Hadoop distributed trademarks retrieval system

Tao Jiang (SID: 1155132038) Lifang Liu (SID: 1155128181) Zichao Dong (SID: 1155129705) Zhende Zhuang (SID: 1155129566)
The Chinese University of Hong Kong
1155132038@link.cuhk.edu.hk

Abstract

This project introduces a trademark retrieval system based on Hadoop. We design and implement the system includes two main parts——trademark search system and trademark update system. In the trademark search system, when users upload their images, the system will return some similar trademarks. The trademark depository can be processed to be a feature library stored in the HDFS, and then use the MapReduce algorithm to match images features with the feature library and receive the results. The trademark update system is used to continually acquire newly trademarks and store them in the database, and then refine the existing classifier by using the algorithm of active learning.

1. Introduction

When people decide to found a company or develop a product, they will create a relevant trademark as their symbol to make other people recognize them easier. However, with the quantities of trademarks increasing, people need to attach importance to the particularity of the trademarks to ensure users will not be confused deliberately. Therefore, we build up the law, which require us to pass the plagiarism test, to protect the trademarks and the privileges of their owner. Then it last for a long time that we need to employ a professional helper like the intellectual property lawyer, that will cost a lot of money and time to help us judge whether our design of trademarks is suitable. That's why we came up an idea to help people search their designs in a platform that help them find out the existing similar trademarks and thus they can judge whether their design is suitable or not by themselves.

1.1. Trademark retrieval system

A trademark is a type of intellectual property consisting of a recognizable sign, design, or expression which identifies products or services of a particular source, which has important economic and social significance. Trademark infringement and squatting can cause huge economic losses to a company. Therefore, many companies spend huge sums of money on trademark protection and monitoring every year. But traditional trademark protection and monitoring relies on intellectual property agencies who hire professional agents to search for similar trademarks in the trademark database, which creates high labor costs and is very inefficient. The traditional trademark search relies on the human eye to judge the trademark similarity, and the Vienna classification[12] is used to assist the judgment of the semantics of the trademark content. An agent takes a long time each day to browse trademarks in the database against the screen to find similar trademarks. People are eager to replace humans with machines to complete the job.

Trademark retrieval is a practicable application of content-based image retrieval (CBIR), which is a long-standing research topic in the computer vision field. In the early 1990s, the study of CBIR truly started. Images were indexed by the visual cues, such as texture and color, and a myriad of algorithms and image retrieval systems have been proposed. In recent years, convolutional neural network (CNN) achieves outstanding performance in specific retrieval tasks, which brings hope to the trademark retrieval problem.

1.2. Active learning

Active learning, which can be regarded as an iterative optimization procedure, plays a key role to construct a refined training set to improve the classification performance[15] in a variety of applications, such as text analysis, image recognition, social network modeling, etc. The past decades have witnessed a rapid development of cheaply collecting huge data, providing the opportunities of intelligently classifying data using machine learning techniques. In classification tasks, a sufficient amount of labeled data is obliged to be provided to a classification model in order to achieve satisfactory classification accuracy. However, annotating such an amount of data manually is time consuming and sometimes expensive.

Hence it is wise to select fewer yet informative samples for labeling from a pool of unlabeled samples, so that a

classification model trained with these optimally chosen samples can perform well on unseen data samples[16]. If we select the unlabeled samples randomly, there would be redundancy and some samples may bias the classification model[17], which will eventually result in a poor generalization ability of the model. Active learning methodologies address such a challenge by querying the most informative samples for class assignments and the informativeness criterion for active sampling[18] has been successfully applied to many data mining and machine learning task. Although active learning has been developed based on many approaches, the dream to query the most informative samples is never changing.

1.3. Motivation

In practical application scenarios, a trademark retrieval system needs to ensure retrieval speed and quality, which leads to huge computation and memory requirements. Therefore, we plan to develop a distributed trademark retrieval system to solve this problem.

