ACTIVE CONVOLUTIONAL NEURAL NETWORKS FOR CANCEROUS TISSUE RECOGNITION

Panagiotis Stanitsas[†] *Anoop Cherian*[‡] *Alexander Truskinovsky*[⋆] *Vassilios Morellas*[†]
Nikolaos Papanikolopoulos[†]

[†] University of Minnesota [‡]Australian National University [⋆]Roswell Park Cancer Institute

ABSTRACT

Deep neural networks typically require large amounts of annotated data to be trained effectively. However, in several scientific disciplines, including medical image analysis, generating such large annotated datasets requires specialized domain knowledge, and hence is usually very expensive. In this work, we present a novel application of active learning to data sample selection for training Convolutional Neural Networks (CNN) for Cancerous Tissue Recognition (CTR). Our main idea is to steer annotation efforts towards selecting the most informative samples for training the CNN. To quantify informativeness, we explore three choices based on discrete entropy, best-vs-second-best, and k-nearest neighbor agreement. Our results on three different types of cancer datasets consistently demonstrate that under limited annotated samples, our proposed training scheme converges faster than classical randomized stochastic gradient descent, while achieving the same (or sometimes superior) classification accuracy.

Index Terms— active learning, cancer detection, uncertainty sampling, deep learning

1. INTRODUCTION

Convolutional Neural Networks (CNN) have revolutionized the domain of computer vision with performances of various applications trending towards human accuracy. One of the main factors that enabled this recent resurgence of CNNs is the availability of large datasets. CNNs usually involve millions of parameters to learn complex real-world tasks, which renders them prone to overfitting. One effective way to reduce overfitting is to increase data diversity, thus providing large annotated datasets for training.

However, there are several applications in which collecting such large amounts of annotated data is either challenging or very expensive. One such domain is medical image analysis, especially Cancerous Tissue Recognition (CTR). In this task, the tissue slides from suspected cancerous regions are examined under a microscope and are classified as benign or malignant – a task that not only requires the expertise of an experienced pathologist, but also is very time consuming. While CNNs may be able to improve the accuracy of diagnosis once they are trained adequately, the training process itself is usually challenging due to the high expenditure of collecting large datasets. To circumvent this issue, we resort to active learning in this paper.

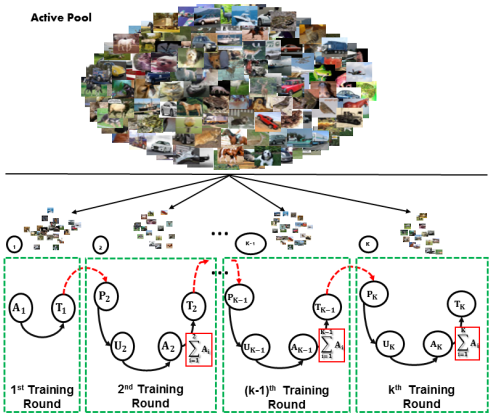

Fig. 1: Outline of the active training scheme. The annotation module $\mathbf{A}_i$ corresponds to the interaction between the training scheme and a human annotator (e.g., surgical pathologist) during training stage i. The training module $\mathbf{T}_i$ corresponds to the training process of the CNN in hand in the presence of the annotation harvested during previous stages $\{1, 2, ..., i\}$. The module $\mathbf{P}_i$ predicts the class labels of future samples in a new batch during stage i based on parameter weights learned in the previous stages. The module $\mathbf{U}_i$ quantifies the uncertainty of the predictions.

Active learning has been very successful in selecting useful data samples in a variety of machine learning and vision applications. Active strategies steer humans' annotation efforts towards data samples that have the highest uncertainty for the classifier being trained. There have been several such uncertainty sampling schemes proposed in the literature geared towards classification (e.g., [1]) and clustering (e.g., [2]) problems.

In this paper, we present a novel application of active learning for the selection of data samples to train a CNN for CTR. However, there are two important challenges to overcome when applying active learning to CNNs, namely (i) to allow learning without overfitting to the limited data given the large number of CNN parameters, and (ii) to score the data samples for selection based on their expected effectiveness in improving the overall CNN training objective. We propose a multi-stage training scheme to bypass these issues. Each stage uses a small number of annotated data samples to train the CNN until it starts overfitting to the validation data. The

CNN trained after every stage, is then used to predict the class labels on unseen data samples (active pool); the predictions are scored using an uncertainty measure. Figure 1 depicts an outline of the proposed framework.

