Title

Deeply Explainable Artificial Intelligence (DExAI): Saliency Driven Retrieval

Team Composition

University of California, Berkeley (UCB), Trevor Darrell, Principal Investigator, trevor@eecs.berkeley.edu
Kitware, Inc., Roddy Collins

Problem Statement

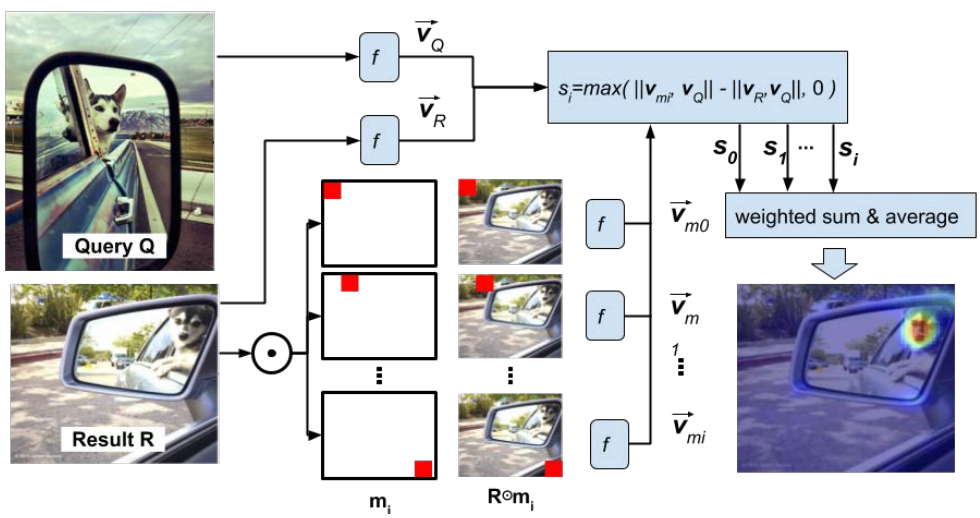
This effort addresses exemplar-based image retrieval, in which the user searches a large image archive for relevant images. Rather than specifying the target via a text string, the user supplies a query image containing the target. The baseline system retrieves similar images and allows the user to supply yes/no relevance feedback, which the system uses to re-rank the results so that higher-ranked results are more relevant to the user.

However, the fundamental process of why some images are retrieved while others are not is opaque to the user. The baseline system converts all images (query and retrievals) into fixed-sized feature vectors, essentially "hashing" the image content into a semantically uninterpretable set of numbers which can be compared but not explained. If a query image and an associated retrieval result both contain a red car next to a blue lake, was the retrieval selected because of the car? Or was the retrieval selected because of the lake? Or neither? There is no way to tell via direct inspection of the feature vectors.

Approach and Technique

Given an query image Q and a retrieved image R, a saliency map can be generated highlighting which regions in R most affect the match to Q. As shown in Figure 1, Q and R are represented by their feature vectors vQ and vR; their initial match score is D==||vQ-vR||. The map is generated by repeatedly masking small blocks in R to yield R', re-computing the feature vector vR', and computing D' == ||vQ-vR'||. If D' is substantially different than D, it can be concluded that the pixels under the mask play a significant role in why R was retrieved. Conversely, if D' is the same as D, it can be concluded that the pixels under the mask are not represented in the feature vector and thus do not contribute to the retrieval. Sliding the mask over the whole image yields the complete saliency map.

