Active learning techniques for user interactive systems: application to image retrieval

Philippe-Henri Gosselin, Matthieu Cord

Active Learning Techniques for User Interactive Systems: Application to Image Retrieval

Philippe Henri Gosselin GOSSELIN@ENSEA.FR
Matthieu Cord CORD@ENSEA.FR
ETIS / CNRS UMR 8051, 6 avenue du Ponceau, 95014 Cergy-Pontoise, France

Abstract

Active learning methods have been considered with an increasing interest for user interactive systems. In this paper, we propose an efficient active learning scheme to deal with this particular context. An active boundary correction is proposed in order to deal with few training data. Experiments are carried out on the COREL photo database.

1. Introduction

Human interactive systems has attracted a lot of research interest in recent years, especially for content-based image retrieval systems. Contrary to the early systems, focused on fully automatic strategies, recent approaches introduce human-computer interaction (Veltkamp, 2002; Vasconcelos & Kunt, 2001).

Starting with a coarse query, the interactive process allows the user to refine his request as much as necessary. Many kinds of interaction between the user and the system have been proposed (Chang et al., 2003), but most of the time, user information consists of binary annotations (labels) indicating whether or not the image belongs to the desired category.

In this paper, we focus on the retrieval of *concepts* within a large document collection. We assume that a user is looking for a set of documents, the query concept, within an existing document database. The aim is to build a fast and efficient strategy to retrieve the query concept.

Performing an estimation of the query concept can be seen as a statistical learning problem, and more precisely as a binary classification task between the relevant and irrelevant classes (Chapelle et al., 1999). The relevant class is the set of documents within the query concept, and the irrelevant class the set of documents out of the query concept. This context defines a very specific learning problem with the following characteristics:

1. *High dimensionality.* The documents used to be represented by vectors of high dimensionality.
2. *Few training data.* At the beginning, the system has to perform a good estimation of the query concept with very few data. Furthermore, the system can not ask user to label thousands of documents, good performances are required using a small percentage of labeled data.
3. *Relevance feedback.* Due to user annotations, the training data set grows step by step during the retrieval session, so the current classification depends on the previous ones.
4. *Unbalanced classes.* The query concept is often a small subset of the database (some hundreds of documents). Thus, the relevant and irrelevant classes are highly unbalanced (up to factor 100), on the contrary to classical classification problems, where the classes have approximatively the same size.
5. *Limited computation time.* The user can not wait several hours between each feedback steps. We assume that a user can wait at most several minutes between each feedback steps.

In this paper, we propose an active learning strategy to deal with these characteristics. In section 2, we present current methods for classification, and motivations for active learning. In section 3, we focus on active learning, and present two well-known approaches: uncertainly-based sampling and error reduction. In section 4, we propose an active learning scheme to enhance the previous methods. In section 5, experiments

Appearing in *Proceedings of the workshop Machine Learning Techniques for Processing Multimedia Content*, Bonn, Germany, 2005.

are carried out on a generalist image database in order to compare the different strategies.

2. Learning for human interactive systems

2.1. Kernels and SVM

The first characteristic to deal with is the high dimensionality of feature vectors. With vectors of high dimensionality (for instance, 100 or more), artifacts appear, known as the result of the curse of dimensionality (Hastie et al., 2001). However, with the theory of kernel functions, one can reduce this curse (Smola & Scholkopf, 2002), especially if one can build a kernel function for a specific application. For instance, when distributions are used as feature vectors, a Gaussian kernel gives excellent results in comparison to distance-based techniques (Gosselin & Cord, 2004a).

Using a kernel function leads to a set of classification methods. For human interactive systems, statistical learning techniques such as nearest neighbors (Hastie et al., 2001), support vector machines (Tong & Chang, 2001; Chapelle et al., 1999; Chen et al., 2001), bayes classifiers (Vasconcelos & Kunt, 2001), have been used. We have previously shown that the SVM classification method is highly adapted to the image retrieval context (Gosselin & Cord, 2004a). Thus, we will use SVM as classification method in the following sections.