To validate the effectiveness of the proposed framework, we apply the scheme for classifying cancerous tissues against benign ones. We experiment with three different types of cancer image patches, namely (i) Breast cancer, (ii) Prostate cancer, and (iii) Myometrium tissue samples. These patches are obtained by imaging Hematoxylin & Eosin (H&E)-stained tissues under a microscope. Our experimental results demonstrate that the proposed active learning setup can consistently lead to better training of the CNN, allowing it to converge much faster at a slightly higher accuracy than using the classical random sampling scheme in a batch-mode stochastic gradient descent training setup.

2. BACKGROUND

Active selection methods were first introduced to the machine learning community in the mid '80's for the text classification problem (e.g., [3]). At the core of active training schemes lies the efficient quantification of prediction uncertainty, which is a reflection of the confidence a model provides on the task. Hanneke et al. [4] were among the first to theoretically demonstrate the positive effects of active learning in a Probably Approximately Correct (PAC) framework.

In a binary classification setup, Tong and Chang [5] derived an active scheme for the recognition of humans in an image, while, Tong and Koller [6] proposed a minimization scheme over decision hyperplanes for active selection. Kapoor et al. [7] derived an active sampling strategy based on the output of a Gaussian process model in a multi-class setup. Holub et al. [1] proposed an entropy based active selection strategy for object classification. Capitalizing on the probabilistic output of Support Vector Machine models, Joshi et al. [8] demonstrated a powerful measure for quantifying the uncertainty via a best-vs-second-best strategy; their scheme outperformed random selection and entropy based active schemes. In addition, Joshi et al. [9] also presented a scalable, cost-aware scheme for active selection. Jain and Kapoor [10] devised a k-nearest neighbor method for active selection in large multi-class problems involving target problems with a large number of classes. Vijayanarasimhan and Grauman [11] proposed a crowdsourcing based active scheme to train object detectors.

In addition, active selection procedures have been combined with multi-layer neural networks to enhance their performance. The first documented attempt of enhancing the performance of deep learning via active selection was presented in the work by Zhou et al. [12] for sentiment classification. An attempt to enhance Deep Belief Networks via active selection was presented by Wang and Shang [13] on a limited set of experiments, while Stark et al. [14] presented an active deep learning scheme for automated public Turing tests using a single best-vs-second-best uncertainty quantification scheme. Finally, Wang et al. [15] provided comparisons on two object recognition benchmarks, using uncertainty sampling measures which capitalized solely on the probability simplex produced by the softmax layer of the CNN. Although there are similarities between the presented work and [15], our works differs in that we consider additionally an uncertainty sampling scheme that treats the CNN as a dimensionality reduction scheme and computes the confidence of the classifier based on the clustering effect that the fully connected layers of the CNN exhibit [16]. Furthermore, we provide a thorough evaluation of this uncertainty sampling framework on different types of cancer to conclude about the feasibility and effectiveness of this approach in the CTR domain.

The coupling of active selection strategies and cancer recognition has appeared in the work of Danziger et al. [17] which derived a Most-Informative-Positive selection scheme. This was applied to discover mutations in a tumor suppressor protein (p53), found in human cancers. Focusing on gene expressions, in the cancer recognition setup, Liu et al. [18] introduced active selection for the classification of colon cancer, lung cancer, and prostate cancer samples.

3. LEARNING SETUP AND METHODOLOGY

We use a CNN as the classifier of choice in this paper due to its impressive performance on a variety of related tasks (e.g., [19, 20]). We adopt a multi-stage training framework for the CNN as depicted in Figure 1 involving multiple stages of training and augmentation of the training set by adding new annotated data; each newly added data sample is selected based on informativeness criteria.

Formally, suppose we have access to a collection of data samples D. Let $f_i : D \rightarrow \Delta_d$ define a CNN trained at the i-th stage that takes a data sample as input and produces a class probability vector (in the simplex Δ_d) as output, where we assume there are d different class labels. Let $S_1 \subset D$ represent a (small) initial set of annotated samples. Our scheme starts by training the CNN using a training set $T = S_1$ for the cross-entropy loss. The training is continued until the model starts overfitting to the training data (as measured using a separate validation set). Once trained, we select a subsequent subset $S_{i+1} \subset D \setminus \bigcup_{j=1}^{i} S_j$ from the training set and apply the current CNN model f_i to generate classifier probabilities for the samples in S_{i+1}. These classifier probabilities are evaluated using an informativeness measure. Suppose $A_{i+1} \subseteq S_{i+1}$ is a subset of this data batch that is deemed to be informative by the measure, then we augment the training set $T = T \cup A_{i+1}$ and use it to train the CNN to generate a better model f_{i+1}. This setup is repeated until the training error plateaus. Note that if the cardinality of A_{i+1} is less than a threshold, we sample more data batches such that we have sufficient training samples for the new training stage. The appropriate amount of annotations for each stage is decided by the size of the stochastic gradient descent training batches, while the number of stage-wise training iterations is guided by the descent in the validation data loss. However, in the absence of large

amounts of initial annotated data that can ensure the proper convergence of training, fine-tuning a pre-trained CNN model could be used. In this case, we use a model that is trained on a very large dataset for a task similar to the one in-hand (but perhaps with a different goal) to initialize the filter weights; the main assumption is that it is cheaper to obtain data annotations for this surrogate task.

