Fine-tuning Convolutional Neural Networks for Biomedical Image Analysis: Actively and Incrementally[*]

Zongwei Zhou[1], Jae Shin[1], Lei Zhang[1], Suryakanth Gurudu[2], Michael Gotway[2], and Jianming Liang[1]

[1]Arizona State University

{zongweiz,sejong,lei.zhang.10,jianming.liang}@asu.edu

[2]Mayo Clinic

{gurudu.suryakanth,gotway.michael}@mayo.edu

Abstract

Intense interest in applying convolutional neural networks (CNNs) in biomedical image analysis is wide spread, but its success is impeded by the lack of large annotated datasets in biomedical imaging. Annotating biomedical images is not only tedious and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible. To dramatically reduce annotation cost, this paper presents a novel method called AIFT (active, incremental fine-tuning) to naturally integrate active learning and transfer learning into a single framework. AIFT starts directly with a pre-trained CNN to seek "worthy" samples from the unannotated for annotation, and the (fine-tuned) CNN is further fine-tuned continuously by incorporating newly annotated samples in each iteration to enhance the CNN's performance incrementally. We have evaluated our method in three different biomedical imaging applications, demonstrating that the cost of annotation can be cut by at least half. This performance is attributed to the several advantages derived from the advanced active and incremental capability of our AIFT method.

1. Introduction

Convolutional neural networks (CNNs) [14] have brought about a revolution in computer vision thanks to large annotated datasets, such as ImageNet [6] and Places [27]. As evidenced by an IEEE TMI special issue [8] and two forthcoming books [28, 17], intense interest in applying CNNs in biomedical image analysis is wide spread,

but its success is impeded by the lack of such large annotated datasets in biomedical imaging. Annotating biomedical images is not only tedious and time consuming, but also demanding of costly, specialty-oriented knowledge and skills, which are not easily accessible. Therefore, we seek to answer this critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging.* In doing so, we present a novel method called AIFT (active, incremental fine-tuning) to naturally integrate active learning and transfer learning into a single framework. Our AIFT method starts directly with a pre-trained CNN to seek "salient" samples from the unannotated for annotation, and the (fine-tuned) CNN is continuously fine-tuned by incrementally enlarging the training dataset with newly annotated samples. We have evaluated our method in three different applications including colonoscopy frame classification, polyp detection, and pulmonary embolism (PE) detection, demonstrating that the cost of annotation can be cut by at least half.

This outstanding performance is attributed to a simple yet powerful observation: To boost the performance of CNNs in biomedical imaging, multiple patches are usually generated automatically for each candidate through data augmentation; these patches generated from the same candidate share the same label, and are naturally expected to have similar predictions by the current CNN before they are expanded into the training dataset. As a result, their *entropy* and *diversity* provide a useful indicator to the "power" of a candidate in elevating the performance of the current CNN. However, automatic data augmentation inevitably generates "hard" samples for some candidates, injecting noisy labels; therefore, to significantly enhance the robustness of our method, we compute entropy and diversity by selecting only a portion of the patches of each candidate according to the predictions by the current CNN.

Several researchers have demonstrated the utility of fine-tuning CNNs for biomedical image analysis, but they only performed one-time fine-tuning, that is, simply fine-tuning a pre-trained CNN once with available training samples involving no active selection processes (e.g., [4, 19, 5, 2, 21, 7, 18, 24]). To our knowledge, our proposed method is among the first to integrate active learning into fine-tuning C-NNs in a continuous fashion to make CNNs more amicable for biomedical image analysis with an aim to cut annotation cost dramatically. Compared with conventional active learning, our AIFT method offers several advantages:

1. Starting with a completely empty labeled dataset, requiring no initial seed labeled samples (see Alg. 1);
2. Incrementally improving the learner through continuous fine-tuning rather than repeatedly re-training (see Sec. 3.1);
3. Naturally exploiting expected consistency among the patches associated for each candidate to select samples "worthy" of labeling (see Sec. 3.2);
4. Automatically handling noisy labels as only a portion (e.g., a quarter) of the patches in each candidate participate in the selection process (see Sec. 3.3);
5. Computing entropy and diversity locally on a small number of patches within each candidate, saving computation time considerably (see Sec. 3.3).

