HOW MANY MORE IMAGES DO WE NEED?
PERFORMANCE PREDICTION OF BOOTSTRAPPING FOR IMAGE CLASSIFICATION

Elisavet Chatzilari[⋆†] *Spiros Nikolopoulos*[⋆] *Yiannis Kompatsiaris*[⋆] *Josef Kittler*[†]

[⋆]Centre for Research & Technology Hellas - Information Technologies Institute
[†]Centre for Vision, Speech and Signal Processing, University of Surrey Guildford, UK

ABSTRACT

Motivated by the recently introduced scalable concept detection challenge that requires classifiers for hundreds or even thousands of concepts, the objective of this work is to predict the cases where the enhancement of an initial classifier with additional training images is not expected to provide significant improvements. To facilitate this objective, we need a model for predicting the performance gain of a bootstrapping process prior to actually applying it. In order to train this model, we propose two features; the initial classifier's *maturity* (i.e. how close is the current hyperplane to the optimal) and the oracle's *reliability* (i.e. how reliable is the oracle in providing the correct labels of new training data). Thus, the contribution of our work is on proposing a method that is able to exploit the correlation between the expected performance boost and these two indicators. As a result, we can considerably improve the scalability properties of such bootstrapping processes by concentrating on the most prominent models and thus reducing the overall processing load.

Index Terms— scalable concept detection, image classification, performance prediction, bootstrapping

1. INTRODUCTION

An important factor that affects the quality of supervised classifiers is the size of the training set. Aiming to improve the performance of the classifiers, the bootstrapping technique was designed to augment the training set with additional training samples [1, 2]. In a similar endeavour, active learning was later proposed aspiring to minimize the annotation cost by enhancing the initial training set with the most informative samples [3]. Relying on the basic principles of active learning but aiming to introduce also the oracle's confidence as an important factor, the most recent works in this field propose to enhance the training set with the most misclassified negatives [4], or with the most prominent positive examples by jointly considering the oracle's confidence and the sample's informativeness when selecting new samples [5].

An interesting aspect of most bootstrapping approaches evaluated so far, is that they have been tested using very few examples to train the initial model, many of which even start with just two [6, 7, 8]. However, with the widespread adoption of crowdsourcing, collecting medium scale datasets with ground truth annotations has become a realistic scenario for a rather high number of concepts. This is particularly important in the context of automatic concept detection, given that the annotations obtained via crowdsourcing has proven to be of comparable quality to the annotations of experts [9]. Prominent examples of such datasets are the 25000 images used for the 2012 imageCLEF photo annotation task [10], which were annotated for 94 concepts by using Amazon's Mechanical Turk (MTurk) service, as well as the 14 million images provided by ImageNET [11], which is currently the largest annotated image database consisting of 21841 concepts.

Considering the scale of such datasets, it is natural to wonder whether bootstrapping techniques could still benefit the cases where the initial training set consists of a few hundreds instances rather than just a couple. More specifically, it becomes particularly important to examine the learning capacity of the initial model with the aim to identify its saturation point, i.e. a point where continuing adding more samples does not really cause the model to perform better. It may be the case that for certain types of concepts the saturation point can already be reached using the few hundred examples included in the initial training set. In this case, the model can be considered to have reached a level of *maturity* that adding more training samples would only result in marginal performance changes. In our work, we define the model *maturity* to be the distance of the current model from the optimal hyperplane. However, since this distance can not be directly calculated, we approximate its value using the classification performance of the model applied on a large set of images with ground truth annotations.

In addition to the model's *maturity*, another critical aspect that is expected to determine whether adding more samples will cause the model's performance to improve is the oracle's *reliability*, which depends on how accurately the oracle can label new training data (i.e. how accurately they have been annotated through active (e.g. MTurk), or passive (e.g. flickr tags) crowdsourcing). This is due to the fact that adding a set of examples, the majority of which has been falsely labelled by an unreliable oracle, will most probably cause the model to deteriorate. The oracle's *reliability* can be considered as

an indicator of how much we trust the oracle's decisions and, among others, depends on the nature of the examined concept. Indeed, there are some inherently ambiguous concepts that are not easy to distinguish using words (e.g. palm-hand and palm-tree) and there are others that can be pretty clearly described by linguistics (e.g. snow). Thus, motivated by the expectation that the oracle will be more accurate when labelling simpler rather than more ambiguous concepts, we formulate the oracle's *reliability* as a function of the concept of interest. More specifically, *reliability* is approximated by the success rate of the oracle in labelling a set of samples with ground truth annotations, which is calculated using the average precision metric.