Meanwhile, since the quantities of trademarks increase steadily, we need to update our trademark depository frequently. In this situation, if we heavily rely on manually annotation, we need to pay more money and time for maintaining our database. Therefore, applying active learning method can help us decrease the cost as much as possible.

2. Related work

2.1. Related work in CBIR

The SIFT-based methods mostly rely on the BoW model. The introduction of the scale-invariant feature transform (SIFT)[4] makes the BoW model feasible.[8]

The CNN-based retrieval models usually compute compact representations and employ the Euclidean distance or some approximate nearest neighbor (ANN) search methods for retrieval[5][14][1], which has outperformed the SIFT-based methods.

After feature extraction, a k-nearest neighbor (KNN)[2] or ANN method can be applied. The KNN method can achieve the highest precision with expensive computation costs hence several ANN-based methods[3][11][6][10] are being used more and more.

2.2. Related work in active learning

Generally speaking, there are two main sampling criteria in designing an effective active learning algorithm, that is, informativeness and representativeness.[19] Informativeness represents the ability of a sample to reduce the generalization error of the adopted classification model and ensures less uncertainty of the classification model in the next iteration. Representativeness decides whether a sample can exploit the structure underlying unlabeled data [20], and many applications have paid much attention on such information [21],[22]. Most popular active learning algorithms deploy only one criterion to query the most desired samples. [23]

The approaches drawing on informativeness attracted more attention in the early research of active learning. Typical approaches include: 1) query-by-committee, in which several distinct classifiers are used, and the samples are selected with the largest disagreements in the labels predicted by these classifiers[24]; 2) max-margin sampling, where the samples are selected according to the maximum uncertainty via the distances to the classification boundaries[25]; and 3) max-entropy sampling, which uses entropy as the uncertainty measure via probabilistic modeling[26]. The common issue of the above active learning methods is that they may not be able to take full advantage of the information of abundant unlabeled data, and query the samples merely relying on scarce labeled data. Therefore, they may be prone to a sampling bias.

3. Methodology

Because the volume of the existing trademark data is huge and is also increasing incessantly which refers to the high velocity of data generation. We decide to propose an trademark retrieval system based on Hadoop (or Spark) and use active learning to strengthen its performance. Also, the data obtained by crawler can be in different size and format etc., which require us preprocess the variety of the data.

3.1. Hadoop distributed trademark retrieval system

We trained a multi-label classification model on the 30-million Chinese trademark database, using the Vienna Classification codes as labels. The model uses densenet161[7] as the backbone and uses SCDA method[13] to obtain better representation capabilities. Features from the hidden layer[9] proposed in this paper are extracted for retrieval, which can do dimensionality reduction and feature aggregation at the same time. Finally, we can use Hadoop to calculate nearest neighbors distributedly.

3.2. Active learning in update system

Initially, even though the volume of our dataset is huge, there are only limited images which are annotated. The idea of our method could be concluded in two aspects: uncertainty and representativeness.

Firstly, uncertainty means that we should find some hard examples to get annotated. To achieve that, we directly use the output of our deep neural network. As mentioned above, our numeral network could do the multi-label classification task while we can get the output score as the degree of confidence of each class. Thus, we can select the images whose score of its real label is lower than the threshold we set. By doing so, we could find the examples which are hard to be classified and then add them in our database after querying their labels by human efforts.

Secondly, representativeness means that the annotated areas need to bear useful characteristics or features for as many unannotated images as possible, since we can only store limited number of images in our database. Our method to solve that is clustering. To be precise, for each class we need to find some positive examples to add in the databases. So firstly, we find the images which are close to the center of the cluster of this certain class. After that, we have found some representative examples from the dataset. Apart from that, different from simple classification tasks, our goal is to find the most similar items in the dataset. So, we also select some images which are relatively far from the center of cluster which intuitively stand for some 'strange' images. To be precise, strange means that image have some special features. That is of vital importance in image retrieval as uncommon features is important in judging whether two images are similar or not. Further, this increasing of feature variety could make our algorithm much robust by refining our database.

