SALIC: Social Active Learning for Image Classification

Elisavet Chatzilari, Spiros Nikolopoulos, Yiannis Kompatsiaris, *Senior Member, IEEE,*
and Josef Kittler, *Life Member, IEEE*

***Abstract*—In this paper we present SALIC, an active learning method that is placed in the context of social networks and focuses on selecting the samples that are most appropriate to expand the training set of a binary classifier. The process of active learning can be fully automated in this social context by replacing the human oracle with the user tagged images obtained from social networks. However, the noisy nature of user-contributed tags adds further complexity to the problem of sample selection since, apart from their informativeness (i.e. how much they are expected to inform the classifier if we knew their label), our confidence about their actual content should also be maximized (i.e. how certain the oracle is on its decision about the contents of an image). The main contribution of this work is in proposing a probabilistic approach for jointly maximizing the two aforementioned quantities with a view to automate the process of active learning. Based on this approach the training set is expanded with samples that maximize the joint probability of selecting a sample given its informativeness and our confidence for its true content. In the examined noisy context, the oracle's confidence is necessary to provide a contextual-based indication of the images' true contents, while the samples' informativeness is required to reduce the computational complexity and minimize the mistakes of the unreliable oracle. We prove the validity and superiority of SALIC over various baselines and state-of-the-art methods experimentally. In addition, we show that SALIC allows us to select training data as effectively as typical active learning, without the cost of manual annotation. Finally, we argue that the speed-up achieved when learning actively in this social context (where labels can be obtained without the cost of human annotation) is necessary to cope with the continuously growing requirements of large scale applications. In this respect, we prove experimentally that SALIC requires 20 times less training data in order to reach the exactly same performance as a straightforward informativeness-agnostic learning approach.**

***Index Terms*—active learning, large scale, user tagged images, social context, image classification, multi-modal fusion.**

I. Introduction

It is commonly accepted that classification models become more robust when generated by high volumes of training data. However, the need for manually labelled training corpus creates an undesirable bottleneck for large-scale classification problems. In an effort to minimize the labelling effort, active learning [1] trains an initial classifier with a very small set of labelled examples and expands the training set by selectively sampling new examples from a much larger set of unlabelled examples (also known as pool of candidates). These examples are selected based on their *informativeness*, i.e. how much they are expected to improve the classifier's performance. They are found in the uncertainty areas of the classifier and, in a typical case, are annotated upon request by an errorless oracle.

On the other hand, the widespread use of on-line social networks has made available large amounts of user tagged images that can be obtained at almost no cost and offer more information than their mere visual content (e.g. tags). If we could leverage these tags to become indicators of the images' actual content, we could potentially remove the need for a human annotator and automate the whole active learning process. However, in this case, where the labels are leveraged from the freely available user tags, actively selecting new samples might seem superfluous, since there is no labelling cost to minimize. Indeed, we could simply just add all the images in the pool instead of actively selecting new ones. The question we should pose though is, *how many more images will we need to reach the same performance?* The computational overload (i.e. storage, memory and processing time) of dealing with training sets several times larger might not seem important in the scale of a few concepts and a few thousand images. However, as we have witnessed in the past years, large scale image benchmarks gain an order of scale almost every year (from the $\sim 10k$ images of the first Pascal-VOC competition at 2007 [2], to the $\sim 200k$ images of NUS-WIDE [3], to the 1 million images of MIRFLICKR [4], to the 14 million images of ImageNet [5] and to the 100 million images of Yahoo [6] nowadays). The situation is similar for the number of concepts, although the growth is not as steep as in the case of images (approximately from 20 concepts at 2007 to 20000 nowadays). By incorporating active learning in the selection process, a smaller part of the pool dataset, which is deemed as informative, is required to be included in the training set. In this way, we only need to add samples that are expected to have a greater impact compared to the straightforward brute-force methodology of adding all images from the pool. This scales down the computational complexity of the proposed method, since it drastically reduces the size of the training set. Besides the gain in computational complexity, incorporating active learning requires to *annotate* (using the tags) only the small *informative* part of the pool dataset, which saves us from the possible mistakes in the annotation process, considering the noisy nature of tagging information.

However, even though tags can be obtained freely, they cannot be considered as accurate labels. While simple string matching could be used to annotate the images, it has been

Elisavet Chatzilari, Spiros Nikolopoulos and Yiannis Kompatsiaris are with the Information Technologies Institute, Centre for Research & Technology Hellas, Thessaloniki, Greece e-mail: ({ehatzi,nikolopo,ikom}@iti.gr).

Josef Kittler is with Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK.

Manuscript received November xx, 2015.

proven that the accuracy of the tags is rather low [7]. This calls for a more sophisticated way to define a measurement that indicates the true contents of the images. Essentially, if we consider the web users as oracles, this measurement can be considered as the *oracle's confidence*, which gauges how probable is an image to depict a specific concept. The novelty of SALIC is, in contrast to what has been considered so far in active learning, a sample selection strategy that maximizes not only the informativeness of the selected samples but also the oracle's confidence about their actual content. In order to achieve this goal we formulate the selection process as a joint optimization problem that maximizes a function conditioned on the samples informativeness and the oracle's confidence. However, since this function cannot be estimated analytically, we propose a probabilistic approximation. Towards quantifying this probability, in this work, we approximate the samples' informativeness by minimizing their distance from the separating hyperplane of the classification model, which is known to be an effective method for finding informative samples [8]. In order to measure the oracle's confidence, we propose the utilization of the popular bag of words approach [9]. The reason for choosing this approach is its ability to decide about the content of an image based on a set of tags, thus capturing important contextual information and reducing the effect of erroneously provided, ambiguous and misleading tags. Joint maximization is then accomplished by ranking the samples based on the probability of a sample being selected given the two aforementioned quantities (see Fig. 1). This probability indicates the benefit that our system is expected to gain if the examined sample is selected and added to the training set.

This work builds on the approach presented in [10], which proposed adding only positive samples in batch mode to training sets that were formed by an already significant number of labelled examples. In this work we introduce a different sample selection approach that incorporates the following novel features compared to [10]. The samples are added to the training set in an iterative manner rather than in batch mode. Negative examples are also included in the selected samples making sure to alleviate the effect of the class imbalance problem. Furthermore, our evaluation study has been extended along the following axes. First, next to the enhanced quantitative evaluation we also provide an auxiliary qualitative evaluation that allows us to grasp intuitively the pros and cons of each sample selection approach. Second, experiments are performed in two feature spaces of different dimensionality and discrimination ability so as to assess the impact of these aspects in the effectiveness of SALIC. Third, we prove experimentally that substituting the human oracle with user tagged images achieves comparable performance to the typical active learning scenario with a human oracle. Finally, we show that the straightforward non-active learning approach, which is equivalent to a random approach, requires 20 times more data to reach the same performance with SALIC.

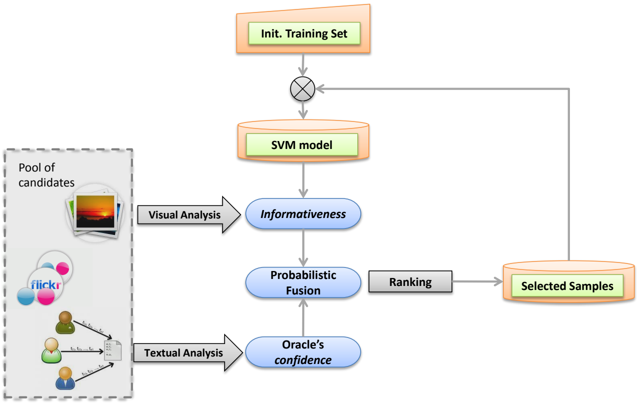


Fig. 1: System Overview.

II. Related Work

During the past decade, various active learning approaches have been developed using different sample selection strategies [8]. In [11], the selective sampling strategy is driven by the requirement to reduce the size of the version space by using an unlabelled sample that halves the version space. In this way the authors aim to cover the full unexplored space with the minimum number of queries. In [12] the authors attempt to attack the insufficient training data problem by initially employing the semi-supervised approach until they collect a sufficient number of reliable training samples. Afterwards they put active learning into force for optimizing the sample selection process. In [13], the authors propose a method to measure the expected change of the model outputs and utilize this measure as the informativeness of the new samples. In [14], the authors analyse different sampling criteria and propose a method that adapts the sampling strategy during training by employing reinforcement learning. In [15], the authors combine the typical uncertainty measure with an information density measure in order to define the *critical* instances to be labelled by the oracle.

There have been quite a few works that investigate active learning with noisy oracles [16], [17], [18]. The objective of the works dealing with noisy oracles is to model the expertise of each oracle, so that the most reliable for a specific instance can be selected. In this direction, the authors of [16] consider active learning in a multiple oracles scenario, where the algorithm not only selects the informative samples but also the oracle to query labels from. The multiple oracles are simulated by clustering the data and assuming each oracle is an expert on one cluster. In a similar scenario, the authors of [17] propose the utilization of transfer learning in order to compute the reliability of each oracle by transfering knowledge from a different domain where labelled data can be abundant instead of depending on a large set of labelled images in the same domain. The authors of [18] extend the Gaussian Processes classification scheme to the multiple annotator scenario by treating the unobserved true labels as latent variables opting to estimate the different levels of expertise of the multiple annotators, obtaining in this way better estimates of the ground truth labels. However, all these works consider that there is at least one oracle that holds the truth, compared to our case where the oracle may never be absolutely confident for its decision. In addition, the sample and oracle selection is performed in two steps (i.e. first select the most informative sample and then the most reliable oracle for the selected sample). This is in contrast to our work, where we *jointly* select the sample which is both informative and our oracle can reliably annotate it.

The idea of combining the benefits of active learning and the knowledge generated from the crowds has recently become the focus of research. In this direction, the authors of [19] propose to use flickr notes in the typical active learning framework (i.e. with a human oracle) with the purpose of obtaining a training dataset for object localization. In a similar endeavour, the authors of [20] introduce the concept of *live learning* where they attempt to combine active learning with crowdsourced labelling. More specifically, rather than filling the pool of candidates with some canned dataset, the system itself gathers possibly relevant images via keyword search on flickr. Then, it repeatedly surveys the data to identify the samples that are most uncertain according to the current model, and generates tasks on MTurk to get the corresponding annotations.

However, although annotations originating from crowdsourcing services are closer to expert's annotation in terms of labelling accuracy [21], they cannot be considered either fully automated or free. In contrast, SALIC relies on data originating from on-line social networks (flickr images and tags) that, although being more noisy, can be used to support a fully automatic learning framework. In the direction of utilizing the free web content, the authors of [22] propose a weakly supervised approach that collects data from web searches, applies clustering and outlier detection and trains a model from every cluster, assuming that it will represent a different characteristic of the concept. Web images are also used as the training set in [23], where the authors treat the associated text as privileged information in a multiple instance learning scenario. Utilizing an extended set of social data (images, tags, users, groups) the authors of [24] propose a method for modeling social strength (i.e. the strength of social relationship between users) by combining a textual, a visual and a friend graph in order to provide personalized recommendations for multimedia content. In a similar vein, the authors of [25] introduce the *need gap*, in addition to the popular semantic gap, which exists between the multimedia semantics and the user needs for multimedia data. Towards automatic creation of a training set, the authors of [26] present a method relying on the results of image search engines, based on the assumption that search engines tend to return very relevant images in the first results of a query. More specifically, the proposed method creates a set of queries, which is given to image search engines and the top images returned by each engine for each query are kept as positive examples. The set of queries is formulated by translating the original concept to 15 different languages, using hyponyms, hypernyms and synonyms from WordNet and finding related terms within the results of the Google text search engine. An extended survey on tag-based image retrieval methods as well as re-ranking with visual content can be found in [27].

