Supplementary material for 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

1. Abalation results

1.1. Interaction of attribute-space with latent-space

We further elaborate upon the abalations done on the baseline, Class+Triplet. When only L_{MLC} is used, we don't get an increment in retrieval accuracy. As also mentioned in the main paper, 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. Next, on adding L_{coral} to L_{MLC}, we see an improvement from 52.89% to 53.56%. It is also observed that improvement is obtained when baseline is trained with just L_{sTrip}. A plausible reason for this is that L_{sTrip} helps push away the semantically-negative datapoints, and enhances the total separation between query and irrelevant points. We observe that the best Rank@1 accuracy of 55.13% is obtained when each component of L_{attr} is used, hence, validating the need for each of them.

Baseline	L_{MLC}	L_{coral}	L_{sTrip}	Similarity Score	R@1	R@5	R@10
Class+Triplet	✗	✗	✗	Latent (L)	52.89	74.01	82.07
	✓	✗	✗	Latent (L)	52.00	74.61	82.26
				Latent+Attibute(LA)	51.75	73.23	81.25
	✗	✗	✓	Latent (L)	52.81	73.85	81.66
				Latent+Attibute(LA)	54.14	74.87	82.89
	✓	✓	✗	Latent (L)	53.44	74.63	82.28
				Latent+Attibute(LA)	53.56	74.32	82.18
	✗	✓	✓	Latent (L)	52.63	74.64	82.41
				Latent+Attibute(LA)	54.42	75.67	83.53
	✓	✓	✓	Latent (L)	53.48	74.77	82.15
				Latent+Attibute(LA)	55.13	76.14	83.77

Table 1: Abalation on Class+Triplet.

2. Attribute extraction

We present the proposed method for extracting attributes for the training set class-ids in (Algo.1). In the algorithm, we use functionality from NLTK [2] and spaCy [1] for noun-phrase extraction (function NOUNPHRASES), parts-of-speech-tagging (function POS-TAGGER), and lemmatization (function LEMMATIZE). The input of the procedure is the text descriptions for each training class-id. A set of words, denoted by 'wordsToDiscard', can also be provided, which are the words to be explicitly deleted from any candidate noun-phrase. We initialize it with all the English stopwords and certain words which are not informative, based on domain-knowledge. For example, we explicitly delete words associated with body-parts (*e.g.* hand, shoulder, head *etc.*). To account for misspelled words, we only keep the words which are in the vocabulary of pre-trained model of Glove embeddings, trained on 2014 Wikipedia [3]. Hence, all the words which are in the train text corpus but out of Glove's vocabulary are also part of 'wordsToDiscard'. We also remove attributes whose frequency is lower than a threshold ν, to reduce noise and to be able to train the attribute classifier with sufficient datapoints.

Algorithm 1 Attribute extraction algorithm.

Input: $Descriptions(classId)$ is given $\forall classId \in \{1 : C\}$, where C is total number of training set classes. Each $Descriptions(classId)$ is a set containing all the text descriptions associated with $classId$. Maximum number of words allowed in phrase is denoted by P. Minimum required phrase frequency is denoted by ν. Set of attributes $wordsToDiscard$, containing words to be explicitly deleted from attributes.
Output: $attr(c)$, $\forall c \in \{1 : C\}$
for $classId \in \{1 \dots C\}$ **do**
  $attrs(classId) \leftarrow \phi$
  for $text \in Descriptions(classId)$ **do**
    Replace punctuations by space
    $candidates(text) \leftarrow$ NOUNPHRASES($text$)
    for $phrase \in candidates(text)$ **do**
      for $word \in phrase$ **do**
        if $word \in wordsToDiscard$ **then**
          Remove $word$ from $phrase$
        end if
      end for
      $posPerWord \leftarrow$ POS-TAGGER($phrase$)
      if ($phrase$ contains atleast one noun) AND (number of words in $phrase \leq P$) **then**
        $phrase \leftarrow$ LEMMATIZE($phrase$)
        $attrs(classId) \leftarrow attrs(classId) \cup \{phrase\}$
      end if
    end for
  end for
  for $phrase \in attrs(classId)$ **do**
    freq($phrase$) $\leftarrow$ freq($phrase$) + 1
  end for
end for
for $classId \in \{1 \dots C\}$ **do**
  for $phrase \in attrs(classId)$ **do**
    if freq($phrase$) $< \nu$ **then**
      Delete $phrase$ from $attrs(classId)$
    end if
  end for
end for

3. Weights for L_{MLC}

We assign $w^p_{l_j}$ and $w^n_{l_j}$ for each attribute $l_j \in V$, where V is the vocabulary of attributes. The weights $w^p_{l_j}$ should be high for rare attributes, hence, it is made to be inversely proportional to frequency of attribute. Similarly, $w^n_{l_j}$ should be high for frequent attributes, hence, it is set to be proportional to frequency of attribute. We denote the frequency of attribute l_j by ν_{l_j}. The following setting for $w^p_{l_j}$ and $w^n_{l_j}$ was used in the experiments:

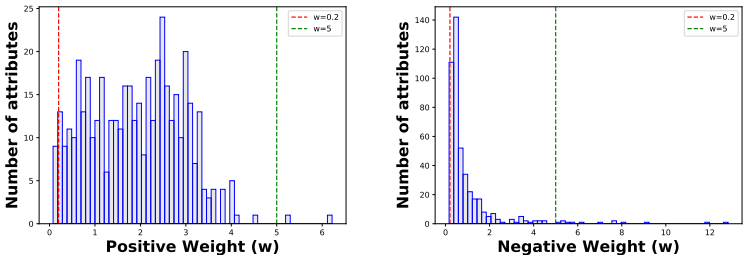
$$f_j = \nu_{l_j}^{0.75} \tag{1}$$

$$f_{avg} = \frac{\sum_{j \in V} f_j}{|V|} \tag{2}$$

$$w^p_{l_j} = \frac{f_{avg}}{f_j} \tag{3}$$

$$w^n_{l_j} = \frac{f_j}{f_{avg}} \tag{4}$$

Finally, we clip the value of each $w^p_{l_j}, w^n_{l_j}$ to be between 0.2 to 5, to reduce the dominance of highly frequent or highly rare words. The plots in Fig.1 show the distribution of weights assigned across the attributes in the vocabulary.



(a) (b)

Figure 1: Weights for attributes in L_{MLC}: Fig.1(a), we show the distribution of weight $w^p_{l_j}$ for each attribute l_j. In Fig.1(b), we show the distribution of weight $w^n_{l_j}$ for each attribute l_j. The histograms are plotted for vocabulary of size 450.

4. CMAAM

We prove the statement that: The cross-entropy loss can be minimized by increasing $\|\mathbf{x}_i\|_2$, while keeping $S_{cos}(W_y, \mathbf{x}_i)$ constant, for classifier with normalized columns.

Proof: Let θ_{ik} denote angle between $\mathbf{x}_i$ and W_k. Let $y_i \in \{1, \ldots, K\}$ denote the class-id of $\mathbf{x}_i$, where K is total number of classes. In following section, θ_{iy_i} denotes angle between $\mathbf{x}_i$ and W_{y_i}. Cross-entropy loss is given as:

$$\begin{aligned} L = L_{XE}(\mathbf{x}_i) &= -\log\left(\frac{\exp(\hat{W}_{y_i}^T \mathbf{x}_i)}{\sum_k \exp(\hat{W}_k^T \mathbf{x}_i)}\right) \\ &= -\log\left(\frac{\exp(\|\mathbf{x}_i\|_2 cos(\theta_{iy_i}))}{\sum_k \exp(\|\mathbf{x}_i\|_2 cos(\theta_{ik}))}\right) \\ &= -\|\mathbf{x}_i\|_2 cos(\theta_{iy_j}) - \log\left(\sum_k \exp(\|\mathbf{x}_i\|_2 cos(\theta_{ik}))\right) \end{aligned} \tag{5}$$

Let the new representation, after increasing the norm but keeping the direction of the vector same, be denoted by $\lambda\mathbf{x}_i$. Note, $\lambda > 1$ as the norm of the vector has been increased and direction is same. Let L_λ denote the cross-entropy loss for $\lambda\mathbf{x}_i$:

$$\begin{aligned} L_\lambda = L_{XE}(\lambda\mathbf{x}_i) &= -\log\left(\frac{\exp(\lambda\hat{W}_{y_i}^T \mathbf{x}_i)}{\sum_k \exp(\lambda\hat{W}_k^T \mathbf{x}_i)}\right) \\ &= -\log\left(\frac{\exp(\lambda\|\mathbf{x}_i\|_2 cos(\theta_{iy_i}))}{\sum_k \exp(\lambda\|\mathbf{x}_i\|_2 cos(\theta_{ik}))}\right) \\ &= -\lambda\|\mathbf{x}_i\|_2 cos(\theta_{iy_j}) - \log\left(\sum_k \exp(\lambda\|\mathbf{x}_i\|_2 cos(\theta_{ik}))\right) \end{aligned} \tag{6}$$

We need to show that for $\lambda > 1, L_\lambda < L$. Assume $\|\mathbf{x}_i\|_2 = \beta_i$.
Note, $\lambda\beta_i \cos(\theta_{ik}) > \beta_i \cos(\theta_{ik})$
$\implies \exp(\lambda\beta_i \cos(\theta_{ik})) > \exp(\beta_i \cos(\theta_{ik}))$
$\implies \sum_k \exp(\lambda\beta_i \cos(\theta_{ik})) > \sum_k \exp(\beta_i cos(\theta_{ik}))$
$\implies \log(\sum_k \exp(\lambda\beta_i \cos(\theta_{ik}))) > \log(\sum_k \exp(\beta_i \cos(\theta_{ik})))$
Similarly, $\lambda\beta_i \cos(\theta_{iy_i}) > \beta_i \cos(\theta_{iy_i})$.
Hence, $\lambda\beta_i \cos(\theta_{iy_i}) + \log(\sum_k \exp(\lambda\beta_i \cos(\theta_{ik}))) > \beta_i \cos(\theta_{iy_i}) + \log(\sum_k \exp(\beta_i \cos(\theta_{ik})))$
$\implies -L_\lambda > -L$
$\implies L_\lambda < L$

4.1. Effect of $L_{norm-reg}$ on Flickr30K

We also study the effect of $L_{norm-reg}$ on Flickr30K [4], a popular image-text cross-modal retrieval dataset. We investigate whether $L_{norm-reg}$ helps over the baseline of $L_{class} + L_{trip}$, as shown in Table 2. It is observed that $L_{norm-reg}$ successfully improves retrieval accuracy over the baseline of $L_{class} + L_{trip}$, hence, further affirming the potential of $L_{norm-reg}$.

