# Information-Theoretic Active Learning for Content-Based Image Retrieval

Björn Barz(✉), Christoph Käding, and Joachim Denzler

Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
bjoern.barz@uni-jena.de
http://www.inf-cv.uni-jena.de

**Abstract.** We propose Information-Theoretic Active Learning (ITAL), a novel batch-mode active learning method for binary classification, and apply it for acquiring meaningful user feedback in the context of content-based image retrieval. Instead of combining different heuristics such as uncertainty, diversity, or density, our method is based on maximizing the mutual information between the predicted relevance of the images and the expected user feedback regarding the selected batch. We propose suitable approximations to this computationally demanding problem and also integrate an explicit model of user behavior that accounts for possible incorrect labels and unnameable instances. Furthermore, our approach does not only take the structure of the data but also the expected model output change caused by the user feedback into account. In contrast to other methods, ITAL turns out to be highly flexible and provides state-of-the-art performance across various datasets, such as MIR-FLICKR and ImageNet.

**Keywords:** Batch-mode active learning · Image retrieval

## 1 Introduction

For content-based image retrieval (CBIR) [25,30], just classifying the query image is, in general, not sufficient. Since images encode complex semantic and stylistic information—sometimes more than text can express—a single query is usually insufficient to comprehend the search interest of the user. A common approach to overcome this issue is enabling the user to provide *relevance feedback* by tagging some retrieval results as relevant or irrelevant [4,6,31]. This way, however, the user will only be able to give feedback regarding images about which the retrieval system is already very confident.

The effect of relevance feedback could hence be significantly improved when the user is not asked to provide feedback regarding the currently top-scoring results, but for those instances of the dataset that are most informative for the

**Electronic supplementary material** The online version of this chapter (https://doi.org/10.1007/978-3-030-12939-2_45) contains supplementary material, which is available to authorized users.

© Springer Nature Switzerland AG 2019
T. Brox et al. (Eds.): GCPR 2018, LNCS 11269, pp. 650–666, 2019.
https://doi.org/10.1007/978-3-030-12939-2_45

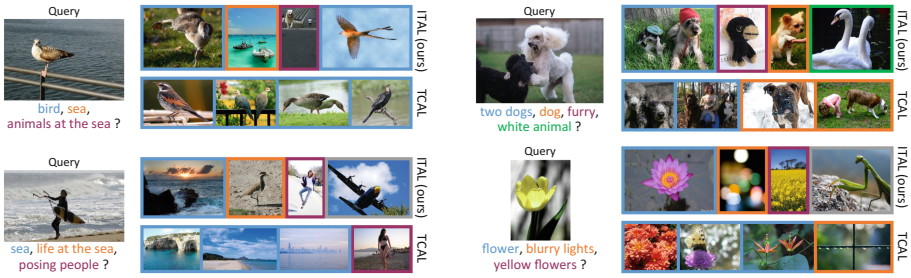


**Fig. 1.** Comparison of candidate batches selected for the second annotation round by ITAL with the selection of TCAL [5] on 4 exemplary queries from the MIRFLICKR [16] dataset. The border colors correspond to different topics that could be associated with the query. Obviously, ITAL explores much more diverse relevant topics than TCAL.

classifier to distinguish between relevant and irrelevant images. Finding such a set of most informative samples is the objective of *batch-mode active learning (BMAL)* [2,12,15,32], which has recently been explored for CBIR [3,5]. However, the performance of existing approaches usually varies substantially between datasets, which is not only observable in our experiments, but also in comparative evaluations in the existing literature (*e.g.*, [15]).

In this work, we propose *Information-Theoretic Active Learning (ITAL)*, a BMAL method for relevance feedback that does not suffer from this instability, but provides state-of-the-art performance across different datasets. This is demonstrated by a comparison of our approach with a variety of competitor methods on 5 image datasets of very different type and structure. Our method for selecting unlabeled samples for annotation by the user ($a$) implicitly maintains both diversity and informativeness of the candidate images, ($b$) employs an explicit model of the user behavior for dealing with the possibility of incorrect annotations and *unnameable instances* [19], *i.e.*, images which the user cannot classify at all, ($c$) takes the model output change caused by the expected user feedback into account, ($d$) can easily be parallelized for processing large datasets, and ($e$) works with as few as a single initial training sample.

The user model allows ITAL to compensate for unreliable users, who are likely to make mistakes or to refuse giving feedback. It acts as an implicit mechanism for controlling the trade-off between redundancy and diversity of the batch of samples selected for annotation. Because care has to be taken not only that all images in the batch of unlabeled samples selected for annotation are *informative* individually, but that they are also *diverse* compared to each other to avoid unnecessary redundant feedback. The majority of existing works on BMAL try to achieve this using a combination of several heuristics to simultaneously maximize the *diversity* within the batch and the *uncertainty* of the selected samples or their *density* in the dataset [2,3,5,12,32,33].

Our proposed ITAL method, in contrast, aims to maximize the mutual information (MI) between the expected user feedback and the relevance model. By

taking the joint distribution of the predictive relevance of the samples in the batch into account, the MI criterion implicitly maintains diversity without the need for any heuristics or manually tuned linear combinations of different criteria. Instead, our method does not only take the structure of the data and the current relevance predictions into account, but also considers the expected impact that annotating the selected samples would have on the predicted relevance after updating the model. This integration of the expected model output change (EMOC) has successfully been used for one-sample-at-a-time active learning [9], but, to the best of our knowledge, not been applied to BMAL yet.

