Supplementary Material for Sketch-Guided Object Localization in Natural Images

Aditay Tripathi[1], Rajath R Dani[1], Anand Mishra[2] and Anirban Chakraborty[1]

[1]Indian Institute of Science, Bengaluru [2]Indian Institute of Technology Jodhpur
http://visual-computing.in/sketch-guided-object-localization/

A Implementation Details

We implemented the *Cross-modal Attention Model* using the Pytorch v1.0.1 framework with CUDA 10.0 and CUDNN v7.1. The model is trained with Stochastic Gradient Descent (SGD) with momentum of 0.9 on one NVIDIA 1080-Ti with a batch-size of 10. The learning rate was initially set at 0.01 but it decays with a rate of 0.1 after every four epochs and it is trained for 30 epochs. The constant $\mathcal{K}$ in eq. (4) is fixed at 256 and $m^+ = 0.3$ and $m^- = 0.7$ in eq. (7) and (8) for all experiments.

For optimal results, *Cross-modal attention* model is trained incrementally. Firstly, the localization model is trained without attention. Then, the attention model is added to it and it is trained again. The training protocol is same as explained before and it is same for both the steps.

For optimal training of multi-query sketch-guided localization model, the single-query sketch-guided model is trained first and it is used to initialize the multi-query sketch-guided model. The training protocol is same as before except the starting learning rate is set at 0.001.

B Additional Results

In this section, we will describe additional results on single query sketch-guided localization as well as multi-query sketch-guided localization with 3 sketch queries on Pascal VOC dataset. The results are reported for seen and unseen sets of disjoint ($C_{train} \cap C_{test} = \phi$) training and test sets as well as the union of both the sets. The results are shown in fig 1 and 2.

For the disjoint train-test classes experiment, the data from 6 classes is used for training and the data from 3 classes is used for evaluation. There is no overlap between training and testing images.

The multi-query sketch-guided localization model has also been evaluated for the case of 3 sketch queries. We evaluated the model for the disjoint train and test classes as well as the common train and test classes. Both the *Query Fusion* and *Attention Fusion* methods show consistent improvement across disjoint classes

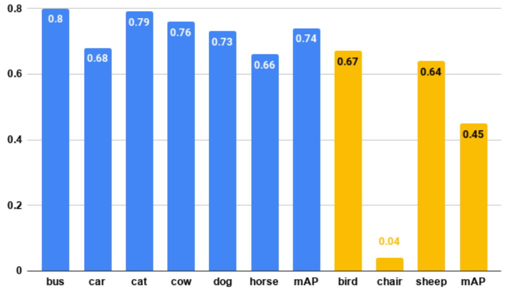

Fig. 1: **Class-wise sketch-guided object localization results of our model on VOC *test2007* dataset are shown in this plot.** In this experiment, training and testing classes are disjoint. Bar-plot in 'blue' color represents results on seen categories and the plots in 'yellow' represents unseen categories. The class-wise AP values are reported and the mAP is also reported separately for seen and unseen categories. [**Best viewed in color**]

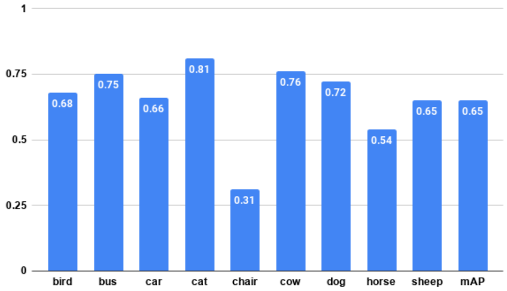

Fig. 2: **Class-wise sketch-guided object localization results of our model on VOC *test2007* dataset are shown in this plot.** The class-wise AP values are reported along with the mAP. [**Best viewed in color**].

Our models	Unseen (D)	Seen (D)	All (C)	mAP (C)
Cross-modal attention	15.0	48.8	50.0	0.30
+Query Fusion(3Q)	17.1	**51.4**	51.9	**0.31**
+Attention Fusion(3Q)	**17.6**	50.9	**52.0**	**0.31**

Table 1: **Results in multi-query common and disjoint train-test categories setting on MS-COCO *Val2017* dataset.** Comparison of the proposed fusion strategies for 3 sketch queries. Here, 3Q, (C) and (D) represents 3 sketch queries, common train and test categories, and disjoint train and test categories respectively. % AP@50 is reported except where mAP is mentioned. For more details refer Section 4.4 and 3.4.

and common classes. The results are reported in table 1. It is evident that both proposed fusion methods are able to effectively combine complementary information present in multiple sketches with as little as 3 sketch queries.

