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.

[*]Center for Vision Technologies, SRI International, Princeton, NJ, U.S.

International Conference on Pattern Recognition 2020

Web-Supervised Image-Text Embedding

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?

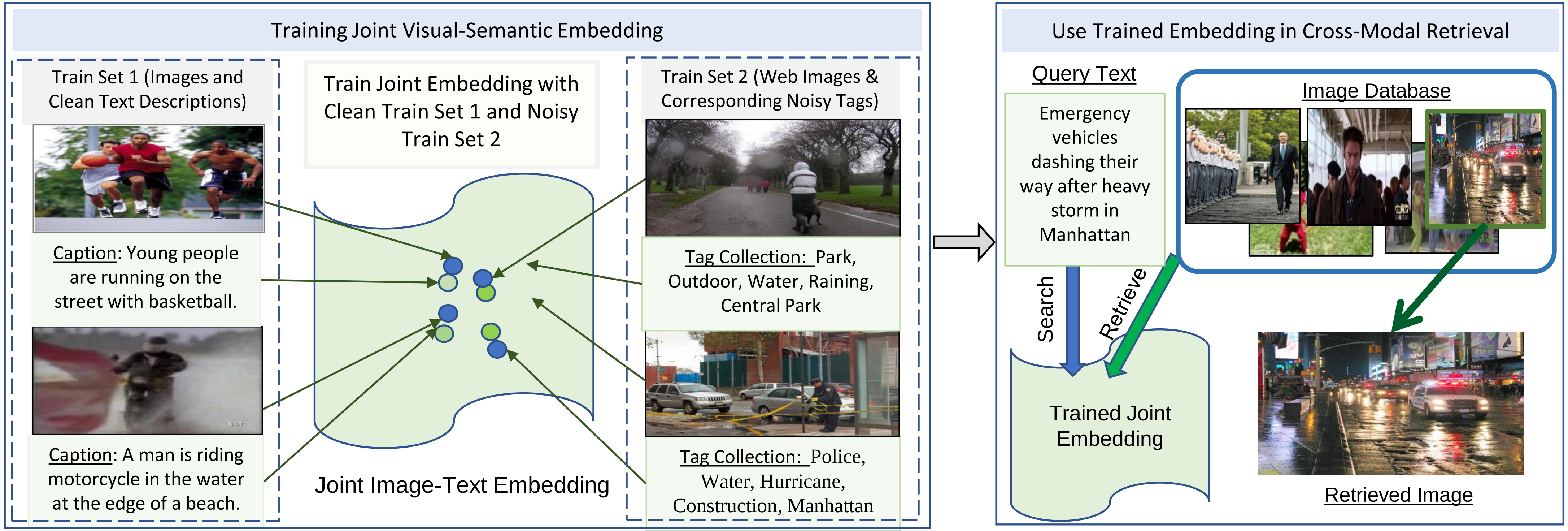
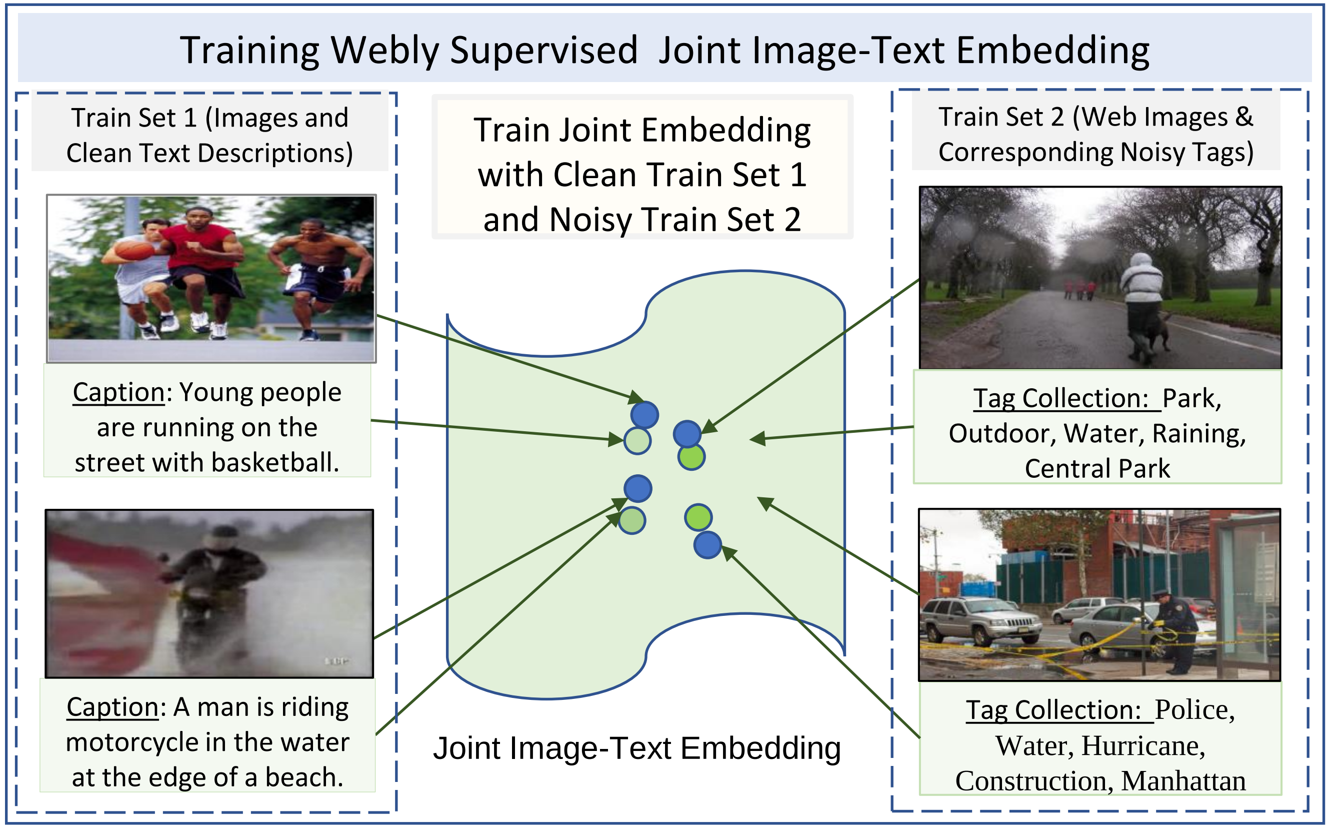