However, computing the expected model output change requires relevance models that can be updated efficiently. In addition, both the relevance model and the active learning technique should be capable of working with as few training data as a single positive query example provided by the user. We achieve both by using a Gaussian process (GP) [26] for classification, which can be fitted to a single training sample and can be updated using a closed-form solution without the need for iterative optimization. This is in contrast to many other works on active learning, which are based on logistic regression [14,23] or support vector machines (SVMs) [2,5,12,31] as classification technique. Moreover, SVMs require a fair amount of both positive and negative initial training data for learning a robust hyperplane. Thus, such an approach is not feasible for image retrieval.

Figure 1 illustrates the advantages of our approach: While existing methods often select images similar to the query, but with high uncertainty (*e.g.*, only dogs for a dog query or birds for a bird query), ITAL additionally explores the different meanings of the query image. The query showing a bird in front of the sea could as well refer to images of the sea or to animals at the sea in general. The dog query, on the other hand, could refer to images showing two dogs, images of dogs in general, or images of white animals. Finally, the user providing the beach image as query could be interested in images of the coast, of creatures at the beach, or also just in images of people in action without necessarily being at the beach. All these various options are explored by ITAL, which actively asks the user for the feedback to resolve these ambiguities.

We will briefly review related methods in the following section and explain our ITAL method in detail in Sect. 3. The experiments mentioned above are presented in Sects. 4 and 5 concludes this paper.

## 2 Related Work

The use of active learning methods is, of course, not limited to information retrieval applications, but also evident in the scenario of manual annotation of large unlabeled datasets: One would prefer spending money and human effort on labeling the most useful samples instead of outliers. Thus, active learning has been extensively studied for several years across various application domains, including binary classification [2,8,9,15,31], multi-class classification [14,17,19,23,32], and regression [13,20,22].

With regard to batch-mode active learning (BMAL), most existing methods employ some combination of the criteria uncertainty, diversity, and density: Brinker [2] proposes to select samples close to the decision boundary, while enforcing diversity by minimizing the maximum cosine similarity of samples within the batch. Similarly, "Sampling by Uncertainty and Density (SUD)" [33] selects samples maximizing the product of entropy and average cosine similarity to the nearest neighbors and "Ranked Batch-mode Active Learning (RBMAL)" [3] constructs a batch by successively adding samples with high uncertainty and low maximum similarity to any other already selected sample. "Triple Criteria Active Learning (TCAL)" [5], on the other hand, first selects a subset of uncertain samples near the decision boundary, divides them into $k$ clusters, and chooses that sample from each cluster that has the minimum average distance to all other samples in the same cluster. Following a more complex approach, "Uncertainty Sampling with Diversity Maximization (USDM)" [32] finds a trade-off between the individual entropy of the samples in the batch and their diversity by formulating this optimization problem as a quadratic program, whose parameters to be determined are the ranking-scores of the unlabeled samples.

In addition, two works use an information-theoretic approach and are, thus, particularly similar to our method:

Guo and Greiner [14] propose to maximize the mutual information between the selected sample and the remaining unlabeled instances, given the already labeled data. They reduce this objective to the minimization of conditional entropy of the predictive label distribution of the unlabeled samples, given the existing labels and a proxy-label for the selected instance. With regard to the latter, they make an optimistic guess assuming the label which would minimize mutual information. If this guess turns out to be wrong, they fall back to uncertainty sampling for the next iteration.

Though the results obtained by this approach called MCMI[min]+MU are convincing, it is computationally demanding and not scalable to real-world scenarios, even though the authors already employed some assumptions to make it more tractable. In particular, they assume that the conditional entropy of a set of samples can be decomposed as a sum of the entropy of individual samples. However, this assumption ignores relationships between unlabeled samples and is hence not suitable for a batch-mode scenario.

In our work, we employ different approximations and Gaussian processes to enable the use of mutual information for BMAL. Using Gaussian processes instead of logistic regression or SVMs also allows us to take the impact of user feedback on the model output into account, since updating a GP does not involve iterative algorithms.

On the other hand, Li and Guo [23] employ mutual information as a measure for the information density of the unlabeled samples and combine it with the conditional entropy of their individual labels as uncertainty measure. Similar to our approach, they use a GP to estimate the mutual information, but then employ logistic regression for the actual classification. Furthermore, their method

cannot be applied to a batch-mode scenario and does not scale to large datasets, so that they need to randomly sub-sample the unlabeled data.

Our ITAL method, in contrast, forms a consistent framework, provides a batch-mode, considers the impact of annotations on the model output, and relies solely on the solid theoretical basis of mutual information to implicitly account for uncertainty, density, and diversity.

## 3 Information-Theoretic Active Learning

We begin with a very general description of the idea behind our ITAL approach and then describe its individual components in more detail. The implementations of ITAL and the competing methods described in Sect. 4.2 are available as open source at https://github.com/cvjena/ITAL/.

### 3.1 Idea and Ideal Objective

Let $\mathfrak{U} = \{x_1, \ldots, x_m\}$ be a set of features of unlabeled samples and $\mathfrak{L} = \{(x_{m+1}, y_{m+1}), \ldots, (x_{m+\ell}, y_{m+\ell})\}$ be a set of features of labeled samples $x_i \in \mathbb{R}^d$ and their labels $y_i \in \{-1, 1\}$. The label 1 is assigned to relevant and $-1$ to irrelevant samples. $\mathfrak{X} = \{x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+\ell}\}$ denotes the set of all $n = m+\ell$ samples. In the scenario of content-based image retrieval, $\mathfrak{L}$ usually consists initially of the features of a single relevant sample: the query image provided by the user. However, queries consisting of multiple and even negative examples are possible as well.

