## <sup>000</sup> Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma

007

011

031

034

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  $[\square]$ , while the text embedding is obtained from a bidirectional LSTM based language model similar to those used in  $[\square, \square]$ . 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.

19	Branch	Lover No.	Input Lovor(c)	Lavar Tuna	Kornal Siza	Strida	Dod	Output
	Branch	Layer NO.	mput Layer(s)	Layer Type	Kerner Size	Suide	r au	Output
20		0	-	Input	-	-	-	$225 \times 225 \times 1$
		1	0	Conv1	$15 \times 15$	3	0	$71 \times 71 \times 64$
21		2	1	Pool1	$3 \times 3$	2	0	$35 \times 35 \times 64$
<b>2</b> 2		3	2	Conv2	$5 \times 5$	1	0	$31 \times 31 \times 128$
66		4	3	Pool2	$3 \times 3$	2	0	$15 \times 15 \times 128$
23	Sketch-photo branch	5	4	Conv3	$3 \times 3$	1	1	$15 \times 15 \times 256$
		6	5	Conv4	$3 \times 3$	1	1	$15 \times 15 \times 256$
24		7	6	Conv5	$3 \times 3$	1	1	$15 \times 151 \times 256$
25		8	7	Pool5	$3 \times 3$	2	0	$7 \times 7 \times 256$
20		9	8	FC6	$1 \times 1$	1	0	$1 \times 1 \times 512$
26		10	-	Input	-	-	-	$1 \times 1 \times 40$ (time stamp)
	Text branch	11	10	Word Embedding	$1 \times 1$	1	0	$1 \times 1 \times 300 \times 40$ (time stamp)
27		12	11	Bidirectional LSTM	$1 \times 512$	1	0	$1 \times 1 \times 1024$ (last output)
28		13	12	FC8	$1 \times 1$	1	0	$1 \times 1 \times 256$
20	Quadruplat loss	14	9	Linear Transform 1	$1 \times 1$	1	0	$1 \times 1 \times 256$
29	Quadruplet loss	15	9	Linear Transform 2	$1 \times 1$	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

<sup>© 2017.</sup> The copyright of this document resides with its authors.

<sup>145</sup> It may be distributed unchanged freely in print or electronic forms.

Sketch + Text Query	Top 5 retrieval result					
Round toe, woven loafers in rust suede with lace up front, suede upper, leather lining, leather sock, rubber sole.	n 19 19 19 19 19 19 19 19 19 19 19 19 19					
Slip on classic espadrilles featuring a coral suede upper, black leather ankle strap with feature buckle by Gaimo, suede/leather upper, canvas sock, rubber sole.	یک کی کی کے ا					
The double strap, sling back sandal by OFFICE features a vibrant pink cow hair and grey leather upper, synthetic upper, leather insole, ruber sole	🍬 🔌 🍋 🍆					
Peep toe, flats with cross strap upper in yellow leather, leather upper, leather sock and lining, synthetic sole.						
Simple slip on in grey canvas with elastic inserts for stretch and a slim blue rubber sole. Light and easy to wear, ideal for sunny holidays, canvas upper, canvas lining, canvas sock, rubber sole.						

Figure 1: Qualitative example of fine-grained image retrieval with both sketch and text query. 062

distance/similarity score for each photo in a gallery set, a simple strategy for fusing the two <sup>063</sup> query modality is to compute a weight sum of the two distances. In our experiments, we <sup>064</sup> give a weight of 0.8 to the sketch modality as it is clearly the strongest out of the two. Table <sup>065</sup> 2 shows that after fusing the two query modalities, the retrieval performance is improved <sup>066</sup> compared to that obtained using each modality alone. This suggests that our model can <sup>067</sup> exploit the complementarity of the two modalities for better retrieval performance. Some <sup>068</sup> qualitative results can also be found in Figure 1. <sup>069</sup>

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%
$(\text{sketch} + \text{text}) \rightarrow \text{photo}$	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

- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A 082 neural image caption generator. In *CVPR*, 2015.
- [2] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving imagetext embeddings. In *CVPR*, 2016.
- [3] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen Change 087 Loy. Sketch me that shoe. In *CVPR*, 2016. 088
  - 089

073 074 075

076

077

081

- 000
- 091