More importantly, our method has the potential to exert important impact on computer-aided diagnosis (CAD) in biomedical imaging, because the current regulations require that CAD systems be deployed in a "closed" environment, in which all CAD results be reviewed and errors if any be corrected by radiologists; as a result, all false positives are supposed to be dismissed and all false negatives supplied, an instant on-line feedback that may make CAD systems self-learning and improving possible after deployment given the continuous fine-tuning capability of our method.

2. Related work

2.1. Transfer learning for medical imaging

Gustavo *et al.* [2] replaced the fully connected layers of a pre-trained CNN with a new logistic layer and trained only the appended layer with the labeled data while keeping the rest of the network the same, yielding promising results for classification of unregistered multiview mammograms. In [5], a fine-tuned pre-trained CNN was applied for localizing standard planes in ultrasound images. Gao *et al.* [7] fine-tuned all layers of a pre-trained CNN for automatic classification of interstitial lung diseases. In [21], Shin *et al.* used fine-tuned pre-trained CNNs to automatically map medical images to document-level topics, document-level sub-topics, and sentence-level topics. In [18], fine-tuned pre-trained CNNs were used to automatically retrieve missing or noisy cardiac acquisition plane

information from magnetic resonance imaging and predict the five most common cardiac views. Schlegl *et al.* [19] explored unsupervised pre-training of CNNs to inject information from sites or image classes for which no annotations were available, and showed that such across site pre-training improved classification accuracy compared to random initialization of the model parameters. Tajbakhsh *et al.* [24] systematically investigated the capabilities of transfer learning in several medical imaging applications. However, they all performed *one-time fine-tuning*—simply fine-tuning a pre-trained CNN just once with available training samples, involving neither active selection processes nor continuous fine-tuning.

2.2. Integrating active learning with deep learning

The literature of general active learning and deep learning is rich and deep [8, 28, 17, 20, 9, 10, 26]. However, the research aiming to integrate active learning with deep learning is sparse: Wang and Shang [25] may be the first to incorporate active learning with deep learning, and based their approach on stacked restricted Boltzmann machines and stacked autoencoders. A similar idea was reported for hyperspectral image classification [15]. Stark *et al.* [22] applied active learning to improve the performance of C-NNs for CAPTCHA recognition, while Al Rahhal *et al.* [1] exploited deep learning for active electrocardiogram classification. All these approaches are fundamentally different from our AIFT approach in that in each iteration they all *repeatedly re-trained the learner from scratch* while we continuously fine-tune the (fine-tuned) CNNs in an incremental manner, offering five advantages as listed in Sec. 1.

3. Proposed method

We present our AIFT method in the context of computer-aided diagnosis (CAD) in biomedical imaging. A CAD system typically has a candidate generator, which can quickly produce a set of candidates, among which, some are *true* positives and some are *false* positives. After candidate generation, the task is to train a classifier to eliminate as many false positives as possible while keeping as many true positives as possible. To train a classifier, each of the candidates must be labeled. We assume that each candidate takes one of $|Y|$ possible labels. To boost the performance of CNNs for CAD systems, multiple patches are usually generated automatically for each candidate through data augmentation; these patches generated from the same candidate inherit the candidate's label. In other words, all labels are acquired at the candidate level. Mathematically, given a set of candidates, $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_n\}$, where n is the number of candidates, and each candidate $\mathcal{C}_i = \{x_i^1, x_i^2, ..., x_i^m\}$ is associated with m patches, our AIFT algorithm iteratively selects a set of candidates for labeling (illustrated in Alg. 1).

Algorithm 1: .