Intuitively, we want to ask the user for relevance feedback for a batch $u \subseteq \mathfrak{U}$ of $k = |u|$ unlabeled samples, whose feedback we expect to be most helpful for classifying the remaining unlabeled instances, *i.e.*, assessing their relevance to the user. Note that these chosen samples are also often referred to as "queries" in the active learning literature. To avoid confusion caused by this conflicting terminology, we will refer to the query image as "query" and to the unlabeled samples chosen for annotation as "candidates".

Ideally, the most informative batch $u$ of candidates can be found by maximizing the conditional mutual information $\mathfrak{I}(R, F \mid u)$ between the relevance $R$ of both labeled and unlabeled samples, which is a multivariate random variable over the space $\{-1, 1\}^n$ of relevance labels, and the user feedback $F$, being a multivariate random variable over the space $\{-1, 0, 1\}^n$ of possible feedbacks. A feedback of 0 represents the case that the user has not given any feedback for a certain candidate. This option is a special feature of our approach, which allows the user to omit candidates that cannot be labeled reliably.

Since the size $n$ of the dataset can be huge, this problem is not solvable in practice. We will show later on how it can be approximated to become tractable. But for now, let us consider the ideal optimization objective:

$$u = \operatorname*{argmax}_{\hat{u} \subseteq \mathfrak{U}} \mathfrak{I}(R, F \mid \hat{u}). \tag{1}$$

Writing the mutual information (MI) in terms of entropy reveals the relationship of our approach to uncertainty sampling by maximizing the entropy $H(R \mid u)$ of the candidate batch [23,32,33] or minimizing the conditional entropy $H(R \mid F, u)$ [14]:

$$\mathfrak{I}(R, F \mid u) = H(R \mid u) - H(R \mid F, u). \tag{2}$$

In contrast to pure uncertainty maximization, we also take into account how the relevance model is expected to change after having obtained the feedback from the user: To select those samples whose annotation would reduce uncertainty the most, we maximize the difference between the uncertainty $H(R \mid u)$ according to the current relevance model and the uncertainty $H(R \mid F, u)$ after an update of the model with the expected user feedback.

In contrast to existing works [23,32,33], we do not assume $H(R \mid u)$ to be equal to the sum of individual entropies of the samples, but use their joint distribution to compute the entropy. Thus, maximizing $H(R \mid u)$ is also a possible novel approach, which will be compared to maximization of MI in the experiments (cf. Sect. 4).

In more detail, the mutual information can be decomposed into the following components (a derivation is provided in the supplemental material):

$$\mathfrak{I}(R, F \mid u) = \sum_{\substack{r \in \{-1,1\}^n \\ f \in \{-1,0,1\}^n}} \Bigg[ P(R = r \mid u) \cdot P(F = f \mid R = r, u) \cdot \log \left( \frac{P(R = r \mid F = f, u)}{P(R = r \mid u)} \right) \Bigg]. \tag{3}$$

The individual terms can be interpreted as follows:

- $P(R \mid u) = P(R)$ is the probability of a certain relevance configuration according to the current relevance model.
- $P(R \mid F, u)$ is the probability of a certain relevance configuration after updating the relevance model according to the user feedback. Thus, $\frac{P(R|F,u)}{P(R|u)}$ quantifies the model output change. Compared to other active learning techniques taking model output change into account, *e.g.*, EMOC [9], we do not just consider the change of the predictive mean, but of the joint relevance probability and hence take all parameters of the distributions into account.
- $P(F \mid R, u)$ is the probability of observing a certain feedback, given that the true relevance of the samples is already known. One might assume that the feedback will always be equal to the true relevance. However, users are not perfect and tend to make mistakes or prefer to avoid difficult samples (so-called *unnameable instances* [19]). Thus, this term corresponds to a *user model* predicting the behavior of the user.

In the following Subsects. 3.2 and 3.3, we will first describe our relevance and user model, respectively. Thereafter, we introduce assumptions to approximate Eq. (1) in the Subsects. 3.4 and 3.5, since finding an optimal set of candidates would require exponential computational effort.

### 3.2 Relevance Model

We fit a probabilistic regression to the training data $\mathfrak{L}$, *i.e.*, the query images and the images annotated so far, using a Gaussian process [26, Chap. 2] with an RBF kernel given by the kernel matrix $K \in \mathbb{R}^{n \times n}$ over the entire dataset $\mathfrak{X}$:

$$K_{ij} = \sigma_{\text{var}}^2 \cdot \exp\left(-\frac{\|x_i - x_j\|^2}{2 \cdot \sigma_{\text{ls}}^2}\right) + \sigma_{\text{noise}}^2 \cdot \delta_{ij} \ , \tag{4}$$

where $\delta_{ij}$ is the Kronecker delta function with $\delta_{ij} = 1 \leftrightarrow i = j$ and zero otherwise, and $\sigma_{\text{var}}$, $\sigma_{\text{ls}}$, and $\sigma_{\text{noise}}$ are the hyper-parameters of the kernel.

The computation of the kernel matrix can be performed off-line in advance and does hence not contribute to the run-time of our active learning method. However, if time and memory required for computing and storing the kernel are an issue, alternative kernels for efficient large-scale Gaussian process inference [27] can be used.