With the recent widespread adoption of deep learning techniques, there have also been works that opt to insert the social context in the deep representations [28], [29], [7]. The authors of [28] propose a method for embedding the user intention learning task in the semantic learning task of a typical convolutional neural network, empowering in this way personalized image recommendations. Focusing also on the social context, the authors of [29] based on the assumption that the features of images in the same category tend to be similar, propose an additional layer in a deep learning network that considers the similarities between the training images. The objective is to find the ideal feature map that has minimum effect from the tag-originating noises by decreasing the contribution of the distant images to the gradients. Utilizing user tagged images in a deep learning scenario, the authors of [7] propose a noise-resistant version of logistic regression that can learn model parameters for any tag. The proposed method, based on stochastic EM, can learn from large volumes of user tagged images. In principle, SALIC falls under the weakly supervised learning algorithms umbrella, since it only uses tagged examples and no human annotator. However, the novelty of this work comes from the joint consideration of the samples' informativeness (active learning) and the oracle's confidence (tag relevance). In this way, the visual content of the user tagged images, apart from being used for training the classifiers as in the typical weakly supervised learning case, additionally defines which samples should be added in the training set in the first place. The advantage of integrating active learning in the selection process is that we can weed out the non-useful part of the dataset. This counters the inclusion of unhelpful and possibly falsely labeled data to the model, which would increase the computational complexity and diminish the performance of the model.

More closely related to SALIC are the works presented in [30], [31], which are examined in the same context as our work (i.e. active learning in the multimedia domain with user tagged images). The first approach [30] is based on the assumption that tags can reliably determine if an image does not include a concept, thus making social sites a reliable pool of negative examples. The selected negative samples are further sampled by a two stage sampling strategy. First, a subset is randomly selected and then, the initial classifier is applied to the remaining negative samples. The examples that are most misclassified (i.e. the examples that received the highest confidence scores from the classification model) are considered as the most informative negatives and are selected to boost the classifier. In [31], the authors present a variation of this work, by initially selecting a set of positive images in the first iteration using tag relevance and then putting into work the negative bootstrapping approach. Compared to these works that also incorporate active learning in their selection process, the novelty of SALIC lies in the proposed probabilistic fusion strategy, unlike the state-of-the-art approaches which rely on a two-step methodology (i.e. using one modality to filter part of the pool dataset and then applying the criterion of the other modality to perform sample selection).

III. Problem formulation

Let us consider the typical active learning scenario where we have a small set of manually labelled instances $\mathfrak{L} = \{l_1, l_2, \dots, l_{N_\mathfrak{L}}\}$, $N_\mathfrak{L} = \|\mathfrak{L}\|$ accompanied by their corresponding labels $Y = \{y_1, y_2, \dots, y_{N_\mathfrak{L}}\}$, $y_i \in \mathbb{D}$, where $\mathbb{D}$

TABLE I: Notation table

Symbol	**Definition**
$\mathfrak{L} = \{l_1, l_2, \ldots, l_{N_\mathfrak{L}}\}$	The set of labelled images.
$Y = \{y_1, y_2, \ldots, y_{N_\mathfrak{L}}\}$	The labels of $\mathfrak{L}$.
$\mathbb{D} = \{+1, -1\}$	The label space.
$\mathfrak{U} = \{u_1, u_2, \ldots, u_{N_\mathfrak{U}}\}$	The set of unlabelled images.
$\mathscr{T} = \{\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{N_\mathfrak{U}}}\}$	The tags of $\mathfrak{U}$, where $\mathbf{t_i}$ is the set of tags that correspond to image u_i.
$\mathbf{u_i}$	The visual descriptor of the image u_i.
$\mathbf{u_i^{text}}$	The textual descriptor of the image u_i.
$\mathscr{H}_m = \{\mathbf{w_m}, b_m\}$	The classifier of iteration m ($\mathbf{w_m}$ is the normal vector to the SVM hyperplane and b_m is the bias term). $\mathscr{H}_0$ is the baseline classifier, trained with the initially labelled data.
$\mathscr{E}(\mathscr{H}_m, u)$	The informativeness of sample u based on classifier H_m (i.e. the classifier at iteration m).
$\mathscr{O}(u, \mathbf{t}, d_k)$	The confidence of the oracle that the sample u belongs to the label d_k based on its tags $\mathbf{t}$.
$S_i \in \{0, 1\}$	The random variable (RV) modelling the event of selecting the image u_i.
$V_i \in [0, 1]$	The RV modelling the probability that u_i is informative.
$T_i \in [0, 1]$	The RV modelling the probability that u_i belongs to the examined concept $d_k \in \mathbb{D}$.

*we use normal letters (e.g. u_i) to indicate individuals of some population and bold face letters (e.g. $\mathbf{u_i}$) to indicate vectors or sets of individuals of the same population

is the label space (the utilized notation is summarized in Table I). In our case, we consider the binary classification problem (i.e. $\mathbb{D} = \{+1, -1\}$), thus the probable labels are positive (i.e. $d_1 = +1$) and negative (i.e. $d_2 = -1$), with respect to the concept that we are trying to learn. In addition to the labelled set, we have a large set of unlabelled instances $\mathfrak{U} = \{u_1, u_2, \ldots, u_{N_\mathfrak{U}}\}$, $N_\mathfrak{U} = \|\mathfrak{U}\|$. Moreover, there is an oracle $\mathscr{O}$ providing labels accurately for the unlabelled set $\mathfrak{U}$ on demand (i.e. $y_i = \mathscr{O}(u_i)$). Initially, a baseline classifier $\mathscr{H}_0$ is trained on the labelled set $\mathfrak{L}$. The objective of active learning is to establish a function $\mathscr{E}(\mathscr{H}_0, \mathfrak{U})$, which defines the informativeness of each instance in the unlabelled set based on the previous classifier $\mathscr{H}_0$. The instance $u^* \in \mathfrak{U}$ maximizing this function is selected to be added in the labelled set $\mathfrak{L}$ along with its label y^*, which is provided by the oracle $\mathscr{O}$ (i.e. $y^* = \mathscr{O}(u^*)$).

In our case, the oracle is substituted with the tags of the web users. Thus, the set of images $\mathfrak{U}$ is not completely unlabelled but is associated with a set of tags $\mathscr{T} = \{\mathbf{t_1}, \mathbf{t_2}, \ldots, \mathbf{t_{N_\mathfrak{U}}}\}$, where $\mathbf{t_i}$ is the set of tags associated with image u_i. In this sense, the oracle has already answered for all the instances in $\mathfrak{U}$ beforehand (i.e. the web users have already tagged the images). However, given the noisy nature of the user tags, we consider the oracle to be of questionable reliability and instead of providing an accurate label for a specific sample u^* as before, it provides its confidence for each sample u_i and label $d_k \in \{+1, -1\}$ (i.e. its confidence $\mathscr{O}(u_i, \mathbf{t_i}, d_k)$ that the sample u_i is positive if $d_k = +1$ and negative if $d_k = -1$) based on the tags $\mathbf{t_i}$ of the image u_i. The objective in our case is not only to find the most informative sample, but also the sample for which the oracle is most confident for its label, so that we do not add falsely annotated samples in the training set. In order to do this, we have to jointly maximize the informativeness of a sample $\mathscr{E}(\mathscr{H}_0, \mathfrak{U})$ and the confidence of the oracle $\mathscr{O}(\mathfrak{U}, \mathscr{T}, d_k)$ in the label d_k. Thus we have to establish a new function $\mathscr{P}$, which defines the benefit of selecting a sample u_i given its informativeness and the oracle's confidence. Maximizing this function, we can select the optimal sample u^* with the maximum informativeness and the higher confidence of the oracle for its label:

$$u^* = \underset{u_i \in \mathfrak{U}}{\operatorname{argmax}} \mathscr{P}(u_i | \mathscr{E}(\mathscr{H}_0, u_i), \mathscr{O}(u_i, \mathbf{t_i}, d_k)) \tag{1}$$

The instance u^* that maximizes this function for the examined label d_k is added to the training set as positive if $d_k = +1$ or negative if $d_k = -1$; The expectation is that the addition of such samples in the training set will allow for the maximum performance gain of the classifier. With this approach, positive and negative instances can be selected independently, by setting $d_k = +1$ for the positive and $d_k = -1$ for the negative in Eq. 1.

Considering that active learning is an iterative method, in the next iterations $m \geq 2$ the informativeness is defined as $\mathscr{E}(\mathscr{H}_m, u)$. For simplicity, in the following, we will show the methodology for the first iteration using the baseline classifier $\mathscr{H}_0$, while the full iterative approach can be seen in Algorithm 1.

IV. Active Learning with an unreliable oracle

Given that the function $\mathscr{P}(u_i | \mathscr{E}(\mathscr{H}_0, u_i), \mathscr{O}(u_i, \mathbf{t_i}, d_k))$ cannot be analytically estimated, we choose to approximate it as a probability. For this reason, let us denote the following random variables (RV); S_i, the RV modelling the event of selecting the image u_i ($S_i \in \{0, 1\}$), V_i, the RV modelling the probability that u_i is informative ($V_i \in [0, 1]$), T_i, the RV modelling the probability that u_i belongs to the examined concept d_k ($T_i \in [0, 1]$). Without loss of generality, we can assume that the function $\mathscr{P}(u_i | \mathscr{E}(\mathscr{H}_0, u_i), \mathscr{O}(u_i, \mathbf{t_i}, d_k))$ is proportional to the probability of selecting an instance ($S_i = 1$) given its informativeness (V_i) and the confidence of the oracle (T_i):

$$\mathscr{P}(u_i | \mathscr{E}(\mathscr{H}_0, u_i), \mathscr{O}(u_i, \mathbf{t_i}, d_k)) \sim P(S_i = 1 | V_i, T_i) \tag{2}$$

Consequently, in order to find the optimal u^*, instead of the function $\mathscr{P}$ in Eq. 1, we can maximize its proportional probability function $P(S_i = 1 | V_i, T_i)$. In order to calculate this probability, we make the reasonable assumption that the probability of an image being informative is conditionally independent from the probability that this image belongs to the examined concept (i.e. V_i and T_i are conditionally independent). Using Bayes rule and based on our assumption that V_i and T_i are independent we can express the probability $P(S|V,T)$ as follows (from now on the subscripts of S, V and T will be omitted):

$$\begin{aligned} P(S|V,T) &= \frac{P(V,T|S)P(S)}{P(V,T)} = \\ &= \frac{P(S|V)P(S|T)P(V)P(T)}{P(V,T)P(S)} \end{aligned} \tag{3}$$

In calculating $P(S|V,T)$ we may encounter two cases:

a) If $\mathbf{P(S=0|V,T)\neq 0}$*::* In order to calculate the probability $P(S=1|V,T)$ and eliminate the probabilities $P(V)$, $P(T)$ and $P(V,T)$, we divide the probability of selecting an image with the probability of not selecting it, following the methodology presented in [32].

$$\frac{P(S=1|V,T)}{P(S=0|V,T)} = \frac{\frac{P(S=1|V)P(S=1|T)P(V)P(T)}{P(V,T)P(S=1)}}{\frac{P(S=0|V)P(S=0|T)P(V)P(T)}{P(V,T)P(S=0)}}$$

Then by replacing $P(S=0|V,T)$, $P(S=0|V)$, $P(S=0|T)$ and $P(S=0)$ with their complements ($1-P(S=1|V,T)$, $1-P(S=1|V)$, $1-P(S=1|T)$ and $1-P(S=1)$ respectively), we get the following equation that computes $P(S=1|V,T)$:

$$\begin{aligned} P(S=1|V,T) = \frac{P(S=1|V)P(S=1|T)}{P(S=1)-P(S=1)P(S=1|T)} \cdots \\ \frac{(1-P(S=1))}{-P(S=1)P(S=1|V)+P(S=1|V)P(S=1|T)} \end{aligned} \tag{4}$$

b) If $\mathbf{P(S=0|V,T)=0}$*::* Assuming that the probabilities $P(V)$ and $P(T)$ cannot be 0 (i.e. there is always an a-priori probability that an image is informative and belongs to the examined concept respectively), from Eq. 3 we have that either $P(S=0|V)=0$ or $P(S=0|T)=0$ (i.e. $P(S=1|V)=1$ or $P(S=1|T)=1$). Note that, in this case, Eq. 4 also produces the same result (i.e. $P(S=1|V,T)=1 \Rightarrow P(S=0|V,T)=0$), so from now on we will use Eq. 4 in all cases.

