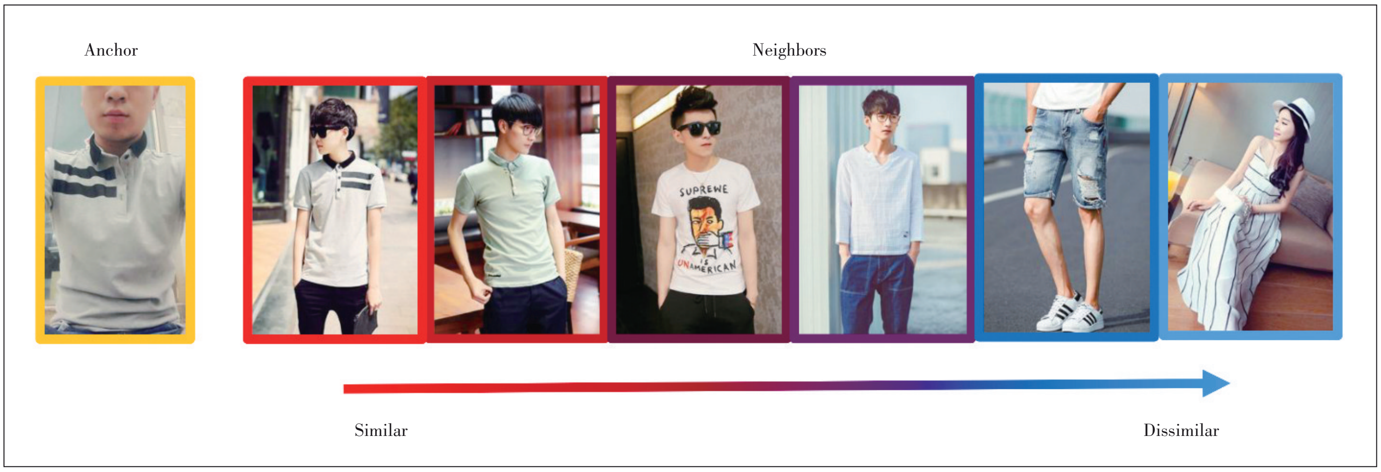
Citation (IEEE Format): B. Yang, C. L. Guo, and Z. Li, "Metric learning for semantic-based clothes retrieval," *ZTE Communications*, vol. 20, no. 1, pp. 76 – 82, Mar. 2022. doi: 10.12142/ZTECOM.202201010.

1 Introduction

Clothes retrieval, commonly associated with visual search, has received a lot of attention recently, which is also an interesting challenge from both commercial and academic perspectives. The task of clothing retrieval is to find the exact or very similar products from the gallery to the given query image. It mainly has two scenarios, namely, in-shop clothes retrieval and consumer-to-shop clothes retrieval. In the former case, query and gallery come from the same domain while it is opposite in the latter case. In consumer-to-shop clothes retrieval, photos in the gallery usually only contain a single product and are taken by professionals using high-quality equipment. On the contrary, the photos taken by the user are usually of low quality, possibly with cluttered backgrounds and multiple unrelated objects[1].

The above uncertain factors have brought great challenges to consumer-to-shop clothes retrieval. In this complex scenario, only the model which can discover the differences between clothes from a semantic point of view will not be affected by environmental factors, so as to obtain better retrieval results. However, most of the current clothing retrieval works belong to instance-based retrieval, and they are all based on the following assumptions: Only the clothes belonging to the same instance are considered relevant, while all the other clothes are irrelevant. However, this assumption appears conflicted with our common sense. For example, as shown in Fig. 1, the semantic similarity between clothes is more than relevant or irrelevant. Instead, it shows a continuous decreasing trend. From a category point of view, when we use a polo shirt as a query, except for the same polo shirt, the other polo shirts should theoretically be the most similar. Ordinary short-sleeved and long-sleeved may follow next, and finally it comes to other different categories such as pants and skirts. However, under the premise of the above assumption, models are more likely to learn the ranking relationship among clothes, rather than their semantic similarities[2]. Moreover, the learned embedding space appears not continuous either, which may lead to worse user experience.

