1. Ablation results

1.1. Interaction of attribute-space with latent-space

We further elaborate upon the ablations 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.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>$L_{MLC}$</th>
<th>$L_{coral}$</th>
<th>$L_{sTrip}$</th>
<th>Similarity Score</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class+Triplet</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Latent (L)</td>
<td>52.00</td>
<td>74.61</td>
<td>82.26</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Latent (L)</td>
<td>51.75</td>
<td>73.23</td>
<td>81.25</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>Latent (L)</td>
<td>52.81</td>
<td>73.85</td>
<td>81.66</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>Latent (L)</td>
<td>54.14</td>
<td>74.87</td>
<td>82.89</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>Latent + Attribute (LA)</td>
<td>53.44</td>
<td>74.63</td>
<td>82.28</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>Latent + Attribute (LA)</td>
<td>53.56</td>
<td>74.32</td>
<td>82.18</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Latent (L)</td>
<td>52.63</td>
<td>74.64</td>
<td>82.41</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Latent + Attribute (LA)</td>
<td>54.42</td>
<td>75.67</td>
<td>83.53</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Latent (L)</td>
<td>53.48</td>
<td>74.77</td>
<td>82.15</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>Latent + Attribute (LA)</td>
<td>55.13</td>
<td>76.14</td>
<td>83.77</td>
</tr>
</tbody>
</table>

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 $\nu$, 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 ∀classId ∈ {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), ∀c ∈ {1 : C}

for classId ∈ {1 . . . C} do
  attrs(classId) ← φ
  for text ∈ Descriptions(classId) do
    Replace punctuations by space
    candidates(text) ← NOUNPHRASES(text)
    for phrase ∈ candidates(text) do
      for word ∈ phrase do
        if word ∈ wordsToDiscard then
          Remove word from phrase
        end if
      end for
      posPerWord ← POS-TAGGER(phrase)
      if (phrase contains atleast one noun) AND (number of words in phrase ≤ P) then
        phrase ← LEMMATIZE(phrase)
        attrs(classId) ← attrs(classId) ∪ {phrase}
      end if
    end for
  end for
  for phrase ∈ attrs(classId) do
    freq(phrase) ← freq(phrase) + 1
  end for
for classId ∈ {1 . . . C} do
  for phrase ∈ attrs(classId) do
    if freq(phrase) < ν then
      Delete phrase from attrs(classId)
    end if
  end for
end for

3. Weights for L_{MLC}

We assign $w_{lj}^p$ and $w_{lj}^n$ for each attribute $l_j \in V$, where $V$ is the vocabulary of attributes. The weights $w_{lj}^p$ should be high for rare attributes, hence, it is made to be inversely proportional to frequency of attribute. Similarly, $w_{lj}^n$ 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 $\nu_j$.

The following setting for $w_{lj}^p$ and $w_{lj}^n$ was used in the experiments:

\[

f_j = \nu_j^{0.75} \quad (1)
\]

\[

f_{avg} = \frac{\sum_{j \in V} f_j}{|V|} \quad (2)
\]

\[

w_{lj}^p = \frac{f_{avg}}{f_j} \quad (3)
\]

\[

w_{lj}^n = \frac{f_j}{f_{avg}} \quad (4)
\]

Finally, we clip the value of each $w_{lj}^p$, $w_{lj}^n$ 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.
We prove the statement that: The cross-entropy loss can be minimized by increasing $\|x_i\|_2$, while keeping $S_{\text{cos}}(W_{yi}, x_i)$ constant, for classifier with normalized columns.

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

$$L = L_{XE}(x_i) = -\log\left(\frac{\exp(W_{yi}^T x_i)}{\sum_k \exp(W_k^T x_i)}\right)$$

$$= -\log\left(\frac{\exp(\|x_i\|_2 \cos(\theta_{iy_i}))}{\sum_k \exp(\|x_i\|_2 \cos(\theta_{ik}))}\right)$$

$$= -\|x_i\|_2 \cos(\theta_{iy_i}) - \log\left(\sum_k \exp(\|x_i\|_2 \cos(\theta_{ik}))\right)$$

(5)

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

$$L_\lambda = L_{XE}(\lambda x_i) = -\log\left(\frac{\exp(\lambda W_{yi}^T x_i)}{\sum_k \exp(\lambda W_k^T x_i)}\right)$$

$$= -\log\left(\frac{\exp(\lambda \|x_i\|_2 \cos(\theta_{iy_i}))}{\sum_k \exp(\lambda \|x_i\|_2 \cos(\theta_{ik}))}\right)$$

$$= -\lambda \|x_i\|_2 \cos(\theta_{iy_i}) - \log\left(\sum_k \exp(\lambda \|x_i\|_2 \cos(\theta_{ik}))\right)$$

(6)

We need to show that for $\lambda > 1$, $L_\lambda < L$. Assume $\|x_i\|_2 = \beta_i$.

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. **Effect of $L_{\text{norm-reg}}$ on Flickr30K**

We also study the effect of $L_{\text{norm-reg}}$ on Flickr30K [4], a popular image-text cross-modal retrieval dataset. We investigate whether $L_{\text{norm-reg}}$ helps over the baseline of $L_{\text{class}} + L_{\text{trip}}$, as shown in Table 2. It is observed that $L_{\text{norm-reg}}$ successfully improves retrieval accuracy over the baseline of $L_{\text{class}} + L_{\text{trip}}$, hence, further affirming the potential of $L_{\text{norm-reg}}$. 
<table>
<thead>
<tr>
<th>Method</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L_{\text{class}} + L_{\text{trip}})</td>
<td>32.96</td>
<td>60.60</td>
<td>71.08</td>
</tr>
<tr>
<td>(L_{\text{class}} + L_{\text{trip}} + L_{\text{norm-reg}})</td>
<td>34.68</td>
<td>61.54</td>
<td>71.46</td>
</tr>
</tbody>
</table>

Table 2: Abalation of \(L_{\text{norm-reg}}\) 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.

The girl has on a grey jacket. She is also wearing a black scarf. She is wearing black tights with grey shorts over them. She has long black boots on.

The man is wearing a black and grey striped tank with green leggings and pink head phones around his neck.

The women is wearing a back pack while holding two other bags, one of which is pink. She has on a shirt and dark pants.

He is wearing black shoes, khaki shorts, and a maroon polo short sleeved shirt with the collar unbuttoned. He is carrying a white bag in his left hand.

The woman is carrying a black back on her back. She has on white shoes and a black knee length skirt. Her top has black and white horizontal stripes.

The man is crossing the street wearing a white shirt, black pants and carrying a black book bag.

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

References