Thus, we only need to estimate three probabilities: $P(S=1)$, $P(S=1|V)$ and $P(S=1|T)$. The first one is set to 0.5 as the probability of selecting an image without any knowledge is the same with the probability of dismissing it. For the probability of selecting an image given its informativeness $P(S=1|V)$, we will approximate it using the informativeness criterion $\mathscr{E}$, which can be any criterion of the Active Learning theory. In our case, we will be using the popular criterion that considers the most informative samples to be the ones lying on the separating hyperplane, since Support Vector Machines (SVMs) will be used to train the classification models. For the probability of selecting an image given the oracle's confidence $P(S=1|T)$, we will approximate it with a textual analysis algorithm that takes as input the tags of an image and provides as output the probability of an image being positive or negative with respect to a concept, as explained below. While any textual analysis algorithm can be used, in this work, we used the popular Bag-of-Words scheme (BoW) due to its ability to additionally consider the context of the tags.

A. Incorporating informativeness in the selection ($P(S|V)$)

The probability $P(S=1|V)$ can be approximated using the informativeness criterion $\mathscr{E}(\mathscr{H}_0, u)$, where $\mathscr{H}_0$ is the baseline classifier. As mentioned above, SVMs were chosen as the classification scheme for this work. For the SVMs, a popular informativeness criterion dictates the selection of the samples closest to the separating hyperplane as the most beneficial samples [11]. While other criteria could be used (e.g. halving the version space by explicitly calculating it when a sample is added), they would require to train a huge number of SVM models in order to calculate the version space. On the other hand, the selected criterion can be calculated very efficiently for linear SVMs since it requires only a dot product to calculate the distance from the hyperplane, making it very attractive for large scale pool of candidates.

Initially, the baseline classifier $\mathscr{H}_0$ is trained using the manually annotated set $\mathfrak{L}$ as a linear SVM classifier (i.e. $\mathscr{H}_0 = \{\mathbf{w}, b\}$, where $\mathbf{w}$ is the normal vector to the hyperplane and b the bias term). Then, for every candidate image in the pool of candidates $u_i \in \mathfrak{U}$, the distance from the hyperplane $V(\mathscr{H}_0, u_i)$ is extracted by applying the SVM classifier ($\mathbf{u_i}$ here denotes the feature vector of the image u_i):

$$V(\mathscr{H}_0, u_i) = \langle \mathbf{w}, \mathbf{u_i} \rangle + b \tag{5}$$

Based on [11], the samples with the minimum distance to the hyperplane are considered as the most informative ones while the samples that lie outside the margin area of the SVM model are not expected to have any impact on the classifier. Thus the function $\mathscr{E}(\mathscr{H}_0, u_i)$, and consequently the probability $P(S|V)$, should be maximized (i.e. $P(S|V)=1$) for the samples lying at the hyperplane and minimized (i.e. $P(S|V)=0$) for the samples that are outside the margin area. Based on the above observation, we approximate the probability $P(S|V)$ with the following equation, which is also visualized in Fig. 2a:

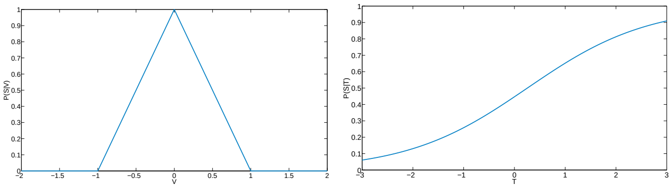
$$P(S|V) \sim \begin{cases} 1-|V| & \text{if } -1 < V < 1 \\ 0 & \text{else} \end{cases} \tag{6}$$

where $|V|$ is the absolute value of the quantity in 5.

B. Measuring oracle's confidence ($P(S|T)$)

In order to measure the oracle's *confidence* $\mathscr{O}(u_i, \mathbf{t_i}, d_k)$ that the image $u_i \in \mathfrak{U}$ belongs to class d_k (i.e. is positive if $d_k = +1$ or negative if $d_k = -1$), we incorporate the associated textual information that is provided in the form of tags. Opting to overcome the noisy nature of social tagging (i.e. lack of structure, ambiguity, redundancy, emotional tagging, etc), we propose the utilization of the popular bag-of-words scheme [9], due to its ability to capture the context of the whole set of tags, instead of only the meaning of each tag independently (e.g. as in the case of tag-to-tag similarity based on WordNet [33]).

The vocabulary is extracted from a large image dataset crawled from flickr. Initially the distinct tags of all images are gathered. The tags that are not included in WordNet are removed and the remaining tags compose the vocabulary. Then, in order to represent each image with a vector, a histogram is calculated by assigning the value 1 to the bins



(a) $P(S|V)$ (b) $P(S|T)$

Fig. 2: Probability of selecting a sample based on (a) visual and (b) textual information.

of the image tags in the vocabulary and 0 to the rest. PCA is applied to reduce the high dimensionality of the initial vocabulary ($\sim 47k$ distinct tags) to $7k$, which was chosen so that 95% of the variance is kept. It is worth noting that, based on internal experiments, PCA has no effect on the performance of the textual analysis algorithm, while reducing significantly the dimensionality and the memory requirements for this step.

Afterwards, a linear SVM model $(\mathbf{w^{text}}, b^{text})$ is trained using the tag histograms as the feature vectors. In order to do this, a training set of images that contain both tags and manual annotations is utilized. The tags are required in order to calculate the feature vectors and the manual annotations to provide the class labels for training the model. In the testing procedure, for every tagged image $u_i \in \mathfrak{U}$ the feature vector $\mathbf{u_i^{text}}$ is calculated as above and the SVM model is applied. This results in a value for each tagged image $T(u_i)$, which corresponds to the distance of $\mathbf{u_i^{text}}$ from the textual hyperplane:

$$T(u_i) = \langle \mathbf{w^{text}}, \mathbf{u_i^{text}} \rangle + b^{text} \tag{7}$$

Thus, the oracle's confidence $\mathscr{O}(u_i, \mathbf{t_i}, d_k = +1)$ that the image u_i is positive, and consequently the probability $P(S|T)$ when selective positive examples, can be calculated using Platt's sigmoid [34] as shown in Fig. 2b:

$$P(S|T) = \mathscr{O}(u_i, \mathbf{t_i}, d_k = +1) = \frac{1}{1 + \exp(AT + B)} \tag{8}$$

The parameters A and B are learnt on the training set using cross validation. The probability of Eq. 8 is for selecting a positive image. In the case that we want to select negative images, the oracle's confidence $\mathscr{O}(u_i, \mathbf{t_i}, d_k = -1)$ that the image u_i is negative, and consequently the probability $P(S|T)$ when selective negative examples, is set to be the complement of Eq. 8:

$$P(S|T) = \mathscr{O}(u_i, \mathbf{t_i}, d_k = -1) = 1 - \mathscr{O}(u_i, \mathbf{t_i}, d_k = +1) \tag{9}$$

C. Re-training

In each iteration, after ranking the images based on the probability $P(S = 1|V, T)$, the top b_+ positive and b_- negative examples are selected to enhance the training set. In the case where both positive and negative samples are selected (as in SALIC), a new classifier is trained using the union of the old data and the newly selected examples as training set. However, in the case where either only positive or only negative examples are selected (as in some baselines we compare with), the classes become imbalanced, which is a typical problem in the machine learning field. In order to cope with the class imbalance problem we also apply a model aggregation method [30]. For the aggregation method, in each iteration m a new *weak* classifier $\mathscr{H}_m^{weak} = \{\mathbf{w_m^{weak}}, b_m^{weak}\}$ is trained on the newly selected examples of class $d_k = +1$ (or $d_k = -1$) and the original examples from the initial training set of the other class $d_k = -1$ (or $d_k = +1$). This classifier is called *weak* since it is not trained on the whole training set, but only on a portion of it. Then, in order to compute the classifier $\mathscr{H}_m = \{\mathbf{w_m}, b_m\}$ for iteration m, all the *weak* classifiers $[\mathscr{H}_1^{weak}, \mathscr{H}_2^{weak}, \cdots, \mathscr{H}_m^{weak}]$ are aggregated by averaging their parameters:

$$\mathscr{H}_m = \frac{1}{m}\sum_{i=1}^{m} \mathscr{H}_i^{weak} = \frac{m-1}{m}\mathscr{H}_{m-1} + \frac{1}{m}\mathscr{H}_m^{weak} \Rightarrow$$
$$\begin{cases} \mathbf{w_m} = \frac{m-1}{m}\mathbf{w_{m-1}} + \frac{1}{m}\mathbf{w_m^{weak}} \\ b_m = \frac{m-1}{m} b_{m-1} + \frac{1}{m} b_m^{weak} \end{cases} \tag{10}$$

By incorporating this method, in each iteration the weak classifier is trained on equally sized sets of positive and negative instances and the final model is extracted by averaging the balanced weak classifiers of all iterations.

The complete algorithmic procedure of SALIC can be seen in Algorithm 1.

Algorithm 1 SALIC

Input: labelled data $\mathfrak{L}$ with their labels Y, unlabelled data $\mathfrak{U}$, an oracle $\mathscr{O}$, the number of iterations N, the number of positive and negative instances selected in each iteration b_+, b_-.