Figure: Weakly Supervised Image-Text Embedding. -- The goal is to utilize a large amount of weakly annotated images with a smaller dataset of fully annotated ones to learn a better image-sentence embedding.

Web Image Tag Refinement

What happens when amount of **noisy and missing tags associated with web images are unexpectedly high** compared to small clean set available?

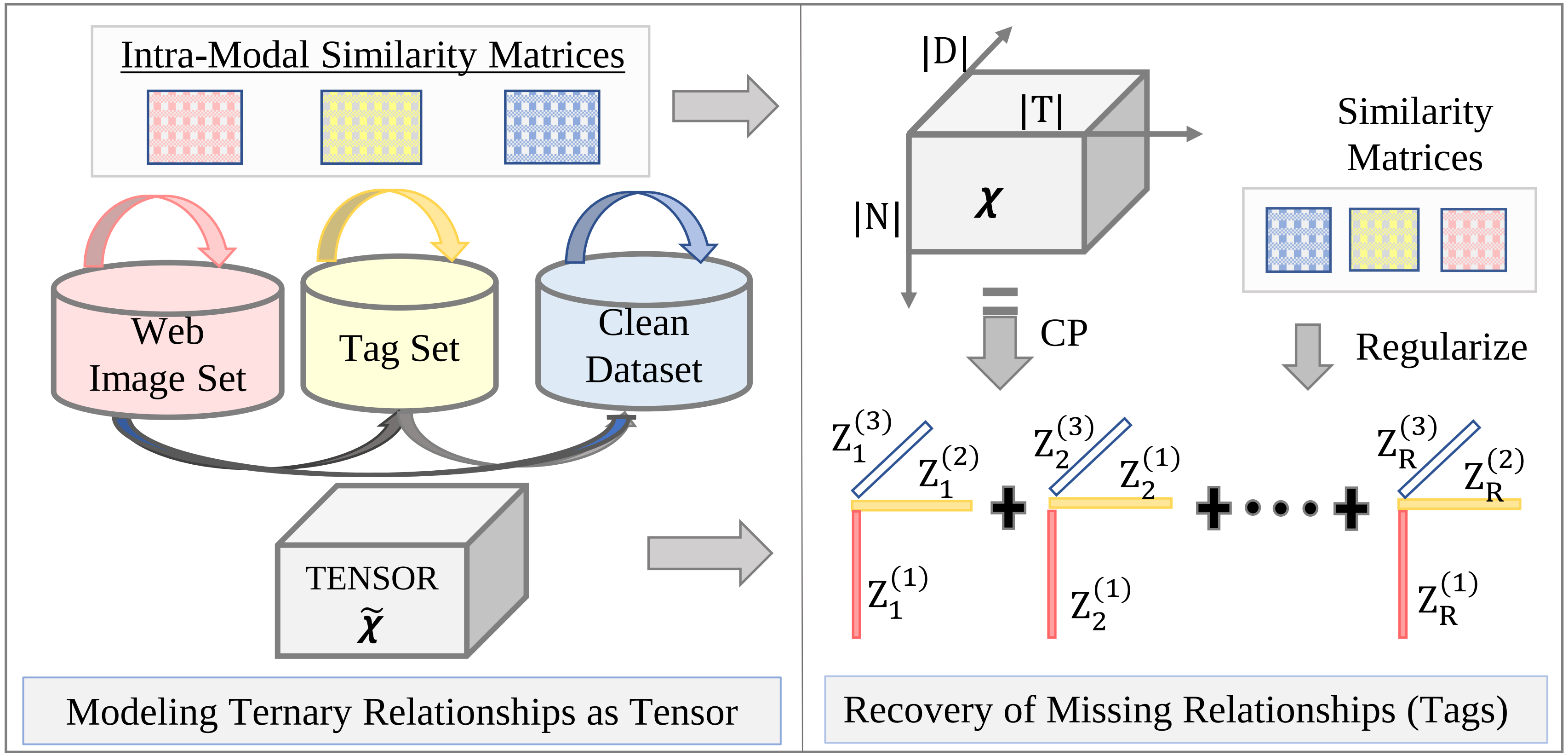
- **Raw tags** associated with web images are often **incomplete and noisy**.
- Using **web data directly in training [1,2] without refinement** may lead to ambiguity and degraded performance.