Based on the above, we propose the utilization of these two features, i.e. the model's *maturity* and the oracle's *reliability*, for predicting the performance gain expected by enhancing the models. Then, based on these predictions, we can select to enhance only the most prominent models, avoiding in this way the computational cost that would be required to enhance the full set of models (Fig. 1). This is particularly useful in the context of recent trends in the image classification domain, where the scalability of methods to numerous concepts is now considered an important element of the proposed solutions. For example, in the ImageCLEF competition [12], the organizers introduced this scalability requirement by adding the concept as an input to the participants' systems rather than giving a pre-defined vocabulary of concepts, while in the ImagneNet competition they had to classify images with respect to a vocabulary of 1000 concepts.

There are only a few works in the literature dealing with the prediction of the expected learning performance. The authors of [13] investigate both theoretically and empirically when effective learning is possible from ambiguously labelled images. They formulate the learning problem as partially-supervised multi-class classification and provide intuitive assumptions under which they expect learning to succeed. On the other hand, we formally formulate the expected performance gain as a function of two pre-computed features and estimate this function using a regression model. More closely related to our approach is the work presented in [14], where the objective is to predict the performance difference between automatically created and manually annotated datasets. On the contrary, our approach is designed for the bootstrapping technique and its scope is to reduce both the annotation effort and the computational complexity, by intelligently selecting the most prominent concepts for which bootstrapping is expected to be beneficial.

2. SELECTIVE MODEL RETRAINING

As already mentioned, the purpose of our work is to examine the correlation of the expected performance gain with the *maturity* of the model and the *reliability* of the oracle, in order to build a classifier trained on these two aspects. However before expressing the performance boost of the initial classifier as a function of the oracle's *reliability* and the classifier's *maturity*, we should first define the approach followed for measuring these quantities.

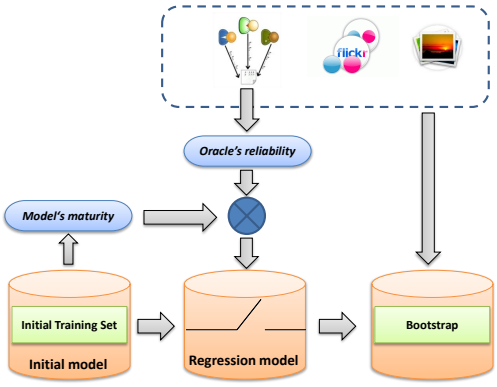

Fig. 1. System Overview

2.1. Oracle reliability

The *reliability* of the oracle R is defined on a per concept basis and indicates the quality of the oracle. A less reliable oracle will tend to make more mistakes feeding the classifiers with wrongly selected images and misleading them from their optimal target. In order to model this property, we quantify the oracle *reliability* to be the performance of the oracle as it is measured by average precision. More specifically, the oracle is asked to rank the images of a manually annotated dataset for the examined concept and the average precision is calculated based on this ranking.

2.2. Model maturity

A more mature classifier, i.e. closer to the optimal model, is expected to exhibit small fluctuations in terms of performance, even if it is guided accurately, since it is closer to its saturation point. On the other hand, an immature model has more potential in increasing its performance although it would need more accurate guidance as it is expected to be highly susceptible to false positives. In this case, the *maturity* of the model M is essentially the quality of the initial classifier, which can be measured by its performance tested on a manually annotated dataset and quantified by the average precision metric.