C Additional Visualizations

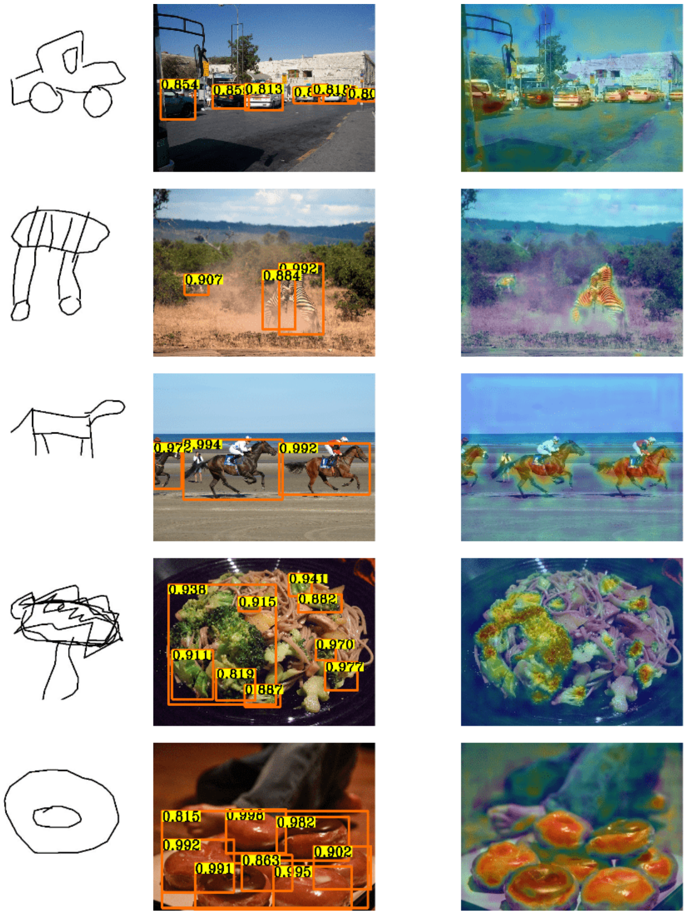

Fig. 3: Sketch query and object localization are shown along with the corresponding attention maps. The images which have multiple instances of the same class as query is shown along with the corresponding attention maps. The image on the left is a sketch query, the image in the center contains the localization generated and the image on the right is the corresponding attention map. As we can see, *cross-modal* attention is able to focus on multiple instance of the same class. [**Best viewed in color**].

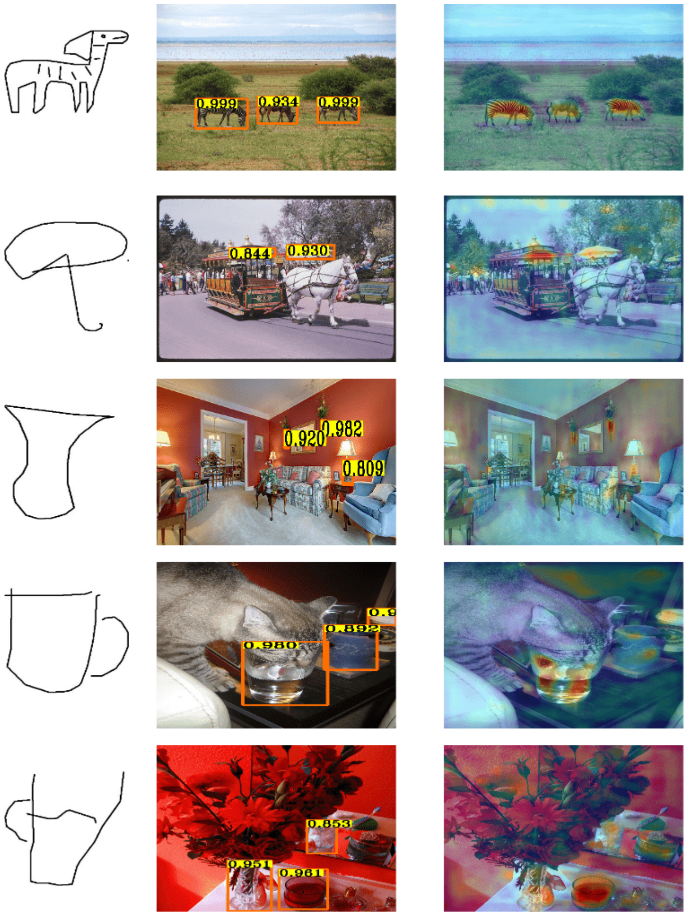

Fig. 4: Some interesting failure cases are illustrated in this figure. The image on the left is a query sketch, the image in the center contains the generated localizations and the image on the right is the corresponding attention map. The model localized objects of wrong classes but upon closer look, the false detection look similar to the objects of the correct class. The query sketches in this case are not visually rich enough the distinguish the correct detection from the false detection. [**Best viewed in color**].