In order to address the above mentioned problems, Refs. [2 – 4] propose several metric learning methods using data with continuous labels, which is called semantic-based retrieval. Our work is inspired by and completed based on the research in these fields. In this paper, we propose a new method called Semantic-Based Clothes Retrieval (SCR) to measure the semantic similarity between clothes, and these



▲Figure 1. Ideal relationship between clothes in different categories

similarities can guide the learning process of continuous embedding space. In this way, we realize the semantic-based clothes retrieval eventually.

Our contributions can be summarized as follows.

- We expose the problem in the instance-based clothes retrieval. The relationship between clothes is simply regarded as relevant or irrelevant, which is not consistent with the actual situation.
- We propose a novel method called SCR to measure the semantic similarity between clothes and design a new adaptive loss based on these similarities. As a result, clothes in a gallery are ranked by their similarity to the query, making multiple clothes relevant.
- We use Recall@k and normalized discounted cumulative gain (nDCG) as the evaluation metrics and conduct experiments on the DeepFashion dataset, and have achieved better results.

2 Related Work

2.1 Instance-Based Clothes Retrieval

Pioneer works[5 – 8] in clothing retrieval utilized traditional features like the scale invariant feature transform (SIFT). Later, due to the wide application of deep neural networks, the development of computer vision was greatly promoted. In Refs. [9 – 16], these methods usually extracted both global features and local features, combining them for similarity calculation and matching. Recently, the authors in Ref. [17] have proposed a graph reasoning network (GRNet), which first represented the multi-scale regional similarities and their relationships as a graph and then performed graph convolutional neural network (CNN) based reasoning over the graph to adaptively adjust both the local and global similarities.

All the aforementioned methods focus on the feature extraction stage of instance-based clothes retrieval. Closest to our work, Ref. [1] transferred the leading ReID model called the ReID strong baseline (RST)[18] to fashion retrieval task, significantly outperforming previous state-of-the-art results despite a much simpler architecture. In this paper, we use this ReID model as our backbone for subsequent semantic-based clothes retrieval.

2.2 Metric Learning for Instance-Based Clothes Retrieval

Metric learning attempts to map data to an embedding space, where similar data are close together while dissimilar data are far apart[1]. In general, metric learning can be achieved by means of embedding and classification losses, and both of them are often utilized at the same time in most retrieval methods. Many state-of-the-art methods combine ID and triplet losses to constrain the same feature f. Combining these two losses always makes the model achieve better performance[18].

The contrastive loss[19] and the triplet loss[20], as two typical embedding losses, provide the foundation of metric learning. Given an image pair, the contrastive loss minimizes their distance in the embedding space if their classes are the same, and separates them with a fixed margin away otherwise. The triplet loss takes triplets of anchor, positive, and negative images, and enforces the distance between the anchor and the positive to be smaller than that between the anchor and the negative image. A wide variety of losses has since been built on these fundamental concepts such as quadruplet loss, n-pair loss, and lifted structured loss.

Although the above mentioned losses substantially improve the quality of the learned embedding space, they are commonly based on binary relations between image pairs, thus they are not directly applicable for metric learning in semantic-based retrieval.

2.3 Metric Learning for Semantic-Based Retrieval Using Continuous Labels

There have been several metric learning methods for semantic-based retrieval using data with continuous labels in some research areas such as image caption[21], place recognition[22] and camera relocalization[23].

Similar to our work, KIM et al.[2] propose a log-ratio loss to learn embedding space in image retrieval. Their work primarily focuses on human poses, in which they use the distance between joints to rank images. They also explore within-modal image retrieval using word mover's distance, as a proxy for semantic similarity. For cross-modal retrieval, ZHOU et al. [3] propose to measure the relevance degree among images and sentences and design a ladder loss to learn coherent embedding space. In the ladder loss, these relevance degree values are divided into several levels, but the relevance in each level is still formulated as a binary variable. WRAY et al.[4] propose the task of semantic similarity video retrieval (SVR), which allows multiple captions to be relevant to a video and vice-versa, and defines non-binary similarity among items. In addition, they also propose several proxies to estimate semantic similarities and introduce nDCG as the evaluation metric.

The above methods all implement semantic-based retrieval by metric learning with continuous labels, and the premise is that the dataset provides continuous labels, or the similarity can be measured by well-designed proxies. While, in the field of clothing retrieval, it is difficult to measure the semantic similarities among clothes because of the semantic gap between visual similarity and semantic similarity. Therefore, semantic-based clothing retrieval has not been realized so far.

