

A Sketch Is Worth a Thousand Words: Image Retrieval with Text and Sketch

Supplementary Material

A Example Sketches at Various Degrees of Complexity

In this section we show examples of sketches at different levels of complexity, as discussed in Section 4.3. For each hand-drawn sketch we collected (100%), we randomly sub-sample and keep only a fraction of strokes in the sketch. These incomplete sketches are only used for investigating the robustness of our model to inaccuracy input sketches (see Section 4.3). Examples can be found in Figure 6.



Fig. 6: For each collected sketch (100%), we randomly sub-sample a fraction of strokes to investigate the robustness of our model to incomplete sketches (see Section 4.3).

B Retrieval Results

We present retrieved images from our model for a number of randomly selected input sketch-text pairs in Figure 7.



Fig. 7: Retrieved images from our model for a number of randomly selected sketch-text pairs.

C Sketch Captioning with Our Model

An interesting application as a result of our caption generation task is caption generation from a sketch. Figure 8 shows examples of these including both success and failure cases. We observe that the captioning component of our network creates a coherent high-level description overall, but appears to struggle to produce an accurate description when presented with a sketch of a non-person object. The network tends to start each caption with “a man” or “a woman” regardless of the content. This may be partly due to the skewed distribution of COCO dataset, where more than half of the training images contain people. Note that caption generation is not the goal of this work. The ability to gain insights into what the network sees can be used to design a better training objective for retrieval. This is an interesting topic for future research.









	a person is riding a horse in the grass .
	a man is sitting on a bench with a dog .
	a man in a suit and tie standing in front of a large building .
	a group of people sitting at a table with food .
	a woman in a black jacket is holding a tennis racquet .
	a man is riding a wave on a surfboard .
	a person is holding a sandwich and a plate of food .
	a person riding a bike with a dog on the back .

Fig. 8: Examples of sketch captions generated from our model. The tendency to start each caption with “a woman” or “a man” stems from the skewness in the label distribution of the training dataset (COCO) where more than half of images contain people. Besides this anticipated issue, our network appears to be able to generate coherent and satisfactorily accurate captions.

D Synthetic Sketches

Examples of synthetically generated sketches of [25] from images in the COCO dataset [27] can be found in Figure 9. Recall from the main text that we use these synthetically sketches to train our model. For evaluation on COCO, we use hand-drawn sketches. Details of how we collect hand-drawn sketches can be found in Appendix F.



Fig. 9: Examples of synthetically generated sketches of [25].

E Training Details

For multi-label classification, we first train a multi-label classifier on top of a pre-trained CLIP embedding (freezing all CLIP pre-trained parameters), and use them to initialize the classifier for fine-tuning. This helps accelerate training as the network is well closer to an optimum than training from scratch with a random initialization. We do train the decoder for the caption generation task from scratch, however. We observe that the presence of caption generation as a helper task helps increase retrieval performance. We start training of the multi-label classifier with a learning rate of 10^{-4} which is then quickly decreased to 10^{-5} . In the final network, we use the weight ratio of 10, 1, and 100 for multi label classification, caption generation, and the contrastive learning term respectively.

F Collecting Hand-Drawn Sketches

While our model is trained with synthetically generated sketches [25], to investigate the retrieval performance on real data, we collect hand-drawn sketches for images in the COCO test set. The collected sketches are drawn by Amazon Mechanical Turk (AMT) workers (Turkers).

The workers draw sketches by going through the following process. An image from the COCO test set is randomly selected as the target image to draw. The image is displayed to the worker for 15 seconds and the worker is asked to memorize crucial details in the image. After 15 seconds, the image disappears and the worker is asked to draw a sketch that represents the image from memory. Drawing takes place on a graphical user interface (GUI) that we provide (see Figure 10). On the provided GUI, the worker can draw, erase, undo a stroke, redo a stroke, or clear the entire canvas. The size of the sketch pad always matches the size of the target image. Some example sketches collected can be found in Figure 11.

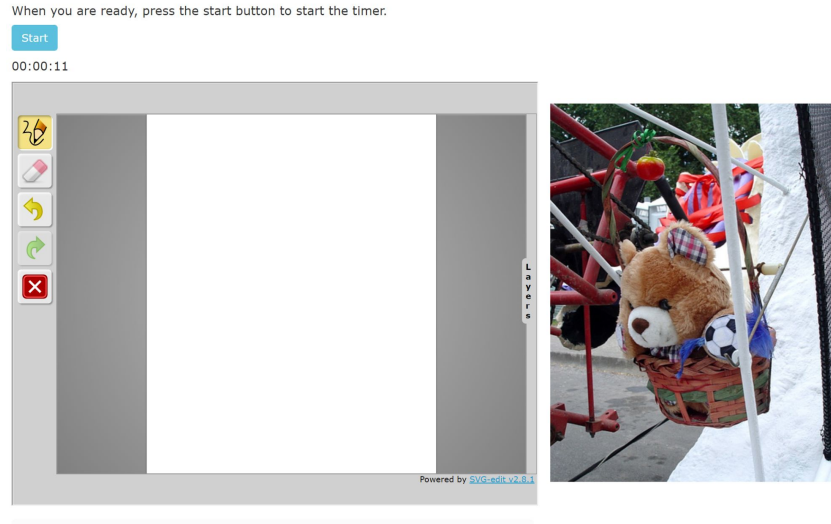


Fig. 10: The graphical user interface that workers on Amazon Mechanical Turk use to draw sketches for images in the COCO test set. Workers can draw, erase, undo/redo, or clear the entire canvas.



Fig. 11: Examples of the hand-drawn sketches collected via Amazon Mechanical Turk.