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?

[1] Yunchao Gong, et al., “Improving image-sentence embeddings using large weakly annotated photo collections”, European Conference on Computer Vision 2014
[2] Niluthpol Mithun et al., “Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval”, ACM Multimedia 2018.

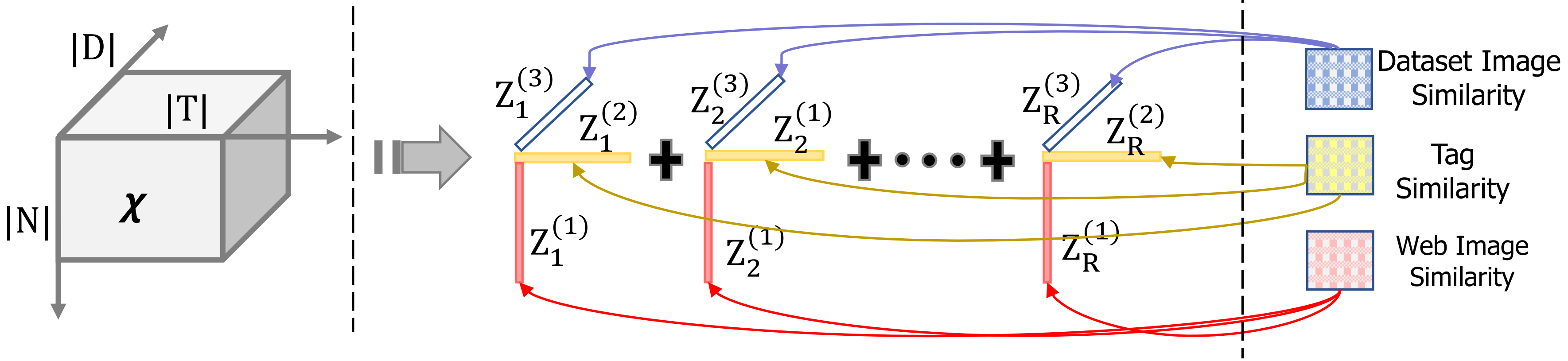
Tensor Completion for Tag Refinement



- 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 side information to regularize CP model



Auxiliary Information

$$L_{AUX} = \sum_{i,j} \Theta(i,j)\ ||Z^{(n)}_{i,:} - Z^{(n)}_{j,:}||^2$$
$$= \sum_{i,j} Z^{(n)^T}_{i,:} \Theta(i,j) Z^{(n)}_{i,:} - \sum_{i,j} Z^{(n)^T}_{i,:} \Theta(i,j) Z^{(n)}_{j,:}$$
$$= tr(Z^{(n)^T} \mathcal{L}\ Z^{(n)})$$