3 Semantic-Based Clothes Retrieval

In this section, we propose to move beyond instance-based clothes retrieval towards that uses semantic similarity among clothes. In our proposed SCR, we first measure the semantic similarity for clothes in Section 3.1. And then we modify classic triplet loss to adaptive loss for clothes retrieval using semantic similarity in Section 3.2. In this way, coherent embedding space can be learned, where candidates with higher semantic similarities are mapped closer to the query than those with lower semantic similarities.

3.1 Semantic Similarity for Clothes

The semantic gap exists between the raw image and the full semantic understanding of the image's content. In our method, we take full advantage of annotation information in DeepFashion to bridge the gap.

Various tags of clothes are labeled in the DeepFashion dataset, in addition to categories, bounding boxes, key points, and the attributes of clothes. Taking the consumer-to-shop benchmark as an example, each piece of clothing has 303 attribute tags. These attribute tags are attached to 18 categories, including clothing length, thickness, material, style and other categories. According to these attributes, clothing can be described from multiple points of view. Previous works used these attribute annotations as supervision for the multi-label classification task, and combined clothing classification and retrieval loss to jointly train a neural network. However, due to the complexity and diversity of attribute tags, the multi-label classification task is considered an extraordinary challenge. In order to fully utilize the semantic information hidden in the annotated labels, we do not use attribute labels as the supervision for attribute classification, but directly use them to guide the learning process of the embedding space.

Specifically, given a set of clothes instances C, we let c_i be a clothes instance. We use a 303-dimensional vector of 0 or 1 to represent the attribute vector of $\boldsymbol{c}_i$, denoted by $\boldsymbol{s}_i$. The similarity between $\boldsymbol{s}_i$ and other clothes $\boldsymbol{s}_j$ can be measured by the inner product of two vectors,

$$S\left(\boldsymbol{c}_i,\boldsymbol{c}_j\right) = \left\langle \boldsymbol{s}_i,\boldsymbol{s}_j \right\rangle. \tag{1}$$

That is to say, if the corresponding position of s_i and s_j is 1, which means two clothes both have a certain attribute, and after accumulation of all positions, the final value is used to represent the semantic similarity between them. The inner product of the vector and itself is always the largest, which is consistent with "clothing is always the most similar to itself". Besides, the attribute labels are characterized by fine granularity as well as high accuracy, so they are able to quantitatively measure the semantic similarities between clothes. We use the calculated similarities to guide the follow-up learning process of embedding space as well as evaluate the performance of the retrieval model.

3.2 Adaptive Loss

The classic triplet loss takes a triplet of an anchor, a positive, and a negative image as input. It is designed to penalize triplets violating the rank constraint. The distance between the anchor and the positive images must be smaller than that between the anchor and the negative images in the embedding space. The loss is formulated as:

$$L_{\text{Triplet}} = \left[\lambda + D\left(c_a,c_p\right) - D\left(c_a,c_n\right)\right]_+, \tag{2}$$

where $\left(c_a,c_p\right)$ is the positive pair, $\left(c_a,c_n\right)$ is the negative pair for a query, $D(\cdot)$ means the squared Euclidean distance of the embedding vector, λ is a fixed margin, and $[\cdot]_+$ denotes the hinge function. Note that the embedding vectors should be L2 normalized since their magnitudes tend to diverge and the margin becomes trivial without such a normalization. For the RST model used in clothes retrieval[1], the mining strategy selects the most difficult positive and negative samples, that is the farthest positive sample and the closest negative sample in embedding space.

The triplet loss tends to treat the relevance between query and candidates in a bipolar way: for a query c_a, only the exactly same clothes c_p are regarded as relevant, and other clothes c_n are all regarded as irrelevant. Therefore, only c_p is pulled closed to c_a, while others are pushed away by a fixed margin equally. However, as mentioned in Section 1, the semantic similarity between samples should not be a binary variable.