2.3. Regression model

Based on the assumption that the performance gain g is correlated both with the *maturity* M of the initial classifier and the *reliability* R of the oracle, we propose to train a regression model using these two features (i.e. M and R):

$$g = f(M, R) \tag{1}$$

In the training phase, we provide pairs $\{g(i), (M(i), R(i))\}$ for every concept c_i and the objective is to map the features (M, R) to the performance gain g by estimating the mapping

function f. The two proposed features, *reliability* and *maturity* are computed for every concept as explained in Sections 2.1 and 2.2 respectively, by applying three fold cross validation on a manually annotated training set. In order to compute the output values $g(i)$, the initial classifiers are trained on the manually annotated training set. Additional training samples are selected by a pool of candidates using the bootstrapping technique and the enhanced models are trained using the initial training set augmented with the additional training samples. Afterwards, both the initial and the enhanced classifiers are applied on a manually annotated evaluation set and their performance, $AP_{init}(i)$ and $AP_{fin}(i)$ respectively, is estimated by the average precision metric. Finally, the performance gain is calculated to be the performance difference between the enhanced and the initial classifiers:

$$g(i) = AP_{fin}(i) - AP_{init}(i) \quad (2)$$

In the testing phase, given a new unseen concept c_j and an initial classifier recognizing this concept, we compute as previously the proposed features $\{M(j), R(j)\}$, while the expected prediction gain $\hat{g}(j)$ is computed by applying the mapping function f. Based on the predicted gain, we can choose whether it is worthwhile to further enhance the classifier for the specific concept or retain the initial classifier.

3. EXPERIMENTS

3.1. Datasets and implementation details

Two datasets were employed for the purpose of our experiments. The imageCLEF dataset IC [10], annotated for 94 concepts, was used as the manually annotated dataset and was split in three parts; $T1$, $T2$ and $Test$, consisting of 5k, 10k and 10k images respectively. The MIRFLICKR-1M dataset S [15] constitutes the pool of user-tagged images out of which 500 images are selected for each concept to act as the positive examples enhancing the initial training set during the bootstrapping approach. The bootstrapping technique which was presented in [5] was employed in our experiments. The code and the data for the following experiments are available at [1, 2].

3.2. Impact of maturity and oracle reliability

In this experiment we investigate empirically how the classifier *maturity* and the oracle *reliability* correlate with the performance gain by artificially simulating different levels of *reliability* for the oracle and examining the susceptibility of classifiers with various levels of *maturity* to noisy examples (i.e. false positive). For this experiment the IC dataset is used. Initially, the classifiers are trained using the 5k images of the $T1$ set. Afterwards, in order to simulate an unreliable oracle, the initial training set is augmented with a combination of true and false positive images from the $T2$ set. The final classifiers are retrained using the augmented dataset and are evaluated on the $Test$ set. We consider five augmented datasets, each one constructed by an oracle that adds samples with $100\%, 80\%, 67\%, 50\%$ and 0% accuracy, simulating this way different levels of the oracle's *reliability*.

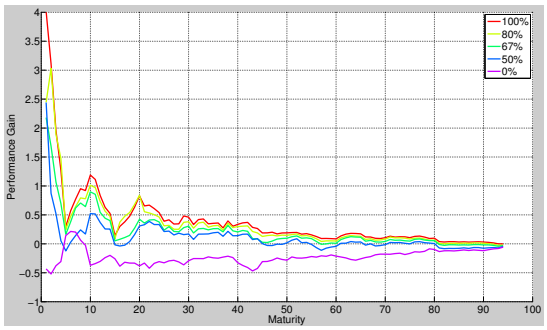

Fig. 2. The effect of the oracle reliability and the classifiers' maturity to the performance gain