The quality of the data samples selected in each stage for training the CNN decides the effectiveness of the resulting model. To this end, we use the probability vector produced by the model f_i trained at the i-th stage and applied on the batch S_{i+1} for the next stage. We deploy two uncertainty measures defined on the probability simplex, namely (i) discrete entropy [1] and (ii) the best-vs-second-best [8] measures. Further, it is well-known that the outputs generated by the fully-connected layers of the CNN can be looked upon as embedding the original high-dimensional data into a low-dimensional feature space [16] – this embedding is often found to have a clustering effect on the data samples. With this intuition, we propose to use an additional uncertainty measure that captures the disagreement between k-NNs for every sample in the active pool.

First we consider the the *discrete entropy* [1] computed on the output class probability vector – each entry of which captures the probability of a data sample to take the associated class label. For a data sample $\mathbf{x} \in S_{i+1}$ from the active pool for stage $i+1$, let $\mathbf{p}(\mathbf{x}) = f_i(\mathbf{x})$ ($\mathbf{p}(\mathbf{x}) \in \Delta_d$) define the probabilistic output of the CNN classifier trained in stage-i. Then, we define the *discrete entropy* of the data sample $\mathbf{x}$ as:

$$H(\mathbf{x}) = -\sum_{j=1}^{d} \mathbf{p}^j(\mathbf{x}) \log(\mathbf{p}^j(\mathbf{x})), \tag{1}$$

where $\mathbf{p}^j$ represents the j-th dimension of the probability vector. We use the output of the softmax output from the last layer of the CNN to compute $\mathbf{p}(\mathbf{x})$.

As is clear, the discrete entropy measures the overall randomness of a data sample. We could explicitly use the confusions in the classifier by quantifying the separability between the data classes as decided by the learned class-decision boundaries. One such heuristic is to use the *difference* between the best and the second-best output class probabilities as suggested in [8] – a smaller difference suggesting a higher confusion between the respective classes. Reusing the notations from above, let $b_1 = \arg\max_{j \in \{1,\cdots,d\}} \mathbf{p}^j(\mathbf{x})$ and $b_2 = \arg\max_{j \in \{1,\cdots,d\} \setminus b_1} \mathbf{p}^j(\mathbf{x})$ be the indices of the best and the second-best classifier probabilities, then the *Best-vs-Second-Best* uncertainty measure is defined as:

$$B(\mathbf{p}(\mathbf{x})) = \mathbf{p}^{b_1}(\mathbf{x}) - \mathbf{p}^{b_2}(\mathbf{x}). \tag{2}$$

Lastly, motivated by similar prior methods such as [10], we define the probability of a sample in the active pool to belong to a class as the *annotation disagreement* among its NNs; these NNs are computed in the embedded lower-dimensional space generated by the fully connected layers of the CNN. To be precise, suppose $\tilde{\mathbf{x}} = \tilde{f}_i(\mathbf{x})$ denotes the output of a given layer of the CNN in stage-i for an input $\mathbf{x} \in S_{i+1}$. Further, let $y \in \{1, 2, \cdots, d\}$ be the class-label associated with the point $\mathbf{x}$. Suppose, there are n_c points in T (which is the training set with annotated samples) with class label c. Then, the *NN agreement* for class-c is defined as:

$$\mathbf{p}_c(\mathbf{x}) = \frac{\frac{1}{n_c} \sum_{\{\mathbf{x}_j \in T \mid y_j = c\}} \mathrm{Dist}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_j)}{\sum_{c=1}^{C} \frac{1}{n_c} \sum_{\{\mathbf{x}_j \in T \mid y_j = c\}} \mathrm{Dist}(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_j)}, \tag{3}$$

where $\mathrm{Dist}(\cdot, \cdot)$ is some suitable similarity measure between the embedded data points $\tilde{\mathbf{x}}$ and its neighbors $\tilde{\mathbf{x}}_j$. In this paper, we use the class label agreement itself as the $\mathrm{Dist}(\cdot, \cdot)$, however, we use the Euclidean distance of the embedded points for computing the nearest neighbors. Specifically, for every unlabelled sample, we use the ground-truth labels for its k-NNs in T. Following that, we construct a normalized histogram on the label occurrences which serves as an approximation of the class membership probability vector. Finally, we compute the discrete entropy of the approximated probability vector as described in (1) towards quantifying the uncertainty associated with every prediction.