In ideal embedding space, the difference in distance between positive/negative samples and anchor should be proportional to the difference in semantic similarities between them. In other words, a negative pair with lower semantic similarity should be pulled farther apart. Therefore, it is beneficial to introduce the semantic similarities to determine how far negative samples will be pushed away. We design a novel adaptive loss for clothes retrieval based on a classic triplet loss as follows:

$$L_{\text{Adaptive}} = \left[M_{an} + D\left(c_a, c_p\right) - D\left(c_a, c_n\right) \right]_+ , \tag{3}$$

$$M_{an} = \left(1 - \frac{S\left(c_a, c_n\right)}{S_{\max}} \right) * \lambda , \tag{4}$$

where $S_{\max}$ is a normalization factor to guarantee that the maximum value of semantic similarity is 1. 1 minus the normalized value, as the coefficient of λ, can allow those negative samples that are very similar to the anchor to be pushed away by a small margin. On the contrary, those negative samples that are not similar to the anchor will be pushed away by a large margin.

In the proposed adaptive loss, the margin between the anchor and the corresponding negative sample is no longer a fixed value, but a dynamic distance varying with semantic similarity computed in Eq. (1). In this way, a model is trained under metric learning beyond binary supervision and a coherent embedding space is learned as a result.

4 Experiments

4.1 Dataset and Experiment Settings

We evaluate our loss on the consumer-to-shop clothes retrieval benchmark of DeepFashion[11]. This benchmark aims at matching consumer-taken photos with their shop counterparts. It contains 33 881 clothing items, 239 557 consumer/shop clothes images, 195 540 cross-domain pairs, and each image is annotated by the bounding box, clothing type and source type. In our experiments, we use 96 708 images for testing and the remaining for training.

In this section, we select the RST model[1] that has obtained the best performance in clothes retrieval as the backbone network for the following experiments. Our loss function consists of two parts, which are obtained by adding the classification loss and the ranking loss (adaptive loss). We keep the identity loss in the original RST model and use it as the classification loss. As for the adaptive loss, we first select the closest negative sample to the anchor and the farthest positive sample in a batch, and then use the dynamic margin calculated by the semantic similarity between the negative sample and anchor to replace the fixed λ in the triplet. The adaptive loss we proposed can be used as an improved version of triplet loss in any backbone. In this paper, RST with the original classic triplet loss is considered as a comparison method.

Except for the loss function, several variants of the RST model were proposed in Ref. [1], and we select one with ResNet50-IBN-A backbone, 320×320 input image size, and no re-ranking setting as our baseline. We follow the warmup learning rate strategy proposed in the original RST model[18]. Models are trained using Adam for 120 epochs, with a batch size of 64. Hyper-parameters are set as $\lambda = 0.3$, and $\epsilon = 0.1$.

For instance-based retrieval, we use the Recall@K (R@K) as the evaluation metric for the task. R@K indicates the percentage of queries for which the model returns the correct item in its top K results.

For semantic-based retrieval, we use nDCG[24] as evaluation metrics. The nDCG has been used previously for information retrieval[25]. It requires similarity scores among all items in the test set. We calculate discounted cumulative gain (DCG) for a query q_i and the set of items Z, ranked according to their distance from q_i in the learned embedding space:

$$\text{DCG@}K\left(q_i\right) = \sum_{j=1}^{K} \frac{2^{S\left(q_i, z_j\right)} - 1}{\log\left(j + 1\right)} , \tag{5}$$

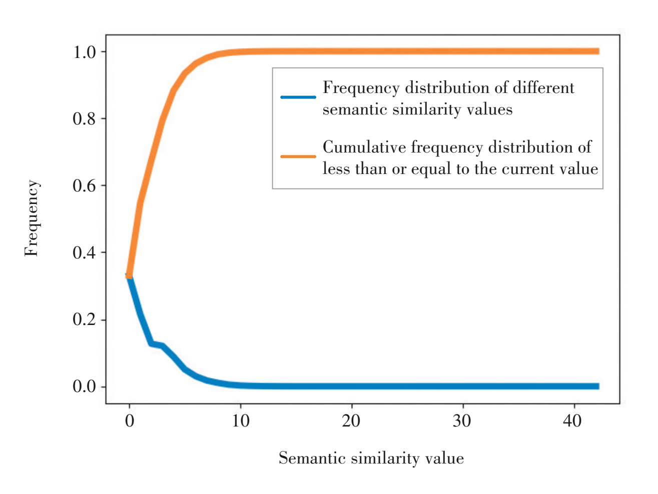
where K is the length of the list returned by the retrieval system. Note that this equation would give the same value when items of the same similarity S_S are retrieved in any order. It also captures different levels of semantic similarity. The nDCG can then be calculated by normalizing the DCG score so that it lies in the range [0, 1]:

$$\text{nDCG@}K\left(q_i\right) = \frac{\text{DCG@}K\left(q_i\right)}{\text{IDCG@}K\left(q_i\right)}, \tag{6}$$

where $\text{IDCG@}K\left(q_i\right)$ is calculated from DCG and Z ordered by relevance to the query q_i.