Input:
$\mathcal{U} = \{\mathcal{C}_i\}, i \in [1, n]$ {$\mathcal{U}$ contains n candidates}
$\mathcal{C}_i = \{x_i^j\}, j \in [1, m]$ {$\mathcal{C}_i$ has m patches}
$\mathcal{M}_0$: pre-trained CNN
b: batch size
α: patch selection ratio
Output:
$\mathcal{L}$: labeled candidates
$\mathcal{M}_t$: fine-tuned CNN model at Iteration t
Functions:
$p \leftarrow P(\mathcal{C}, \mathcal{M})$ {outputs of $\mathcal{M}$ given $\forall x \in \mathcal{C}$}
$\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1})$ {fine-tune $\mathcal{M}_{t-1}$ with $\mathcal{L}$}
$a \leftarrow mean(p_i)$ {$a = \frac{1}{m} \sum_{j=1}^m p_i^j$}
Initialize:
$\mathcal{L} \leftarrow \varnothing$, t$\leftarrow$ 1

1 **repeat**
2   **for** each $\mathcal{C}_i \in \mathcal{U}$ **do**
3       $p_i \leftarrow P(\mathcal{C}_i, \mathcal{M}_{t-1})$
4   **if** $mean(p_i) > 0.5$ **then**
5          $\mathcal{S}_i' \leftarrow$ top α percent of the patches of $\mathcal{C}_i$
6   **else**
7          $\mathcal{S}_i' \leftarrow$ bottom α percent of the patches of $\mathcal{C}_i$
8   **end**
9   Build matrix R_i using Eq. 3 for $\mathcal{S}_i'$
10   **end**
11   Sort $\mathcal{U}$ according to the numerical sum of R_i
12   Query labels for top b candidates, yielding $\mathcal{Q}$
13    $\mathcal{L} \leftarrow \mathcal{L} \bigcup \mathcal{Q}; \quad \mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{Q}$
14    $\mathcal{M}_t \leftarrow F(\mathcal{L}, \mathcal{M}_{t-1}); t \leftarrow t + 1$
15 **until** *classification performance is satisfactory*;

3.1. Continuous fine-tuning

At the beginning, the labeled dataset $\mathcal{L}$ is empty; we take a pre-trained CNN (e.g., AlexNet) and run it on $\mathcal{U}$ to select b number of candidates for labeling. The newly labeled candidates will be incorporated into $\mathcal{L}$ to continuously fine-tune the CNN incrementally until the performance is satisfactory. Several researchers have demonstrated that fine-tuning offers better performance and is more robust than training from scratch. From our experiments, we have found that continuously fine-tuning the CNN, which has been fine-tuned in the previous iteration, with enlarged datasets converges faster than repeatedly fine-tuning the original pre-trained CNN. We also found that continuously fine-tuning the CNN with only newly labeled data demands careful meta-parameter adjustments.

3.2. Active candidate selection

In active learning, the key is to develop a criterion for determining the "worthiness" of a candidate for annotation.

Our criterion is based on an observation: All patches generated from the same candidate share the same label; they are expected to have similar predictions by the current C-NN. As a result, their *entropy* and *diversity* provide a useful indicator to the "power" of a candidate in elevating the performance of the current CNN. Intuitively, entropy captures the classification certainty—higher uncertainty values denote higher degrees of information; while diversity indicates the prediction consistency among the patches within a candidate—higher diversity values denote higher degrees of prediction inconsistency among the patches within a candidate. Therefore, candidates with higher entropy and higher diversity are expected to contribute more in elevating the current CNN's performance. Formally, assuming the prediction of patch x_i^j by the current CNN is p_i^j, we define its entropy as:

$$e_i^j = -\sum_{k=1}^{|Y|} p_i^{j,k} \log p_i^{j,k} \tag{1}$$

and diversity between patches x_i^j and x_i^l of candidate $\mathcal{C}_i$ as:

$$d_i(j, l) = \sum_{k=1}^{|Y|} (p_i^{j,k} - p_i^{l,k}) log \frac{p_i^{j,k}}{p_i^{l,k}} \tag{2}$$

