Text-based Person Search via Attribute-aided Matching

Surbhi Aggarwal R. Venkatesh Babu Anirban Chakraborty
Indian Institute of Science, Bangalore, India
surbhia@iisc.ac.in, venky@iisc.ac.in, anirban@iisc.ac.in

Abstract

Text-based person search aims to retrieve the pedestrian images that best match a given text query. Existing methods utilize class-id information to get discriminative and identity-preserving features. However, it is not well-explored whether it is beneficial to explicitly ensure that the semantics of the data are retained. In the proposed work, we aim to create semantics-preserving embeddings through an additional task of attribute prediction. Since attribute annotation is typically unavailable in text-based person search, we first mine them from the text corpus. These attributes are then used as a means to bridge the modality gap between the image-text inputs, as well as to improve the representation learning. In summary, we propose an approach for text-based person search by learning an attribute-driven space along with a class-information driven space, and utilize both for obtaining the retrieval results. Our experiments on benchmark dataset, CUHK-PEDES, show that learning the attribute-space not only helps in improving performance, giving us state-of-the-art Rank-1 accuracy of 56.68%, but also yields humanly-interpretable features.

1. Introduction

Person search is a class of video surveillance problems that addresses image retrieval from diverse and large databases of pedestrian images. The details of the person to be searched can be provided as an image, or a list of attributes or as a natural language description. Based on the modality of the query, the problem can be broadly categorized into image-based, attribute-based and text-based person search [33, 18, 11, 12].

In this paper, we focus on text-based person search. Formally, the task of text-based person search is: given a text description of the person to be searched and a large gallery of images, it is required to rank the images according to their relevance to the query text, and return the best matching images [11]. The problem of text-based person search is interesting as it involves challenges associated with person re-id as well as those associated with cross-modal retrieval.

One of the main challenges of person search is to learn image features that are robust to the changes in viewing conditions, such as pose, illumination and camera viewpoints. Furthermore, the pedestrian images are generally of low resolution and may also be occluded, which demands additional focus on better image representations. In the task of cross-modal retrieval, due to the intrinsic difference in image and text domain, features of different modalities are not directly comparable [32, 24]. Hence, cross-modal retrieval introduces an additional challenge of reducing the heterogeneity in the input caused by the modality difference. Moreover, in text-based person search, the identities of persons present in test time are disjoint from the identities used in training. Hence, we need to ensure that the features are modality-invariant as well as transferable from the seen classes to the unseen classes.

To address the above challenges, we propose an approach based on learning a semantics-driven image-text common embedding and we utilize attributes as a means to represent the semantics of the data. Attributes are humanly-understandable semantic concepts, such as gender, age, clothing description, and are used as soft biometrics in human search [13, 12, 23]. Since, for a given person, attributes are modality-invariant, they can help in bridging the modality gap. Additionally, unlike class labels which are known only for seen classes, attributes are universal concepts, applicable to both seen and unseen classes. This suggests that attributes are good representatives for the semantics of data and can help in addressing the challenges associated with text-based person search. Based on these reasons, we hypothesize that through an additional task of learning to identify attributes, the model will get aid in the task of retrieval.

In text-based person search, image-text pairs are not annotated with the corresponding attributes. Hence, we automatically mine the annotation for the training datapoints from their corresponding text descriptions (Fig. 1), thereby circumventing the need for manual annotation. To utilize these attributes for improving the image-text representations, we present a novel hierarchical model in which each text/image datapoint is projected onto two levels of common embedding spaces: a) a mid-level, semantics-driven,

attribute-space and b) a high-level, class-driven, latent-space. In the attribute-space, the features are trained using the mined attributes, such that they are capable of predicting the correct attributes. Unlike in attribute-based re-id, which uses attributes for learning image features, we use attributes for improving both image and text features in attribute-space. In addition, we also learn discriminative, identity-preserving high-level features. Intuitively, the attributes-driven features help the model to rank the data-points using semantic similarity, while the class-driven features help to rank using latent, identity-based information. We show that both these mid-level semantics-preserving features and high-level discriminative features, when used in combination, improve retrieval performance in the task of text-based person search. Our contributions, therefore, can be summarized as follows:

- We propose a novel framework to learn common representations for image and text, such that semantics are explicitly preserved in the features. For this, we first automatically extract attributes for the classes in the training data from their text descriptions. Then, a hierarchy of features is learnt, such that the mid-level features are attributes-driven, and the high-level features are identity- preserving. We further propose a novel semantic triplet loss, which provides an adaptive margin to triplet loss. We also motivate and propose a novel norm-regularization constraint to further improve the learning of feature spaces.
- We validate our approach through extensive experimentation on CUHK-PEDES [11], and show that our method achieves the state-of-the-art performance.

2. Related work

Cross-modal retrieval. Many research areas such as image captioning, visual question answering, cross-modal retrieval [15, 8, 16, 25] involve learning relation between image and text data. Since cross-modal retrieval [32, 25] is closely related to text-based person search, we review the works related to it. A popular method for cross-modal retrieval is to learn an image-text joint space in which the two modalities can be compared. Canonical Correlation Analysis(CCA) and DeepCCA [28] learn a linear and non-linear projection respectively such that the image and text vectors have high correlation in the joint space. Weston *et al*. and Frome *et al*. [27, 5] used uni-directional ranking loss, while Wang *et al*. and Faghri *et al*. [25, 4] used bi-directional ranking loss to learn a common representation. Recently, adversarial training based methods have also been proposed [22].