The prediction of the Gaussian process for any finite set of $k$ samples consists of a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ over continuous values $\hat{y} \in \mathbb{R}^k$, where $\mu \in \mathbb{R}^k$ is a vector of predictive means of the samples and $\Sigma \in \mathbb{R}^{k \times k}$ is their predictive joint covariance matrix. Let $p(\hat{y}) = \mathcal{N}(\hat{y} \mid \mu, \Sigma)$ denote the probability density function of such a distribution.

We use this probabilistic label regression for binary classification by considering samples $x_i$ with $\hat{y}_i > 0$ as relevant. The probability of a given relevance configuration $r \in \{-1, 1\}^n$ for the samples in $\mathfrak{X}$ is hence given by

$$P(R = r) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(y_1, \ldots, y_n) \, \mathrm{d}y_n \cdots \mathrm{d}y_1, \tag{5}$$

with

$$a_i = \begin{cases} 0, & r_i = 1, \\ -\infty, & r_i = -1, \end{cases} \qquad b_i = \begin{cases} \infty, & r_i = 1, \\ 0, & r_i = -1, \end{cases} \tag{6}$$

for $i = 1, \ldots, n$. This is a multivariate normal distribution function, which can be efficiently approximated using numerical methods [11].

The posterior probability $P(R \mid F, u)$ can be obtained in the same way after updating the GP with the expected feedback. For such an update, it is not necessary to re-fit the GP to the extended training data from scratch, which would involve an expensive inversion of the kernel matrix of the training data. Instead, efficient updates of the inverse of the kernel matrix [24] can be performed to obtain updated predictions at a low cost. This is an advantage of our GP-based approach compared with other methods relying on logistic regression (*e.g.*, [14,23]), which requires expensive iterative optimization for updating the model.

### 3.3 User Model

We employ a simple, but plausible user model for $P(F \mid R, u)$, which comes along with a slight simplification of the optimization objective in Eq. (1): First

of all, we assume that if we already know the true relevance $r = [r_1, \ldots, r_n]^\top$ of all samples, the feedback $f_i$ given by the user for an individual sample $x_i$ is conditionally independent from the feedback provided for the other samples. More formally:

$$P(F = f \mid R = r, u) = \prod_{i=1}^{n} P(F_i = f_i \mid R_i = r_i, u). \tag{7}$$

Clearly, if a sample $x_i$ has not been included in the candidate batch $u$, the user cannot give feedback for that sample, *i.e.*, $x_i \notin u \rightarrow f_i = 0$. Furthermore, we assume that the user will, on average, label a fraction $p_{\text{label}}$ of the candidate samples. For each labeled sample, the user is assumed to provide an incorrect label with probability $p_{\text{mistake}}$. In summary, this user model can be formalized as follows:

$$P(F_i = f_i \mid R_i = r_i, u) = \begin{cases} 0, & x_i \notin u \wedge f_i \neq 0, \\ 1, & x_i \notin u \wedge f_i = 0, \\ 1 - p_{\text{label}}, & x_i \in u \wedge f_i = 0, \\ p_{\text{label}} \cdot p_{\text{mistake}}, & x_i \in u \wedge f_i \neq r_i, \\ p_{\text{label}} \cdot (1 - p_{\text{mistake}}), & x_i \in u \wedge f_i = r_i. \end{cases} \tag{8}$$

The fact that $\left(\exists_{i \in \{1,\ldots,n\}} : x_i \notin u \wedge f_i \neq 0\right) \rightarrow P(F = f \mid R = r, u) = 0$ allows us to adjust the sum in Eq. (3) to run over only $3^k$ instead of $3^n$ possible feedback vectors, where $k \ll n$ is the batch size and independent from the size $n$ of the dataset. This is not an approximation, but an advantage of our user model, that decreases the complexity of the problem significantly.

Modeling the user behavior can enable the active learning technique to find a trade-off between learning as fast as possible by asking for feedback for very diverse samples and improving confidence regarding existing knowledge by selecting not extremely diverse, but slightly redundant samples. The latter can be useful for difficult datasets or tasks, where the user is likely to make mistakes or to refuse to give feedback for a significant number of candidates. Nevertheless, the assumption of a perfect user, who labels all samples in the batch and never fails, is an interesting special case since it results in a simplification of the MI term from Eq. (3) and can reduce computation time drastically:

$$(p_{\text{label}} = 1 \wedge p_{\text{mistake}} = 0) \rightarrow$$
$$\mathfrak{I}(R, F \mid u) = \sum_{r \in \{-1,1\}^n} \left[ P(R = r \mid u) \cdot \log \left( \frac{P(R = r \mid F = r, u)}{P(R = r \mid u)} \right) \right]. \tag{9}$$

### 3.4 Approximation of Mutual Information

Even with the perfect user assumption, evaluating Eq. (9) still involves a summation over $2^n$ possible relevance configurations, which does not scale to large datasets. To overcome this issue, we employ an approximation based on the

assumption, that the probability of observing a certain relevance configuration depends only on the samples in the current candidate batch:

$$P(R = r \mid u = \{x_{i_1}, \ldots, x_{i_k}\}) \;=\; P(R_{i_1} = r_{i_1}, \ldots, R_{i_k} = r_{i_k}). \tag{10}$$

This means that we indeed condition $P(R \mid u)$ on the current batch $u$, though actually $P(R \mid u) = P(R)$ holds for the original problem formulation.