Entropy e_i^j denotes the information furnished by patch x_i^j of candidate $\mathcal{C}_i$ in the unlabeled pool. Diversity $d_i(j, l)$, captured by the symmetric Kullback Leibler divergence [13], estimates the amount of information overlap between patches x_i^j and x_i^l of candidate $\mathcal{C}_i$. By definition, all the entries in e_i^j and $d_i(j, l)$ are non-negative. Further, $d_i(j, j) = 0$, $\forall j$, therefore, for notational simplicity, we combine e_i^j and $d_i(j, l)$ into a single matrix R_i for each candidate $\mathcal{C}_i$:

$$R_i(j, l) = \begin{cases} \lambda_1 e_i^j & \text{if } j = l, \\ \lambda_2 d_i(j, l) & \text{otherwise} \end{cases} \tag{3}$$

where λ_1 and λ_2 are trade-offs between entropy and diversity. We use two parameters for convenience, so as to easily turn on/off entropy or diversity during experiments.

3.3. Handling noisy labels via majority selection

Automatic data augmentation is essential to boost CNN's performance, but it inevitably generates "hard" samples for some candidates as shown in Fig. 1 and Fig. 2 (c), injecting noisy labels; therefore, to significantly enhance the robustness of our method, we compute entropy and diversity by selecting only a portion of the patches of each candidate according to the predictions by the current CNN. Specially, for each candidate $\mathcal{C}_i$ we first compute the average probabilistic prediction of all of its patches:

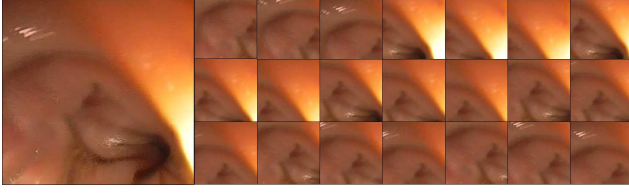$$a_i = \frac{1}{m} \sum_{j=1}^m p_i^j \tag{4}$$

Figure 1: Experts label a frame based on the overall quality: if over 75% of a frame (i.e., candidate in this application) is clear, it is considered "informative". For example, the whole frame (the leftmost image) is labeled as "informative", but not all the patches associated with this frames are "informative", although they inherit the "informative" label. This is the main motivation for the majority selection in our AIFT method.

where m is the number of patches within candidate $\mathcal{C}_i$, p_i^j is the prediction probability of patch x_i^j. If $a_i > 0.5$, we select the top α percent patches; otherwise, the bottom α percent patches. Based on the selected patches, we then use Eq. 3 to construct the score matrix R_i of size $\alpha m \times \alpha m$ for each candidate $\mathcal{C}_i$ in $\mathcal{U}$. Our proposed majority selection method automatically excludes the patches with noisy labels because of their low confidences. We should note that the idea of combining entropy and diversity was inspired by [3], but there is a fundamental difference because they computed R across the whole unlabeled dataset with time complexity $\mathcal{O}(m^2)$, which is very computational expensive, while we compute $R_i(j, l)$ locally on the selected patches within each candidate, saving computation time considerably with time complexity $\mathcal{O}(\alpha^2 m^2)$, where $\alpha = 1/4$ in our experiments.

3.4. An illustration of prediction patterns

Given unlabeled candidates $\mathcal{U} = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_n\}$ with $\mathcal{C}_i = \{x_i^1, x_i^2, ..., x_i^m\}$, assuming the prediction of patch x_i^j by the current CNN is p_i^j, we call the histogram of p_i^j for $j \in [1, m]$ the prediction pattern of candidate $\mathcal{C}_i$. As shown in Column 1 of Tab. 1, there are seven typical prediction patterns:

- Pattern A: The patches' predictions are mostly concentrated at 0.5, with a higher degree of uncertainty. Most active learning algorithms [20, 9] favor this type of candidate as it is good at reducing the uncertainty.
- Pattern B: It is flatter than Pattern A, as the patches' predictions are spread widely from 0 to 1, yielding a higher degree of inconsistency. Since all the patches belonging to a candidate are generated via data argumentation, they (at least the majority of them) are expected to have similar predictions. This type of candidate has the potential to contribute significantly to

enhancing the current CNN's performance.