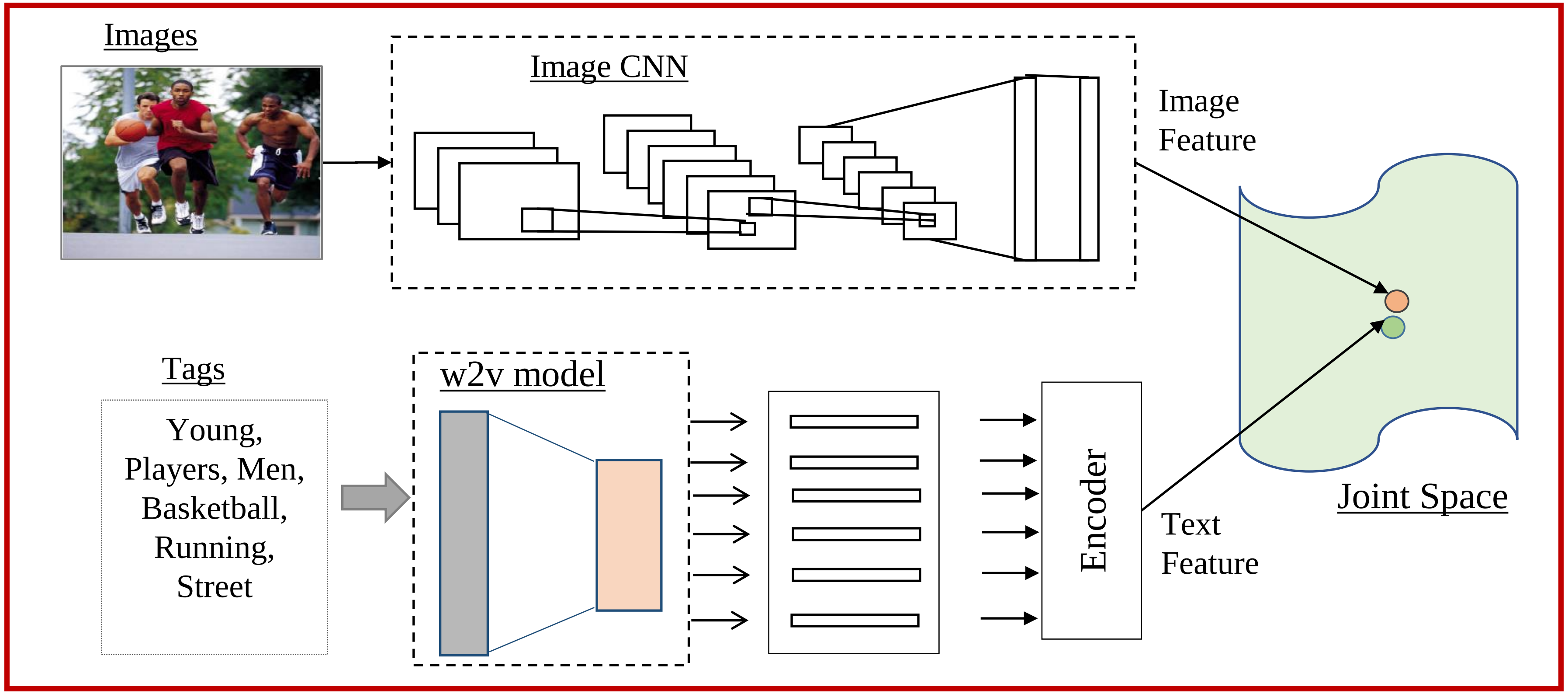
Similarity Matrix :

$$\Theta_D(i,j) = \frac{d_i^T d_j}{||d_i||_2\ ||d_j||_2}$$

Integrating Auxiliary Information in CP Optimization Problem

$$\min_{Z^{(n)},\boldsymbol{\mathcal{X}}} \frac{1}{2}||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3} ||Z^{(n)}||_F^2$$
$$+ \sum_{n=1}^{3} \alpha_n tr(Z^{(n)^T} \mathcal{L}_n\ Z^{(n)});$$
$$\text{s.t. } \boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}, \mathbf{Z^{(n)}} = \mathbf{U^{(n)}} \geq \mathbf{0}$$

Training Image-Text Embedding Model



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

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

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)

Missing	Flickr30K 30%	Flickr30K 50%	Flickr30K 70%	MSCOCO 30%	MSCOCO 50%	MSCOCO 70%
Observed	0.563	0.721	0.839	0.534	0.703	0.838
Predicted (Proposed)	0.514	0.649	0.762	0.463	0.635	0.751
Improvement (%) by Proposed	**9.53%**	**11.09%**	**10.10%**	**15.33%**	**10.71%**	**11.58%**
Predicted Tensor by Baselines						
Proposed (without Regularization)	0.533	0.705	0.826	0.516	0.689	0.822
Matrix Refinement	0.546	0.709	0.834	0.521	0.686	0.828

- 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

	Missing (%) = 30			Missing (%) = 50			Missing (%) = 70		
	R@1	R@10	MedR	R@1	R@10	MedR	R@1	R@10	MedR
Actual (No Missing)	9.7	40.6	17	9.7	40.6	17	9.7	40.6	17
Observed (Missing(%) of Actual)	8.8	37.5	20	8.6	33.7	27	3.8	19.3	136
Predicted (Proposed)	9.7	40	19	9.2	35.4	25	6.8	28.9	34

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!