4.2 Analysis of Semantic Similarity

In this paper, we follow the method introduced in Section 3.1 to obtain the inner product of two attribute vectors as the semantic similarity among clothes. After statistics, the minimum value of similarity is 0 and the maximum is 42, as shown in Fig. 2, where the blue curve represents the frequency distribution of different semantic similarity values and the sum of these values is 1; The orange line shows the cumulative frequency distribution of less than or equal to the current value. In this range, the greater the similarity, the lower the frequency of occurrence. Among them, the proportion of similarity of 0 or 1 is 54%, and the proportion of similarity of less



▲**Figure 2. Frequency statistical graph of semantic similarity**

than or equal to 8 is as high as 99%. In other words, the similarity among most clothes is generally very low, and the similarity of only a few clothes can be close to or equal to their own similarity. Our goal is to use these high similarity samples to learn their common features from a semantic point of view, so as to optimize the learning quality of the embedding space.

Fig. 3 shows the normalized semantic similarities calculated by the inner product of the 303-dimensional attribute vectors. Note that similarities here are normalized by S_{ap} instead of $S_{\max}$ for better understanding. If clothes possess all the attribute tags of anchor, their semantic similarity with anchor will come to 1. While, in the training process, we use the semantic similarity normalized by $S_{\max}$ to determine the distance that negative samples are pushed away from the query, that is, the lower the semantic similarity, the farther will be pushed away from the query.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares the proposed SCR with state-of-the-art methods, including FashionNet[11], Visual Attention Model (VAM) and its variants (VAM+ImgDrop, VAM+Product, and VAM+Nonshared)[26], DREML[27], KPM[28] and GRNet[17]. The proposed SCR obtains competitive results. Specifically, it obtains an accuracy of 29.2, 51.0 and 61.4 and outperforms the existing methods at R@1.

4.4 Comparison with Triplet Loss and Quadruplet Loss

Since the adaptive loss we proposed is an improved version of the triplet loss, in order to verify whether the proposed adaptive loss is effective, we compare it with the triplet loss in both instance-based and semantic-based retrieval. Moreover, a comparison with quadruplet loss is also revealed below.

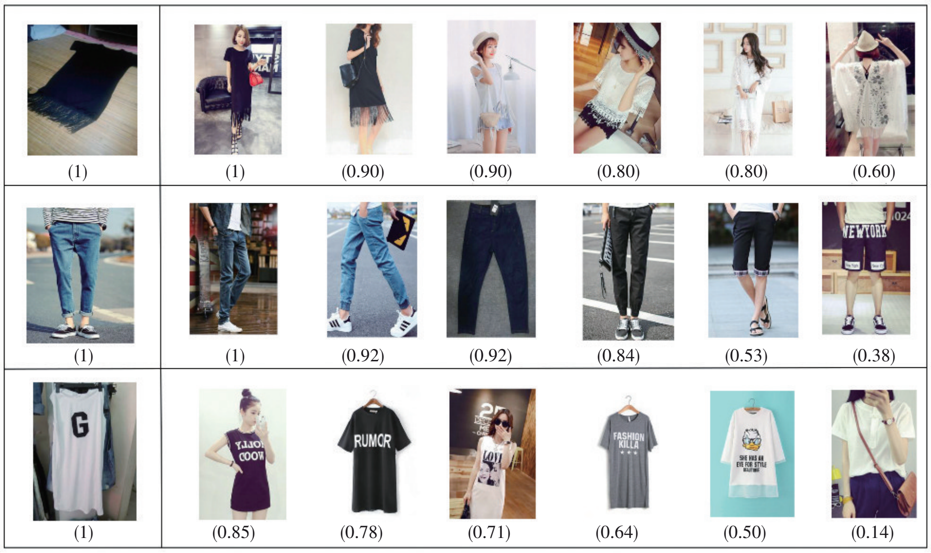
1) Instance-based retrieval results

As shown in Table 2, the adaptive loss has a significant improvement compared with a triplet loss function. This means that the information provided by the calculated semantic similarity can improve the quality of the learned embedding space and performance better in the retrieval process. However, Recall@k only cares about the appearance of the first k results, and does not care about their sorting positions, so the measurement results for our optimization based on semantic similarity are limited.