Text-based person search. Li *et al*. [11] proposed GNA-RNN, which aims to find an affinity score between the query text and each image in the gallery set. Image-sentence affinity score is calculated using each word in the sentence via attention. As an improvement over GNA-RNN, Chen *et al*. [2] proposed a patch-wise word matching model to find the image-word affinity. GLA [1] learns local association of image regions and the noun phrases in the text via attention and enforces the attented vector to predict the noun phrases. Other methods such as [10, 35, 29] learn a joint embedding of image-text. Unlike the CNN-RNN approaches for image-text matching, a CNN-CNN architecture was proposed by Zheng *et al*. [35] to learn the common representation. Recently, Jing *et al*. [7] used pose as body part information to guide visual feature extraction. Their model uses hard attention to find the image regions most relevant to the corresponding text description.

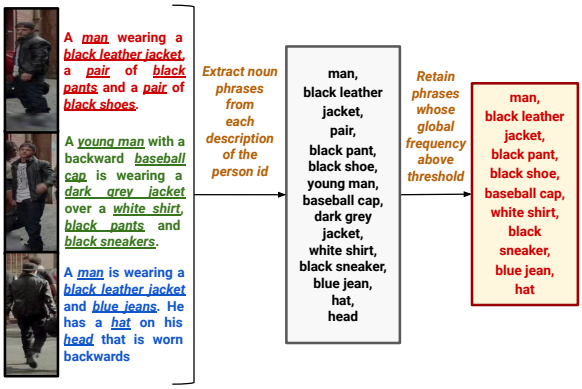


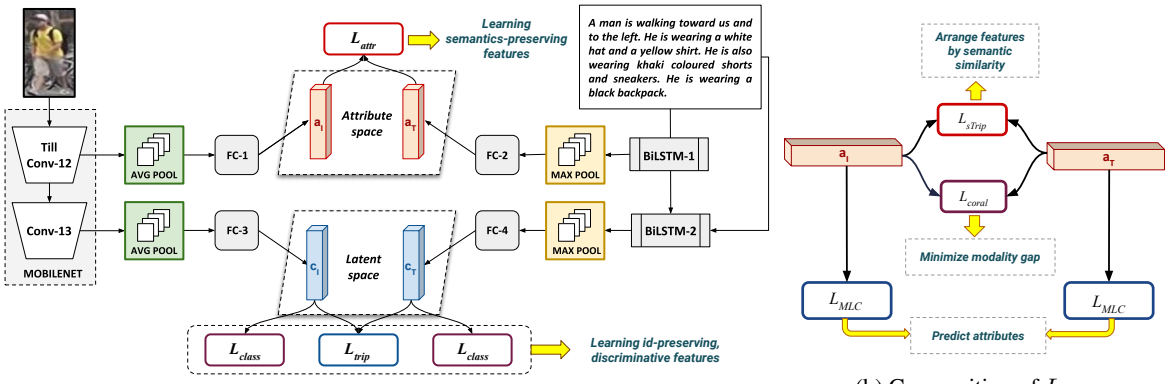
Figure 1: Extracting attributes for a person class from text.

Attribute recognition for person search. Attribute recognition is a multi-label classification task, which aims at selecting a relevant subset of attributes from a set of predefined list of attributes. With the advent of deep-learning based methods, many works have shown improvement in attribute recognition [9, 36, 20, 13, 30]. In [12], Lin *et al*. provide manually annotated Market-1501 and DukeMTMC-reID datasets.

3. Proposed method

3.1. Extraction of attributes

Text descriptions can be considered as a unstructured form of annotation as compared to a manually assigned set of attributes. In general, most attributes fall in the category of noun phrases. Hence, as shown in Fig. 1, noun phrases from the training set are extracted using NLTK [14]. For each person class in the training data, we collect the candidate noun phrases from the descriptions associated with any of its image. This was done to get the entire information of the person. Next, phrases which have very low frequency in the corpus are discarded, and the remaining ones are marked as the attributes for the person class. This process gives us a mapping from class identities in training set to the corresponding attributes.



(a) Proposed architecture

(b) Composition of L_{attr}

Figure 2: In the proposed architecture, the features are extracted in a hierarchical manner. First, a semantics-driven attribute-space is learnt which ensures that the attributes can be predicted from the embedding. Next, a discriminative, latent-space is learnt using classification and triplet loss. Retrieval is done using similarity scores from both the spaces.

3.2. Proposed architecture and losses

We propose an architecture (Fig. 2(a)) such that two levels of features are extracted from each image/text input: a) d_a-dimensional semantics-driven features, b) d_c-dimensional class-driven features. We denote the space of the semantics-driven features as 'attribute-space', while the space of the class-information driven features as 'latent-space'. The feature extraction pipeline is as follows:

Visual features. A Mobilenet model [6], pretrained on ImageNet [3] is used for encoding the images. For each image I_k, the average-pooled output of Conv-12 is transformed into a d_a-dimensional attribute-space embedding, $\mathbf{a}_{I_k}$. Similarly, the average-pooled output of Conv-13 is transformed to a d_c-dimensional latent-space representation, $\mathbf{c}_{I_k}$.

Text features. A bi-directional LSTM (Bi-LSTM) [19] is used for encoding the text. The forward and backward hidden states of each word are concatenated to get the word representation. Finally, the word representations are max-pooled to get the text embedding. We create two such Bi-LSTMs with a shared word-embedding matrix. For each text T_k, the output of the first Bi-LSTM is transformed to a d_a-dimensional attribute-space representation, $\mathbf{a}_{T_k}$. Similarly, output of the second Bi-LSTM is transformed into a d_c-dimensional latent-space representation, $\mathbf{c}_{T_k}$.

Next, we elaborate upon each of the proposed spaces: a) Attribute-space and b) Latent-space.