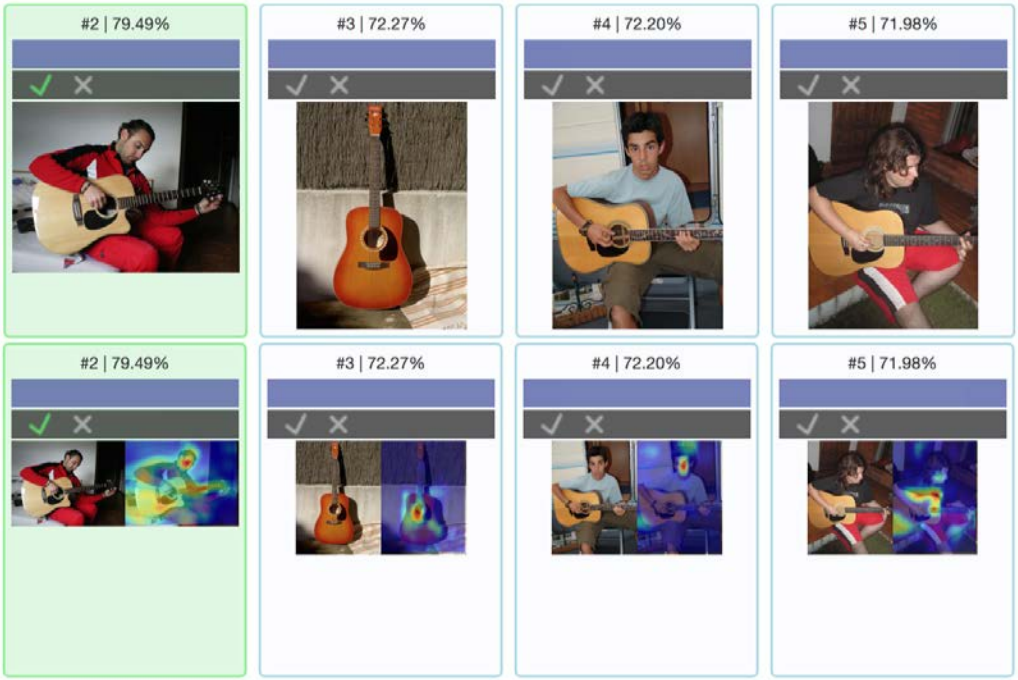
Figure 1. Overview of computing a saliency map between query Q and result R



Approved for Public Release, Distribution Unlimited

These saliency maps are presented to the user as the explanation of why any particular image was retrieved. The baseline system has been modified to compute saliency maps; the user study aims to determine if saliency maps allows the user to more efficiently search the archive by allowing them to more efficiently give feedback; the hypothesis is that the saliency maps allow the user to identify images whose relevant content is actually reflected in the underlying feature vectors which are used during retrieval. Figure 2 illustrates the query and system results both without and with saliency maps.

Figure 2. Query and results without (top) and with (bottom) saliency maps. The query is #2 on left. The saliency maps suggest #3 and #5 were retrieved due to the guitar, while #4 was matched due to the face.



Key Objectives and Current Status

The team has implemented saliency maps on their baseline system (the open-source Social Media Query Toolkit, SMQTK) and adapted it for use with Amazon Mechanical Turk (AMT) to conduct user studies measuring how retrieval accuracy improves when saliency maps are available to guide how feedback should be given. To date, a small pilot study with 24 users has been conducted (see Table 1).

Table 1. Measures, Objectives, and Actuals to Date

Measures	Objectives	Actuals to Date
User-Machine Task Performance	Increase retrieval accuracy	Results suggest saliency maps help when the target object is not too small
Explanation Satisfaction	Measure if user believes the retrieved image was matched to the target object in the query	Likert scale