This assumption allows us to restrict the sum in Eq. (3)to $2^k$ instead of $2^n$ possible relevance configurations, leading to the approximate mutual information

$$\tilde{\mathfrak{I}}(R, F \mid u) = 2^{n-k} \sum_{\substack{r \in \{-1,1\}^k \\ f \in \{-1,0,1\}^k}} \Bigg[ P(R_u = r) \cdot P(F_u = f \mid R_u = r) \cdot \log \left( \frac{P(R_u = r \mid F_u = f)}{P(R_u = r)} \right) \Bigg] \tag{11}$$

with $u = \{x_{i_1}, \ldots, x_{i_k}\}$ and, by an abuse of notation, $R_u = [R_{i_1}, \ldots, R_{i_k}]^\top$ (analogously for $F_u$). The number of involved summands now does not depend on the size of the dataset anymore, but only on the number $k$ of candidates chosen at each round for annotation.

On the other hand, this assumption also restricts the estimation of the expected model output change, expressed by the term $P(R_u \mid F_u)/P(R_u)$, to the current batch, which is probably the most severe drawback of this approximation. However, estimating the model output change for the entire dataset would be too expensive and the experiments in Sect. 4.4 show that our approach can still benefit from the expected model output change. Future work might explore the option of taking a tractable subset of context into account additionally.

### 3.5 Greedy Batch Construction

Although the computational effort required to calculate the approximate MI given in Eq. (11) is independent from the size of the dataset, finding the exact solution to the optimization problem from Eq. (1) would still require computing the MI for all possible $2^m$ candidate batches $u \subseteq \mathfrak{U}$ of unlabeled samples. Since we do not want to confront the user with an unlimited number of candidates anyway, we set the batch size to a fixed number $k$, which leaves us with a number of $\binom{m}{k}$ possible candidate batches.

Assessing them all would involve a polynomial number of subsets and is, thus, still too time-consuming in general. On first sight, one might think that this problem can be solved more efficiently using dynamic programming, but this is unfortunately not an option since the MI is not adequately separable.

Thus, we follow a linear-time greedy approach to approximate the optimal batch by successively adding samples to the batch [13], taking their relationship to already selected samples into account: We first select the sample $x_{i_1}$ with

maximum $\tilde{\mathfrak{I}}(R, F \mid u = \{x_{i_1}\})$. The second sample $x_{i_2}$ is chosen to maximize MI together with $x_{i_1}$. This continues until the batch contains $k$ samples.

At each iteration, the unlabeled samples eligible for being added to the current batch can be treated completely independently from each other, allowing for straightforward parallelization.

## 4 Experiments

We demonstrate the performance of our ITAL approach on five image datasets of varying type and structure, described in Sect. 4.1, and compare it against several existing active learning techniques briefly explained in Sect. 4.2. The quantitative results in Sect. 4.4 show that ITAL is the only method that can provide state-of-the-art performance across all datasets. Qualitative examples are shown in Fig. 1 and failure cases are included in the supplemental material.

### 4.1 Datasets

All datasets used in our experiments consist of multiple classes and are divided into a training and a test set. We define image retrieval tasks for each dataset as follows: Pick a single random instance from the training set of a certain class as query image and consider all other images belonging to that class as relevant, while instances from other classes are irrelevant. Batch-mode active learning is performed for 10 successive rounds with a batch-size of $k = 4$ candidates per round and retrieval performance is evaluated after each round by means of average precision on the test set. This process is repeated multiple times with different random queries for each class and we report the mean average precision (mAP) over all repetitions.

Note that our goal is not to achieve state-of-the-art performance in terms of classification accuracy, but with respect to the active learning objective, *i.e.*, obtaining better performance after fewer feedback rounds.

The smallest dataset used is the **Butterflies** dataset [18], comprising 1,500 images of 5 different species of butterflies captured over a period of 100 years. We use the CNN features provided by the authors and, following their advice, reduce them to 50 dimensions using PCA. A random stratified subset of 20% of the dataset is used as test set.

Second, we use the **USPS** dataset [10] consisting of 9,300 gray-scale images of handwritten digits, scanned from envelopes by the U.S. Postal Service. The number of images per class is very unevenly distributed. All images have a size of $16 \times 16$ pixels and are used without further feature extraction as 256-dimensional feature vectors. We use the canonical training-test split provided with the dataset.

As a more real-word use-case, we perform evaluation on the **13 Natural Scenes** dataset [7] and the **MIRFLICKR-25K** dataset [16]. The former consists of more than 3,400 images from 13 categories of natural scenes such as forests, streets, mountains, coasts, offices, or kitchens. The latter comprises

25,000 images, each assigned to a subset of 14 very general topics such as "clouds", "tree", "people", "portrait" etc. Thus, query images can belong to multiple categories and will be ambiguous. Asking the user the right questions is hence of great importance. There are also "wide-sense annotations" assigning images to categories if they could be related to a small degree. If a candidate image is annotated in this way, we consider it as unnameable during our simulation. For both datasets, we extracted image features from the first fully-connected layer of the VGG-16 convolutional neural network [29], pre-trained for classification on ImageNet, and reduce their dimensionality to 512 using PCA. Experiments in the supplemental material show that the relative performance of the different methods is not very sensitive w.r.t. the dimensionality of the features. 25% of the natural scenes and 20% of the MIRFLICKR dataset are used as test set.