3.2.1 Attribute Space

We introduce a semantics-driven space, named *attribute-space*, to explicitly incorporate semantics during representation learning and to learn transferable and humanly-understandable features. In the attribute-space, we want to learn a mapping from the embeddings to the corresponding attributes. In addition, we want to minimize the image-text modality gap and ensure that similarity in attribute-space represents semantic similarity. Hence, we impose losses, as shown in Fig. 2(b), to achieve each of these desired properties.

Multi-label classification. To incorporate semantics during training, we enforce that the attributes relevant to a datapoint are predictable from the its attribute-space representation. For this, we employ 'multi-label classification loss' (L_{MLC}), to ensure that the predicted attributes match the mined ground-truth attributes.

In general, attribute prediction is trained using binary cross-entropy loss (BCE). Let x_i be an image/text datapoint; that is, it can replaced by I_i to denote the i^{th} image, or T_i to denote the i^{th} text. The attribute-space representation for x_i is denoted by $\mathbf{a}_{x_i} \in \mathbb{R}^{d_a}$. Let $\mathcal{V} = \{l_1, \ldots, l_{|\mathcal{V}|}\}$ be the mined vocabulary of attributes, $\mathcal{P}_i$ be the set of attributes marked positive for x_i, and $\mathcal{N}_i$ be the set $\mathcal{V} \backslash \mathcal{P}_i$. We refer to attributes in $\mathcal{P}_i$ as positive attributes for x_i, and those in $\mathcal{N}_i$ as negative attributes. A binary indicator vector $\mathbf{b}_i \in \{0,1\}^{|\mathcal{V}|}$ is obtained for each x_i, such that $b_{ik} = 1$ if $l_k \in \mathcal{P}_i$, else it is 0. An attribute classifier matrix $W \in \mathbb{R}^{|\mathcal{V}| \times d_a}$ with biases $\mathbf{z} \in \mathbb{R}^{|\mathcal{V}|}$ is constructed. For each attribute $l_j \in \mathcal{V}$, we use $S(l_j|x_i)$ to denote the score of l_j for given datapoint x_i, and $P(l_j|x_i)$ to denote the probability that l_j is a positive attribute for x_i. Using these notations, the binary cross-entropy loss for attribute prediction is given by:

$$L_{BCE}(x_i) = -\sum_{l_j \in \mathcal{P}_i} \log P(l_j|x_i) - \sum_{l_j \in \mathcal{N}_i} \log(1 - P(l_j|x_i)) \quad (1)$$

$$P(l_j|x_i) = \sigma(S(l_j|x_i)) \quad (2)$$

$$S(l_j|x_i) = \beta_j S_{cos}(\mathbf{a}_{x_i}, \mathbf{W_j}) + z_j \quad (3)$$

where σ is sigmoid activation, $\beta \in \mathbb{R}^{|\mathcal{V}|}$ is a learnt scaling vector and S_{cos} denotes cosine similarity. However, this loss is not directly usable due to two types of imbalances that are observed in the attributes. First, there is an imbalance in the frequency of different attributes. For instance, attributes like 'man' or 'woman' are highly frequent, while attributes like 'red hat' or 'yellow shoe' are rare. To ensure that rare concepts are not ignored, we use weighted binary cross entropy loss (Eq. 4) [9]. The weights $w^p_{l_j}$, $w^n_{l_j}$ are computed such that $w^p_{l_j}$ is high for rare attributes and $w^n_{l_j}$ is high for frequent attributes. Second, there is an imbalance across positive and negative attributes present on each datapoint. This is because the attribute vocabulary is large, but on average, a small fraction of attributes is applicable on a datapoint. Hence, there would be more loss on attributes which are absent, than on attributes which are present. We refer to the loss over negative attributes as 'negative loss' and loss over positive attributes as 'positive loss', and this imbalance as the 'positive-negative loss imbalance'. To address this imbalance on each x_i, we first compute the average negative loss and then scale it by $\alpha|\mathcal{P}_i|$, where $|\mathcal{P}_i|$ is the cardinality of $\mathcal{P}_i$ and $\alpha > 0$ is a hyper-parameter, denoting the relative weight of 'negative loss' over the 'positive loss'. Thus, the multi-label classification after addressing the imbalances is:

$$L_{MLC}(x_i) = -\sum_{y_j \in \mathcal{P}_i} w^p_{l_j} \log P(l_j|x_i) - \frac{\alpha|\mathcal{P}_i|}{|\mathcal{N}_i|} \sum_{y_j \in \mathcal{N}_i} w^n_{l_j} \log(1 - P(l_j|x_i)) \quad (4)$$

Reducing modality-gap. Cross-modal retrieval suffers from modality gap between image and text spaces. To alleviate this, we propose to use Deep CORAL loss [21] in the attribute-space. Deep CORAL loss aims at minimizing the distance between the second-order statistics of source and target activations. Hence, this loss shall help in improving the alignment of image and text features. For an image/text datapoint x_i, let the vector $\mathbf{v}_{x_i} = [S(l_1|x_i), \ldots, S(l_{|\mathcal{V}|}|x_i)] \in \mathbb{R}^{|\mathcal{V}|}$ denote the per-attribute score vector. We use v_{x_i} as the activation on which Deep CORAL loss is applied. Given a batch $\mathcal{B} = (\mathcal{I}, \mathcal{T})$, consisting of n image-text pairs, $\{(I_1, T_1), \ldots, (I_n, T_n)\}$, let $\mathbf{v}_{I_j}$ be the activation for the image I_j and $\mathbf{v}_{T_j}$ be the activation for the text T_j. Let C_I and C_T denote the covariance matrix of image activations and text activations respectively. Using these notations, the Deep CORAL loss [21] is applied as:

$$L_{coral}(\mathcal{I}, \mathcal{T}) = \frac{1}{4|\mathcal{V}|^2} \|C_I - C_T\|^2_F \quad (5)$$

Semantics-based feature alignment. We want to arrange the attribute-space such that neighbours are semantically similar. This is a meaningful feature alignment for the task of retrieval, as it will ensure that the images retrieved for a query are semantically consistent to the required text description. It can be observed that class-level dissimilarity doesn't necessarily imply semantic dissimilarity. This is because, a pair of datapoints from different identities can have similar descriptions, which means semantically they are not very different. Hence, we need an alternative indicator for semantic similarity.