2.2. Semi-supervised learning

A natural choice for dealing with the second characteristic – the few training data – is to use semi-supervised learning techniques. Semi-supervised techniques uses labeled and unlabeled documents to compute a classification function. For instance Transductive SVM (Joachims, 1999), semi-supervised Gaussian mixtures (Najjar et al., 2003), and semi-supervised Gaussian fields (Zhu et al., 2003). However, TSVM and SSGM do not lead to significant improvements (Chang et al., 2003; Gosselin et al., 2004). Furthermore, these techniques have high computational needs in comparison to inductive techniques, and sometimes untractable. For instance, SSGF needs the inversion of a $N \times N$ matrix, where N is the size of the database. For now, semi-supervised learning techniques do not seem to be adapted to the context we are focusing on.

2.3. Active learning

Active learning is another solution to deal with few training data. The interaction between the user and the system can be exploited. The user is able to label any document in the database. The only constraint is that this user will not label a lot of documents. However, even a small labelling lead to significant improvements with active learning.

3. Active learning strategies

In this paper, we focus on the active learning scheme where a pool unlabeled examples is available. We suppose that we have a set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of documents, a set of labels $\mathbf{y} = (y_1, \ldots, y_N)$ (1 relevant, -1 irrelevant, 0 unknown), a relevance function $f_{\mathbf{y}} : \mathbf{X} \to [-1, 1]$ trained with $\mathbf{y}$, and a teacher $\tau : \mathbf{X} \to \{-1, 1\}$ that labels documents as -1 or 1. We also denote by I the set of indexes of labeled documents.

The aim of an active learning within this context is to choose the unlabeled document $\mathbf{x}$ that will enhance the most the relevance function trained with the label $\tau(\mathbf{x})$ added to the previous labeling $\mathbf{y}$. We propose to formalize this choice as the minimization of a cost function $g(\mathbf{x})$ over all unlabeled documents. Thus, according to a particular active learning method, the chosen document to label is the argument of the minimum of $g(\mathbf{x})$. We also denote by J the set of candidates, *i.e.* the indexes of unlabeled documents evaluated by $g(\mathbf{x})$.

We present here two active learning strategies: uncertainly-based sampling, which selects the documents for which the relevance function is most uncertain about, and error reduction, which aims at minimizing the generalization error of the classifier. We also present a strategy for batch selection.

3.1. Uncertainly-based sampling

This strategy aims at selecting unlabeled documents that the learner of the relevance function is most uncertain about. The first solution is to compute a probabilistic output for each documents, and select the unlabeled documents with the probabilities closest to 0.5 (Lewis & Catlett, 1994). Similar strategies have been also proposed with SVM classifier (Park, 2000), with a theoretical justification (Tong & Koller, 2001), and with nearest neighbor classifier (Lindenbaum et al., 2004).

In all cases, a relevance function may be computed. This function can be a distribution, a fellowship to a class (distance to the hyperplane for SVM), or a utility function. Thus, with some adaptation of each approach, a relevance function $f_{\mathbf{y}} : \mathbf{X} \to [-1, 1]$ is trained, where the most uncertain documents have an output close to 0. The cost function to minimize is then $g(\mathbf{x}) = |f(\mathbf{x})|$.

With such a strategy, the efficiency of a method depends on the accuracy of the relevance function estimation close to 0. This is the area where it is the most difficult to perform a good evaluation[1]. In this particular context, statistical techniques are not always the best ones, and we propose in the next section an heuristic-based correction to the estimation of $f_{\mathbf{y}}$ close to 0.

3.2. Error Reduction

Active learning strategies based on error reduction select documents that, once added to the training set, minimize the error of generalization (Roy & McCallum, 2001).

Let $P(c|\mathbf{x})$ the (unknown) probability of a document $\mathbf{x}$ to be in class c, and $P(\mathbf{x})$ the (also unknown) distribution of the documents. A training set $\mathcal{A}$ with pairs $(\mathbf{x}, c)$ sampled from $P(\mathbf{x}), P(c|\mathbf{x})$ provides the estimation $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ of $P(c|\mathbf{x})$. The expected error of generalization can be written as:

$$E_{\hat{P}_{\mathcal{A}}} = \int_{\mathbf{x}} L(P(c|\mathbf{x}), \hat{P}_{\mathcal{A}}(c|\mathbf{x}))dP(x)$$

with L a loss function which evaluates the loss between the estimation $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ and the true distribution $P(c|\mathbf{x})$.