2) Semantic-based retrieval results

Table 3 summarizes the results evaluated by nDCG. The proposed loss function has an obvious improvement. In terms of nDCG@5, nDCG@10, and nDCG@50, the adaptive loss separately increases by 4.3%, 7.6%, 6.7%, compared with the classical triplet loss. Since the nDCG metric can simultaneously consider the relevance degree and ranking position, it reflects semantic-based retrieval results more accurately.

Fig. 4 shows the qualitative comparison between the triplet loss and adaptive loss on the consumer-to-shop benchmark using RST. For each retrieval, given an image query, we show the top-5 retrieved products. The correct retrieval items for each query are outlined in green boxes. Compared with the triplet loss, the correct query result ranks higher. Moreover, the search results based on the



The number in brackets is the normalized semantic similarity between the query and candidates that calculated by their attribute vectors

▲**Figure 3. Normalized semantic similarity visualization** adaptive loss possess higher semantic similarity with the query. For example, in the first query results based on the adaptive loss, although the first two candidates are not considered ground truth, they are indeed indistinguishable from the query, which also proves that it is not in line with the actual retrieval requirements to only consider ground truth to be relevant.

In addition, even if it is an incorrect query result, it will be semantically closer to the query. For example, in the third query result, when using a shirt with a distinctive pattern as a query, the network trained by the triplet loss pays too much attention to the local similarity because four of the top five search results are pants with similar patterns. From the perspective of the clothing category, the search results are completely contrary to the user' s intention. The network trained by the adaptive loss focuses on the overall semantic similarity so the searching scope tends to be restricted to the same category, which is crucial to the user experience. This benefits from the fact that the margin we use in the adaptive loss can adjust dynamically according to different samples. In this way, the model is easier to explore the semantic relationship among samples and finally a more semantic coherent embedding space is learned.

▼Table 1. Comparison with state-of-the-art methods on DeepFashion consumer-to-shop benchmark

Methods	R@1	R@20	R@50
FashionNet[11]	7.0	18.8	22.8
VAM+Nonshared[26]	11.3	38.8	51.5
VAM+Product[26]	13.4	43.6	56.7
VAM+ImageDrop(192, 48)[26]	13.7	43.9	56.9
DREML[27]	18.6	51.0	59.1
KPM[28]	21.3	54.1	65.2
GRNeT[17]	25.7	64.4	75.0
SCR (Adaptive)	29.2	51.0	61.4

GRNet: graph reasoning network
KPM: Kronecker-product matching
SCR: semantic-based clothes retrieval
VAM: visual attention model

▼Table 2. Instance-based retrieval results on DeepFashion

Loss Functions	R@1	R@5	R@1	mAP	Mean
Triplet	26.9	35.9	41.0	33.9	1 275
Quadruplet	26.3	35.1	40.3	33.3	1 348
Adaptive	29.2	38.6	44.0	36.6	1 091

▼Table 3. Semantic-based retrieval results on DeepFashion

Loss Functions	NDCG@1	NDCG@10	NDCG@50
Triplet	22.2	21.3	16.4
Quadruplet	21.8	20.9	16.1
Adaptive	23.9	22.8	17.5