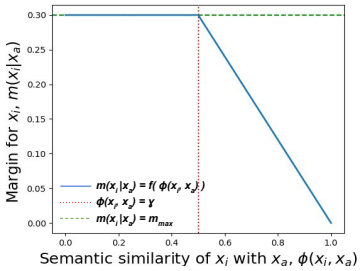


Figure 3: Adaptive margin based on semantic similarity: Blue line represents the margin applied as a function of semantic similarity with anchor (Eq. 6).

Let $\mathbf{b}_i$ be the ground-truth binary attribute vector for datapoint x_i. We use $\phi(x_i, x_j) = S_{cos}(\mathbf{b}_i, \mathbf{b}_j)$ to denote the semantic similarity for data-pair (x_i, x_j), where S_{cos} represents cosine similarity. Note, since $\mathbf{b}_i$ and $\mathbf{b}_j$ are binary vectors, $\phi \in [0, 1]$. Having obtained a measure of semantic similarity ϕ, we then modify triplet loss [17], such that negatives are mined based on ϕ instead of class similarity. That is, for an anchor x_a, the candidate negative datapoints are $\mathcal{N}(x_a) = \{x_j : \phi(x_a, x_j) < 1\}$. We further use an adaptive margin to push different datapoints based on their semantic dissimilarity with the anchor. Thus, given an anchor x_a, and candidate negative datapoints x_1, x_2 such that $\phi(x_a, x_1) > \phi(x_a, x_2)$, we ensure that the margin imposed on x_1 is less than that imposed on x_2. This is achieved by setting the margin for each $x_i \in \mathcal{N}(x_a)$ as follows:

$$m(x_i|x_a) = \begin{cases} m_{max}, \text{if } \phi(x_i, x_a) \leq \gamma \\ \frac{m_{max}}{1-\gamma}(1 - \phi(x_i, x_a)), \text{otherwise} \end{cases} \quad (6)$$

where γ is a threshold on ϕ, such that margin remains equal to m_{max} even on decreasing ϕ. It is introduced because below a certain degree of semantic similarity, all the points are irrelevant to the query. Hence, it is sufficient to keep them away from the query, and not needed to rank the entire space. Next, for creating the triplets, when text T_k is the anchor, we select the corresponding image I_k as the positive and the negative $\widetilde{I^{sn}_k}$ is mined as follows:

$$\widetilde{I^{sn}_k} = \underset{I_i \in \mathcal{N}(T_k)}{\text{argmax}} \; S_a(I_i, T_k) + m(I_i|T_k) \quad (7)$$

$$S_a(I_i, T_k) = S_{cos}(\mathbf{a}_{I_i}, \mathbf{a}_{T_k}) \quad (8)$$

where $S_a(.,.)$ is named as 'attribute similarity'. Similarly, when an image I_k is the anchor, we mine $\widetilde{T_k^{sn}}$ as the negative and T_k as the positive. We name this proposed adaptation of triplet loss as 'semantic triplet loss', as the embeddings are arranged based on semantic similarity instead of class membership. The semantic triplet loss is calculated as follows, where $[x]_+ = \max(x, 0)$:

$$L_{sTrip}(\mathcal{I},\mathcal{T}) = \sum_{I_k\in\mathcal{I}} [m(\widetilde{T_k^{sn}}|I_k) + S_a(I_k,\widetilde{T_k^{sn}}) - S_a(I_k,T_k)]_+ + \sum_{T_k\in\mathcal{T}} [m(\widetilde{I_k^{sn}}|T_k) + S_a(T_k,\widetilde{I_k^{sn}}) - S_a(T_k,I_k)]_+ \quad (9)$$

Total attribute-space loss. For a mini-batch $\mathcal{B} = (\mathcal{I},\mathcal{T})$ of n image-text pairs, the total loss for attribute-space is:

$$L_{attr}(\mathcal{I},\mathcal{T}) = L_{sTrip}(\mathcal{I},\mathcal{T}) + \lambda_c L_{coral}(\mathcal{I},\mathcal{T}) + \frac{\lambda_b}{n}\sum_i (L_{MLC}(I_i) + L_{MLC}(T_i)) \quad (10)$$

3.2.2 Latent Space

Unlike the attribute-space, which is driven by semantics, the *latent space* is learnt using class-information to ensure that embeddings of different identities are well-separated. Similar to Zheng *et al*. [35], we utilize a combination of triplet loss and classification loss to learn this space. In addition, a novel feature norm-regularization is used, which further helps in improving the cosine similarity between datapoints of the same class. In the latent space, $\mathbf{c}_{I_k}$ is used for denoting the latent-space embedding of k^{th} image, and $\mathbf{c}_{T_k}$ for that of k^{th} text.

