Ranking user-annotated images for multiple query terms

Moray Allan
http://lear.inrialpes.fr/people/allan/
Jakob Verbeek
http://lear.inrialpes.fr/people/verbeek/

INRIA Grenoble Rhône-Alpes
655 avenue de l'Europe
38330 Montbonnot, France

This paper examines the task of searching photographs shared on the Flickr website (`http://www.flickr.com/`) to find images which contain objects matching user queries. If, for example, the user enters the search terms 'cat' and 'dog', we can quickly find potentially-relevant images using the textual tags associated with the images. Our aim is to rank those potentially-relevant images and return to the user the images which we are most confident do in fact contain a cat and a dog. The system identifies what is consistent across images to learn the visual meaning of user queries and return relevant images to the user. We improve on previous approaches to this task by taking into account the relationships between photographs and the users who uploaded them, to avoid being overly influenced by individual users' inaccurate or idiosyncratic use of textual tags. We focus on user searches containing multiple terms, and show that image ranking performance for multiple query terms can be improved by learning a combined model for the query and adding images which match only a subset of the query terms as negative training examples.

Search queries are answered by using a mixture of kernel density estimators to rank the visual content of web images from the Flickr website whose noisy tag annotations match the given query terms. We estimate the probability that each image in the query set was generated from the query-related class:

$$P(C|\mathbf{X}_j) = \frac{P(\mathbf{X}_j|C)P(C)}{P(\mathbf{X}_j)} = \frac{P(\mathbf{X}_j|C)P(C)}{P(\mathbf{X}_j|C)P(C)+P(\mathbf{X}_j|O)P(O)}. \qquad (1)$$

As a baseline we can take a leave-one-out approach, considering each tagged image in turn as the test image, with the rest of the image set treated as training data. We use kernel density estimation to model the probability that an image $\mathbf{X}_j$ was generated from the non-query-specific class O containing images $\mathbf{Y}_i$:

$$P(\mathbf{X}_j|O) = \frac{1}{N_O}\sum_{i=1}^{N_O}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{d(\mathbf{X}_j,\mathbf{Y}_i)^2}{2\sigma^2}}, \qquad (2)$$

where $d(\mathbf{X}_i,\mathbf{X}_j)$ is an appropriate distance function that measures the similarity of two images' visual content. For each image we also fit a class-specific kernel density estimator to the rest of the image set.

We enhance the baseline model by taking into account user information relating to the images in the query set. Different users use the same textual tags to mean different things, and different users have very different levels of accuracy when the tags are interpreted as describing the visual content of the images. The kernel density estimator in a baseline leave-one-out approach may be excessively influenced by a collection of similar images which a user has labelled inaccurately or idiosyncratically. To avoid this, we can change from 'leave-one-out' to 'leave-some-out'. We first compile information about what images come from the same source – for the experiments below we retrieved from Flickr the username for the uploader of each image, and cross-referenced this information to make a list of images for each user within the query set. We then model the image's probability of being generated from the query-specific class C as:

$$P(\mathbf{X}_j|C) = \frac{1}{Z_j}\sum_{i=1}^{N_C}\delta'_{ij}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{d(\mathbf{X}_i,\mathbf{X}_j)^2}{2\sigma^2}}, \qquad (3)$$

where δ'_{ij} is zero for images belonging to the same user and one otherwise, and Z_j is a normalising factor. This means that images from a user cannot reinforce our confidence in other images from the same user. Instead before we will believe an image label we require agreement from the labels provided by other users. Additionally, if we cross-reference image and user information, we are likely to discover that some users are responsible for many images which our model judges to have inaccurate tags. Our image rankings may be improved if we avoid trusting such users' images even when our model would otherwise rank them highly. The paper includes experimental results which show that requiring agreement between images from different users allows a better model of the visual class to be learnt, and that precision can be increased by rejecting images from 'untrustworthy' users.

If the user searches for two query terms together, one approach would be to learn a separate model for images matching each of the query terms, and then evaluate our confidence that both relevant objects appear in a test image as the product of the individual classifiers' confidence, by the approximation $P(C_1,C_2|\mathbf{X}_j) \approx P(C_1|\mathbf{X}_j)P(C_2|\mathbf{X}_j)$. In our experiments we compare this approach with using a single model for the combination of query terms.

We also look at improving the ranking performance of the model for a combination of query terms by making use of class-specific negative data, in addition to the images sampled from the overall distribution for other classes. While in the product-of-classifiers approach images matching individual query terms are taken as positive training examples for the separate classifiers, our experiments demonstrate improved performance by using a single classifier and treating these images as negative training data.

Figure 1 shows some filtered Flickr images matching two example queries, and the highest-ranking search results given by our model.

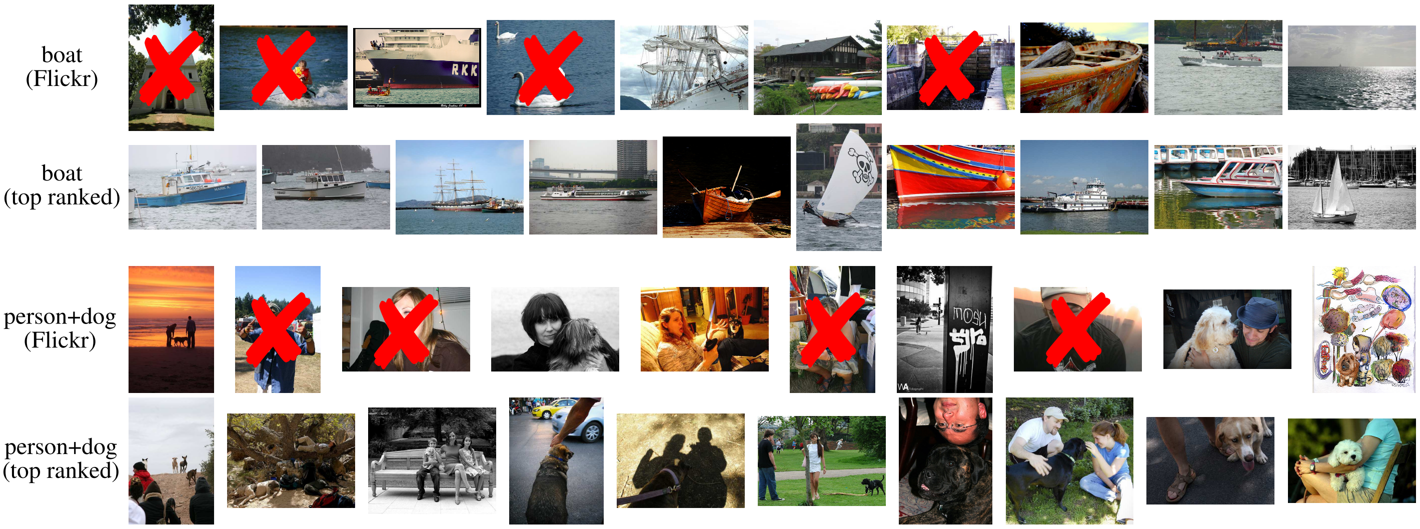

Figure 1: Sample of filtered Flickr images and our top-ranked results.