Supplementary Material

1. The Shuttersong Dataset

1.1. Favorite Count

Apart from the song clip, image, and mood, we also collect the favorite count for each image-song pair from the Shuttersong application. The favorite counts vary from 1 to 8,964, which could be used to estimate the quality of image-song pairs as a reference. The specific statistics can be found in Fig. 1. There are 6,043 (image, music clip, lyric) triplets owning at least 3 favorite counts, which are considered to jointly show better expressions compared with the others.

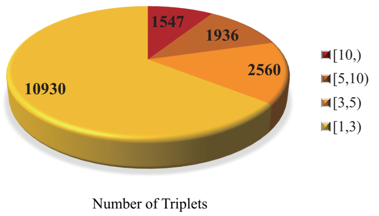

Figure 1. The statistics of triplet number in favorite counts. There are 1,547 triplets owning at least 10 favorite counts, which could be considered as the image-song pair with high quality.

1.2. Lyric Refinement

As there are some abnormal lyrics existing in the automatically searched set, it is necessary to verify each of them. Hence, we ask twenty participants to refine the lyrics, and the corresponding flow char of the refinement is shown in Fig. 2. First, the participants judge whether the song is in English or not. Then they select the mismatch ones and conduct manual searching for the filtered English songs. The websites used for searching in this paper are *www.musixmatch.com* and *search.azlyrics.com*. Finally, both the correct matching and successfully updated ones constitute the refined lyric set. And the rest lyrics are the abnormal ones, *e.g.* non-English songs, unfound lyrics.

2. Additional Experiments

We have shown the specific comparison results of the 28 songs with more 50 times occurrence in the paper, The following subsections show more results of our models with these songs, as well as other compared models.

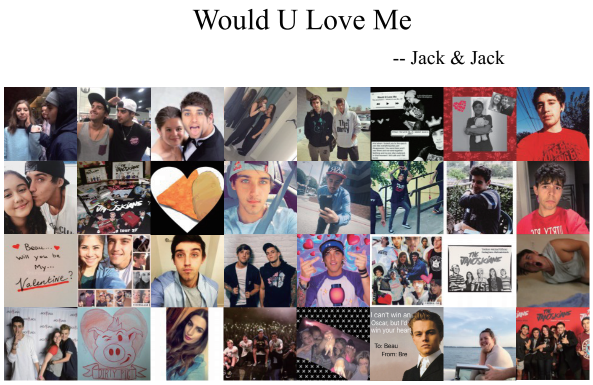

Figure 3. Examples of songs with high frequency appearance in the Shuttersong dataset. Multiple corresponding images are also shown for each of them.

2.1. More Retrieval Results

Apart from the lyric words and image features, we also take consideration of the mood information, which is combined with the encoded lyric representation, but only 18.6% is available. As shown in Table 1, the extra mood information indeed strengthens the correlation between image and lyric, which even outperforms the attention model in some cases. This is because the mood tag directly points out the core information of the shared image-song pair and therefore makes the pair become closer.

2.2. Pooling Operation

The tag attention is obtained by performing the pooling operation over the tag matrix, which plays an important role in establishing the correlation between image and lyric. In view of this, the average and max pooling strategy are compared to evaluate their performances in remaining effective image content. Table. 2 shows the comparison results. It is clear that using average pooling is much better than max pooling. The potential reason is that the average pooling could extract more tag semantic values from the tag matrix, so that more tag values provide a more complete description for images.

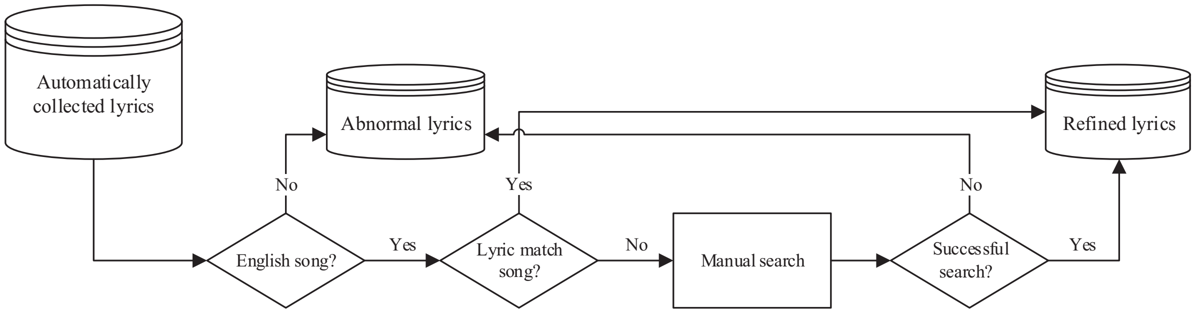

Figure 2. The flow chart of manual lyric refinement. The automatically collected lyrics are divided into two parts, one is the abnormal ones that contains non-English song and undetected lyric, while the other is the refined lyrics used to constitute the final Shuttersong dataset.