Output: a classifier $\mathscr{H}_N$.

```
$\mathscr{H}_0 = \{\mathbf{w_0}, b_0\} = train(\mathfrak{L}, Y)$
for $m = 1 : N$ do
    // Get Positive Instances
    $d_k = +1$;
    for $j = 1 : b_+$ do
        Choose $u^* = \operatorname{argmax}_{u \in \mathfrak{U}} \mathscr{P}(u|\mathscr{E}(\mathscr{H}_0, u), \mathscr{O}(u, \mathbf{t}, d_k))$
        $\mathscr{P}$ is calculated using equations 2, 4, 6 and 8
        $\{\mathfrak{L}, Y\} \leftarrow \{\mathfrak{L}, Y\} \cup \{u^*, y^* = d_k\}$
        $\mathfrak{U} \leftarrow \mathfrak{U} \setminus u^*$
    end for
    // Get Negative Instances
    $d_k = -1$;
    for $j = 1 : b_-$ do
        Choose $u^* = \operatorname{argmax}_{u \in \mathfrak{U}} \mathscr{P}(u|\mathscr{E}(\mathscr{H}_0, u), \mathscr{O}(u, \mathbf{t}, d_k))$
        $\mathscr{P}$ is calculated using equations 2, 4, 6 and 9
        $\{\mathfrak{L}, Y\} \leftarrow \{\mathfrak{L}, Y\} \cup \{u^*, y^* = d_k\}$
        $\mathfrak{U} \leftarrow \mathfrak{U} \setminus u^*$
    end for
    $\mathscr{H}_m = \{\mathbf{w_m}, b_m\} = train(\mathfrak{L}, Y)$
end for
```

V. Experimental Setup

A. Datasets

Three datasets were employed for the purpose of our experiments. The imageCLEF dataset [35] ($Imageclef$) consists of 25000 manually labelled images and was split into two parts, $Imageclef_{train}$ and $Imageclef_{test}$ consisting of 15k train and 10k test images respectively. The dataset was annotated by a vocabulary of 94 concepts which belong to 19 general categories (*age, celestial, combustion, fauna, flora, gender, lighting, quality, quantity, relation, scape, sentiment, setting, style, time of day, transport, view, water, weather*). The images of this dataset originate from flickr and the tags were also provided. From this dataset we obtained the manually labelled dataset $\mathfrak{L}$ and the evaluation set ($Imageclef_{test}$)

The MIRFLICKR-1M dataset ($Mirflickr$) [4] consists of one million user tagged images harvested from flickr. The images of $Mirflickr$ were tagged with 862115 distinct tags of which 46937 were meaningful (included in WordNet). The tags that were not included in WordNet were removed and the images ending up with no tags were removed from this dataset (out of the 1 million, 131302 images had no tags). In addition, given that the $Imageclef$ dataset is a subset of $Mirflickr$, the images that are included in both sets were removed from $Mirflickr$. In our experiments, this dataset constitutes the pool of loosely tagged images.

The third dataset ($Imagenet$) includes images from the manually labelled ImageNET database [5]. The vocabulary of $Imagenet$ consists of the 34 concepts that were included in both the synsets of ImageNet and the vocabulary of $Imageclef$. For every concept in the vocabulary of $Imagenet$, the images in the corresponding ImageNet synset were downloaded. This dataset consists of approximately $500k$ images.

B. Implementation details

For the visual representation of the images, we have used two popular approaches in order to verify the effectiveness of SALIC in different conditions; one that results in very high dimensional features that are of medium performance and one for low dimensional features that have shown remarkable performance. Our objective is to examine whether this difference in the discrimination ability of the two feature spaces will lead to particular requirements with respect to the sample selection process.

First, for the high dimensional features we have used the approach that was shown to perform best in [36]. More specifically gray SIFT features were extracted at densely selected key-points at four scales, using the vl-feat library [37]. Principal component analysis was applied on the SIFT features, decreasing their dimensionality from 128 to 80. Then, Fisher vector encoding (256 GMM components) and spatial pyramids ($1 \times 1, 3 \times 1, 2 \times 2$ regions) were applied, resulting in a $327680 - dimensional$ feature vector per image.

Second, the low dimensional features are extracted from Convolutional Neural Networks (CNNs), which have shown remarkable performance in the past years in both image annotation and object detection [38]. The implementation and the pre-trained CNN models of [39] were used. More specifically, the models for extracting the lowest dimensional feature vectors were utilized (model CNN M 128 for the utilized implementation [39] or vgg-m-128 for the MatConvNet implementation [37]), resulting in only 128 dimensions.

Linear SVM models were trained for both feature spaces using the LIBSVM library [40] and were evaluated by mean Average Precision (mAP). The code for the implementation of the presented approach and the data to reproduce the results can be downloaded from[1].

Considering that the utilized datasets consist of images labelled for multiple concepts, in order to conform with the binary classification requirements of SALIC, we transformed the multi label scheme to binary using the one-vs-all approach. More specifically, for every concept, as positive samples were considered all images including this concept in their list of annotated labels, leaving all the rest of the images as negative samples. Thus, the same image could serve as a positive sample for more than one concepts. In our experimental study, 100 positive and 100 negative images from the 15k training images of $Imageclef_{train}$ were randomly selected to train the baseline classifiers (from now on called the $Imageclef_{init}$ dataset). The same initial examples were used for all experiments to allow a fair comparison. Then in each iteration, 100 images were selected from the pool of candidates to enhance the training set. Based on our experiments in Section VII-D, a good choice for the size of the initial training set is 100 positive and 100 negative manually labeled examples and for the size of the batch 100 examples (i.e. in each iteration we add either i) 100 positive or ii) 100 negative or iii) 50 positive and 50 negative depending on the training set expansion strategy). Finally, 50 iterations were conducted in all cases, i.e. in total, 5k images were added for each independent binary classification problem (i.e. concept).

VI. Expanding with positive, negative or both

Our motivation to search whether positive or negative examples should be selected is based on the following intuitive and experimental observations. Intuitively, negative examples are expected to be more informative than the positive ones, since they tend to cover a larger part of the concept space and as a consequence of the feature space (i.e. they depict a larger variety of objects compared to the positive examples which depict just one). In order to verify our intuitive observation experimentally, we developed the following experimental set up; Initially, by training a classifier on the $Imageclef_{init}$ and applying it on $Imageclef_{test}$ we achieved the performance of 17.43% in terms of mAP by averaging the performance of the 34 binary classification problems (i.e. one for each concept). Adding all the remaining positive examples of the $Imageclef_{init}$ dataset, so essentially using all positive vs 100 negative examples for each concept, the performance only increased to 18.38%. Adding all negative examples instead to the training set, i.e. using 100 positive vs all negative examples for each concept, the performance increased to 23.62%. On the other hand, if we use the full training set, the performance of

[1] http://mklab.iti.gr/research/salic

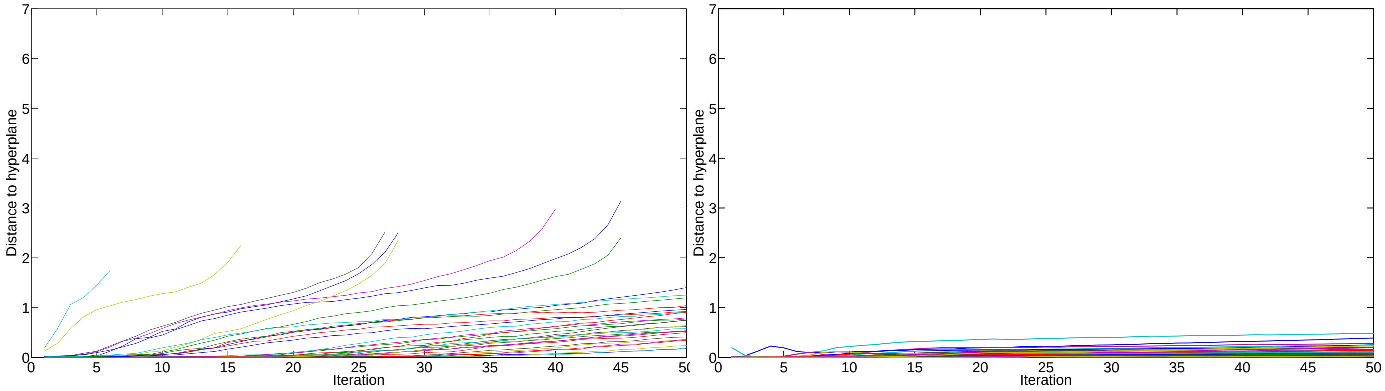
the classifiers reaches the value of 29.67%. We can see that although the addition of negative examples is indeed more beneficial than the positive examples, it is far from achieving the performance obtained by using both positive and negative examples. The previous numerical results are for the Fisher based features. Similar conclusions can be reached for the CNN based features.

In order to test the contribution of each example type (i.e. positive or negative) in an iterative active learning scenario, we apply the active learning paradigm using the $Imageclef_{init}$ dataset to train the initial models and the $Imagenet$ dataset as the pool of candidates. The reason is that $Imagenet$ is manually annotated, removing the additional factor of how accurately the new samples are annotated. Only the 34 concepts of the $Imagenet$ dataset were tested in this section. Then for every concept, 100 images were selected to enhance the initial training set; positive images were selected from the manually annotated $Imagenet$ dataset and negative examples from the remaining negatives of the $Imageclef_{train}$ dataset. In each iteration, a total of 100 images are selected based on their *informativeness*. In Fig. 3 the following five configurations are compared; i) 50 positive and 50 negative images were added in the training set, ii) 100 positive images were added in the training set, iii) 100 negative images were added in the training set, iv) 100 positive images were selected and the model aggregation method described in Section IV-C was used, and v) 100 negative images were selected and the model aggregation method described in Section IV-C was used. The aggregation method was used only with the addition of only one type of examples to alleviate the class imbalance problem. On the contrary, this was not required for the addition of both positive and negative examples since in this case class imbalance is not an issue.

The results are shown in Fig. 3a for the Fisher based features and in Fig. 3b for the CNN based features. We can see that, in both cases, the addition of negative examples is more beneficial than that of positive ones which confirms our previous findings as well as our intuitive explanation. Moreover, tackling the imbalance issue with the aggregation method increases the performance for the low dimensional CNN based features while it does not improve the performance for the Fisher based features. On the other hand, we can see that adding negative examples without aggregation (red line with circles) increases the classifier's performance at the initial iterations (i.e. $\sim$ up to 5) when the class imbalance problem is mild, but deteriorates significantly later as the classes get severely imbalanced. The addition of both positive and negative examples (blue line) outperforms greatly all the other variations for the Fisher based features, as shown in Fig. 3a, demonstrating the importance of adding examples from both classes. For the CNN based features, the addition of negative examples with aggregation performs similarly with adding both positive and negative examples (Fig. 3b). At the initial iterations, adding only negative examples with aggregation provides a higher boost in the classifiers' performance. However, after the 15^{th} iteration their performance starts to deteriorate, while adding both positive and negative provides a more stable increase in the performance of the classifiers.

This difference in performance between the two feature spaces can be attributed to their discrimination ability. The CNN based features are expected to be much more descriptive than the Fisher based features, judging by their relative performance. However, a better performing classifier is expected to be confident for more samples from the pool of candidates and thus leave a smaller area in the sample space for which it is uncertain. This evidently means that there will be fewer informative samples in this case. In order to examine if this is the case, we plot the average distance to the hyperplane for each iteration, for both approaches, i.e. positive and negative and negative with aggregation, for both feature spaces. The average distance to the hyperplane is an indication of the average informativeness of the selected samples (i.e. when the selected samples have an average distance from the hyperplane less than 1, it means that most of them are within the margin area and therefore they are informative). Thus, we would expect that the distance from the hyperplane in the case of the Fisher based features to stay below 1 for most concepts and iterations, compared to the CNN based features, where we expect the distance to grow quickly beyond 1 for the positive examples and stay below or around 1 for the negative examples.