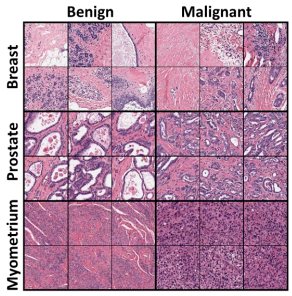

Fig. 2: H&E-stained samples for three types of tissue; Breast (1st and 2nd row), Prostate (3rd and 4th row) and Myometrium (5th and 6th row).

4. EXPERIMENTS

Due to the lack of widely accepted CTR benchmarks, we present experiments on three private CTR datasets to evaluate our proposed active learning framework for training CNNs. We consider the problem of CTR based on H&E stained tissue samples. Hematoxylin stains the nuclei in blue or dark purple color, while Eosin imparts a pink or lighter purple color to the cytoplasm, as depicted in Figure 2.

4.1. Datasets

First, for the case of carcinomas of the breast, 21 annotated images of carcinomas and 19 images of benign tissue, taken from 21 patients, are combined towards deriving a 17,497 sample dataset. 3,913 samples depicted benign tissue, while 13,584 patches corresponded to cancerous tissue. Second, 39 myometrial leiomyomas were combined with 41 images of leiomyosarcomas to construct our second dataset for the myometrium from 39 patients. We randomly selected 1539 cancerous image patches and combined them with 1782 benign patches to derive a dataset of 3321 samples. Finally, for prostate cancer, 31 images of carcinomas and 8 images from

benign regions are annotated, taken from 10 patients. A 3500 image patches dataset was created with 1750 patches depicting cancerous regions, with the other 1750 corresponding to benign regions. A more detailed description on the utilized datasets, as well as alternative feature representations, can be found in [21]. We present our experimental validation based on patches of size 150×150 pixels, while the test set of each dataset remained fixed throughout all training stages and consisted of 20% of the original datasets.

4.2. CNN Training

For our experiments, the *BVLC Caffe* [16] framework was utilized on a machine with a single graphics card (NVIDIA TITAN X), a quad-core Intel $i7$ processor and 32Gb of memory. For this section, we assume some basic familiarity of the reader with the core CNN terminology. LeCunn et al. provides an introduction to the different CNN layer types in [22] which the reader can also refer to. The Caffenet topology, distributed with the *Caffe* framework, was used for fine-tuning on the collected datasets, while weight initializations were taken from training the network on the 1M image database of the ILSVRC challenge. Furthermore, we reduced the weights of all intermediate layers of the network to 15% of the original values and trained for a binary classification objective. We set the base learning rate to 0.0001, while we selected a step strategy that decreases the rate every 2.5K iterations, and we also set the weight decay to 0.005. Ten thousand iterations were performed for the first training stage, and 5K iterations were performed for all subsequent stages. Finally, for the uncertainty measure based on NN-agreement, we found that working with 41-NNs is the most effective.

4.3. Results

Figure 3 presents the results obtained on the breast cancer dataset for 16 training stages. All three active schemes were found to be consistently more accurate when compared to the random selection scheme. For the first training stage, 3.5K annotated samples were selected and remained the same for all the sampling strategies. For all subsequent training stages, 500 additional annotations were provided. Active schemes reached a 2.2% increase in performance on the test when compared to random selection after the 6^{th} training stage. Furthermore, active schemes, for the case that 5.5K annotated samples were provided, achieved a performance as high as random selection when 11K samples were provided for training; this is 50% decrease in the number of queries, which strongly supports the merits of the proposed framework for CTR.

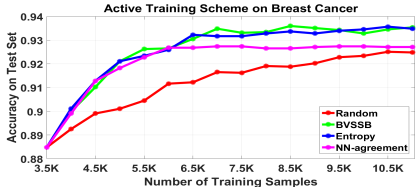

Fig. 3: Results on the Breast Cancer dataset.

For the myometrial leiomyomas dataset, Figure 4 presents the results for 12 training stages. For the first training stage, 540 annotated samples were provided, while the training set was augmented by 150 samples for the subsequent training stages. The largest performance gains for active schemes was achieved for the case that 1140 annotated samples were provided and reached 2.1%. Furthermore, interestingly, we found that similarly to the case of breast cancer, we achieved higher performance (94%) with 50% of the annotated samples that the random selection required to reach an equivalent performance (93.2%).