Finally, we derive further challenging image retrieval tasks from the **ImageNet Large Scale Visual Recognition Challenge (ILSVRC)** [28], which comprises more than 1,2 million images from 1,000 classes. Following Freytag et al. [9], we obtain binary classification tasks by randomly choosing a single positive and 19 negative classes. This is repeated 25 times and 10 random queries are chosen for each task, leading to a total of 250 image retrieval scenarios. We use the bag-of-words (BoW) features provided with ImageNet instead of CNN features, mainly for two reasons: First, the BoW features are public and hence facilitate reproduction. Second, most neural networks are pre-trained on ImageNet, which could bias the evaluation.

The number of random repetitions per class for the natural scenes and the MIRFLICKR dataset has been set to 10 as well, while we use 25 queries per class for USPS and 50 for the butterflies dataset.

The features of all datasets were scaled to be in [0, 1].

### 4.2 Competitor Methods

We compare ITAL with a variety of baselines and competing methods, including **SUD** [33], **TCAL** [5], **RBMAL** [3], and the method of Brinker [2] referred to as "**border_div**" in the following. All these native BMAL methods have been described in Sect. 2.

In addition, we evaluate the following successful one-by-one active learning techniques in the BMAL scenario by selecting the $k$ samples with the highest selection scores: Uncertainty sampling for SVM active learning by choosing samples close to the decision boundary (**border**) [31], uncertainty sampling for Gaussian processes (**unc**) [21], where uncertainty is defined as the ratio between absolute predictive mean and predictive standard deviation, and sample selection by maximizing the expected model output change (**EMOC**) [9].

All methods have to compete against the baselines of **random** selection, selecting the **topscoring** samples with maximum predictive mean, resembling the standard retrieval scenario [1], and variance sampling (**var**) by maximizing the difference of the sum of variances and the sum of covariances in the batch.

Finally, we also investigate maximizing the **joint entropy** $H(R_{i_1}, \ldots, R_{i_k}) = -\sum_{r \in \{-1,1\}^k} P(R_u = r) \cdot \log(P(R_u = r))$ of candidate batch $u = \{x_{i_1}, \ldots, x_{i_k}\}$. Being a component of our ITAL method, this is also a novel approach, but lacks the model output change term and the user model (cf. Eq. (2)).

We also tried applying **USDM** [32], **MCMI[min]** [14] and **AdaptAL** [23], especially since the latter two also maximize a mutual information criterion. However, all these methods scale so badly to datasets of realistic size, that they could not be applied in practice. AdaptAL, for example, would require 14 h for composing a single batch on MIRFLICKR, which is clearly intractable. Our ITAL method, in contrast, can handle this dataset with less than a minute per batch. These three competitors could, thus, only applied to USPS, MIRFLICKR, and ImageNet by randomly sub-sampling 1000 candidates to choose from, as suggested by Li et al. [23]. This usually leads to a degradation of performance, as can be seen from the results reported in the supplemental material.

### 4.3 Hyper-parameters

The hyper-parameters of the RBF kernel, *i.e.*, $\sigma_{\text{ls}}$, $\sigma_{\text{var}}$, and $\sigma_{\text{noise}}$ (cf. Sect. 3.2), potentially have a large impact on the performance of the active learning methods. However, the overall goal is to eventually obtain a classifier that performs as well as possible. Therefore, we determine the optimal kernel hyper-parameters for each dataset using tenfold cross-validation on the training set and alternating optimization to maximize mean average precision. This optimization aims only for good classification performance independent of the active learning method being used and the same hyper-parameters are used for all methods.

With regard to the hyper-parameters of the user model used by ITAL, we employ the perfect user assumption for being comparable to competing methods that do not model the user. An experiment evaluating the effect of different user model parameters is presented in Sect. 4.5.

In case that other methods have further hyper-parameters, we use the default values provided by their authors.

### 4.4 Results

Figure 2 depicts the average precision obtained on average after 10 feedback rounds using the different BMAL methods. On the Butterflies dataset, ITAL obtains perfect performance after the least number of feedback rounds. TCAL and border_div perform similar to ITAL on USPS, but ITAL learns faster at the beginning, which is important in interactive image retrieval scenarios. While sampling candidates based on batch entropy behaves almost identical to ITAL on Butterflies, USPS, and MIRFLICKR, it is slightly superior on the Natural Scenes dataset, but fails to improve after more than 4 rounds of feedback on the ImageNet benchmark, where ITAL is clearly superior to all competitor methods. This indicates that taking the effect of the expected user feedback on the model output change into account is of great benefit for datasets as diverse as ImageNet.

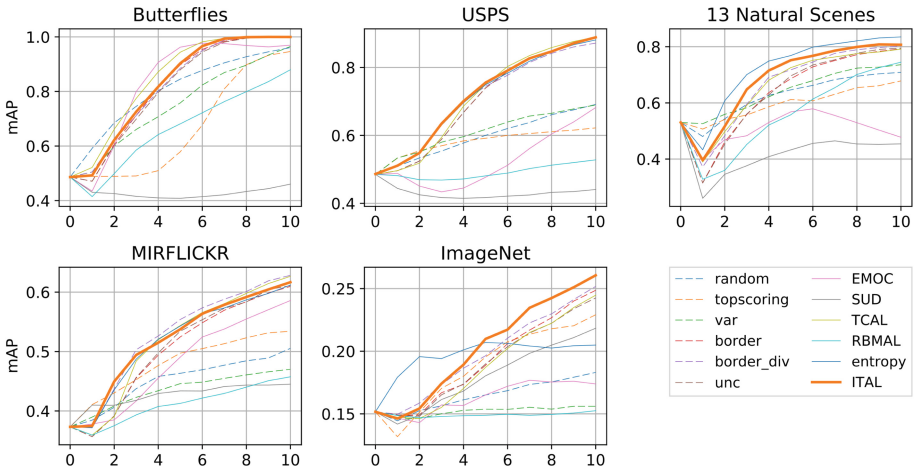


**Fig. 2.** Comparison of retrieval performance after different numbers of feedback rounds for various active learning methods. The thick, orange line corresponds to our proposed method. Figure is best viewed in color. (Color figure online)