4. Prospective results

We give the prospective results as follow including the overall results and the phased outcome.

4.1. Overall results

A distributed trademark search system includes two important parts, one of which is the main structure that we call it search system, the other of which is the update structure using algorithm of active learning.

In the system, when users input an image, our system will output the trademarks which are similar to the input image.

4.2. Main structure outcome

The working steps can be listed as follow:

a) The trademark depository can be processed in the master machine to be a feature database
b) The database can be preprocessed by Hadoop
c) The database can be distributed to all the slave machines
d) The users can upload an image
e) The image can be changed into feature map using the pre-trained model
f) The feature map can be distributed to all the slave machines
g) The slave machines can use the received feature map to compare with the database and output the similar images which are now in the format of feature
h) The search results on each slave machines can be aggregated to the master machine by using the MapReduce algorithm
i) The master machine can return the search result to users

The deliverables can refer the table 1.

Table 1: The work procedures and their prospective results of the main structure.

Main structure	
Work	Prospective results
Put the trademark depository in the pre-trained model	A feature vector database of all the trademarks
Use Hadoop to preprocess the database	Change the database format Distributed database in the slave machines
preprocess the image from the users	A feature map of the input image
Distribute the feature map to the slave machines	Distributed feature map in the slave machines
The slave machines compare the input feature map with the database	Find out the suitable results in each slave machine
Summary the suitable results and return them to the master machine	Aggregate the search results to the master machine
Output the search results	Return the trademarks which are similar to the input image

4.3. Update structure outcome

Periodically, we will collect some new trademarks by crawler or other methods. After the collection, we will label them by active learning whose steps can be listed as follow:

a) Preprocess the new data (like resize, unduplicated, compress etc.)
b) Use the existing classifier to label them and find out the uncertain and representative points
c) Manually annotate the uncertain points
d) Use the new dataset to train a new classifier
e) Use the new classifier to find out the uncertain and representative points
f) Repeat step c, d, e until satisfied the setting condition
g) Output a new classifier

The deliverables can refer table 2

Table 2: The work procedures and their prospective results of the update structure.

Update structure	
Work	Prospective results
Preprocess the new data	The data contrast between the raw data and the processed data
Use the existing classifier to label them and find out the uncertain and representative points	The uncertain and representative points (show in image)
Manually annotate the uncertain points and use the new dataset to train a new classifier	An updated classifier
Repeat the above procedure to find a latest classifier	A latest classifier

5. Evaluation

Initially, the procedures listed above should be finished and work properly. We will set the prospective input and output to check every procedure and also give the essential checkpoint.

Then, we will test the structure by the following evaluation methods and the test will be conducted in the form of comparative experiments.

5.1. Experiments in trademark retrieval system

5.1.1 Dataset

Our system is based on the Chinese Trademark database consisting of 30 million trademarks, and due to the limitation of computing and memory resources, we randomly extract 5,000 trademarks as dataset for experiments.

5.1.2 Time cost

Situation 1: different query

Use different image to test the system and record the search time and the quantity of the results

Situation 2: different quantity of slave machines

Use 3, 4, 5 slave machines to test the same query and record their search time.

5.1.3 Recall and accuracy

When we get the results, we will compare them with our setting results which would be labeled before the search, and thus we can calculate the recall of the system

5.1.4 Adaptability of increasing data

Increase the data until the existing slave machines cannot handle them and try to add more machine to test whether it can work

5.2. Experiments in active learning update system

5.2.1 Accuracy and performance analysis

Compare the accuracy between the previous model and the updated model

5.2.2 Comparing to stochastic subsample update

Give some fix queries and search them in each update which can support us give a statistic analysis like the accuracy and recall change after each update

6. Schedule

1. 16 Sep, 2019 to 21 Sep, 2019——Determine the research content of the project and write a project proposal.
2. 22 Sep, 2019 to 28 Sep 2019——Understand and learn the tools and materials needed to be used in this project.
3. 29 Sep, 2019 to 5 Oct, 2019——Set up the data running environment and preprocess the training data.
4. 6 Oct, 2019 to 1 Nov, 2019——Implement the trademarks search system.
5. 2 Nov, 2019 to 22 Nov, 2019——Finish writing code for the trademarks update system and implement it.
6. 23 Nov, 2019 to 25 Nov, 2019——Prepare project presentation.
7. 26 Nov, 2019 to 6 Dec, 2019——Write final report and submit report file, presentation file and source code.