The optimal pair $(\mathbf{x}^\star, c^\star)$ is the one which minimizes this expectation:

$$\forall (\mathbf{x}, c) \quad E_{\hat{P}_{\mathcal{A}^\star}} < E_{\hat{P}_{\mathcal{A}+(\mathbf{x},c)}}$$

with $\mathcal{A}^\star = \mathcal{A} + (\mathbf{x}^\star, c^\star)$.

Roy and McCallum propose to estimate the probability $P(c|\mathbf{x})$ with the relevance function provided by the classifier, and estimate $P(\mathbf{x})$ over $\mathbf{X}$. With a maximum loss function, the estimation of the expectation becomes, with J the set of unlabeled documents:

$$\hat{E}_{\hat{P}_{\mathcal{A}^\star}} = \frac{1}{|J|} \sum_{\mathbf{x} \in J} \left(1 - \max_{c \in \{-1,1\}} \hat{P}_{\mathcal{A}^\star}(c|\mathbf{x})\right)$$

As We don't know the label of each candidate. Roy and McCallum compute the expectation for each possible label, which finally gives the following cost function:

$$g(\mathbf{x}) = \sum_{c \in \{-1,1\}} E_{\hat{P}_{\mathcal{A}+(\mathbf{x},c)}} \hat{P}_{\mathcal{A}}(c|\mathbf{x})$$

with $\hat{P}_{\mathcal{A}}(c|\mathbf{x})$ estimated with the relevance function $f_{\mathbf{y}}(\mathbf{x})$:

$$\hat{P}_{\mathcal{A}}(c|\mathbf{x}) = \frac{c}{2}(f_{\mathbf{y}}(\mathbf{x}) + c)$$

with $f_{\mathbf{y}}(\mathbf{x})$ such as $\mathbf{y}$ encodes the training set $\mathcal{A}$.

[1] In the context of human interactive system, where only few training data is available, this is a major problem.

3.3. Batch selection

In human interactive systems, it is often necessary to select batches of new training examples. A lot of active learning strategies are made to select only one new training example. With no particular extension, these strategies can select several instances very close in the feature space. Considering the power of current classification techniques, labeling a batch of very close documents or only one of them always gives the same classification.

In the version space reduction scheme, (Tong & Koller, 2000) propose to select batches yielding minimum worst-case version space volume. However, this method requires a lot of computations making it infeasible in practice. (Brinker, 2003) proposes a fast approximation of this strategy, based on the diversity of angles between the hyperplanes in the version space. The method selects documents close to the SVM boundary one far from another, and also far from the current training data:

$I^\star = 0$
repeat
  $t = \underset{i \in J}{\operatorname{argmin}} \; (\lambda |f(\mathbf{x}_i)| + (1-\lambda) \max_{j \in I \cup I^\star} k^\star(\mathbf{x}_i, \mathbf{x}_j))$
  $I^\star = I^\star \cup \{\mathbf{x}_t\}$
until $|I^\star| = l$

with $\lambda \in [0,1]$ and $k^\star(\mathbf{x}_i, \mathbf{x}_j)$ the angle between two instances:

$$k^\star(\mathbf{x}_i, \mathbf{x}_j) = \frac{|k(\mathbf{x}_i, \mathbf{x}_j)|}{\sqrt{k(\mathbf{x}_i, \mathbf{x}_i) k(\mathbf{x}_j, \mathbf{x}_j)}}$$

The λ parameter can be used to adjust the diversity strategy contribution; $\frac{1}{2}$ is chosen as default value [2].

4. Active learning scheme

For both active learning strategies, the estimation of the relevance function is decisive. We propose in the following subsection an active correction to deal with very few training data (less than 1%). We also propose an active learning scheme with diversity for any cost function-based active learning method, and a practical solution to reduce the computation time.

4.1. Active Boundary Correction

We propose to perform the following correction to the relevance function:

$$f^\star(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}_{O_s})$$

[2] If additional knowledge is available (for instance, keywords), it can be used to tune this parameter.