Discriminative alignment of features. Triplet loss [17, 25] is popularly used to learn a discriminative alignment of features. Further, to make the embeddings robust to inter-class similarity and intra-class variance, hard negative and positive mining is used for creating the triplets [31]. Hard negative is a datapoint from a different class which is nearest to the anchor in the latent-space. Similarly, hard positive is a datapoint from the same class which is farthest from the anchor. For anchor text T_k, we use $\widetilde{I_k^n}$ to denote the hard negative image and $\widetilde{I_k^p}$ for hard positive image. Finally, the triplet loss is applied on the latent-space as follows, where $S_l(.,.)$ is named as 'latent similarity' and Δ is the margin:

$$L_{trip}(\mathcal{I},\mathcal{T}) = \sum_{I_k\in\mathcal{I}} [\Delta + S_l(I_k,\widetilde{T_k^n}) - S_l(I_k,\widetilde{T_k^p})]_+ + \sum_{T_k\in\mathcal{T}} [\Delta + S_l(T_k,\widetilde{I_k^n}) - S_l(T_k,\widetilde{I_k^p})]_+ \quad (11)$$

$$S_l(I_j,T_k) = S_{cos}(\mathbf{c}_{I_j},\mathbf{c}_{T_k}) \quad (12)$$

Identity-classification. We train an identity-classifier to train the latent embeddings to be identity-preserving. A common classifier, $W \in \mathbb{R}^{d_c\times C}$, is used to help bridge the modality gap between image and text features [35], where C is total number of train identities. We denote the classifier column corresponding to the k^{th} class by $\mathbf{W_k}$. The columns of the classifier are normalized, because otherwise, a datapoint x_i is classified into the class k which has maximum score of $\mathbf{W_k}^T\mathbf{c}_{x_i} = \|\mathbf{W_k}\|_2\|\mathbf{c}_{x_i}\|_2 S_{cos}(\mathbf{c}_{x_i},\mathbf{W_k})$. However, this score should not depend upon $\|\mathbf{W_k}\|_2$, as we want to align vectors using cosine similarity. Post normalization, $\mathbf{c}_{x_i}$ will be classified to the class y with maximum $\|\mathbf{c}_{x_i}\|_2 S_{cos}(\mathbf{c}_{x_i},W_k)$. This is same as class ranked highest based on $S_{cos}(\mathbf{c}_{x_i},\mathbf{W_k})$, which justifies the need to normalize the columns. Finally, the classification loss used is as follows:

$$L_{class}(x_i) = -\log\left(\frac{\exp(\mathbf{c}_{x_i}^T\hat{\mathbf{W}}_{\mathbf{y(x_i)}})}{\sum_{j=1:C}\exp(\mathbf{c}_{x_i}^T\hat{\mathbf{W}}_{\mathbf{j}})}\right) \quad (13)$$

where $y(x_i)$ is the class of x_i, $\hat{\mathbf{W}}_{\mathbf{j}}$ denotes the classifier column $\mathbf{W_j}$ after normalization.

Feature norm-regularization. Since we use cosine similarity for ranking, the norm of the features plays no role in retrieval. However, the score used for classification, $\|\mathbf{c}_{x_i}\|_2 S_{cos}(\mathbf{W_y},\mathbf{c}_{x_i})$, depends on $\|\mathbf{c}_{x_i}\|_2$. It can be observed that after classifier columns have been normalized, the L_{class} can be minimized by increasing $\|\mathbf{c}_{x_i}\|_2$, while keeping $S_{cos}(\mathbf{W_y},\mathbf{c}_{x_i})$ constant. We prove this statement in the supplementary. Hence, to focus the training on improving cosine similarity, we add a regularization on $\|\mathbf{c}_{x_i}\|_2$.

Furthermore, a novel constraint to reduce the variance in the norms of latent-space features is added. This is because, for two vectors of the same class and pointing in the same direction but having different norms, the loss will be different [34]. This is undesirable because as per cosine similarity, both the points are equivalent. To address this discrepancy due to difference in norms, we penalize variance in $\|\mathbf{c}_{x_i}\|_2$, thus ensuring that the model learns features of almost constant norms. Formally, for a mini-batch $\mathcal{B} = (\mathcal{I},\mathcal{T})$ of n image-text pairs, let $\mathbf{v_{norm}}$ contain the norm of each latent-space feature in the batch, σ^2 be the variance of $\mathbf{v_{norm}}$. Then the loss is given as:

$$L_{norm-reg}(\mathcal{I},\mathcal{T}) = \lambda_{reg}\|\mathbf{v_{norm}}\|_2 + \lambda_{var}\sigma^2 \quad (14)$$

Total latent-space loss. For a mini-batch $\mathcal{B} = (\mathcal{I},\mathcal{T})$, of n image-text pairs, the total loss for latent-space is given as:

$$L_{latent}(\mathcal{I},\mathcal{T}) = L_{trip}(\mathcal{I},\mathcal{T}) + \frac{1}{n}\sum_i L_{class}(I_i) + L_{class}(T_i) \quad (15)$$

$$L_{reg-latent}(\mathcal{I},\mathcal{T}) = L_{latent} + L_{norm-reg} \quad (16)$$

CMAAM. We summarize the losses for our proposed approach. We use CMAAM (**C**ross **M**odal **A**ttribute-**A**ided **M**atching) to denote the method which utilizes the attribute-space losses (Eq. 10) and latent-space losses along with norm-regularization (Eq. 16), and α-CMAAM for the method which utilizes the attribute-space and latent-space losses without norm-regularization (Eq. 15). The net loss for our model is:

$$L_{\alpha-CMAAM} = L_{attr} + \lambda_t(L_{latent}) \quad (17)$$

$$L_{CMAAM} = L_{attr} + \lambda_t(L_{reg-latent}) \quad (18)$$

In both α-CMAAM and CMAAM, the retrieval results are ranked by adding both latent similarity and attribute similarity. This similarity measure is named as 'latent+attribute similarity' and is given by:

$$S_{la}(I_i, T_k) = S_a(\mathbf{a}_{I_i}, \mathbf{a}_{T_k}) + S_l(\mathbf{c}_{I_i}, \mathbf{c}_{T_k}) \quad (19)$$

4. Experiments and Analysis

4.1. Experiment setup

Dataset. The dataset used is CUHK-PEDES [11], which is the lone benchmark dataset available for the task. In total, CUHK-PEDES is a collection of 5 person-reidentification datasets and has 13,003 unique person identities. For each person, there are on average 3.1 images. Each image has been annotated by 2 sentence descriptions. The average length of the sentences is 23.5 words. We follow the data splitting scheme of [11] which uses 11,003 identities for training, and 1,000 identities for validation and remaining 1,000 identities for testing.

