Deep Multi-task Attribute-driven Ranking for Fine-grained Sketch-based Image Retrieval

Jifei Song[1]
j.song@qmul.ac.uk
Yi-Zhe Song[1]
yizhe.song@qmul.ac.uk
Tao Xiang[1]
t.xiang@qmul.ac.uk
Timothy Hospedales[1]
t.hospedales@qmul.ac.uk
Xiang Ruan[2]
ruanxiang@gmail.com

[1] School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, E1 4NS
United Kingdom

[2] TIWAKI Corporation, Ltd.
Japan

With touch-screen devices becoming ever more ubiquitous, sketch holds great promise as an intuitive and efficient mode of input compared to classic alternatives. This has motivated a major revival of interest in vision-based analysis of sketches, notably in sketch-based image retrieval (SBIR). Superior to classic SBIR methods, fine-grained SBIR (FG-SBIR) methods [1] are proposed to make fine-grained retrieval in category-level.

In this work, we introduce a multi-task learning (MTL) model for FG-SBIR (as illustrated in Fig. 1), where the main task is a retrieval task with triplet-ranking objective similar to [1], and attributes are detected and exploited in two additional side tasks: The *first* side task is to predict the attributes of the input sketch and photo images. By optimising this task at training, we encourage the learned representation to more meaningfully encode the semantic properties of the photo/sketch; The *second* side-task is to perform retrieval ranking based on the attribute predictions themselves. At test time, this means that the retrieval ordering is explicitly driven by semantic attribute-level similarity as well as the similarity of the internally learned representation. The multi-task loss is formulated as Eq. 1 (full details can be found in our paper).

$$\begin{aligned} L\left(s,p^{+},p^{-}\right) &= L_{\theta}\left(s,p^{+},p^{-}\right)+\lambda_{a}L_{a}\left(s,p^{+},p^{-}\right) \\ &+\lambda_{s}L_{p}\left(s,t^{s}\right)+\lambda_{p^{+}}L_{p}\left(p^{+},t^{p^{+}}\right) \\ &+\lambda_{p^{-}}L_{p}\left(p^{-},t^{p^{-}}\right)+\lambda_{\theta}\|\theta\|_{2}^{2} \end{aligned} \tag{1}$$

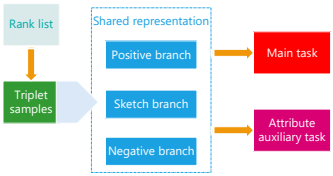

Figure 1: Diagram of the proposed deep multi-task fine-grained SBIR model.

By introducing multiple tasks in the network, the model generalises better and further can rely less on expensive human ranking annotation. Specifically, we show that the highly non-scalable step of triplet annotation required by the model in [1] can now be avoided and an automatic attribute-based strategy is developed instead to focus on the most informative 'hard' training samples for more efficient learning of the model.

Contributions Our contributions are two-fold: (1) A novel deep MTL model is proposed to exploit two attribute-based auxiliary tasks for learning semantically meaningful and domain-invariant representation for FG-SBIR. (2) A new attribute-based triplet generation and sampling strategy is developed to boost the effectiveness of the deep MTL model.

Experiments Extensive experiments are carried out on two benchmarks and the results demonstrate that the proposed model significantly outperforms the state-of-the-art while simultaneously requiring less costly annotation. Partial results are shown in Table 1 (full comparisons can be found in our paper).

Table 1: Comparative results against state-of-the-art retrieval performance.

Shoe Dataset	top 1	top 10	trip-acc	Chair Dataset	top 1	top 10	trip-acc
Triplet model [1]	39.13%	87.83%	69.49%	Triplet model [1]	69.07%	97.94%	72.30%
Ours	**50.43%**	**91.30%**	**70.59%**	Ours	**78.35%**	**98.97%**	**73.13%**

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