Summary

- Participation in all 5 runs
- A different algorithm for each feature type
- A common criterion for model selection: best CR@10 calculated using leave-one-(location)-out cross-validation on the devset locations
- A simple visualization tool for getting more familiar with the problem at hand!

Run 1: Visual-only features

A simple feature: VLAD+SURF vectors with multiple vocabulary aggregation (k=128) and joint dimensionality reduction (to 1024d) with PCA and whitening. Implementation publicly available at: https://github.com/socialsensor/socialsensor

Relevance & Diversity (RD) method

A greedy optimization algorithm that selects a fixed-size subset $S$ of the set of images $I = \{i_1, \ldots, i_m\}$ that is (approximately) optimal with respect to the following criterion $U(S)$ that accounts for both relevance to the query location and diversity within $S$:

$$U(S) = \sum_{i \in S} R(i) - \sum_{i \in S} D(i|m) + (1 - w) \cdot D(i|m|S)$$

Relevance: The definition of $R(i|m)$ is $1 - d(i|m, i_m)$ would not work especially when using only visual information.

Diversity in $S$: $D(i|m, i_m) = \frac{1}{|S|} \sum_{i \in S} d(i|m, i_m)$. This definition is not ideal because a small image $i_m$ that has a high similarity with $i_m$ suffices to reduce the diversity of the set. Thus, we define diversity as $D(i|m|S) = \min_{j \in S} d(i|m, i_j)$, i.e. the dissimilarity of $i_m$ to the most similar image in $S$.

Optimization Algorithm: First adds the image with the highest relevance score in $S$ and then sequentially adds the image which has the highest RD score among the remaining images.

Experiments:
- With different classifiers using cross-validation and AUC for model selection, best results obtained with linear SVMs.
- We used the $w$ that gave the best results for CR@10 on the devset ($\geq 0.56$), for producing the test set predictions.

References


Run 2: Text-only features

- Image relevance: We built a forest of 100 random decision trees using most of the textual descriptors available in the datasets. We used both direct image features, such as number of comments and views, and also derived features from the description, tag and title image fields separately, such as the number of words in the field and the normalized sum of tf-idf, social tf-idf and probabilistic values of each word. All continuous variables were discretized.

- Diverse images: We used heuristic-based clustering to find 15 clusters for each location.

Within each cluster, images are ranked by the predicted relevance using the random forest. We then stepped through the clusters iteratively selecting the most relevant remaining image until (up to) 50 had been selected.

Run 3: Visual-text fusion

A simple late fusion scheme: The union of the images returned for each location by Run 1 & 2, ordered in ascending average rank.

Run 4: Human-machine hybrid approach

Task: To improve computer-generated short-lists of 15 images by filtering out 5 images as being either poor-quality or near-duplicates with any of the remaining images, leaving 10 images per location. Short-lists were generated using the text-only method.

Human participants

Not expected to be familiar with any of the locations, nor allowed to consult other sources. Two participants carried out the annotation on a total of 46 locations, around 12% of the total test set.

Run 5: Device and local weather data

The following data sources are combined to get pictures that are diverse in terms of distance from the landmark, angle of the shot, weather conditions and time of the day:

1. Date and time the photo was taken, generally reliable at the granularity of one day.
2. f-stop (aperture size of the shutter) and the exposure time (shutter speed), that can be combined as $EV=f\cdot exp$ exposure, used previously to differentiate indoor from outdoor pictures.
3. Geo-location of the device when the photo was taken, from which we compute the angle and distance to the photographed landmark.
4. We also query a public database of historical weather data to get the weather of the day the picture was taken, which indicates the main weather conditions (e.g. sun, fog, rain, snow, haze, thunderstorm, tornado).

We input the features to the $k$-means algorithm $(k=10)$. Inside each cluster, when multiple candidates photos are available, we select the photo with the highest number of Flickr favorites. We verified that including the number of favorites as an additional feature to the $k$-means is beneficial for the selection of diverse images.

Results

- Best performance in terms of CR@10 and F1@10 for our visual run (run 1)
- Human-machine hybrid (run 4) run improves the textual run (run 2)

Acknowledgements

This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.