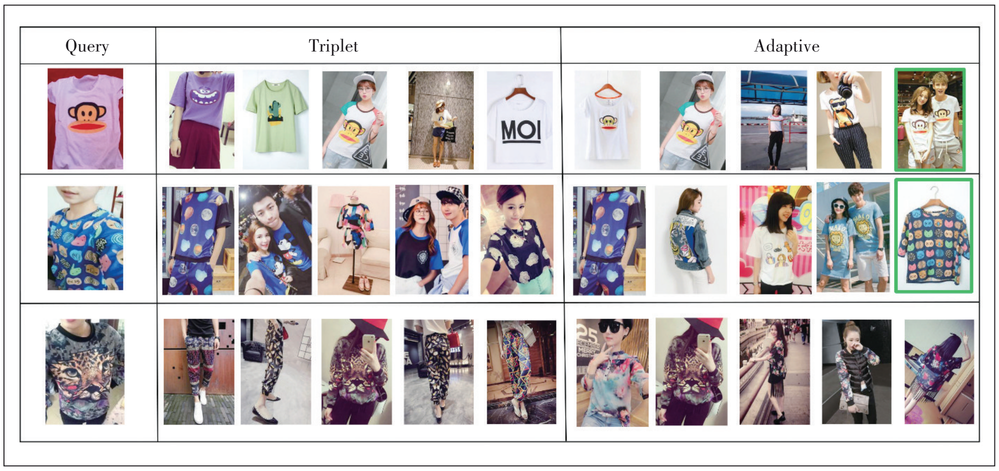
nDCG: normalized discounted cumulative gain

5 Conclusions

This paper focuses on semantic-based clothes retrieval and proposes a novel method called SCR to measure the semantic similarity between clothes. Motivated by metric learning with continuous labels in other research areas, we modify the classic triplet loss using semantic similarity and design an adaptive loss for clothes retrieval. As a result, more reasonable embedding space is learned, where candidates with higher semantic similarities are mapped closer to the query than those with lower similarities, which is more in line with the actual user experience of the retrieval system. Our method outperforms the baseline and obtains competitive semantic-based retrieval results on consumer-to-shop retrieval benchmarks of DeepFashion.

Acknowledgment

We would like to acknowledge LIU Xuean, HE Baolin, ZHANG Yuanjian from ZTE Corporation for their contribution to this paper.



▲Figure 4. Qualitative retrieval comparison between triplet loss and our adaptive loss on consumer-to-shop benchmark

References

[1] WIECZOREK M, MICHALOWSKI A, WROBLEWSKA A, et al. A strong baseline for fashion retrieval with person re-identification models [EB/OL]. (2020-03-09) [2021-11-05]. https://www. arxiv-vanity. com/papers/2003.04094/. DOI: 10.1007/978-3-030-63820-7_33

[2] KIM S, SEO M, LAPTEV I, et al. Deep metric learning beyond binary supervision [EB/OL]. (2019-04-21) [2021-11-05]. https://arxiv. org/abs/1904.09626. DOI: 10.1109/CVPR.2019.00239

[3] ZHOU M, NIU Z X, WANG L, et al. Ladder loss for coherent visual-semantic embedding [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(7): 13050 – 13057. DOI: 10.1609/aaai.v34i07.7006

[4] WRAY M, DOUGHTY H, DAMEN. On semantic similarity in video retrieval [EB/OL]. [2021-11-05]. https://arxiv.org/abs/2103.10095

[5] WANG X W, ZHANG T. Clothes search in consumer photos via color matching and attribute learning [C]//MM '11: Proceedings of the 19th ACM international conference on Multimedia. ACM, 2011: 1353 – 1356. DOI: 10.1145/2072298.2072013

[6] DI W, WAH C, BHARDWAJ A, et al. Style finder: fine-grained clothing style detection and retrieval [C]//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2013: 8 – 13. DOI: 10.1109/CVPRW.2013.6

[7] FU J L, WANG J Q, LI Z C, et al. Efficient clothing retrieval with semantic preserving visual phrases [C]//Asian Conference on Computer Vision. ACCV, 2012: 420 – 431

[8] GARCIA N, VOGIATZIS G. Dress like a star: Retrieving fashion products from videos [C]//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. IEEE, 2017: 2293 – 2299. DOI: 10.1109/ICCVW.2017.270

[9] HUANG J, FERIS R S, CHEN Q, et al. Cross-domain image retrieval with a dual attribute-aware ranking network [C]//Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2015: 1062-1070. DOI: 10.1109/ICCV.2015.127

[10] KIAPOUR M H, HAN X F, LAZEBNIK S, et al. Where to buy it: matching street clothing photos in online shops [C]//Proceedings of 2015 IEEE International Conference on Computer Vision. IEEE, 2015: 3343 – 3351. DOI: 10.1109/ICCV.2015.382

[11] LIU Z W, LUO P, QIU S, et al. DeepFashion: powering robust clothes recognition and retrieval with rich annotations [J]. 2016 IEEE conference on computer vision and pattern recognition (CVPR), 2016: 1096 – 1104. DOI: 10.1109/CVPR.2016.124