In Figure 2, we plot the performance gain between the enhanced and initial classifiers with respect to the *maturity* of the initial classifier. Initially, for each concept c_i of IC the *maturity* $M(i)$ is calculated. Then the examined oracle proposes new training samples, the enhanced classifiers are trained and the performance gain $g(i)$ is calculated. Finally, for every level of reliability we have a set of 94 points described by $(M(i), g(i))$ pairs. For better visualization, we applied a smoothing filter on the data points and produced an interpolated line for each oracle. The expected correlation becomes obvious if we make the following observations; (a) as the percentage of noisy data included in the augmentation dataset increases, the classifiers' performance deteriorates (higher decrease in performance for the magenta line, i.e. adding 100% false positive examples, than the black line, i.e. adding 50% true and 50% false positive examples), and (b) the classifiers that exhibit a high level of *maturity*, are not affected by the addition of the augmentation sets, neither positively when the oracle is perfectly reliable (i.e. red line) nor negatively when adding only false positive examples (i.e. magenta line). More specifically, there are only small fluctuations of the performance gain when the *maturity* of the classifier is high (e.g. over 50%). All the above verify our expectation that the performance gain is correlated with both the *maturity* of the classifier and *reliability* of the oracle. This justifies the selection of these two features to train the proposed regression model for predicting the expected performance gain.

3.3. Performance gain prediction

Our goal in this section is to verify whether the proposed regression model can effectively predict the performance gain of bootstrapping. For this purpose we learn the parameters of function f as specified in Section 2 using Support Vector Regression. The initial classifiers are trained using the com-

[1] http://mklab.iti.gr/project/PerformancePrediction
[2] https://github.com/ehatzi/PerformancePrediction

bination of $T1$ and $T2$ datasets and afterwards, classifiers are enhanced by the images of the S dataset using the approach presented in [5]. The different concepts (i.e the 94 concepts of IC) constitute the instances for training the regression model. In order to predict the expected gain $\hat{g}(i)$ for an instance i, the leave one out protocol is used (i.e. the regression model is trained on the 93 concepts and it is used to predict the expected gain $\hat{g}(i)$ of the remaining concept i). The leave-one-out estimate is selected to ensure that the proposed approach can generalize to different concepts than the ones used to train the regression model f. In order to fit the parameters of the function f, we tested two different modelling approaches, an e-SVR and a nu-SVR regression model, while both linear and RBF kernels were considered. The best performing approach, the e-SVR regression model with an RBF kernel, was chosen using cross validation.

In order to visualize the results, the concepts are ranked based on the predicted gain $\hat{g}$, which is computed by applying the regression function f. Then, the cumulative actual gain g is calculated for every concept in the following way. If we denote as $c'_1, c'_2, ..., c'_N$ the sorted concepts so that $\hat{g}(c'_k) > \hat{g}(c'_{k+1})$, we define the cumulative gain function $Cg(k)$ as:

$$Cg(k) = \sum_{i=1}^{k} g(c'_i) \quad (3)$$

This function indicates the total actual gain of the bootstrapping algorithm if the classifiers representing the top k concepts are enhanced, while the initial classifiers are maintained for the rest $N - k$ cases. In the optimal scenario, the predicted k concepts yield the highest improvement in the bootstrapping process (i.e. $g(c'_1) > g(c'_2) > \cdots > g(c'_k)$). In Fig. 3, the function of Eq. 3 is plotted for every k. The proposed approach is compared to three baselines; (a) **Random:** The instances are ranked randomly, (b) **Upper Baseline:** The instances are ranked based on the actual gain $g(i)$ simulating the best possible regression model (i.e. best case scenario). (c) **Lower Baseline:** The instances are inversely ranked based on the actual gain $g(i)$ simulating the worst regression model (i.e. worst case scenario). It is obvious that the proposed regression model significantly outperforms the random baseline and lies quite close to the upper baseline.

In order to provide an indication of the benefits that could be gained in terms of processing load by the employment of the proposed approach, we provide Table 1. In this table, we can see the achieved performance gain if we decide to enhance the top 10, 20, 40, 60, 80 and 94 (all) concepts as they were ranked by the proposed approach and the random baseline. The first column (*Abs.*) refers to the absolute performance gain achieved, while the second column (*Perc.*) is the percentage of this value with respect to the maximum possible achieved gain, which occurs if we enhance all the 94 concepts. We can see that using the proposed approach we can achieve the same performance gain with significantly less processing load. For example, if we decide to enhance the 40 most prominent concepts as ranked by the proposed approach,

Table 1. Prediction performance comparison between the proposed approach and the random baseline