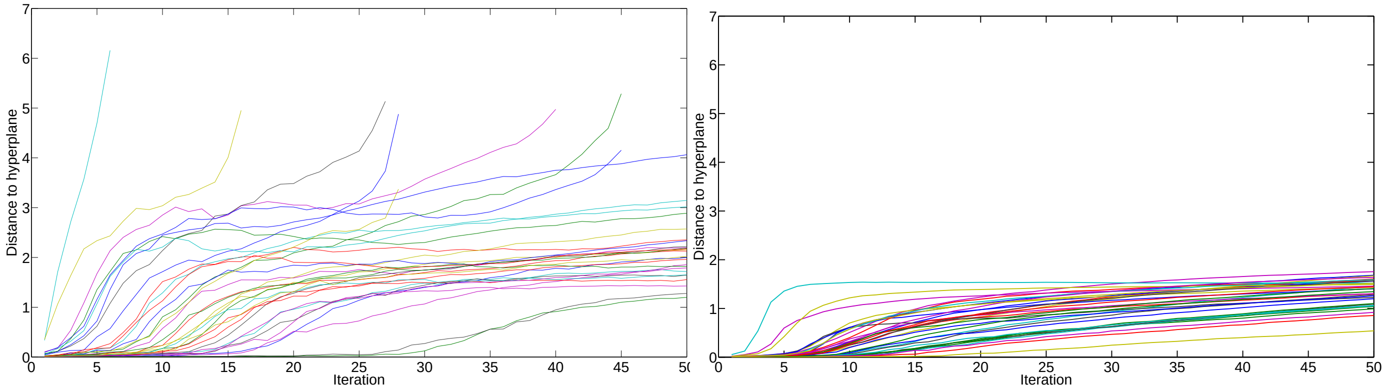
The results can be seen in Fig. 4 for the Fisher based features and in Fig. 5 for the CNN based features. Each line in all plots corresponds to a different concept, thus there are 34 lines in each plot. We can see that in the case of the CNN based features and for the negative with aggregation method (Fig. 5b), the average distance stays below 1 for the first iterations and around 1 later on, which means that we continuously select informative samples (i.e. samples close to the hyperplane). In the case of positive and negative method (Fig. 5a), we can see that for many concepts, we start selecting samples far from the hyperplane even from the early iterations (i.e. the average distance to the hyperplane grows beyond 1 quickly). Practically, this means that after the positive informative samples end, we only add half informative samples for the blue line (i.e. adding both positive and negative examples) compared to the magenta line (i.e. adding negative examples with aggregation). Indeed, if one looks closer at the results in Fig. 3b, the blue line shows great performance increase and potential in the first 5 iterations and suddenly converges. On the other hand, this happens much later in the case of the high dimensional Fisher features (Fig. 4a), which explains the superiority of the blue line compared to the magenta line in Fig. 3a. From the above, it is evident that, for more discriminant and low dimensional feature spaces, it is sufficient to just add negative images with aggregation, while for the less discriminant and high dimensional feature spaces adding both positive and negative images provides a significant performance increase. For the next experiments the first configuration was selected (i.e. both positive and negative examples are added in the initial training set), since a more robust performance is achieved with this configuration across the two feature spaces.



(a) Positive and negative approach.

(b) Negative with aggregation approach.

Fig. 4: Average distance from the hyperplane, at each iteration, of the selected samples - Fisher. Training set: $Imageclef_{init}$ for 34 concepts, Pool of candidates: $Imagenet$ for positive and $Imageclef_{train} \setminus Imageclef_{init}$ for negative, Test set: $Imageclef_{test}$ for 34 concepts.



(a) Positive and negative approach.

(b) Negative with aggregation approach.

Fig. 5: Average distance from the hyperplane, at each iteration, of the selected samples - CNNs. Training set: $Imageclef_{init}$ for 34 concepts, Pool of candidates: $Imagenet$ for positive and $Imageclef_{train} \setminus Imageclef_{init}$ for negative, Test set: $Imageclef_{test}$ for 34 concepts.

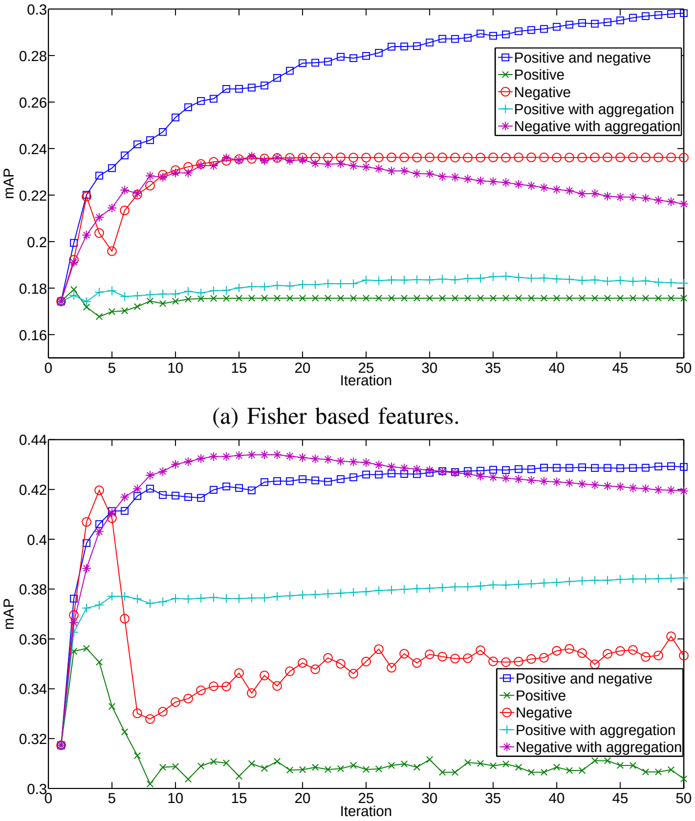
VII. Experimental results

The objective of this section is to compare the proposed active sample selection strategy against various baselines and state-of-the-art approaches. For the experiments of this section, all 94 concepts of $Imageclef$ were used. The first iteration is based on the initial models that were generated using $Imageclef_{init}$. Afterwards, in each iteration, the initial models are enhanced with 50 positive and 50 negative samples from $Mirflickr$.

A. Comparing with sample selection approaches

1) Quantitative evaluation of the proposed selective sampling approach: In this section, we want to compare the performance boost that is achieved by SALIC with two selective sampling baselines; a) self learning [41], where the images that maximize the certainty of the SVM model (Eq. 5) trained on visual information are selected to expand the training set and b) a text-based approach, where the images that maximize the oracle's *confidence* (Eq. 8) are selected. In our case, where the pool of candidates consists of user tagged images instead of the typical manually labelled images, the text based approach is equivalent to the *random* baseline that every active learning approach compares with, since it neglects completely the samples' informativeness. It was favoured over a completely random approach (i.e. adding samples completely randomly without taking into account if they are actually positive or negative) in order to avoid adding false positives/negatives.

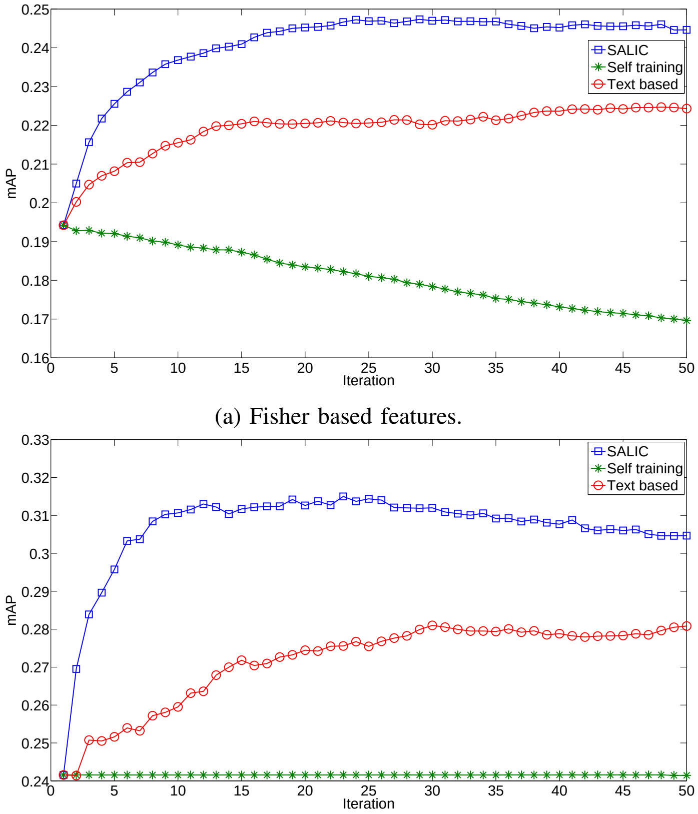
The results can be seen in Fig. 6. We can see that self learning does not provide any benefit to the models; this can be explained by the fact that adding examples far away from the hyperplane does not add any information to the model. Moreover, the fact that SALIC outperforms both the text based and the self training approaches, indicates the importance of incorporating the informativeness of new images in the selection strategy. With respect to the two utilized feature spaces, we can see that the results are very similar, with the only difference of self learning. In the case of the CNN features, the updated models do not gain in performance compared to the baseline ones. This can be attributed to the high discrimination ability of these features that forces self learning to select images far away from the hyperplane and thus not producing additional support vectors (i.e. the hyperplane is the same in all iterations). This also agrees



(a) Fisher based features.

(b) CNN based features.

Fig. 3: Iteratively adding positive, negative or both examples. Training set: $Imageclef_{init}$ for 34 concepts, Pool of candidates: $Imagenet$ for positive and $Imageclef_{train} \setminus Imageclef_{init}$ for negative, Test set: $Imageclef_{test}$ for 34 concepts.



(a) Fisher based features.

(b) CNN based features.

Fig. 6: Comparing with sample selection baselines. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$

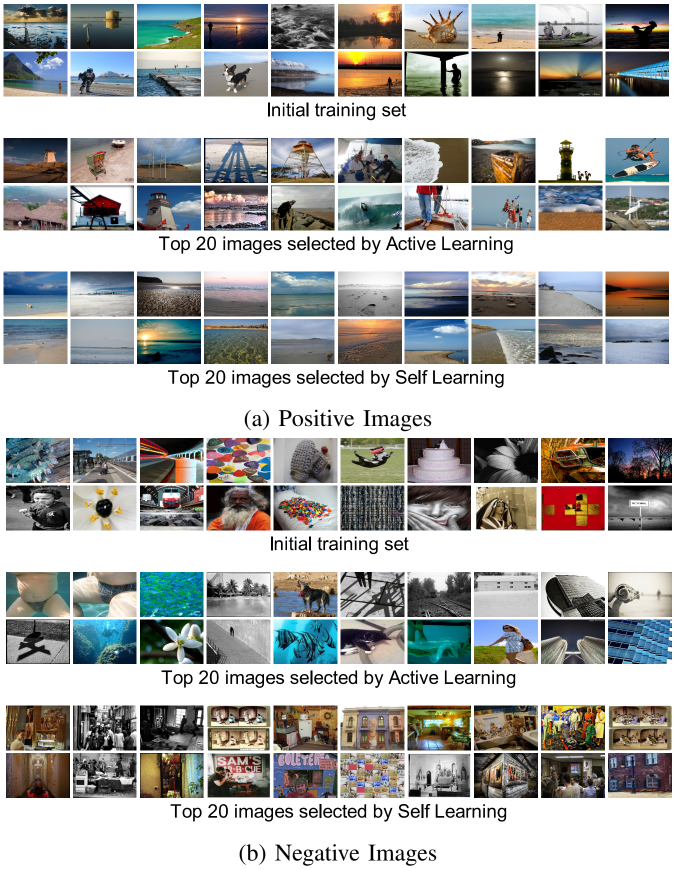
with our observations in Section VI. Furthermore, in order to demonstrate the effectiveness of SALIC, it is worth noting that the fully supervised result (i.e. using all the 15000 images of $Imageclef_{train}$ to train the classification models) achieves a performance of 27.74% in terms of mAP, while SALIC yields a maximum performance of 31.5%.