[12] JI X, WANG W, ZHANG M H, et al. Cross-domain image retrieval with attention modeling [C]//Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017: 1654 – 1662. DOI: 10.1145/3123266.3123429

[13] SONG Y, LI Y, WU B, et al. Learning unified embedding for apparel recognition [EB/OL]. (2017-07-19) [2021-11-05]. https://arxiv.org/abs/1707.05929

[14] CORBIÈRE C, BEN-YOUNES H, RAMÉ A, et al. Leveraging weakly annotated data for fashion image retrieval and label prediction [C]//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. IEEE, 2017: 2268 – 2274. DOI: 10.1109/ICCVW.2017.266

[15] CHENG Z Q, WU X, LIU Y, et al. Video2Shop: exact matching clothes in videos to online shopping images [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 4169 – 4177. DOI: 10.1109/CVPR.2017.444

[16] ZHANG Y H, PAN P, ZHENG Y, et al. Visual search at Alibaba [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018: 993 – 1001. DOI: 10.1145/3219819.3219820

[17] KUANG Z H, GAO Y M, LI G B, et al. Fashion retrieval via graph reasoning networks on a similarity pyramid [C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 3066 – 3075. DOI: 10.1109/ICCV.2019.00316

[18] LUO H, JIANG W, GU Y Z, et al. A strong baseline and batch normalization neck for deep person re-identification [J]. IEEE transactions on multimedia, 2020, 22(10): 2597 – 2609. DOI: 10.1109/TMM.2019.2958756

[19] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping [C]//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2006: 1735 – 1742. DOI: 10.1109/CVPR.2006.100

[20] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification [J]. Journal of machine learning research, 2009, 10(2): 207 – 244

[21] GORDO A, LARLUS D. Beyond instance-level image retrieval: leveraging captions to learn a global visual representation for semantic retrieval [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2017: 5272 – 5281. DOI: 10.1109/CVPR.2017.560

[22] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 5297 – 5307. DOI: 10.1109/CVPR.2016.572

[23] BALNTAS V, LI S D, PRISACARIU V. RelocNet: continuous metric learning relocalisation using neural nets [C]//Computer Vision. ECCV, 2018: 751 – 767. DOI: 10.1007/978-3-030-01264-9_46

[24] JÄRVELIN K, KEKÄLÄINEN J. Cumulated gain-based evaluation of IR techniques [J]. ACM transactions on information systems, 2002, 20(4): 422 – 446. DOI: 10.1145/582415.582418

[25] LIU T Y. Learning to rank for information retrieval [M]. Berlin, Germany: Springer. 2011

[26] WANG Z H, GU Y J, ZHANG Y, et al. Clothing retrieval with visual attention model [C]//Proceedings of 2017 IEEE Visual Communications and Image Processing. IEEE, 2017: 1 – 4. DOI: 10.1109/VCIP.2017.8305144

[27] XUAN H, SOUVENIR R, PLESS R. Deep randomized ensembles for metric learning [C]//Computer vision. ECCV, 2018: 723 – 734. DOI: 10.1007/978-3-030-01270-0_44

[28] SHEN Y, XIAO T, LI H, et al. End-to-end deep kronecker-product matching for person re-identification [C]//Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. IEEE, 2018: 6886 – 6895. DOI: 10.48550/arXiv.1807.11182

Biographies

YANG Bo received the B.S. degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2019. He is currently pursuing the M.S. degree in information and communication engineering at BUPT. His current research interests include computer vision and image retrieval.

GUO Caili (guocaili@bupt.edu.cn) received the Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunication (BUPT), China in 2008. She is currently a professor in the School of Information and Communication Engineering, BUPT. Her general research interests include machine learning and statistical signal processing, with current emphasis on semantic communications, deep learning, and intelligence visual computing. In the related areas, she has published over 200 papers and holds over 30 granted patents. She won Diamond Best Paper Award of IEEE ICME 2018 and Best Paper Award of IEEE WCNC 2021.

LI Zheng received the B.S. degree in telecommunication engineering from Shandong University, China in 2016, and the M.S. degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2019. He is currently pursuing the Ph.D. degree in information and communication engineering at BUPT. His current research interests include computer vision and multimedia retrieval.