# concepts	Proposed		Random	
	Abs.	Perc.	Abs.	Perc.
10	0,93	31,39	0,26	8,71
20	1,36	45,58	0,62	21
40	1,81	**60,87**	1,21	40,71
60	2,19	73,47	1,71	**57,38**
80	2,7	90,81	2,35	79,04
94 (All)	*2.98*	*100*	*2.98*	*100*

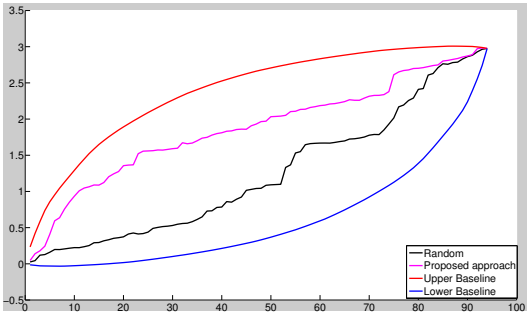

Fig. 3. Actual cumulative gain

we can achieve more than half of the total performance boost, while we need to enhance around 60 concepts to achieve similar boost if the random model decides for the ranking.

4. CONCLUSIONS

In this paper, in an effort to improve the scalability properties of the computationally expensive approaches that follow the bootstrapping paradigm, we investigate the correlation of two new features, i.e. the model's *maturity* and the oracle's *reliability*, with the expected performance gain. This correlation can be exploited to devise mechanisms appropriate for ruling out the cases that are not expected to substantially benefit from augmenting the training set. For example, when a new concept is examined (i.e., which was not included in the training instances/concepts), the regression function f is applied and the expected performance gain of adding new images is estimated. Having this knowledge, one can decide if it is considered worthwhile adding more training data to achieve the predicted performance improvement (e.g. for certain concepts even a small improvement might be considered important). Our experiments have shown that by utilizing this regression function f we can achieve approximately 60% of the performance gain by enhancing less than half of the concepts. Our plans for future work include the investigation of additional features for predicting the expected performance gain.

Acknowledgements

This work was supported by the EU 7th Framework Programmes under grant numbers IST-FP7-288815 and FP7-ICT-2011-9 in projects Live+Gov (www.liveandgov.eu) and i-treasures (www.i-treasures.eu) respectively.

5. REFERENCES

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

[2] Elisavet Chatzilari, Spiros Nikolopoulos, Yiannis Kompatsiaris, and Josef Kittler, "Multi-modal region selection approach for training object detectors," in *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, New York, NY, USA, 2012, ICMR '12, pp. 5:1–5:8, ACM.

[3] David Cohn, Les Atlas, and Richard Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, May 1994.

[4] X. Li, C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders, "Bootstrapping visual categorization with relevant negatives," *IEEE Transactions on Multimedia*, vol. In press, 2013.

[5] Elisavet Chatzilari, Spiros Nikolopoulos, Yiannis Kompatsiaris, and Josef 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.

[6] Colin Campbell, Nello Cristianini, and Alex J. Smola, "Query learning with large margin classifiers," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, CA, USA, 2000, ICML '00, pp. 111–118, Morgan Kaufmann Publishers Inc.

[7] Shouxian Cheng and Frank Y. Shih, "An improved incremental training algorithm for support vector machines using active query," *Pattern Recogn.*, vol. 40, no. 3, pp. 964–971, Mar. 2007.

[8] A.J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, june 2009, pp. 2372 –2379.

[9] Stefanie Nowak and Stefan 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*, New York, NY, USA, 2010, MIR '10, pp. 557–566, ACM.

[10] Bart Thomee and Adrian Popescu, "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.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[12] Mauricio Villegas, Roberto Paredes, and Bart Thomee, "Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask," in *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, Valencia, Spain, September 23-26 2013.

[13] T. Cour, B. Sapp, C. Jordan, and B. Taskar, "Learning from ambiguously labeled images," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 919–926.

[14] Lyndon S. Kennedy, Shih-Fu Chang, and Igor V. Kozintsev, "To search or to label?: Predicting the performance of search-based automatic image classifiers," in *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, 2006, MIR '06.

[15] B. Thomee Mark J. Huiskes and Michael S. Lew, "New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative," in *MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2010, pp. 527–536, ACM.