Since it is often desirable to compare different methods by means of a single value, we report the area under the learning curves (AULC) in 1, divided by the number of feedback rounds so that the best possible value is always 1.0. In all cases, our method is among the top performers, achieving the best of all results in 3 out of 5 cases. The improvement over the second-best method on 13 Natural Scenes and ImageNet is significant on a level of <1% and on a level of 7% on USPS, according to Student's paired t-test.

The performance of the competing methods, on the other hand, varies significantly across datasets. ITAL, in contrast, is not affected by this issue and provides state-of-the-art performance independent from the characteristics of the data. To make this more visible, we construct a ranking of the tested methods for each dataset and report the average rank in 1 as well. ITAL achieves the best average rank and can thus be considered most universally applicable.

This is of high importance because, in an active learning scenario, labeled data is usually not available before performing the active learning. Thus, adaptation of the AL method or selection of a suitable one depending on the dataset is difficult. A widely applicable method such as ITAL is hence very desirable.

### 4.5 Effect of the User Model

To evaluate the effect of the user model integrated into ITAL, we simulated several types of users behaviors on the 13 Natural Scenes dataset: ($a$) an aggressive user annotating all images but assigning a wrong label in 50% of the cases, ($b$) a conservative user who always provides correct labels but only annotates 25% of the images on average, and ($c$) a blend of both, labeling 50% of the candidate images on average and having a 25% chance of making an incorrect annotation.

**Table 1.** Area under the learning curves from Fig. 2. The numbers in parentheses indicate the position in the ranking of all methods. The best value in each column is set in bold face, while the second-best and third-best values are underlined.

| Method | Area under Learning Curve (AULC) | | | | | Avg. Rank |
|---|---|---|---|---|---|---|
| | Butterflies | USPS | Nat. Scenes | MIRFLICKR | ImageNet | |
| random | 0.7316 (8) | 0.5416 (8) | 0.5687 (8) | 0.4099 (9) | 0.1494 (9) | 8.4 |
| topscoring | 0.5991 (11) | 0.5289 (9) | 0.5419 (9) | 0.4358 (7) | 0.1708 (7) | 8.6 |
| var | 0.6800 (9) | 0.5550 (7) | 0.5831 (5) | 0.3957 (10) | 0.1383 (11) | 8.4 |
| border | 0.7434 (6) | 0.6393 (5) | 0.5775 (7) | 0.4559 (6) | 0.1743 (4) | 5.6 |
| border_div | 0.7456 (5) | <u>0.6465</u> (3) | <u>0.6031</u> (3) | **0.4795** (1) | <u>0.1791</u> (3) | <u>3.0</u> |
| unc | 0.7373 (7) | 0.6391 (6) | 0.5793 (6) | 0.4585 (5) | 0.1725 (5) | 5.8 |
| EMOC | <u>0.7561</u> (2) | 0.4723 (10) | 0.4654 (11) | 0.4357 (8) | 0.1483 (10) | 8.2 |
| SUD | 0.3887 (12) | 0.3903 (12) | 0.3766 (12) | 0.3883 (11) | 0.1626 (8) | 11.0 |
| TCAL | **0.7720** (1) | 0.6459 (4) | 0.6016 (4) | 0.4688 (4) | 0.1708 (6) | 3.8 |
| RBMAL | 0.6023 (10) | 0.4457 (11) | 0.5046 (10) | 0.3732 (12) | 0.1356 (12) | 11.0 |
| entropy (ours) | <u>0.7512</u> (3) | <u>0.6484</u> (2) | **0.6547** (1) | <u>0.4703</u> (3) | <u>0.1793</u> (2) | <u>2.2</u> |
| ITAL (ours) | 0.7511 (4) | **0.6522** (1) | <u>0.6233</u> (2) | <u>0.4731</u> (2) | **0.1841** (1) | **2.0** |

The same active learning methods as in the previous sections are applied and the parameters $p_{\text{label}}$ and $p_{\text{mistake}}$ of ITAL are set accordingly.

The results presented in the supplemental material show that the user model helps ITAL to make faster improvements than with the perfect user model. A possible reason for this effect is that the parameters of the user model control an implicit trade-off between diversity and redundancy used by ITAL: For an imperfect user, selecting samples for annotation more redundantly can help to reduce the impact of wrong annotations. However, ITAL performs reasonably well even with the perfect user assumption. This could hence be used to speed-up ITAL noticeably with only a minor loss of performance.

## 5 Conclusions

We have proposed information-theoretic active learning (ITAL), a novel batch-mode active learning technique for binary classification, and applied it successfully to image retrieval with relevance feedback. Based on the idea of finding a subset of unlabeled samples that maximizes the mutual information between the relevance model and the expected user feedback, we propose suitable models and approximations to make this NP-hard problem tractable in practice. ITAL does not need to rely on manually tuned combinations of different heuristics, as many other works on batch-mode active learning do, but implicitly trades off uncertainty against diversity by taking the joint relevance distribution of the instances in the dataset into account.

