Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma

BMVC 2017 Submission # 121

1 Network Architecture of the Proposed Model

Our model is an unified multi-modal learning framework, as illustrated in Fig. 2 of the main paper. Here a detailed description of the network architecture is provided in Table 1. Note that both the sketch and photo branches of the network follow the modified Sketch-a-Net architecture in [3], while the text embedding is obtained from a bidirectional LSTM based language model similar to those used in [1, 2]. The weights between the sketch and photo branches are tied. The formulated quadruplet loss is applied to constrain the feature learning on the linear transform of sketch, photo and text embeddings.

Branch	Layer No.	Input Layer(s)	Layer Type	Kernel Size	Stride	Pad	Output
Sketch-photo branch	0	-	Input	-	-	-	$225 \times 225 \times 1$
	1	0	Conv1	15×15	3	0	$71 \times 71 \times 64$
	2	1	Pool1	3×3	2	0	$35 \times 35 \times 64$
	3	2	Conv2	5×5	1	0	$31 \times 31 \times 128$
	4	3	Pool2	3×3	2	0	$15 \times 15 \times 128$
	5	4	Conv3	3×3	1	1	$15 \times 15 \times 256$
	6	5	Conv4	3×3	1	1	$15 \times 15 \times 256$
	7	6	Conv5	3×3	1	1	$15 \times 151 \times 256$
	8	7	Pool5	3×3	2	0	$7 \times 7 \times 256$
	9	8	FC6	1×1	1	0	$1 \times 1 \times 512$
Text branch	10	-	Input	-	-	-	$1 \times 1 \times 40$ (time stamp)
	11	10	Word Embedding	1×1	1	0	$1 \times 1 \times 300 \times 40$ (time stamp)
	12	11	Bidirectional LSTM	1×512	1	0	$1 \times 1 \times 1024$ (last output)
	13	12	FC8	1×1	1	0	$1 \times 1 \times 256$
Quadruplet loss	14	9	Linear Transform 1	1×1	1	0	$1 \times 1 \times 256$
	15	9	Linear Transform 2	1×1	1	0	$1 \times 1 \times 256$

Table 1: The detailed configuration of each branch of the proposed model.

2 Experiments on Fine-grained Image Retrieval with Sketch-Text Query

In this work we focus on the application scenario where both the text and sketch modalities are available for learning a photo retrieval model; yet during testing, only one modality is used for to conduct retrieval, *i.e.*, we assume that the user of our model would only provide either sketch or text, but not both as the query input. In this experiment, we investigate a different application scenario where a user provides both a sketch and a text description as input to our model for photo retrieval. Note that the same trained model for single modality query is used here for multi-modality query. Since each modality can be used to compute a

© 2017. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

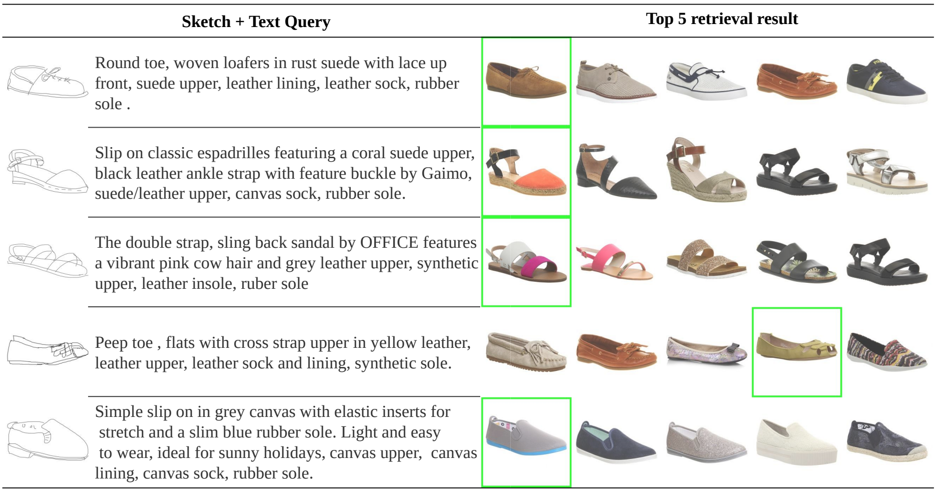

Figure 1: Qualitative example of fine-grained image retrieval with both sketch and text query.

distance/similarity score for each photo in a gallery set, a simple strategy for fusing the two query modality is to compute a weight sum of the two distances. In our experiments, we give a weight of 0.8 to the sketch modality as it is clearly the strongest out of the two. Table 2 shows that after fusing the two query modalities, the retrieval performance is improved compared to that obtained using each modality alone. This suggests that our model can exploit the complementarity of the two modalities for better retrieval performance. Some qualitative results can also be found in Figure 1.

Query	Model	Top 1 acc	Top 10 acc
sketch $\rightarrow$ photo	Our full model	50.38%	84.73%
text $\rightarrow$ photo	Our full model	12.60%	37.40%
(sketch + text) $\rightarrow$ photo	Our full model	**52.67%**	**87.02%**

Table 2: The performance of fine-grained image retrieval when both sketch and text is available as input.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[2] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[3] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016.