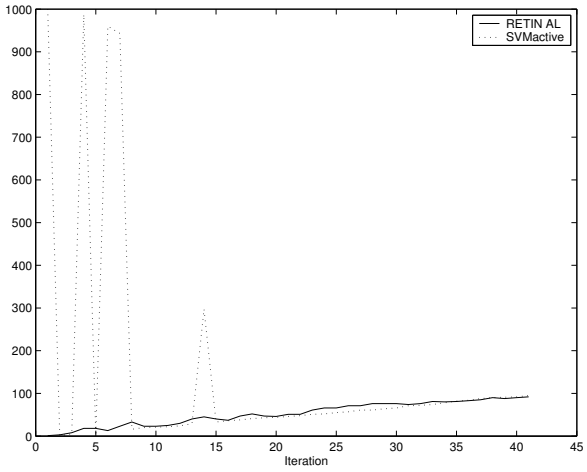

Figure 1. Values of s(t) according to feedback steps.

where $O = \text{argsort } f$, and s the correction index.

The aim of this approach is to compute s such as the "ideal" relevance function is zero at $\mathbf{x}_{O_s}$. To perform this, we propose to use the interaction with the user. The idea is to ensure that the user labels as many relevant as irrelevant documents. Then, the selected area is the most uncertain one. If the user provides a lot of relevant labels, we assume that we are close to the heart of the relevant class. Then we move the selected area further from the heart of the relevant class. If the user provides a lot of irrelevant labels, we assume that we are far from the relevant class, and then move the selected area closer to the heart of the relevant class.

We define a document $\mathbf{x}$ as close to the heart of the relevant class as $f(\mathbf{x})$ is close to 1. We change the correction index s after a sort O of the documents according to the relevance function $f(\mathbf{x})$. Small values of s means that the zero of the ideal relevance function is close to the heart of the relevant class, and vice-versa. At the feedback iteration t, we assume that the "ideal" relevance function is zero at $\mathbf{x}_{O_{s(t)}}$. We compute the new correction index $s(t+1)$ according to the labels given by the user:

$$s(t+1) = s(t) + h(\text{pos}(t), \text{neg}(t))$$

with $\text{pos}(t)$ (resp. $\text{neg}(t)$) is the number of relevant (resp. irrelevant) labels provided by the user at the feedback iteration t, and $h(a,b)$ an heuristic function. In order to get the desired behavior, we propose the following heuristic function: $h(a,b) = 2 \times (a-b)$.

At step $t = 0$, because we have no idea of the level of complexity of the searched concept, we set $s(0) = 0$.

This method is especially interesting in a context with training data, where the estimation of $f(\mathbf{x})$ is difficult. We compared this method with SVM_{active} on an image database, with 5 labels per iteration (see Experiments Section for further details). The curves in Figure 1 shows the values of $s(t)$ according to feedback steps. For the SVM_{active} method, $g(\mathbf{x}) = |f(\mathbf{x})|$ and s is such as $f(\mathbf{x}_{O_s})$ is closest to 0. Both methods have the same behavior, but SVM_{active} is very unstable during the first iterations.

This correction can be used with the active learning methods presented in the previous section. For uncertainly-based, this is simple. For error reduction, the correction is applied each time a classifier is computed. The same correction is applied, according to a new ranking of the new relevance function. A new value $f_{\mathbf{y}}(\mathbf{x}_{O_s})$ is computed for each new training set $\mathbf{y}$, and each case the relevance function is such as $f^{\star}_{\mathbf{y}}(\mathbf{x}_{O_s}) = 0$.

4.2. Incorporating diversity

In order to select batches with diversity, we propose to use the angle diversity scheme (with $g(\mathbf{x})$ instead of $|f(\mathbf{x})|$):

$I^{\star} = 0$
repeat
  $t = \underset{i \in J}{\text{argmin}} \ (\lambda g(\mathbf{x}_i) + (1-\lambda) \max_{j \in I \cup I^{\star}} k^{\star}(\mathbf{x}_i, \mathbf{x}_j))$
  $I^{\star} = I^{\star} \cup \{\mathbf{x}_t\}$
until $|I^{\star}| = l$

We normalize the cost function $g(\mathbf{x})$ before performing this step, in order to get values in the same interval than the cosines value interval. We observe that a diversity technique allows to select documents for labeling which are not close one to another. It is decisive in image retrieval context.