Our method also features an explicit user model that enables it to deal with unnameable instances and the possibility of incorrect annotations. This has been demonstrated to be beneficial in the case of unreliable users.

We evaluated our method on five image datasets and found that it provides state-of-the-art performance across datasets, while many competitors perform well on certain datasets only. Moreover, ITAL outperforms existing techniques on the ImageNet dataset, which we attribute to its ability of taking the effect of the expected user feedback on the model output change into account.

**Acknowledgements.** This work was supported by the German Research Foundation as part of the priority programme "Volunteered Geographic Information: Interpretation, Visualisation and Social Computing" (SPP 1894, contract number DE 735/11-1).

## References

1. Ayache, S., Quénot, G.: Evaluation of active learning strategies for video indexing. Sig. Process.: Image Commun. **22**(7), 692–704 (2007)
2. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: International Conference on Machine Learning (ICML), pp. 59–66 (2003)
3. Cardoso, T.N., Silva, R.M., Canuto, S., Moro, M.M., Gonçalves, M.A.: Ranked batch-mode active learning. Inf. Sci. **379**, 313–337 (2017)
4. Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V., Yianilos, P.N.: The Bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. IEEE Trans. Image Process. **9**(1), 20–37 (2000)
5. Demir, B., Bruzzone, L.: A novel active learning method in relevance feedback for content-based remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. **53**(5), 2323–2334 (2015)
6. Deselaers, T., Paredes, R., Vidal, E., Ney, H.: Learning weighted distances for relevance feedback in image retrieval. In: International Conference on Pattern Recognition (ICPR), pp. 1–4. IEEE (2008)
7. Fei-Fei, L., Perona, P.: A Bayesian hierarchical model for learning natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 524–531. IEEE (2005)
8. Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Labeling examples that matter: relevance-based active learning with Gaussian processes. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 282–291. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40602-7_31
9. Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: active learning with expected model output changes. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 562–577. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_37
10. Friedman, J., Hastie, T., Tibshirani, R.: Example: ZIP code data (Ch. 11.7). In: The Elements of Statistical Learning. Springer Series in Statistics, New York (2001)
11. Genz, A.: Numerical computation of multivariate normal probabilities. J. Comput. Graph. Stat. **1**(2), 141–149 (1992)
12. Giang, N.T., Tao, N.Q., Dung, N.D., The, N.T.: Batch mode active learning for interactive image retrieval. In: 2014 IEEE International Symposium on Multimedia (ISM), pp. 28–31. IEEE (2014)

13. Guestrin, C., Krause, A., Singh, A.P.: Near-optimal sensor placements in Gaussian processes. In: International Conference on Machine Learning (ICML), pp. 265–272. ACM (2005)
14. Guo, Y., Greiner, R.: Optimistic active-learning using mutual information. In: IJCAI, vol. 7, pp. 823–829 (2007)
15. Guo, Y., Schuurmans, D.: Discriminative batch mode active learning. In: Advances in Neural Information Processing Systems (NIPS), pp. 593–600 (2008)
16. Huiskes, M.J., Lew, M.S.: The MIR flickr retrieval evaluation. In: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval, MIR 2008. ACM, New York (2008)
17. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 762–769. IEEE (2009)
18. Johns, E., Mac Aodha, O., Brostow, G.J.: Becoming the expert - interactive multi-class machine teaching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2616–2624. IEEE (2015)
19. Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4343–4352 (2015)
20. Käding, C., Rodner, E., Freytag, A., Mothes, O., Barz, B., Denzler, J.: Active learning for regression tasks with expected model output changes. In: British Machine Vision Conference (BMVC) (2018)
21. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with Gaussian processes for object categorization. In: IEEE International Conference on Computer Vision (ICCV), pp. 1–8. IEEE (2007)
22. Krause, A., Guestrin, C.: Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach. In: International Conference on Machine Learning (ICML), pp. 449–456. ACM (2007)
23. Li, X., Guo, Y.: Adaptive active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 859–866 (2013)
24. Lütz, A., Rodner, E., Denzler, J.: I want to know more–efficient multi-class incremental learning using Gaussian processes. Pattern Recogn. Image Anal. **23**(3), 402–407 (2013)
25. Niblack, C.W., et al.: QBIC project: querying images by content, using color, texture, and shape. In: Storage and Retrieval for Image and Video Databases, vol. 1908, pp. 173–188. International Society for Optics and Photonics (1993)
26. Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning, vol. 1. MIT Press, Cambridge (2006)
27. Rodner, E., Freytag, A., Bodesheim, P., Fröhlich, B., Denzler, J.: Large-scale Gaussian process inference with generalized histogram intersection kernels for visual recognition tasks. Int. J. Comput. Vis. **121**(2), 253–280 (2017)
28. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) **115**(3), 211–252 (2015)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
30. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **22**(12), 1349–1380 (2000)
31. Tong, S., Chang, E.: Support vector machine active learning for image retrieval. In: ACM International Conference on Multimedia, pp. 107–118. ACM (2001)

32. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. Int. J. Comput. Vis. **113**(2), 113–127 (2015)
33. Zhu, J., Wang, H., Yao, T., Tsou, B.K.: Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In: International Conference on Computational Linguistics, vol. 1 (2008)