On the other hand, in the case of the Fisher features, although the hyperplane of the self learning approach changes through the iterations, showing that the selected images are not necessarily far from the hyperplane, we can see that instead of increasing its performance it deteriorates. This can be attributed to the fact that the confidence of the oracle is not taken into account and thus the algorithm selects false positives/negatives to add to the training set. Finally, in this case, the fully supervised result achieves a performance of 28.06% in terms of mAP, while SALIC yields a maximum performance of 25%, requiring only a small portion of annotated samples ($\sim 1.3\%$ of the total training set).

2) Qualitative evaluation of the proposed selective sampling approach: In this section, we want to compare SALIC with the typical self learning approach qualitatively. In order to do so, we show visually the positive and negative samples that have been selected by the two approaches for the concept *coast*. The CNN based features are used for this visualization (Fig. 7). In each figure we show 20 randomly positive (or negative) examples from the initial training set and the top 20 positive and negative images that were chosen by each strategy. For the positive examples (Fig. 7a), we can see that the initial training set mostly consists of calm beaches depicting the sea, the sun and the sky, without many additional objects. As expected the self learning strategy also selects very similar images to the training set, whereas the proposed approach selects images that depict rocky beaches or wavy seas with more objects (e.g. house, boat, lifeguard tower, surf boards and people). By adding images with larger content variety, the generalization ability of the enhanced classifiers is maximized. Similarly for the negative examples (Fig. 7b), we can see that the proposed approach selects more relevant negatives (e.g. underwater images, blue buildings, skies with fields, etc.), whereas the self learning approach mostly selects indoor images that are completely irrelevant to the examined concept.

B. Comparing with fusion approaches

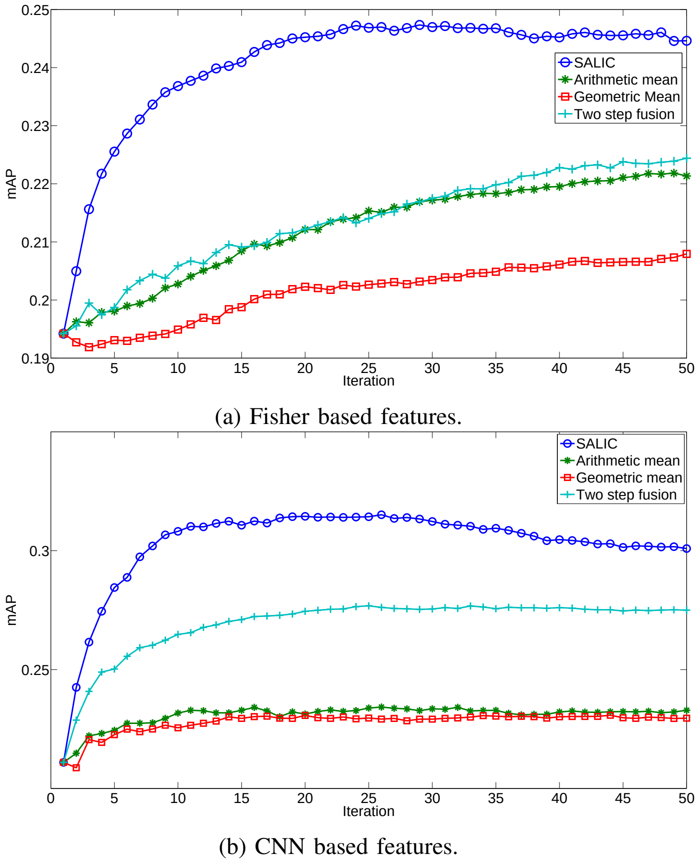
In this section, in order to show the benefit of utilizing the proposed probabilistic approach presented in Section IV for fusing the informativeness of the samples ($P(S|V)$) and the oracle's confidence ($P(S|T)$), we compare it with three


Initial training set
Top 20 images selected by Active Learning
Top 20 images selected by Self Learning
Initial training set
Top 20 images selected by Active Learning
Top 20 images selected by Self Learning

(a) Positive Images

(b) Negative Images

Fig. 7: Selected positive and negative examples by active learning and self training for the concept coast. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$.



(a) Fisher based features.

(b) CNN based features.

Fig. 8: Comparing with fusion baselines. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$.

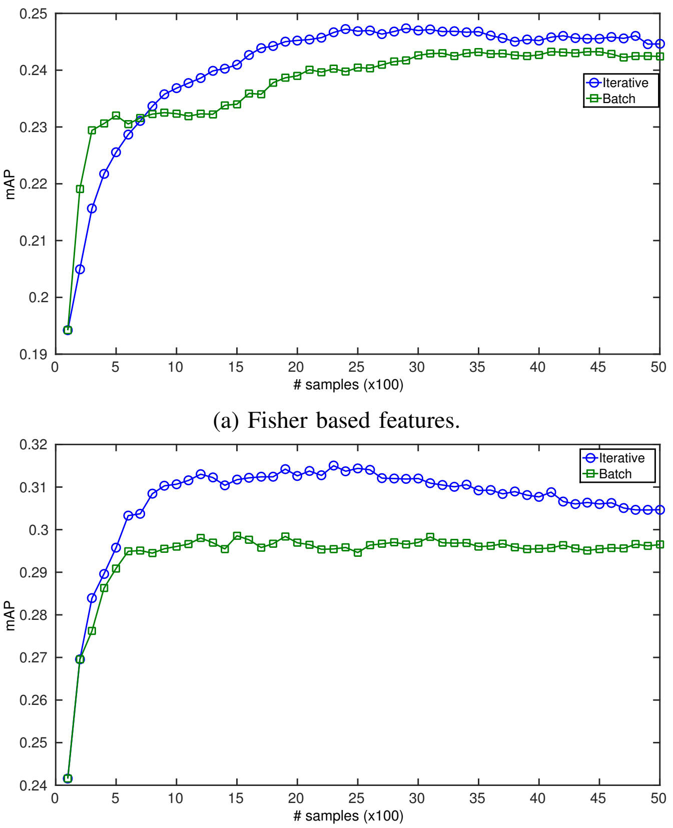
baseline fusion strategies: a) the arithmetic mean, where the arithmetic mean is used to combine the two probabilities $P(S|V)$ and $P(S|T)$ into the selection probability $P(S|V,T)$, b) the geometric mean, where the geometric mean is used to combine the two probabilities $P(S|V)$ and $P(S|T)$ into the selection probability $P(S|V,T)$ and c) a two step approach that simulates the typical way that active learning is performed, while also keeping the classes balanced. In this case, in the first step, the tagged images in the pool of candidates are annotated as positive or negative based on the oracle's confidence ($P(S|T)$). In the second step, the most informative of the positive and negative images are selected (i.e. the images that maximize the probability $P(S|V)$). The results are shown in Fig. 8 both for Fisher (8a) and CNN (8b) based features. We can see that using the simplistic arithmetic and geometric mean approaches the initial models do not improve significantly, showing that the selected samples are either not informative or false positives/negatives. On the other hand, the two step approach yields a higher performance boost, which is probably due to the strict filtering it applies to the tagged images so that false positives/negatives are not selected, while keeping the notion of informativeness in the selection process. On the contrary, SALIC greatly outperforms all the baseline fusion strategies by selecting the samples that will have a higher impact when added in the training set, showing the importance of maximizing the joint probability $P(S|V,T)$.

C. Comparing with batch mode AL

In this section, our objective is to investigate the effect of iteratively selecting examples compared to the simple batch mode (followed by [10]), where all the samples are selected in one iteration. In order to examine this effect in a fair way, we compare two variations of SALIC; a) selecting samples iteratively by adding 100 samples in each iteration (i.e. a total of $100 \times m$ samples in the m^{th} iteration) and b) selecting the same number of $100 \times m$ examples in a single iteration. In Fig. 9 we plot the mAP for both approaches with respect to the total number of examples added ($\#samples$). We can see the benefit of iterative active learning in both feature spaces. More specifically, in the case of the Fisher based vectors (Fig. 9a), the iterative approach gives a performance of $\sim 1\%$ higher compared to the batch mode, with the exception of when adding very few examples (up to 400 samples), while in the case of the CNN based vectors (Fig. 9b), the iterative approach consistently outperforms the batch mode one with a difference of 2% in mAP.

D. Parameter sensitivity investigation

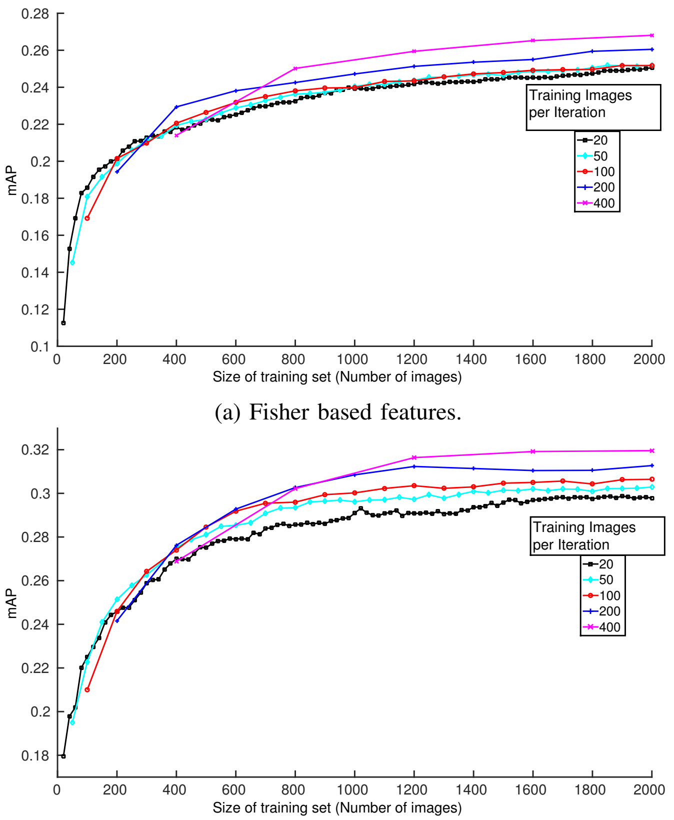
In this section, we want to investigate the sensitivity of the proposed approach to various parameters. More specifically,



(a) Fisher based features.

(b) CNN based features.

Fig. 9: Comparing iterative with batch mode AL. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$.



(a) Fisher based features.