Method	R@1	R@5	R@10
$L_{class} + L_{trip}$	32.96	60.60	71.08
$L_{class} + L_{trip} + L_{norm-reg}$	34.68	61.54	71.46

Table 2: Abalation of $L_{norm-reg}$ on Flickr30K

5. Qualitative results

In Fig.2, we present some more qualitative retrieval results. We observe that in the Top-6 results, most of the positive and the negative retrieved images are semantically similarity to the query. Few examples are partially semantically consistent, for instance, in the fourth row, the second image misses the concept of 'bag'. In the last row, we present a result in which each of the retrieved image is of an incorrect identity. However, the images are in accordance with the required description.

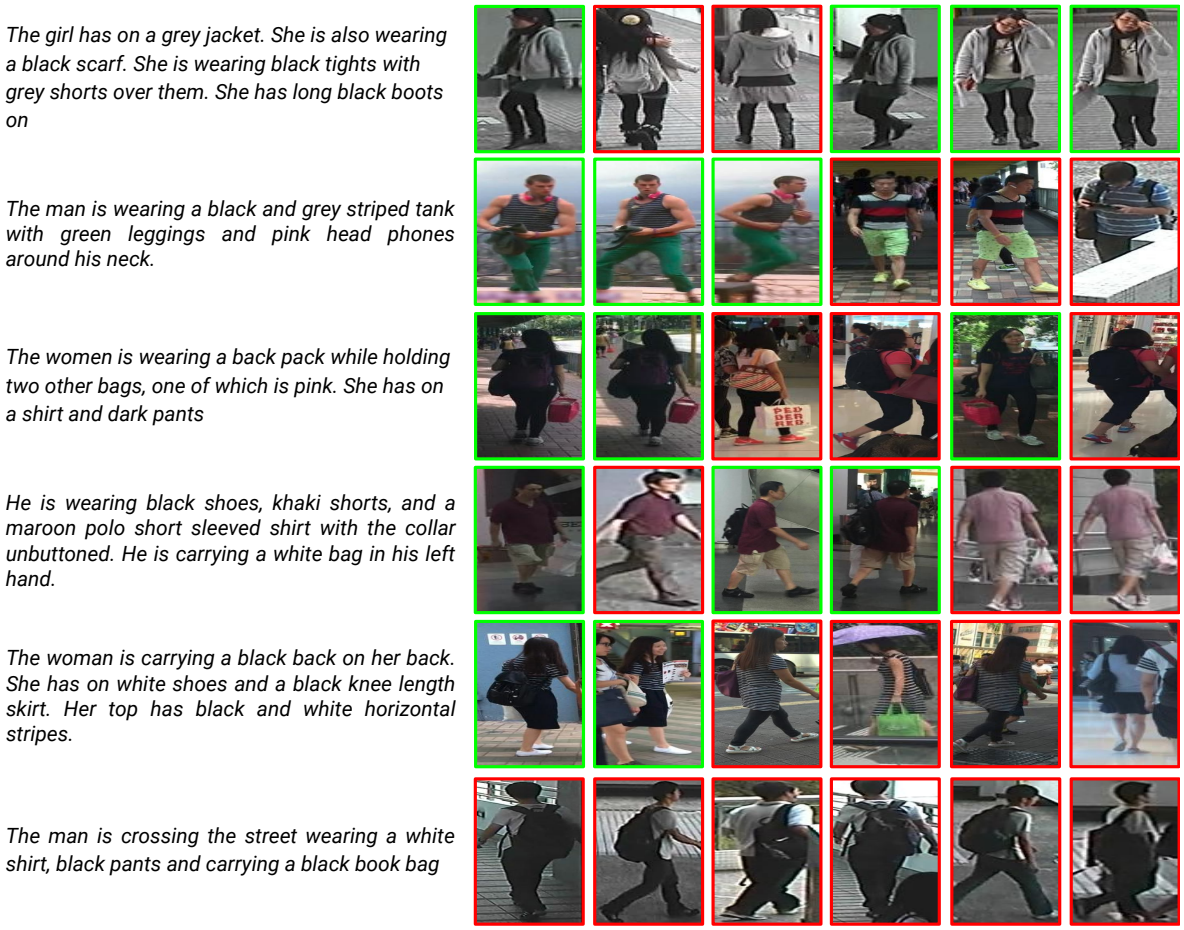


Figure 2: Top 6 retrieval results by CMAAM. Green highlight represents successful match, while red represents incorrect retrieval (best seen in color).

References

[1] M. Honnibal and I. Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[2] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[3] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[4] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014.