Information-Theoretic Active Learning for Content-Based Image Retrieval

Supplemental Material

1 Performance of MCMI[min] and AdaptAL

Since MCMI[min] [1] and AdaptAL [2] also maximize a mutual information criterion and are, thus, similar to our method, we also tried to apply those methods to our benchmark datasets. Even though we replaced the expensive logistic regression with Gaussian process inference for being comparable to our method, they could only be applied to the Butterflies and 13 Natural Scenes dataset within reasonable time. For the remaining 3 datasets, we randomly sub-sampled 1000 candidates from the entire dataset, as suggested by [2].

Table 1. Comparison of ITAL with MCMI[min] and AdaptAL in terms of AULC.

Method	Butterflies	USPS	Nat. Scenes	MIRFLICKR	ImageNet
random	0.7316	0.5416	0.5687	0.4099	0.1494
MCMI[min]	0.6846	0.5293	0.4554	0.4087	0.1413
AdaptAL	**0.7716**	0.6487	0.6424	0.4643	0.1746
entropy (ours)	0.7512	0.6484	**0.6547**	0.4703	0.1793
ITAL (ours)	0.7511	**0.6522**	0.6233	**0.4731**	**0.1841**

The results in Table 1 show that MCMI[min] does not work well in a batch-mode scenario and performs worse than random.

AdaptAL, on the other hand, is the top performer on the Butterflies dataset and the second-best method on Natural Scenes, directly behind our batch-entropy approach. These are the two datasets where it could be applied in reasonable time on the entire dataset. The sub-sampling that is necessary on the remaining three datasets, however, negatively impacts performance, especially on ImageNet.

To the best of our knowledge, our method is the first one that makes an information-theoretic approach to batch-mode active learning applicable in realistic scenarios without sub-sampling the dataset.

2 Simulation of Imperfect Users

As described in section 4.5 of the paper, we have investigated the effect of three different extreme user behavior models on the performance of the tested BMAL methods. With regard to our approach, we have evaluated both ITAL with the user model parameters p_{label} and p_{mistake} set according to the simulated user and ITAL with the perfect user assumption, which is faster.

We have selected batches of 4 images for annotation at each round.

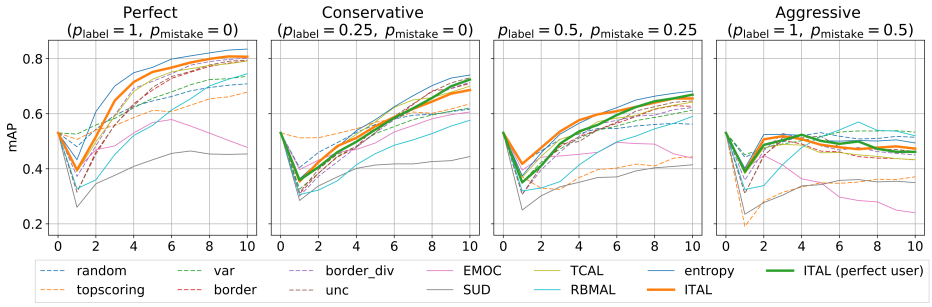

Fig. 1. Comparison of different user behavior models on Natural Scenes.

As expected, all methods suffer from imperfect user feedback compared to a perfect user. While an adequate user model helps ITAL to learn faster during the first rounds, the difference is small enough to justify the use of the perfect user assumption even if it is not true in order to gain a significant speed-up. The case of overly aggressive but error-prone users obviously cannot be handled by the active learning method alone, but also requires adequate handling of such scenarios by the classifier.

3 Sensitivity of Results regarding Feature Dimensionality

To assess to which extent the results presented in the paper are affected by certain transformations applied to the features, we experimented with different dimensionalities of the feature space on the MIRFLICKR dataset. To this end, we have applied PCA to the features extracted from the first fully-connected layer of VGG16, which comprise 4096 dimensions, and projected them onto spaces with 64, 128, 256, 512, and 1024 features. Experiments with all BMAL methods have been conducted on those features for 10 rounds of user feedback and the area under the learning curve (AULC) for the various dimensionalities is reported in Fig. 2.