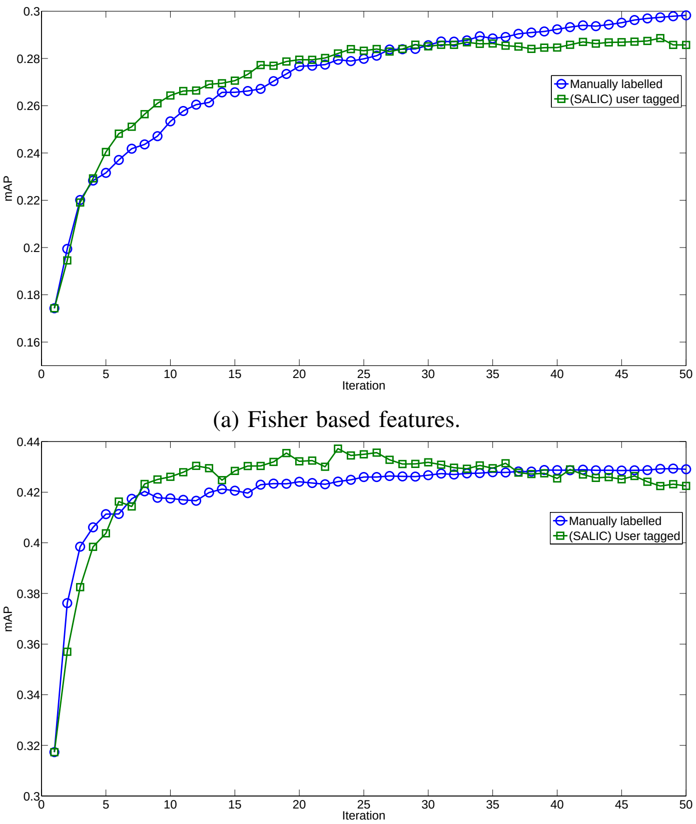
(b) CNN based features.

Fig. 10: Parameter Sensitivity. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$.

we want to examine how the size of the initial training set and the number of the images added in each iteration affects the performance of the resulting models. In order to investigate this, we run SALIC using 20, 50, 100, 200 and 400 manually labeled images in the first iteration (half positive and half negative) and add 20, 50, 100, 200 and 400 images in each iteration respectively (half positive and half negative). In order to visualize the results, we plot the performance of each run with respect to the total number of samples used in the training set. The results can be seen in Fig. 10. It is obvious that, as it would be expected, selecting more manually labeled images in the first iteration the performance of the initial models is significantly higher. For example, in the case of CNN-based features (Fig. 10b), the models created with 20 labeled examples - black line - have a performance of 18% mAP, while when we have 400 labeled examples - magenta line - the performance rises to 26%. However, by adding training data with the help of SALIC, all the lines tend to converge to a similar performance (i.e. the black line grows from 18% mAP to 30% mAP, while the magenta line grows from 26% mAP to 32% mAP). Moreover, as expected, each classifier trained with a larger number of initially manually labeled data converges to a slightly higher performance and is much faster since only a few iterations are required. However, we also have to consider that these models require much more effort in order to acquire manually the increased number of labels that are needed to train the initial classifier. We can see that selecting 100 samples for the initial and the batch size seems a good compromise between training time, performance and annotation cost. Similar conclusions can be drawn for the Fisher based features (Fig. 10a)

E. Comparing with an errorfree oracle

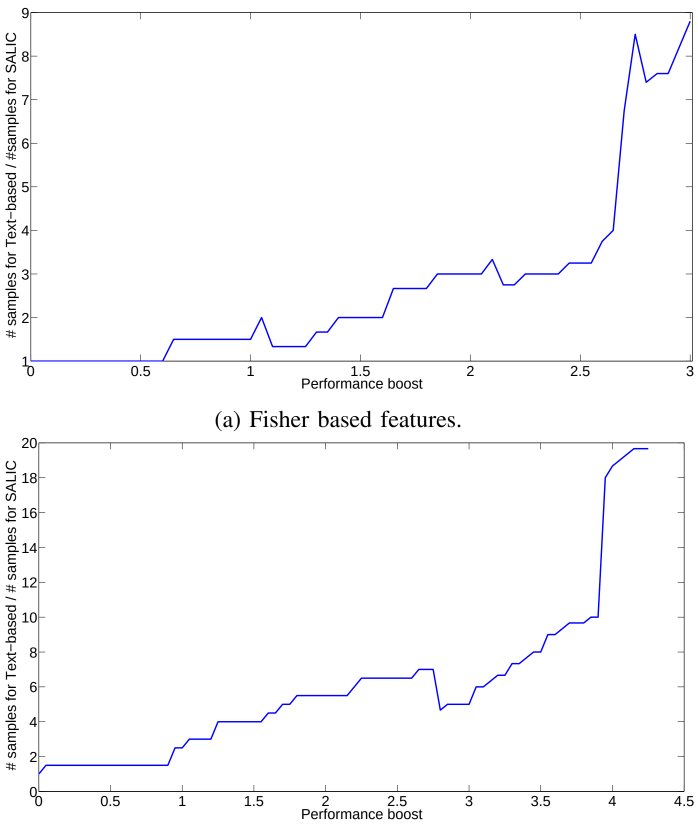
In order to investigate the impact of replacing the errorfree human oracle with web users, we compare the results of SALIC, which employs user tagged images as the pool of candidates, with the ones obtained in Section VI using the manually annotated $Imagenet$ as the pool of candidates. Note that these numbers correspond to the 34 concepts used in Section VI so that the results of the two approaches are directly comparable. In Fig. 11, we plot the performance of the classification models for all iterations, a) when enhanced by a human oracle (blue line) and b) when enhanced by the proposed approach (green line). In Fig. 11a the results for the Fisher based features are shown, while in Fig. 11b the results for the CNN based features are shown. We can see that, for both feature spaces, SALIC performs comparably well with the typical case of having a human oracle. This means that both the selection of bag-of-words for reducing the noise of tags and the approximation of the sample selection function with the probability of Eq. 2 are appropriate choices.



(a) Fisher based features.

(b) CNN based features.

Fig. 11: The impact of substituting the human oracle with web users. Training set: $Imageclef_{init}$ for 34 concepts, Pool of candidates: a) for SALIC (green line) $Mirflickr$ and b) for manually labelled (blue line) $Imagenet$ for positive and $Imageclef_{train} \setminus Imageclef_{init}$ for negative, Test set: $Imageclef_{test}$ for 34 concepts



(a) Fisher based features.

(b) CNN based features.

Fig. 12: Ratio of samples, required to achieve the same performance between the text-based approach and SALIC. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$.

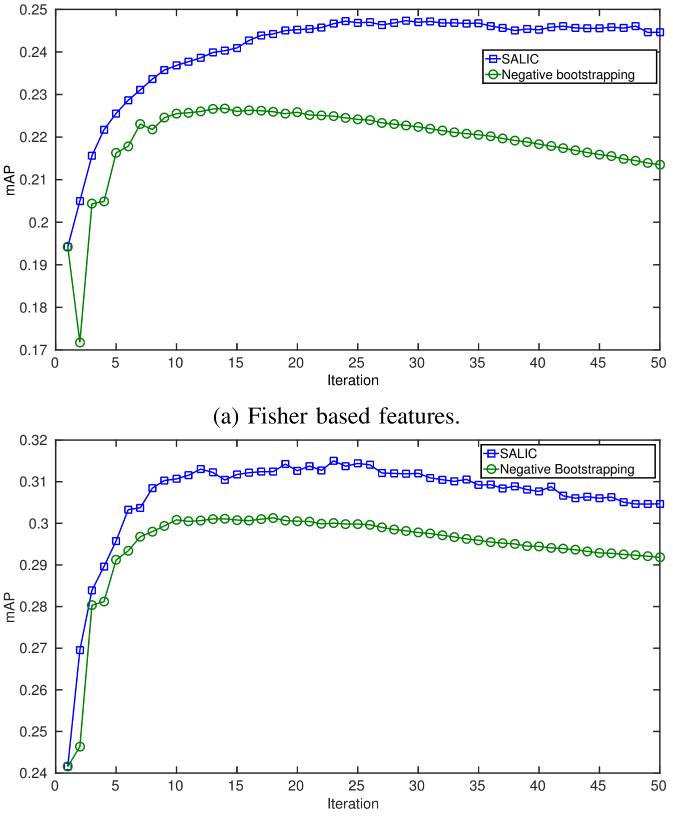
F. Why active learning when tags are free?

In this section, we want to demonstrate the gain in scalability achieved by actively forming the training set compared to adding positive and negative images in an informativeness-agnostic manner. For this experiment, the text-based approach plays the role of the informativeness-agnostic learning algorithm. For demonstrating the gain in scalability, we compute the ratio of the images that are required by the text based approach to the images that are required by SALIC in order to achieve the exact same performance. In Fig. 12, the plot of this ratio with respect to the achieved performance boost is shown. We can see that, for the CNN based features, in order to achieve a 4% boost in mAP, the text based methods requires 20 times more images than SALIC (Fig. 12b). Similarly, for the Fisher based features, the informativeness-agnostic approach requires 9 times more instances to reach the same performance with SALIC for a 3% boost in mAP (Fig. 12a). The fact that we need an order of scale more data instances to achieve the same performance gain, shows the importance of active learning even in the case where labels can be obtained for free. Moreover, it is interesting to note that the text based approach was not able to reach the maximum performance of SALIC, even after 200 iterations (within the first 50 iterations, it approximately achieved half of the performance boost compared to SALIC as it can be seen in Fig. 6).

G. Comparing with state-of-the-art

1) Comparing with Active Learning: In this section we want to compare SALIC with the negative bootstrapping approach presented in [30] (i.e. selecting in each iteration the 100 negative images that are most misclassified and utilizing the model aggregation approach presented in Section IV-C to cope with the class imbalance problem). The results can be seen in Fig. 13. We can see that SALIC continuously increases the performance of the classifiers and constantly outperforms the other approach in both cases.

It is also worth viewing in parallel the comparison between SALIC and [30] along with the comparison in Fig. 3b (Section VI) between adding positive+negative (blue line) and negative with aggregation (magenta line), which is also used by [30]. While in Fig. 3b using positive and negative produces similar performance to only using negative with aggregation, SALIC performs much better compared to [30] (Fig. 13b). This can be attributed to the proposed probabilistic fusion strategy between the two modalities, i.e. the samples'



(a) Fisher based features.

(b) CNN based features.

Fig. 13: Comparing with active learning. Training set: $Imageclef_{init}$, Pool of candidates: $Mirflickr$, Test set: $Imageclef_{test}$.

informativeness and the oracle's confidence, compared to [30], which is closer to a two step fusion.

2) Comparing with Weakly Supervised Learning: In this section we want to compare with weakly supervised methods that also learn classification models based on freely available images on the web. More specifically, we compare with two types of methods; one that selects images by querying search engines [26] (SE) and another that uses flickr images and tags in a deep learning scenario (DL). The first approach is based on the observation that search engines tend to provide accurate results for the top retrieved images. The proposed method constructs a set of queries by i) translating the original query to 15 different languages, ii) using hyponyms, hypernyms and synonyms from WordNet and iii) finding related terms within the results of the Google text search engine. Then, they query image search engines (i.e. Google, Bing and flickr) with each term in the previously constructed set and retrieve the top 24 images of each query (24 was found to be optimal in [26]). The second method [7] proposes a noise-resistant version of logistic regression that can learn model parameters for any tag, called robust logistic regression (RLR). This is accomplished by introducing a parameter, which models the probability that a user will supply a tag, conditioned on the tag being true for the image. This layer of robust logistic regression is added after the last fully connected layer of a deep neural network. In order to learn the parameters of the proposed network the model weights are initialized from AlexNet, a network trained previously on the manually labeled images from ImageNet and then they are refined using the user tagged images.