Evaluation setup. Recall@K, for K=1, 5 and 10 is used as the evaluation metric, where Recall@K defines the percentage of times the image of the correct person is returned amongst the top-K results. During retrieval, for ranking the gallery images with a query text, we explore three variants of similarities between image and text: a) latent-space similarity (Eq. 12), b) Attribute similarity (Eq. 8) and c) Latent+Attribute similarity (Eq. 19). In the following section, we use 'L', 'A' or 'LA' to denote latent similarity, attribute similarity, latent+attribute similarity respectively.

Training procedure. The model is trained in a single-stage, with both the image encoder and text encoder being trainable. We set the dimension of latent-space features as 512 and of attribute-space features as 300. The word embedding matrix and attribute embedding matrix are randomly initialized. Training is done for 50 epochs using batch size of 16 and Adam optimizer with learning rate of 2e-4. We used α = 2 in Eq.2. The hyperparameters are set as: λ_a = 0.1, λ_b = 0.25, λ_c = 50, λ_{reg} = 1e-3, λ_{var} = 1 and λ_t = 3. For the losses L_{sTrip} and L_{trip}, we used $\gamma = 0.5$, and $m_{max}, \Delta = 0.3$. To create the attribute vocabulary, we retain the noun phrases which contain maximum of 3 words and whose frequency across the training text corpus is not less than 40, yielding attribute vocabulary of size 450[1].

Method	R@1	R@5	R@10
Prior works			
GNA-RNN [11]	19.05	-	53.64
IATV [10]	25.94	-	60.49
PWM [2]	27.14	49.45	61.02
DPCE [35]	44.40	66.26	75.07
GLA [1]	43.58	66.93	76.26
CMPC+CMPM [29]	49.37	-	79.27
MCCL [26]	50.58	-	79.06
GALM [7]	54.12	75.45	82.97
Baselines			
Class (Eq. 13)	21.31	41.28	51.09
Triplet (Eq. 11)	46.30	70.50	79.03
Class+Triplet (Eq. 15)	52.89	74.01	82.07
Proposed			
α-CMAAM (Eq. 17)	55.13	76.14	83.77
CMAAM (Eq. 18)	**56.68**	**77.18**	**84.86**

Table 1: Text-to-image retrieval on CUHK-PEDES: We compare our proposed method with three baselines as well as the state-of-the-art methods. Our framework significantly outperforms the rest on all metrics.

Baselines. We first establish 3 baselines: a) only L_{class} (Class) (Eq. 13), b) only L_{trip} (Triplet) (Eq. 11), c) $L_{class} + L_{trip}$ (Class+Triplet) (Eq. 15). We observe that Class+Triplet is the best performing baseline, as shown in Table 1.

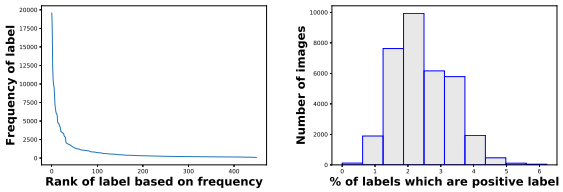
4.2. Quantitative analysis

First we compare our results with prior works and state-of-the-art methods in this domain. Next, to quantitatively establish that learning an attribute-space helps, we study the performance of attribute-space alone and then the interaction of the attribute-space with each of the baselines.

Comparison with other methods. In Table 1, we compare our method with the other related works on CUHK-PEDES. It can be observed that our model outperforms each of the works by good margin under all three metrics. We can also see that considerable improvement is achieved on each of our baselines.

Effect of addressing imbalance. To highlight the presence of imbalances in the attributes, we plot the attribute statistics in Fig. 4. Fig. 4(a), shows that the frequency of the attributes follow the Zipf's law, which validates the imbalance in frequency of attributes. Fig. 4(b) shows that on average only 2-3% of the attributes are positively associ-

[1]More details on attribute extraction and weights in L_{MLC} are given in Supplementary.



(a) (b)

Figure 4: Attributes statistics: Fig. 4(a) shows the frequency of each attribute, where attributes are ranked in descending order of frequency. In Fig. 4(b), the bar chart shows that on average, only 2-3% of attributes are positively associated to a datapoint.

Baseline	Balance-1	Balance-2	R@1	R@5	R@10
L_{BCE}	✓	✗	11.29	27.66	36.74
	✗	✓	23.15	45.18	56.82
	✓	✓	**24.14**	**47.60**	**57.83**

Table 2: Abalation on L_{MLC} (Eq. 4): Retrieval in attribute-space using attribute-similarity. 'Balance-1' refers to effect of using weighted binary-cross entropy, 'Balance-2' refers to balancing done by α.

ated to a datapoint. This implies that for each datapoint, the loss on the attributes which are absent is much more than that on the attributes which are present. Next, the ablations for addressing the frequency-based imbalance and 'positive-negative loss imbalance' have been presented in Table 2. We observe that handling each imbalance shows improvement, justifying that the need to address each of them.