Image tags	obj-tags				attr-tags				obj-attr-tags			
Models	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
BoW [1]	10.71	31.21	52.62	9.34	9.32	30.03	51.34	10.06	9.42	34.51	55.73	9.15
CONSE [4]	10.44	30.93	52.42	9.50	9.13	29.61	51.19	10.20	9.39	34.24	55.19	9.35
Attentive-Reader [3]	11.45	32.81	52.02	9.47	9.13	30.26	51.47	9.91	12.95	37.16	61.79	8.62
Our	11.34	32.52	51.44	9.61	8.92	29.82	51.18	9.81	10.95	36.31	57.51	8.87
Our-mood	12.13	34.52	54.60	8.83	**9.70**	31.31	**52.84**	9.13	12.13	37.46	61.85	8.23
Our-attention	**12.71**	**35.14**	**57.37**	**8.37**	9.26	**33.64**	52.26	**8.97**	**13.10**	**38.38**	**62.50**	**7.82**

Table 1. Image2song retrieval experiment result in R@K and Med r. Three kinds of image representation are considered, *e.g.*, object (obj), attribute (attr), and both them (obj-attr).

Pooling	R@1	R@3	R@5	R@10	Med r
Average	13.10	28.30	38.38	62.50	7.82
Max	12.08	26.54	35.40	59.74	8.37

Table 2. The performance of the proposed model with different pooling strategies over the tag matrix.

2.3. Loss Comparison

In addition to the *Mean Squared Error* (MSE) loss function employed in the paper, *Cosine Proximity Loss* (CPL) and *Marginal Ranking Loss* (MRL) are also considered. C-PL is based on the cosine distance, which is commonly used in vector space model and written as follow,

$$l_{cpl} = -\sum_{i=1}^{T} \cos\left(v_i,\ \tilde{l}_i\right). \tag{1}$$

As for MRL, it takes consideration of both positive and negative samples with respect to the images query and is more prevalent in retrieval tasks. It belongs to the hinge loss and is written as,

$$l_{mrl} = \sum_{i=1}^{T} \max\left\{0, 1 + \cos\left(v_i,\ \tilde{l}_i^-\right) - \cos\left(v_i,\ \tilde{l}_i^+\right)\right\}, \tag{2}$$

where $\tilde{l}_i^+$ is the ground truth lyric for current image representation v_i, and $\tilde{l}_i^-$ is a negative one that is randomly selected from the entire lyric database.

Table. 4 shows the comparison results among the three introduced loss functions. It is obvious that MSE performs the best in both Recall@K and Med r metric, while MRL has the worst performance. We consider that the main reason comes from the diversity of images, *e.g.* the examples in Fig. 3. The images related to the same lyrics have high variance in the appearance, which makes these two modalities lack the content correspondence to each other. Hence, it becomes more challenging to deal with the positive and negative samples simultaneously. Such conditions can be also found in the image-to-text retrieval task [2].

Loss	R@1	R@3	R@5	R@10	Med r
MRL	9.90	22.70	36.04	57.84	8.94
CPL	11.29	26.25	37.07	60.92	8.29
MSE	13.10	28.30	38.38	62.50	7.82

Table 4. The retrieval performance of our model with distinct loss functions.

2.4. Attribute Property

In our paper, the attribute tags perform worse than the object ones, one of the potential reasons is due to the im-

Attributes	white	black	blue	brown	red	green	pink	blonde	smiling	···
Average Probabilities	0.30	0.25	0.20	0.19	0.14	0.12	0.09	0.09	0.08	···

Table 3. The top 9 detected attributes with corresponding prediction probabilities.

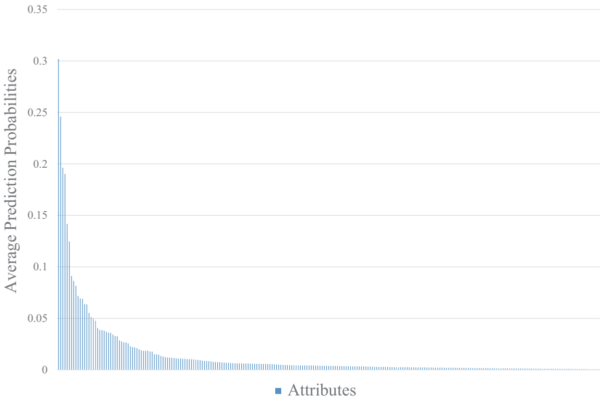

Figure 4. The average attribute prediction results over all the images in dataset†. The results are sorted in the descend order.

balanced attributes. We perform a statistical analysis with the attribute prediction probabilities, where all the images whose corresponding lyrics appear at least 5 times are considered. There are 249 attribute types employed in this paper, and Fig. 4 shows the average prediction results. It is clear to find that only a few types have high value, while most remain the low probabilities, which is actually a kind of long-tailed distribution. The imbalanced results could make it difficult to distinguish the images that belong to different songs. More importantly, the top 9 attributes are almost color-related, as shown in Table. 3. These attributes commonly appear in colorful images, and therefore become weaker in describing the specific image appearance compared with other ones, *e.g. happy*, *messy*. Hence, only employing attribute tags may suffer from the aforementioned problems and result in the unreliable correlation.

References

[1] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, C. Cortes, and M. Mohri. Polynomial semantic indexing. In *Advances in Neural Information Processing Systems*, pages 64–72, 2009. 2

[2] J. Dong, X. Li, and C. G. Snoek. Word2visualvec: Cross-media retrieval by visual feature prediction. *arXiv preprint arXiv:1604.06838*, 2016. 2

[3] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015. 2

[4] 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. 2