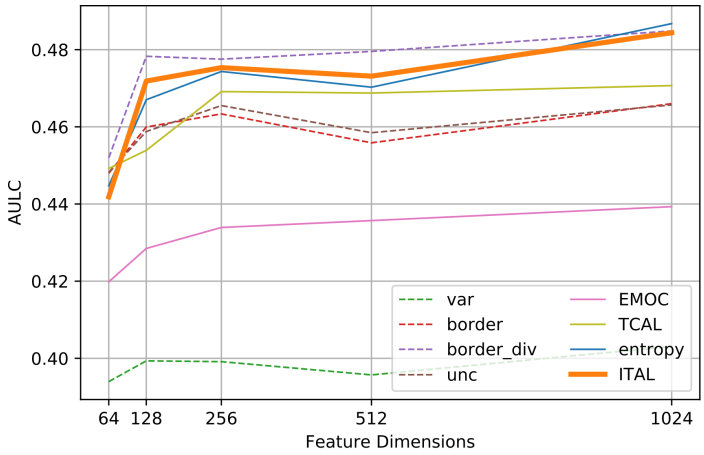

Fig. 2. Area under Learning Curve (AULC) of various BMAL methods on MIRFLICKR with varying feature dimensionality.

The results show that the relative performance of the different methods compared to each other is largely insensitive to the number of features. The performance of ITAL is stable up to as few as 128 dimensions, while some other methods such as TCAL and EMOC already degrade after reducing the number of features to less than 256. When using 1024 features, ITAL is even able to catch up to border_div, which is the best performing method on this particular dataset. However, we have used 512 features for our experiments in the paper due to the increased computational cost incurred by higher-dimensional feature spaces.

4 Examples for Failure Cases

To analyze the possible shortcomings of our method, we have picked four queries from the MIRFLICKR dataset where ITAL had the worst AULC score. These are depicted in Fig. 3, along with the candidate images selected for annotation over 4 rounds of feedback and the top results retrieved by the relevance model after each round.

The first query could be interpreted in multiple ways: The user could be searching for images of people, of babies, or of adults with babies. All these options are covered by the candidate images selected by ITAL. Only one of those image shows a baby alone, which is the actual search objective in this example. That image, however, has not been annotated confidently as showing a baby in the MIRFLICKR dataset, so that it remains unnameable here.

The second query shows a swarm of birds on a power pole, but the simulated user actually searches for birds. The features used in our experiment are apparently not sufficient to capture the semantics of this image well enough for recognizing that it is about birds. Thus, the selected candidates do not contain any image of a bird in a different scene and the classifier cannot abstract away from power poles.

The "night" query, on the other hand, is again an example of erroneous annotations in the dataset: Several images of night scenes have been selected as candidates, but have been annotated either as unnameable or even as irrelevant.

Finally, the last query image shows a river and the candidates are actually quite suitable to identify whether the user is more interested in mountain scenes, water scenes, river scenes, or natural scenes in general. However, either the features or the small number of annotated images seem to be insufficient in this case for distinguishing between rivers and other bodies of water.