4.3. Reduce the computation time

In order to propose labels to a human expert in a reasonable time, all unlabelled documents can not be evaluated. We propose to restrict the evaluation of $g(\mathbf{x})$ to a set of *candidates*. We denote by J the set of the indexes of these candidates. We propose to reduce the set of unlabeled documents to the m closest documents to the boundary. For methods using boundary correction, the correction is made before the selection of the candidates. Thus, the boundary correction also changes the choice of the candidates.

5. Experiments

5.1. Evaluation Protocol

Tests are carried out on the generalist COREL photo database, which contains more than 50,000 pictures. To get tractable computation for the statistical evaluation, we randomly selected 77 of the COREL folders, to obtain a database of 6,000 images. To perform interesting evaluation, we built from this database 50 concepts[3]. Each concept is built from 2 or 3 of the COREL folders. The concept sizes are from 50 to 300. The set of all the concepts covers the whole database, and each image of the database is at least in one of the concepts, and at most in 5 different concepts.

Color and texture distributions are used as feature vectors, the kernel is a Gaussian kernel with a χ^2 distance, and the classification method is SVM.

We simulate the use of a image retrieval system. For one retrieval session, we assume that the user chooses one picture in the database for the concept he is looking for. A concept and a picture from this concept are randomly chosen for each new simulated retrieval session. In practice, this is done when the user brings one picture of its own. Then, the system computes features of this picture, and labels as relevant the closest picture. Other techniques could be used for this, for instance using keywords.

Thus, the simulated retrieval session starts with one relevant picture. Next, the system asks the active learner for 5 images to label. These images are labeled according to the desired concept, and the system asks again the active learning for 5 other images to label, using the 6 current labels. These feedback steps are iterated 10 times, and at the end of the retrieval session, the training set has 51 labels. Using these labels, a classification of the database is performed. The error of classification and the number of pictures in the concept within the 100 most relevant ones (top-100) are computed. We simulate 1,000 retrieval sessions for each active learning method. The error of classification and the top-100 are averaged over all retrieval sessions.

5.2. Comparison

Results are reported in Figure 2 with a full set of candidates (all unlabeled documents), and in Figure 3 with a reduced set of 100 candidates. The first line shows the active learning method. The "None" method means that no classification is performed, only the distance between an image and all other ones is computed. The second line shows the Top-100 for each method. For the "None" method, this result means that the average probability to find an image within the same concept than the considered image in the 100 nearest neighbors is 16%. The third line shows the average classification error. The last line shows the average computation time for a retrieval session.

The error reduction method (ER) gives better results than the uncertainly-based method (UB) (*cf.* Fig. 2). However, much more computation time is required (*cf.* Fig. 2) for ER, and it does not well support the reduction of the set of candidates in terms of classification error (*cf.* Fig. 3). The angle diversity improvement (AD) increases performances in all cases. This shows that, even if this method was built especially for UB, it can be used with others active learning methods. Furthermore, its costs in terms of computation time is small. The active boundary correction (BC) also increases performances in all cases. It has also a negligible cost in terms of computation time, and well supports the reduction of the set of candidates. Note that the improvement is much more significant for UB than for ER. Globally, the reduction of the set of candidates is interesting for all strategies, expect for ER without BC. For comparable performances, the computation time is divided by 10. Finally, the most efficient strategy is the BC+UB+AD strategy, which combines boundary correction, uncertainly-based, and angle diversity.

6. Conclusion

In this paper, we proposed active learning strategies for interactive search systems. We introduced an algorithm to correct the boundary of a classifier function, in order to improve the active learning efficiency. We proposed an active learning scheme combining different techniques, and a method to reduce the computation time. These strategies have been compared on a generalist image database. Results show that the efficiency of the proposed combinations, especially our strategy using boundary correction, uncertainly-based, and angle diversity. These results also show that the computation time can be significantly reduced using the proposed method without dramatical degradation of performances.

[3] A description of this database and the 50 concepts can be found at: http://www-etis.ensea.fr/~cord/data/mcorel50.tar.gz. This archive contains lists of image file names for all the concepts.

