Supplementary Material for Fast Zero-Shot Image Tagging

Yang Zhang, Boqing Gong, and Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, FL 32816
yangzhang@knights.ucf.edu, bgong@crcv.ucf.edu, shah@crcv.ucf.edu

Due to the page limit, we have omitted some details and experiments from the main text.. We use this document to supplement the discussions in the main text.

Section A presents the formulation of ranking SVM [9, 8] used in Sections 3 and 4.2 of the main text.

Section B shows some additional experimental results on the IAPRTC-12 dataset [6] following the same protocol in Section 5 of the main text.

Section C includes more qualitative results obtained by our Fast**0**Tag and other methods.

A. The formulation of ranking SVM

Ranking SVM plays a vital role in Section 3 of the main text to verify our hypothesis and in Section 4.2 to develop our linear Fast**0**Tag model. We use the implementation of solving ranking SVM in the primal [2] with the following formulation:

$$\min_{\boldsymbol{w}} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{\boldsymbol{y}_i \in Y_m} \sum_{\boldsymbol{y}_j \in \overline{Y_m}} \max(0, 1 - \boldsymbol{w}\boldsymbol{y}_i + \boldsymbol{w}\boldsymbol{y}_j) \tag{1}$$

where λ is the hyper-parameter controlling the trade-off between the objective and the regularization. In the main text, Figure 3 shows how λ could impact the existence and generalization of the principal directions inferred from the ranking SVMs.

B. Experiments on IAPRTC-12

We present another set of experiments conducted on the widely used IAPRTC-12 [6] dataset. We use the same tag annotation and image training-test split as described in [7] for our experiments.

There are 291 unique tags and 19627 images in IAPRTC-12. The dataset is split to 17341 training images and 2286 testing images. We further separate 15% from the training images as our validation set.

Table 1: Comparison results of the **conventional** image tagging with 291 tags on IAPRTC-12.

Method %	MiAP	$K=3$			$K=5$		
		P	R	F1	P	R	F1
TagProp [7]	52	54	29	38	46	41	43
WARP [5]	48	50	27	35	43	38	40
FastTag [3]	48	53	28	36	44	39	41
Fast**0**Tag (lin.)	46	52	28	37	43	38	40
Fast**0**Tag (net.)	**56**	**58**	**31**	**41**	**50**	**44**	**47**

B.1. Configuration

Just like the experiments presented in the main text, we evaluate our methods in three different tasks: **conventional** tagging, **zero-shot** tagging, and **seen/unseen** tagging.

Unlike NUS-WIDE where a relatively small set (81 tags) is considered as the groundtruth annotation, all the 291 tags of IAPRTC-12 are usually used in the previous work to compare different methods. We thus also use all of them conventional tagging.

As for zero-shot and seen/unseen tagging tasks, we exclude 20% from the 291 tags as unseen tags. At the end, we have 233 seen tags and 58 unseen tags.

The visual features, evaluation metrics, word vectors, and baseline methods remain the same as described in the main text.

B.2. Results

Table 1 and 2 show the results of all the three image tagging scenarios (conventional, zero-shot, and seen/unseen tagging). The proposed Fast**0**Tag still outperforms the other competitive baselines in this new IAPRTC-12 dataset.

A notable phenomenon, which is yet less observable on NUS-WIDE probably due to its noisier seen tags, is that the gap between LabelEM+ and LabelEM is significant. It indicates that the traditional zero-shot classification methods are not suitable for either zero-shot or seen/unseen image tagging task. Whereas we can improve the performance by tweaking LabelEM and by carefully removing the terms in its formulation involving the comparison of identical images.

Table 2: Comparison results of the **zero-shot** and **seen/unseen** image tagging tasks with 58 unseen tags and 233 seen tags.

Method %	**Zero-shot** image tagging							**Seen/unseen** image tagging						
	MiAP	$K=3$			$K=5$			MiAP	$K=3$			$K=5$		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1
Random	8.1	2.0	4.5	2.8	2.2	2.2	8.1	3.5	2.2	1.2	1.5	1.9	1.7	1.8
Seen2Unseen	15.6	6.1	13.5	8.4	5.3	19.5	8.4	7.2	3.6	1.9	2.5	4.2	3.7	3.9
LabelEM [1]	11.5	3.6	7.9	4.9	3.6	13.3	5.7	13.8	3.1	1.7	2.2	4.4	3.9	8.7
LabelEM+ [1]	17.6	7.3	16.1	10.0	6.4	23.4	10.0	20.1	13.9	7.4	9.7	13.2	11.8	12.5
ConSE [10]	**24.1**	9.7	21.3	13.3	8.9	32.5	13.9	32.5	38.8	20.6	26.9	31.1	27.6	29.2
Fast**0**Tag (lin.)	23.1	**11.3**	**24.9**	**15.6**	**9.0**	**33.2**	**14.2**	42.9	**50.6**	**27.0**	**35.2**	40.8	36.2	38.4
Fast**0**Tag (net.)	20.3	8.5	18.6	11.6	7.2	26.4	11.3	**45.9**	48.2	25.7	33.5	**42.2**	**37.4**	**39.7**
RankSVM	21.6	10.2	22.6	14.1	8.6	31.7	13.6	–	–	–	–	–	–	–

(a) (b)

Figure 1: The top five tags for exemplar images in [4](a) and [6](b) returned by Fast**0**Tag on the conventional, zero-shot, seen/unseen and 4,093 zero-shot image tagging tasks, and by TagProp for conventional tagging. (Correct tags: green; mistaken tags: red and *italic*)

C. Qualitative Results

In this section, we provide more qualitative results of different tagging methods on both the NUS-WIDE, shown in Figure 1.(a) supplementing Figure 5 in main text, and the IAPRTC-12, shown in Figure 1.(b).

Due to incompletion and noise of tag groundtruth, many actually correct tag predictions are often evaluated as mistaken predictions since they mismatch with groundtruth. This phenomenon becomes especially apparent in 4k zero-shot tagging results in Figure 1.(a) where plentiful diverse tag candidates are considered.

References

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of Output Embeddings for Fine-Grained Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[2] O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with SVMs. *Information Retrieval*, 13(3):201–215, 2010.

[3] M. Chen, A. Zheng, and K. Weinberger. Fast image tagging. In *Proceedings of the 30th international conference on Machine Learning*, pages 1274–1282, 2013.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

[5] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.

[6] M. Grubinger, P. Clough, H. Mller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006.

[7] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316. IEEE, 2009.

[8] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

[9] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[10] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.