References

[1] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, “Deep image retrieval: Learning global representations for image search,” in ECCV, 2016.

[2] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.

[3] Bentley, J. L. (1975). "Multidimensional binary search trees used for associative searching". Communications of the ACM. 18 (9): 509–517. doi:10.1145/361002.361007.

[4] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,”International journal of computer vision, vol. 60, no. 2,pp. 91–110, 2004.

[5] G. Tolias, R. Sicre, and H. Jégou, “Particular object retrieval with integral max-pooling of cnn activations,” in ICLR, 2016

[6] H. Jegou, M. Douze, and C. Schmid, “Product Quantization for Nearest Neighbor Search,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117-128, Jan. 2011.

[7] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[8] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in ICCV, 2003
[9] Lin K, Yang H F, Hsiao J H, et al. Deep learning of binary hash codes for fast image retrieval[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 27-35.
[10] Malkov, Yury; Yashunin, Dmitry (2016). "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs". arXiv:1603.09320
[11] Rajaraman, A.; Ullman, J. (2010). "Mining of Massive Datasets, Ch. 3".
[12] Vienna Agreement Establishing an International Classification of the Figurative Elements of Marks (Authentic text) (as amended on October 1, 1985)
[13] Wei X S, Luo J H, Wu J, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2868-2881.
[14] Y. Kalantidis, C. Mellina, and S. Osindero, “Cross-dimensional weighting for aggregated deep convolutional features,” in ECCV,2016.
[15] M.Yu, L.Liu, and L.Shao, “Structure preserving binary representations for RGB-D action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., Doi: 10.1109/TPAMI.2015.2491925.
[16] Q. Zhu, J. Mai, and L. Shao, “A fast single image haze removal algorithm using color attenuation prior,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 3522–3533, Nov. 2015.
[17] T. Liu and D. Tao, “Classification with noisy labels by importance reweighting,” IEEE Trans. Pattern Anal. Mach. Intell., Doi: 10.1109/TPAMI.2015.2456899.
[18] C. Xu, D. Tao, and C. Xu, “Multi-view intact space learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 12, pp. 2531–2544, Dec. 2015.
[19] S. J. Huang, R. Jin, and Z. H. Zhou, “Active learning by querying informative and representative examples,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 10, pp. 1936–1949, Oct. 2014.
[20] B. Settles, “Active learning literature survey,” Dept. Comput. Sci., Univ. Wisconson–Madison, Madison, WI, USA, Tech. Rep. 1648, 2009.
[21] X. Lu, Y. Wang, and Y. Yuan, “Sparse coding from a Bayesian perspective,” IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 6, pp. 929–939, Jun. 2013.
[22] X. Lu, H. Wu, and Y. Yuan, “Double constrained NMF for hyper- spectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 5, pp. 2746–2758, May 2014.
[23] X. Lu, Y. Wang, and Y. Yuan, “Graph-regularized low-rank representation for destriping of hyperspectral images,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 7, pp. 4009–4018, Jul. 2013.
[24] D. Vasisht, A. Damianou, M. Varma, and A. Kapoor, “Active learning for sparse Bayesian multilabel classification,” in Proc. ACM SIGKDD, New York, NY, USA, 2014, pp. 472–481.
[25] M. Wang and X.-S. Hua, “Active learning in multimedia annotation and retrieval: A survey,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 2, pp. 1–21, 2011.
[26] J. Zhu and M. Ma, “Uncertainty-based active learning with instability estimation for text classification,” ACM Trans. Speech Lang. Process., vol. 8, no. 4, pp. 1–21, 2012.