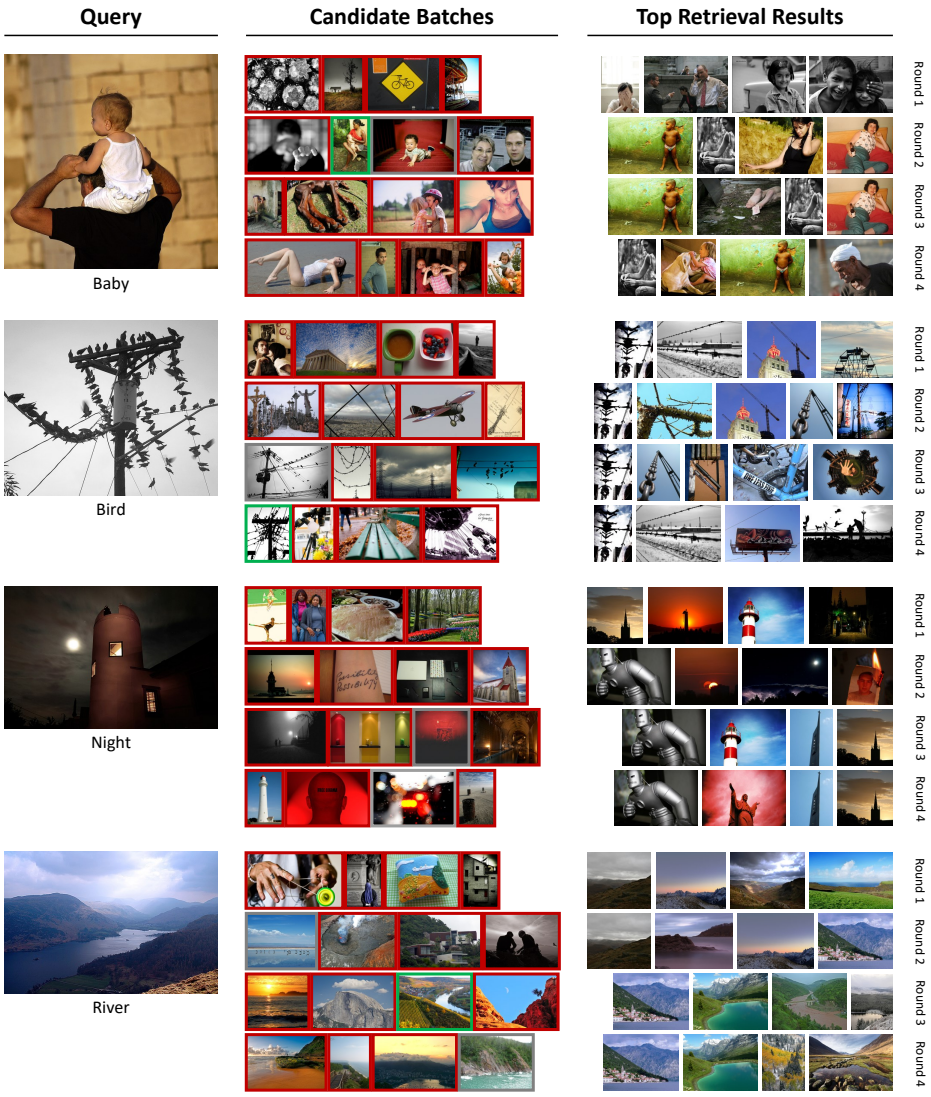

Fig. 3. Four queries from MIRFLICKR where ITAL performed worst.

5 Derivation of Eq. (3)

Plugging in the definitions of entropy and conditional entropy into the definition of mutual information given in eq. (2) leads to the following:

$$\mathfrak{I}(R, F \mid u) = -\left[\sum_{r \in \{-1,1\}^n} P(R = r \mid u) \cdot \log P(R = r \mid u)\right] + \left[\sum_{\substack{r \in \{-1,1\}^n \\ f \in \{-1,0,1\}^n}} P(F = f \mid u) \cdot P(R = r \mid F = f, u) \cdot \log P(R = r \mid F = f, u)\right].$$

Expressing $P(R = r \mid u)$ in the first sum as the marginalization

$$P(R = r \mid u) = \sum_{f \in \{-1,0,1\}^n} P(F = f \mid u) \cdot P(R = r \mid F = f, u)$$

allows us to merge the two sums:

$$\mathfrak{I}(R, F \mid u) = \sum_{\substack{r \in \{-1,1\}^n \\ f \in \{-1,0,1\}^n}} \left[P(F = f \mid u) \cdot P(R = r \mid F = f, u) \cdot \log\left(\frac{P(R = r \mid F = f, u)}{P(R = r \mid u)}\right)\right].$$

Using Bayes' Theorem we can substitute

$$P(F = f \mid u) \cdot P(R = r \mid F = f, u) = P(R = r \mid u) \cdot P(F = f \mid R = r, u)\,,$$

finally leading to eq. (3) from the main paper.

References

1. Guo, Y., Greiner, R.: Optimistic active-learning using mutual information. In: IJCAI. vol. 7, pp. 823–829 (2007)
2. Li, X., Guo, Y.: Adaptive active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 859–866 (2013)