In order to compare SALIC with [26], [7], we applied our method to the dataset used in [26], which consists of 40 ImageNet concepts. As there was no manually labelled set for these concepts, we selected the initial 100+100 images to train $\mathscr{H}_0$ from the $Mirflickr$ dataset using the text-based approach. In Fig. 14, we compare between the proposed approach (SALIC), the search engine-based method presented in [26] (SE) and the deep learning based method presented in [7] (DL). In order to obtain comparable results, all methods used deep learning-based features with a fully sized penultimate layer (i.e. 4096). In addition, the number of iterations in the DL method were set to 8000 so that all four approaches require similar computational time (i.e. ~ 6 hours for the whole training procedure, including feature extraction for SALIC and SE). With respect to SE, we can see that SALIC outperforms this approach for 28 out of 40 concepts, yielding a mAP of 63.03%, compared to 60.3% for SE. With respect to DL, we can see that SALIC compares slightly favorably to DL (63.03% for SALIC versus 62.7% for DL), outperforming it for 23 concepts out of 40. It is also worth noting that there are three concepts, namely animal, rhino and vehicle, where SALIC fails to gather good quality training samples. For the concepts animal and vehicle, this can be attributed to their generic nature partly, since they include many diversiform sub-concepts (e.g. dogs, cats, insects, fish, etc for animal and boats, buses, cars, airplanes for vehicle). In addition, $Mirflickr$ seems to be unfit for these concepts. This can be verified if we note the SE outperforms significantly both SALIC and DL for these concepts, which is the only method of the three based on search engines rather than $Mirflickr$.

VIII. Conclusions

In this paper, we propose SALIC, an automatic active learning approach for image classification, where the oracle is replaced with web users and the pool of candidates with user tagged images. The contribution of SALIC is a probabilistic approach for *joint* maximization of the samples' informativeness and the oracle's confidence. Experimental results show the superiority of SALIC compared to baselines and state-of-the-art approaches. Moreover, we argue that despite the abundant availability of user tagged images that can be labelled with no cost, the existence of methods like SALIC is necessary to ensure the scalability of image classification applications to the constantly increasing demands. In the future, we plan to investigate the utilization of ensemble and on-line learning in the training process, in order to better accommodate the current scalability challenges.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreements no FP7-ICT-600676 "i-Treasures: Intangible Treasures - Capturing the Intangible

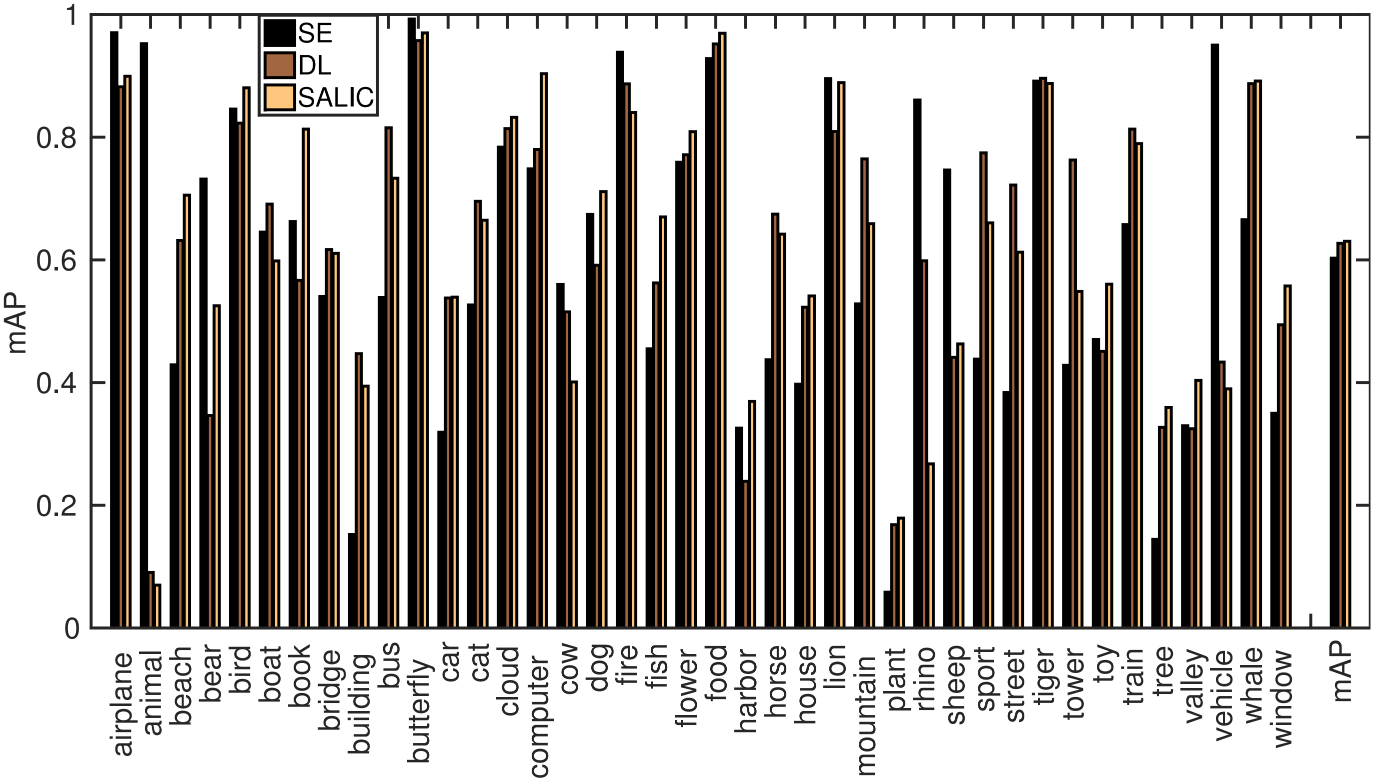


Fig. 14: Comparing with weakly supervised learning. Training set: a) $Mirflickr$ for SALIC and DL and b) search engine images for SE, Pool of candidates: $Mirflickr$ (only for SALIC), Test set: dataset from [26].

Cultural Heritage and Learning the Rare Know-How of Living Human Treasures" and FP7-601138 PERICLES Digital Preservation.

References

[1] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, May 1994.

[2] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[3] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: A real-world web image database from national university of singapore," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009, pp. 48:1–48:9. [Online]. Available: http://doi.acm.org/10.1145/1646396.1646452

[4] M. J. Huiskes, B. Thomee, and M. S. Lew, "New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative," in *Proceedings of the International Conference on Multimedia Information Retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 527–536. [Online]. Available: http://doi.acm.org/10.1145/1743384.1743475

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 248–255.

[6] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016. [Online]. Available: http://doi.acm.org/10.1145/2812802

[7] H. Izadinia, B. C. Russell, A. Farhadi, M. D. Hoffman, and A. Hertzmann, "Deep classifiers from image tags in the wild," in *Proceedings of the 2015 Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions*, ser. MMCommons '15. New York, NY, USA: ACM, 2015, pp. 13–18.

[8] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. N?dellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142. [Online]. Available: http://dx.doi.org/10.1007/BFb0026683

[10] E. Chatzilari, S. Nikolopoulos, Y. Kompatsiaris, and J. Kittler, "Active learning in social context for image classification," in *9th International Conference on Computer Vision Theory and Applications (VISAPP)*, Lisbon, Portugal, January 5-8 2014.

[11] E. Chang, S. Tong, K. Goh, and C.-W. Chang, "Support Vector Machine Concept-Dependent Active Learning For Image Retrieval," *IEEE Transactions on Multimedia*, 2005.

[12] S. C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 302–309. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.44

[13] A. Freytag, E. Rodner, and J. Denzler, "Selecting influential examples: Active learning with expected model output changes," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8692. Springer, 2014, pp. 562–577.

[14] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] X. Li and Y. Guo, "Adaptive active learning for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013, pp. 859–866.

[16] Y. Yan, R. Rosales, G. Fung, and J. Dy, "Active learning from crowds," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ser. ICML '11, L. Getoor and T. Scheffer, Eds. New York, NY, USA: ACM, June 2011, pp. 1161–1168.

[17] M. Fang, J. Yin, and D. Tao, "Active learning for crowdsourcing using knowledge transfer," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 2014, pp. 1809–1815.

[18] F. Rodrigues, F. C. Pereira, and B. Ribeiro, "Gaussian process classification and active learning with multiple annotators," in *Proceedings of*

the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, 2014, pp. 433–441.

[19] L. Zhang, J. Ma, C. Cui, and P. Li, "Active learning through notes data in flickr: an effortless training data acquisition approach for object localization," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 46:1–46:8.

[20] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 97–114, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11263-014-0721-9

[21] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 557–566.

[22] E. Golge and P. Duygulu, "Conceptmap: Mining noisy web data for concept learning," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8695. Springer, 2014, pp. 439–455.

[23] W. Li, L. Niu, and D. Xu, "Exploiting privileged information from web data for image categorization," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 437–452.

[24] J. Zhuang, T. Mei, S. C. Hoi, X.-S. Hua, and S. Li, "Modeling social strength in social media community via kernel-based learning," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 113–122.

[25] P. Cui, W. Zhu, T. S. Chua, and R. Jain, "Social-sensed multimedia computing," *IEEE MultiMedia*, vol. 23, no. 1, pp. 92–96, Jan 2016.

[26] O. Papadopoulou and V. Mezaris, "Exploiting multiple web resources towards collecting positive training samples for visual concept learning," in *Proceedings of the 5th ACM International Conference on Multimedia Retrieval*, ser. ICMR '15. ACM, 2015, pp. 531–534.

[27] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 38:1–38:38, Jan. 2014.

[28] S. Liu, P. Cui, W. Zhu, and S. Yang, "Learning socially embedded visual representation from scratch," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: ACM, 2015, pp. 109–118.

[29] J. Fu, Y. Wu, T. Mei, J. Wang, H. Lu, and Y. Rui, "Relaxing from vocabulary: Robust weakly-supervised deep learning for vocabulary-free image tagging." IEEE International Conference on Computer Vision, December 2015.

[30] X. Li, C. G. M. Snoek, M. Worring, D. Koelma, and A. W. M. Smeulders, "Bootstrapping visual categorization with relevant negatives," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 933–945, June 2013. [Online]. Available: http://www.science.uva.nl/research/publications/2013/LiITM2013

[31] X. Li and C. G. Snoek, "Classifying tag relevance with relevant positive and negative examples," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 485–488.

[32] S. Kordumova, X. Li, and C. G. Snoek, "Best practices for learning video concept detectors from social media examples," *Multimedia Tools and Applications*, vol. 74, no. 4, pp. 1291–1315, 2014.

[33] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May 1998. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp;path=ASIN/026206197X

[34] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[35] T. Bart and P. Adrian, "Overview of the clef 2012 flickr photo annotation and retrieval task. in the working notes for the clef 2012 labs and workshop," Rome, Italy, 2012.

[36] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *British Machine Vision Conference*, 2011.

[37] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[39] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.

[40] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[41] V. Ng and C. Cardie, "Bootstrapping coreference classifiers with multiple machine learning algorithms," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, ser. EMNLP '03, 2003, pp. 113–120.