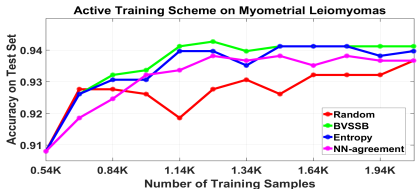

Fig. 4: Results on the Myometrial Leiomyomas dataset.

Finally, for the prostate cancer dataset, Figure 5 illustrates the extracted performance curves for 14 training stages. For the first training stage, 560 annotated samples were used, while 150 annotations were provided for every subsequent training stage. For the case that 1.01K annotations were provided, random selection performed significantly less than active schemes (entropy) with a 2.9% difference in the obtained performance. An instance that highlights the annotation gains of the proposed framework is illustrated by the fact that random selection requires 40% more annotated samples to reach accuracy of 89.3% when compared to the entropy based active selection scheme. The best accuracy is attained by the BVSSB scheme for the case that 2.21K samples were provided for training, reaching 89.6%.

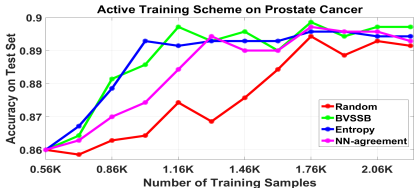

Fig. 5: Results on the Prostate Cancer dataset.

4.4. Discussion

Our results clearly show that active learning is beneficial and leads to faster training of the CNN, while achieving similar (or sometimes slightly superior) accuracy than randomized sampling schemes. For all three uncertainty sampling schemes the achieved performance was comparable. Finally, the observed query reductions reached 50%, while the absolute performance on the CTR datasets reached 93.4%, 94.1% and 89.6% for breast cancer, myometrial leiomyomas and prostate cancer respectively.

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation through grants #IIP-0934327, #CNS-1039741, #SMA-1028076, #CNS-1338042, #IIP-1439728, #OISE-1551059, and #CNS-1514626. Dr. Cherian is funded by the Australian Research Council Centre of Excellence for Robotic Vision (#CE140100016)..

6. REFERENCES

[1] A. Holub, M. Welling, and P. Perona, "Exploiting unlabelled data for hybrid object classification," in *NIPS 2005 Workshop on Inter-Class Transfer*, 2005.

[2] P. Stanitsas, A. Cherian, V. Morellas, and N. Papanikolopoulos, "Active constrained clustering via non-iterative uncertainty sampling," in *International Conference on Intelligent Robots and Systems*. IEEE, 2016.

[3] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, pp. 1134–1142, 1984.

[4] S. Hanneke, "A bound on the label complexity of agnostic active learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 353–360.

[5] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, 2001.

[6] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning*, vol. 2, pp. 45–66, 2001.

[7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with Gaussian Processes for object categorization," in *International Conference of Computer Vision*, 2007.

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

[9] A. J Joshi, F. Porikli, and N. Papanikolopoulos, "Scalable active learning for multiclass image classification," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2259–2273, 2012.

[10] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[11] 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.

[12] S. Zhou, Q. Chen, and X. Wang, "Active deep learning method for semi-supervised sentiment classification," *Neurocomputing*, vol. 120, pp. 536–546, 2013.

[13] D. Wang and Y. Shang, "A new active labeling method for deep learning," in *2014 International Joint Conference on Neural Networks*. IEEE, 2014, pp. 112–119.

[14] F. Stark, C. Hazırbas, R. Triebel, and D. Cremers, "Captcha recognition with active deep learning," in *Workshop New Challenges in Neural Computation 2015*. Citeseer, 2015, p. 94.

[15] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *International Conference on Multimedia*. ACM, 2014.

[17] S. Danziger, R. Baronio, L. Ho, L. Hall, K. Salmon, G. Hatfield, P. Kaiser, and R. Lathrop, "Predicting positive p53 cancer rescue regions using most informative positive (mip) active learning," *PLoS Computational Biology*, vol. 5, no. 9, pp. e1000498, 2009.

[18] Y. Liu, "Active learning with support vector machine applied to gene expression data for cancer classification," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 6, pp. 1936–1941, 2004.

[19] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732, IEEE.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] P. Stanitsas, A. Cherian, X. Li, A. Truskinovsky, V. Morellas, and N. Papanikolopoulos, "Evaluation of feature descriptors for cancerous tissue recognition," in *International Conference of Pattern Recognition*. IAPR/IEEE, 2016.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.