Test and Evaluation

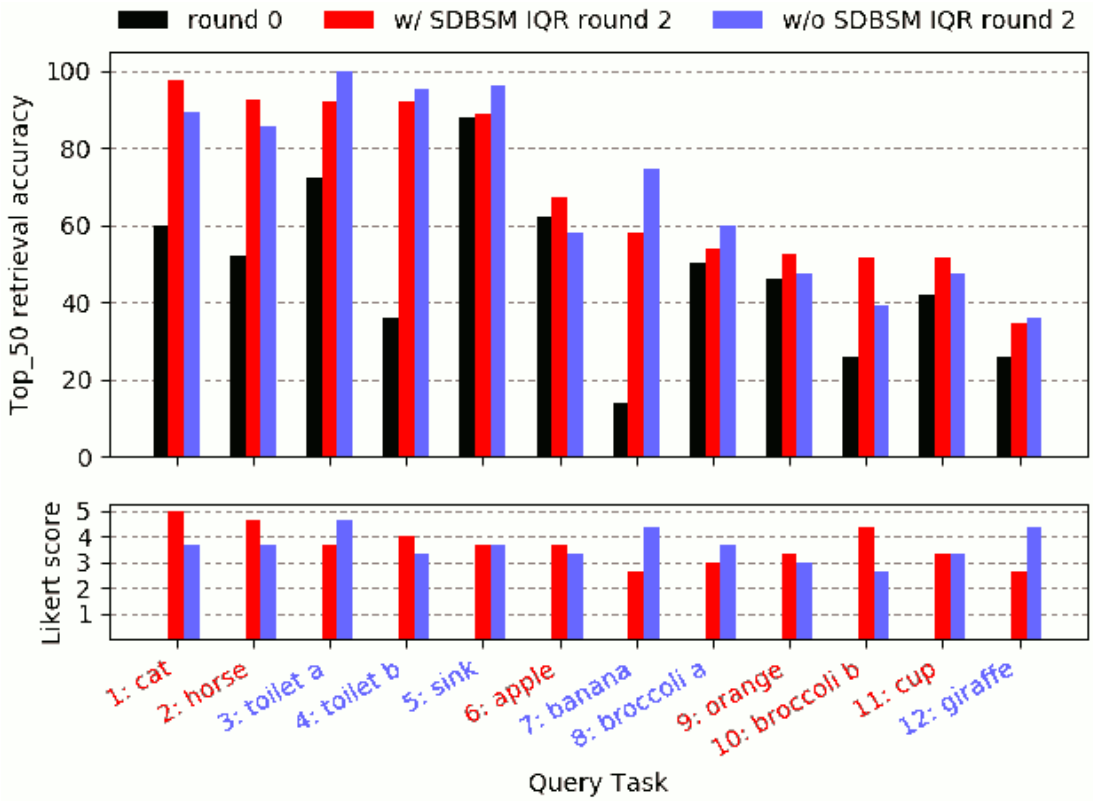
The user study consisted of 12 query images, each presented to a set of three users twice: once with saliency maps and once without (the set of three users was different between the two cases). Each

Approved for Public Release, Distribution Unlimited

query image (and the underlying image archive) was selected from the MS-COCO dataset; the query contained two concepts (e.g., "cat" and "sink"). One of the two was randomly selected as the target; the task was to retrieve as many instances of the target from the archive as possible in two rounds of retrieval + feedback (keep in mind that the user can only query with an image; there is no way to directly communicate the target concept to the system). The hypothesis is that the users with saliency maps could more effectively judge how to give feedback, and thus would achieve higher performance. Additionally, all users, after giving feedback, were asked to rank the statement *"I have high confidence the correct object in the retrieved image is matched to the target object in the query image"* on a Likert scale of 1 - 5 (1 == strongly disagree, 5 == strongly agree).

Figure 3 (top) shows the retrieval accuracy for the twelve concepts; the red bar indicates retrieval accuracy when saliency maps are shown; the blue bar indicates when saliency maps where not shown. The bottom graph shows the Likert scores for each concept and with/without condition. Half of the concepts had higher accuracy with saliency; half had higher accuracy without. The Likert scores strongly correlated with relative accuracy (i.e., if the with-saliency map had a higher score, then the with-saliency Likert score tended to also be higher, and vice versa). Analysis of the users' feedback demonstrated some issues in the protocol, in particular, to avoid negligent or trivial responses, it was required that feedback be given on every result. This is artificially constraining, because the underlying system does not require feedback on every image, and there is no way to positively award neutral feedback. This forced users to make a relevant/not relevant case even when ambiguity may have been more appropriate (for example, a result might have three cats, but the saliency map only highlights one). It is expected that refinement of the protocol, as well as expanding both the number of users and the number of query images, will yield less ambiguous results.

Figure 3. (Top) Concept retrieval accuracy; (Bottom) Likert scores



Approved for Public Release, Distribution Unlimited

Bo Dong, Roddy Collins, and Anthony Hoogs. *Explainability for Content-Based Image Retrieval*, in submission to the CVPR 2019 Explainable AI Workshop.

https://github.com/Kitware/SMQTK

Approved for Public Release, Distribution Unlimited