Webly Supervised Image-Text Embedding with Noisy Tag Refinement

Niluthpol Chowdhury Mithun*, Ravdeep Pasricha†, Evangelos Papalexakis†, Amit K. Roy–Chowdhury†

†University of California, Riverside, CA, U.S.
Can web images with noisy annotations be leveraged upon with a fully annotated dataset of images with textual descriptions to learn better joint Image-Text embedding models?

Figure: Weakly Supervised Image-Text Embedding. -- The goal is to utilize a large amount of weakly annotated images with a smaller dataset of fully annotated ones to learn a better image-sentence embedding.
What happens when amount of noisy and missing tags associated with web images are unexpectedly high compared to small clean set available?

- **Raw tags** associated with web images are often **incomplete and noisy**.
- Using **web data directly in training [1,2]** without **refinement** may lead to ambiguity and degraded performance.

Based on a limited fully annotated set of images with textual descriptions, is it possible to refine the tags of web image and utilize them in boosting the performance of joint image-text embedding models?

---


Tensor Completion for Tag Refinement

- Inter-relation between web image collection and clean dataset images (based on associated tags) is modeled as a tensor
- A tensor completion based approach to refine tags
- Intra-modal similarity is used side information to regularize CP model
Tensor Completion for Tag Refinement

- Intra-modal similarity is used as side information to regularize CP model

\[ L_{AUX} = \sum_{i,j} \Theta(i, j) ||Z_{i, :}^{(n)} - Z_{j, :}^{(n)}||^2 \]

Auxiliary Information

\[ = \sum_{i,j} Z_{i, :}^{(n)T} \Theta(i, j) Z_{i, :}^{(n)} - \sum_{i,j} Z_{i, :}^{(n)T} \Theta(i, j) Z_{j, :}^{(n)} \]

\[ = \text{tr}(Z^{(n)T} \mathcal{L} Z^{(n)}) \]

\[
\min_{Z^{(n)}, \mathcal{X}} \frac{1}{2} ||\mathcal{X} - [Z^{(1)}, Z^{(2)}, Z^{(3)}]||_F^2 + \frac{\lambda}{2} \sum_{n=1}^{3} ||Z^{(n)}||_F^2 \\
+ \sum_{n=1}^{3} \alpha_n \text{tr}(Z^{(n)T} \mathcal{L}_n Z^{(n)}); \]

Integrating Auxiliary Information in CP Optimization Problem

s.t. \( \Omega \ast \mathcal{X} = \mathcal{T}, Z^{(n)} = U^{(n)} \geq 0 \)
Training Image-Text Embedding Model

- Image-text pairwise ranking loss objective is used for training the joint image-text embedding

$$\mathcal{L}_{IT} = \sum_{(i,t)} \left\{ \sum_{t^-} \max[0, \Delta - f(i, t) + f(i, t^-)] + \sum_{i^-} \max[0, \Delta - f(t, i) + f(t, i^-)] \right\}$$
Experiments

Data Preparation:

- Create **synthetic clean image-tag dataset** from datasets (Flickr30K, MSCOCO) by collecting the unique nouns and verbs as image tags from the associated sentences.
- Create **noisy image-tag datasets (Observed)** from the synthetic clean set based on the missing ratio of tags (e.g., 30%, 50%, 70%)

### Table: Relative errors for recovering missing tags (before and after tensor completion)

<table>
<thead>
<tr>
<th>Missing</th>
<th>Flickr30K</th>
<th>MSCOCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>Observed</td>
<td>0.563</td>
<td>0.721</td>
</tr>
<tr>
<td>Predicted (Proposed)</td>
<td>0.514</td>
<td>0.649</td>
</tr>
<tr>
<td>Improvement (%) by Proposed</td>
<td>9.53%</td>
<td>11.09%</td>
</tr>
</tbody>
</table>

- Average 11% improvement over the observed tensor
- Proposed without regularization shows drop in performance
- Matrix Refinement approach is on par with Observed.
Experiments

Table: Image to Text Retrieval Performance on MSCOCO Sets

<table>
<thead>
<tr>
<th></th>
<th>Actual (No Missing)</th>
<th>Observed (Missing(%) of Actual)</th>
<th>Predicted (Proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing (%) = 30</td>
<td>R@1  9.7 R@10 40.6 MedR 17</td>
<td>R@1  8.8 R@10 37.5 MedR 20</td>
<td>R@1  9.7 R@10 40 MedR 19</td>
</tr>
<tr>
<td>Missing (%) = 50</td>
<td>R@1  9.7 R@10 40.6 MedR 17</td>
<td>R@1  8.6 R@10 33.7 MedR 27</td>
<td>R@1  9.7 R@10 40 MedR 17</td>
</tr>
<tr>
<td>Missing (%) = 70</td>
<td>R@1  9.7 R@10 40.6 MedR 17</td>
<td>R@1  3.8 R@10 19.3 MedR 136</td>
<td>R@1  6.8 R@10 28.9 MedR 34</td>
</tr>
</tbody>
</table>

Actual – Initial Synthetic Clean Image-Tag Set Created by Extracting Unique Noun and Verbs from Captions Associated with Images as Tags.

Observed - Synthetic Noisy Web Image-Tag Set Constructed by Removing Tags based on a Given Missing (%)

Predicted - Refined Image-Tag Set by Refining the Observed Set Applying Proposed Tensor Completion Approach

Qualitative example of tag refinement

(a) Original Tags: airport
Refined Tags: airport, airplane

(b) Original Tags: Cat, pet
Refined Tags: Cat, pet, water
Thank You!