Summary

- All runs produced by a different instantiation (features, parameter configuration) of the ReDiv method.
- The runs (fully automated, no external data):
  1. visual-only using VLAD+CSURF features for relevance and diversity
  2. text-only using BoW features for relevance and diversity
  3. visual+textual variations using early fusion of VLAD+CSURF and BoW features for relevance and VLAD+CSURF features for diversity
- A common criterion for model selection: best F1@20 calculated using leave-one-(location)-out cross-validation on the devset locations.

The ReDiv Method

ReDiv casts the dual relevance and diversity goal of a diversification algorithm into the following optimization problem:

$$\arg \max_{S \subseteq U} I(S) = w R(S) + (1 - w) D(S),$$

where $I$ is the initial set of images and $S$ is a $k$-sized subset of $I$ that has maximum utility $U(S)$, defined as a weighted combination of the relevance and the diversity of $S$.

Relevance

Relevance in $R(S) = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j d(i_m, i_m_j)$.

This definition can be problematic, especially when one relies only on visual information. Dissimilar images can be relevant (e.g. inside views) and vice versa (e.g. people in focus).

Solution: Learn what is relevant from the ground truth! A distinct model is built for each location using relevant/irrelevant images of other locations as positive/negative examples. The Wikipedia images of each location as also added in the training set and used as positive examples with a large weight. Thus in our case $R(i_m, i_m_j)$ is the output of a probabilistic classification model.

Diversity

Assuming a ranking $i_M, \ldots, i_1$ of the images in $S$, $D$ defines diversity as:

$$D(S) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j d(i_m, i_m_j) \text{ [high average dissimilarity $\rightarrow$ high diversity]}$$

Problem: Image sets that contain highly similar image pairs (probably belonging to the same cluster) can receive high diversity scores $\rightarrow$ negatively impacts Cluster Recall!

Solution: A more strict definition of diversity:

$$D(S) = \min_{i_M, i_m, i_m_j} d(i_M, i_m)$$

The diversity of a set $S$ is defined as the dissimilarity between the most similar pair of images in $S$.

Optimization

The exact optimization of Equation 3 is infeasible!

We perform a greedy, approximate optimization:

Start with an empty set $S$ and sequentially expand it by adding at each step $j$ the image(s) $i^m$ that scores highest (among the unselected images), to the criterion:

$$U(i^m) = w R(i^m) + (1 - w) \min_{i_M, i_m} d(i^m, i_m) \text{, where } S^j \text{ represents } S \text{ at step } j-1$$

A less greedy version keeps $M > 1$ highest scoring image subsets in each step.

Experimental Protocol

ReDiv allows using different representations for relevance and diversity.

- To reduce complexity of experiments, representations were first evaluated in terms of relevance detection and only the best performing were used for diversity.
- AUC (ability to rank relevant images higher than irrelevant) was used to measure relevance detection performance.
- L2-regularized logistic regression was used as classification algorithm.
- For each combination of relevance detection model and diversity representation, ReDiv was applied with different values for $w$, $n$ (the number of most relevant images to consider) and $M$.
- The best performing setup for each type of features (visual, textual, visual+textual) was used to produce the final runs.

Runs

Visual (Run 1)

- We tested all precomputed visual features made available by the task organizers as well as our own features.
- Best results were obtained using VLAD+CSURF vectors $k = 128$, $d = 128$ for both relevance and diversity.
- Cosine distance was used as dissimilarity measure.
- The parameters used to produce the 1st run are: $w = 0.4$, $n = 75$, $M = 1$.

Textual (Run 2)

- A parsed version of the Wikipedia page in place of the Wikipedia images.
- Flickr images are substituted by a concatenation of the words in their titles ($\times 3$), description ($\times 2$) and tags ($\times 1$).
- A bag-of-words representation with the 20K/7.5K most frequent words was used for the relevance/diversity component.
- Again, cosine distance was used as dissimilarity measure.
- The parameters used to produce the 2nd run are: $w = 0.95$, $n = 110$, $M = 1$.

Visual+Textual (Runs 3 and 5)

- An early fusion (concatenation) of the visual and textual features used in Runs 1 and 2 was used for relevance.
- The visual features used in Runs were used for diversity.
- The parameters used to produce the 3rd run are: $w = 0.75$, $n = 90$, $M = 5$.
- The 5th run differs from the 3rd run only in the value used for $n$ ($= 95$).

Results

Better performance obtained by visual+textual runs, followed by visual-only runs.

Textual features alone are less helpful but powerful in combination.

Acknowledgements

This work is supported by the SocialSensor FP7 project (http://www.socialsensor.eu), partially funded by the EC under contract number 287975.

References