Retrieval using attribute-space only. We first investigate the utility of the novel attribute-space for text-based person search, as shown in Table 3. We observe that R@1 of 24.14% is obtained from just attribute classification (Eq. 4). Additional improvement is observed by using L_{coral}. This can be attributed to the better alignment in the modalities, which in turn helps the attribute classifier. On adding the semantic triplet loss (Eq. 9), we get 49.24% R@1 accuracy in the attribute-space. These experiments validate our hypothesis that an attributes-driven space can help in the task of text-based person search.

Interaction of attribute-space with latent-space. We next study whether learning an attribute-space is beneficial in conjunction with a class-driven latent-space (Table 4). To evaluate this, we show interaction of attribute-space with each of our baseline for latent-space, which are: a) L_{class} (Class), b) L_{trip} (Triplet), and c) $L_{class} + L_{trip}$ (Class+Triplet).

When used along with id-classification, we observe that through the simultaneous training of latent and attribute-

Baseline	L_{coral}	L_{sTrip}	R@1	R@5	R@10
L_{MLC}	✗	✗	24.14	47.60	57.83
	✓	✗	35.23	58.56	68.97
	✓	✓	**49.24**	**72.04**	**79.92**

Table 3: Abalation on L_{attr} (Eq. 10): Retrieval in attribute-space using attribute-similarity. 'L_{coral}' refers to (Eq. 5), 'L_{sTrip}' refers to (Eq. 9)

Baseline	L_{MLC}	L_{coral}	L_{sTrip}	Sim	R@1	R@5	R@10
Class (L_{class})	✗	✗	✗	L	21.31	41.28	51.09
	✓	✗	✗	L	26.97	47.77	57.62
				LA	35.02	57.34	67.37
	✓	✓	✗	L	29.03	51.04	61.40
				LA	41.13	64.00	73.57
	✓	✓	✓	L	31.40	54.06	63.66
				LA	49.11	70.78	79.30
Triplet (L_{trip})	✗	✗	✗	L	46.30	70.50	79.03
	✓	✗	✗	L	47.27	71.02	79.19
				LA	48.44	70.78	79.50
	✓	✓	✗	L	47.48	70.26	79.08
				LA	49.42	71.56	79.78
	✓	✓	✓	L	47.22	70.94	79.38
				LA	49.38	72.64	80.86
Class+Triplet ($L_{class} + L_{trip}$)	✗	✗	✗	L	52.89	74.01	82.07
	✓	✗	✗	L	52.00	74.61	82.26
				LA	51.75	73.23	81.25
	✓	✓	✗	L	53.44	74.63	82.28
				LA	53.56	74.32	82.18
	✓	✓	✓	L	53.48	74.77	82.15
				LA	**55.13**	**76.14**	**83.77**

Table 4: Abalation on baselines.'Sim' stands for similarity score used for retrieval, which can be either: a) 'L' which denotes 'latent similarity' (Eq. 12), or b) 'LA' which denotes 'latent+attribute similarity' (Eq. 19).

Method	Baseline	L_{attr}	Norm	Reg	Var	Sim	R@1	R@5	R@10
-	Class	✗	✓	✗	✗	L	21.31	41.28	51.09
-				✓	✗		36.89	58.15	67.32
-				✓	✓		39.41	61.29	70.89
-	Class+Triplet	✗	✓	✗	✗	L	52.89	74.01	82.07
-				✓	✗		53.44	74.77	82.47
-				✓	✓		54.11	74.68	81.71
-	Class+Triplet	✓	✗	✗	✗	LA	55.52	76.32	83.85
-			✗	✓	✗		55.47	76.62	83.64
α-CMAAM			✓	✗	✗		55.13	76.14	83.77
-			✓	✓	✗		56.30	**77.57**	84.58
CMAAM			✓	✓	✓		**56.68**	77.18	**84.86**

Table 5: Effect of $L_{norm-reg}$: 'Var' indicates inclusion of variance minimization, 'Reg' indicates inclusion of feature norm-regularization, as described in Eq. 14. 'Norm' indicates normalization of columns of classifier.

space, the retrieval using L-similarity itself improves. This is indicative of the benefits of attribute-space. Abalation over the components of L_{attr} shows that each of them are necessary. Best improvement is seen by training along L_{attr} and using LA-similarity for retrieval.

In class-based triplet loss driven model, after training

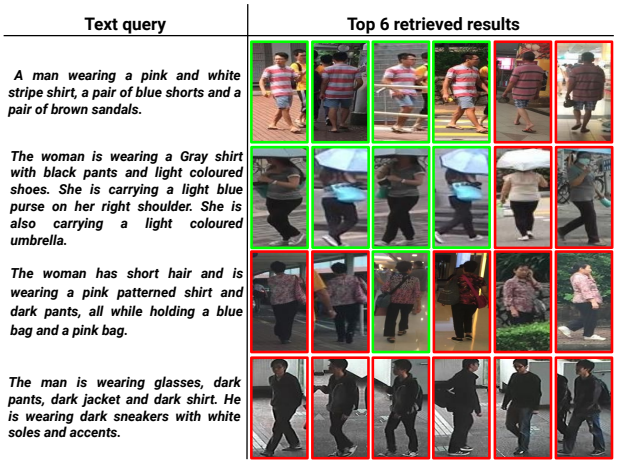


Figure 5: Top 6 retrieval results by CMAAM. Green highlight represents successful match, while red represents incorrect retrieval (best seen in color).