Method	None	UB	ER	UB+AD	ER+AD	BC+UB+AD	BC+ER+AD
Top-100	16	28	33	31	34	**36**	35
Classification Error	–	8.2%	6.7%	6.7%	4.3%	**2.5%**	3.0%
Time	0.07s	0.41s	600s	60s	700s	60s	700s

Figure 2. Average Top-100, classification error and execution time for each active learning method, 10 feedbacks steps, 5 labels per step, with full set of candidates (all unlabeled documents). Legend: UB = Uncertainly-Based, ER = Error Reduction, AD = Angle diversity, BC = Boundary Correction.

Method	None	UB	ER	UB+AD	ER+AD	BC+UB+AD	BC+ER+AD
Top-100	16	28	32	31	34	**36**	35
Classification Error	–	8.4%	19.3%	6.9%	13.7%	**2.6%**	3.1%
Time	0.07s	0.41s	5.4s	2.4s	6.8s	2.4s	6.8s

Figure 3. Same protocol as Fig.2 with a reduced set of 100 candidates.

References

Brinker, K. (2003). Incorporating diversity in active learning with support vector machines. *International Confenrence on Machine Learning* (pp. 59–66).

Chang, E., Li, B. T., Wu, G., & Goh, K. (2003). Statistical learning for effective visual information retrieval. *IEEE International Conference on Image Processing*. Barcelona, Spain.

Chapelle, O., Haffner, P., & Vapnik, V. (1999). Svms for histogram based image classification. *IEEE Transactions on Neural Networks*, *9*.

Chen, Y., Zhou, X., & Huang, T. (2001). One-class svm for learning in image retrieval. *International Conference in Image Processing (ICIP'01)* (pp. 34–37). Thessaloniki, Greece.

Gosselin, P., & Cord, M. (2004a). A comparison of active classification methods for content-based image retrieval. *International Workshop on Computer Vision meets Databases (CVDB), ACM Sigmod*. Paris, France.

Gosselin, P., & Cord, M. (2004b). RETIN AL: An active learning strategy for image category retrieval. *IEEE International Conference on Image Processing*. Singapore.

Gosselin, P., Najjar, M., Cord, M., & Ambroise, C. (2004). Discriminative classification *vs* modeling methods in CBIR. *IEEE Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Brussel, Belgium.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The element of statistical learning*. Springer.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. *Proc. 16th International Conference on Machine Learning* (pp. 200–209). Morgan Kaufmann, San Francisco, CA.

Lewis, D., & Catlett, J. (1994). Heterogenous uncertainty sampling for supervised learning. *International Conference on Machine Learning* (pp. 148–56).

Lindenbaum, M., Markovitch, S., & Rusakov, D. (2004). Selective sampling for nearest neighbor classifiers. *Machine Learning*, 54(2):125–152.

Najjar, N., Cocquerez, J., & Ambroise, C. (2003). Feature selection for semi supervised learning applied to image retrieval. *IEEE ICIP*. Barcelena, Spain.

Park, J. (2000). On-line learning by active sampling using orthogonal decision support vectors. *IEEE Neural Networks for Signal Processing*.

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *International Confenrence on Machine Learning*.

Smola, A., & Scholkopf, B. (2002). *Learning with kernels*. MIT Press, Cambridge, MA.

Tong, S., & Chang, E. (2001). Support vector machine active learning for image retrieval. *ACM Multimedia*.

Tong, S., & Koller, D. (2000). Support vector machine active learning with applications to text classification. *International Conference on Machine Learning*, 999–1006.

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to classification. *The Journal of Machine Learning Research*, 2:46–66.

Vasconcelos, N. (2000). *Bayesian models for visual information retrieval*. Doctoral dissertation, Massachusetts Institute of Technology.

Vasconcelos, N., & Kunt, M. (2001). Content-based retrieval from image databases: current solutions and future directions. *International Conference in Image Processing* (pp. 6–9). Thessaloniki, Greece.

Veltkamp, R. (2002). *Content-based image retrieval system: A survey* (Technical Report). University of Utrecht.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *International Conference on Machine Learning*.