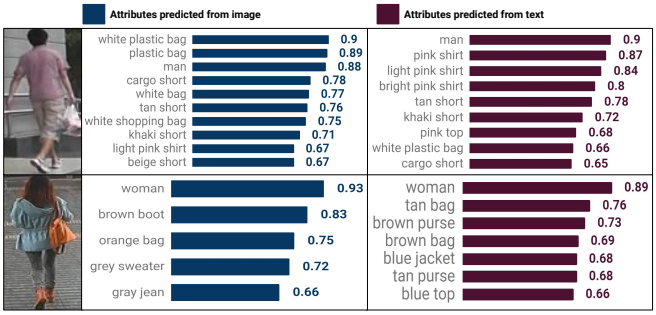


Figure 6: Attribute prediction for unseen images and text: We show all attributes predicted with probability ≥ 0.65 for unseen images and their corresponding text.

with L_{attr}, R@1 obtained using L-similarity itself improves by about 1%. Further improvement is seen by using LA-similarity for retrieval. Here too, abalation over each component of L_{attr} justifies the need for each loss.

When the baseline is trained on both classification and triplet losses, just L_{MLC} does not show improvement. A plausible reason is that the latent embeddings have been trained in a highly discriminative manner, while the attribute-space is comparatively less discriminative of identities. On training with L_{attr}, the retrieval based on LA-similarity gives us R@1 of 55.13%, showing the benefits of the attribute-space. We also show in Table 4 that all the three losses are necessary to achieve this improvement. A more comprehensive abalation is provided in the supplementary. These experiments show that each loss involved in L_{attr} is essential, and that training along with L_{attr} and using LA-similarity for retrieval brings improvement to each baseline; hence, validating our hypothesis that semantics-preserving losses can improve retrieval.

Effect of feature-norm regularization. We next study the effect of $L_{norm-reg}$ (Eq. 14). It is observed that adding norm-regularization helps in each baseline, as shown in (Table 5). Furthermore, we show in Class+Triplet+L_{attr}, effect of regularization is more when the columns of the classifier are normalized, giving us R@1 accuracy of 56.30%. Additional improvement is seen by adding variance reduction constraint, yielding in state-of-art accuracy of 56.68%.

4.3. Qualitative analysis

In Fig. 5, we show examples of retrieval by our proposed method, CMAAM using LA-similarity. In the first-two rows, all the correct images obtain the top-4 ranking (green highlight). Further, the negative retrievals (red highlight) that follow the correctly predicted images are semantically similar to the query. In the third row, the correct prediction comes at rank-3. Here too, the other images are to a great extent similar to the query. The last row is a failure case, in which none of the images is correct. The retrieved images indeed exhibits some attributes, as mentioned in the text description, *e.g.* dark clothings and spectacles. However, they are incorrect, as it can be observed that some images which contain 'bag' are appearing, and in one image the person is wearing a 'sweatshirt' instead of 'jacket'. In Fig. 6, we further illustrate the attributes predicted for test images and text queries. It can be observed that the attribute predictions from a text and its corresponding image have high consensus, and the model is able to provide reasonable predictions. Hence, this shows that we are also able to get good humanly-understandable features.

5. Conclusion

In this work, we proposed a novel method for text-based person search. We devised a hierarchical architecture, such that middle-level features are semantically arranged and high-level features are discriminative. We performed extensive experiments on CUHK-PEDES, and showed that explicitly ensuring that semantics are retained not only gave us state-of-the-art performance of 56.68% Rank-1 accuracy, but also provided humanly-understandable features.

Acknowledgments. This work is partially supported by Microsoft Data Science Fellowship (to Surbhi Aggarwal) and a young investigator award by Pratiksha Trust, Bangalore (to Anirban Chakraborty). The authors would also like to gratefully thank Mr. Devraj Mandal for the discussions.

References

[1] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *European Conference on Computer Vision*. Springer, 2018.

[2] T. Chen, C. Xu, and J. Luo. Improving text-based person search by spatial matching and adaptive threshold. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[4] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*, 2018.

[5] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013.

[6] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[7] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, and T. Tan. Pose-guided joint global and attentive local matching network for text-based person search. *arXiv preprint arXiv:1809.08440*, 2018.

[8] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[9] D. Li, X. Chen, and K. Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *IAPR Asian Conference on Pattern Recognition*, 2015.

[10] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang. Identity-aware textual-visual matching with latent co-attention. In *IEEE International Conference on Computer Vision*, 2017.

[11] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[12] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019.

[13] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, 2017.

[14] E. Loper and S. Bird. Nltk: The natural language toolkit. In *ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, ETMTNLP '02. Association for Computational Linguistics, 2002.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *ICLR*, 2015.

[16] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[18] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[19] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 1997.

[20] P. Sudowe, H. Spitzer, and B. Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *IEEE International Conference on Computer Vision Workshop*, 2015.

[21] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016.

[22] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. Adversarial cross-modal retrieval. In *ACM international conference on Multimedia*. ACM, 2017.

[23] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision*, 2017.

[24] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[25] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE conference on computer vision and pattern recognition*, 2016.

[26] Y. Wang, C. Bo, D. Wang, S. Wang, Y. Qi, and H. Lu. Language person search with mutually connected classification loss. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[27] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *International Joint Conference on Artificial Intelligence*, 2011.

[28] F. Yan and K. Mikolajczyk. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[29] Y. Zhang and H. Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.

[30] X. Zhao, L. Sang, G. Ding, Y. Guo, and X. Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *International Joint Conference on Artificial Intelligence*, 2018.

[31] Y. Zhao, Z. Jin, G.-j. Qi, H. Lu, and X.-s. Hua. An adversarial approach to hard triplet generation. In *European Conference on Computer Vision*, 2018.

[32] L. Zhen, P. Hu, X. Wang, and D. Peng. Deep supervised cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, 2015.

[34] Y. Zheng, D. K. Pal, and M. Savvides. Ring loss: Convex feature normalization for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[35] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.

[36] J. Zhu, S. Liao, D. Yi, Z. Lei, and S. Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics. In *International Conference on Biometrics*, 2015.