- Pattern C: The patches' predictions are clustered at both ends, resulting in a higher degree of diversity. This type of candidate is most likely associated with noisy labels at the patch level as illustrated in Fig. 1, and it is the least favorable in active selection because it may cause confusion in fine-tuning the CNN.
- Patterns D and E: The patches' predictions are clustered at one end (i.e., 0 or 1) with a higher degree of certainty. The annotation of these types of candidates at this stage should be postponed because the current CNN has most likely predicted them correctly; they would contribute very little to fine-tuning the current CNN. However, these candidates may evolve into different patterns worthy of annotation with more fine-tuning.
- Patterns F and G: They have higher degrees of certainty in some of the patches' predictions and are associated with some outliers in the patches' predictions. These types of candidates are valuable because they are capable of smoothly improving the CNN's performance. Though they may not make significant contributions, they should not cause dramatic harm to the CNN's performance.
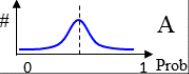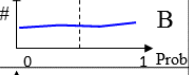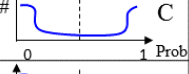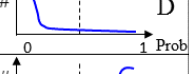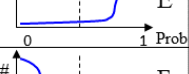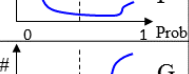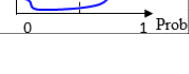
4. Applications

In this section, we apply our method to three different applications including colonoscopy frame classification, polyp detection, and pulmonary embolism (PE) detection. Our AIFT algorithm is implemented in the Caffe framework [11] based on the pre-trained AlexNet model [12]. In the following, we shall evaluate six variants of AIFT (active incremental fine-funing) including Diversity$^{1/4}$ (using diversity on 1/4 of the patches of each candidate), Diversity (using diversity on all the patches of each candidate), Entropy$^{1/4}$, Entropy, (Entropy+Diversity)$^{1/4}$, (Entropy+Diversity), and compare them with IFT Random (incremental fine-tuning with random candidate selection) and Learning from Scratch in terms of AUC (area under ROC curve).

4.1. Colonoscopy Frame Classification

Objective quality assessment of colonoscopy procedures is vital to ensure high-quality colonoscopy. A colonoscopy video typically contains a large number of non-informative images with poor colon visualization that are not ideal for inspecting the colon or performing therapeutic actions. The larger the fraction of non-informative images in a video, the lower the quality of colon visualization, thus the lower the quality of colonoscopy. Therefore, one way to measure the quality of a colonoscopy procedure is to monitor the quality of the captured images. Technically, image quality assessment at colonoscopy can be formulated as an image clas-

Table 1: Relationships among seven prediction patterns and six AIFT methods in active candidate selection. We assume that a candidate has 11 patches, and their probabilities predicted by the current CNN are listed in Column 2. AIFT Entropy$^\alpha$, Diversity$^\alpha$, and (Entropy+Diversity)$^\alpha$ operate on the top or bottom α percent of the candidate's patches based on the majority prediction as described in Sec. 3.3. In this illustration, we choose α to be 1/4, meaning that the selection criterion (Eq. 3) is computed based on 3 patches within each candidate. The first choice of each method is highlighted in yellow and the second choice is in light yellow.

| Prediction Pattern	Example	Entropy	Entropy$^{1/4}$	Diversity	Diversity$^{1/4}$	(Entropy+ Diversity)	(Entropy+ Diversity)$^{1/4}$
A	{0.4 0.4 0.4 0.5 0.5 0.5 0.5 0.5 0.5 0.6 0.6}	7.52	2.02	4.38	0.00	11.90	2.02
B	{0.0 0.1 0.2 0.3 0.4 0.4 0.6 0.7 0.8 1.0 1.0}	4.57	0.83	1237.21	20.79	1241.77	21.62
C	{0.0 0.0 0.0 0.0 0.1 0.1 0.9 0.9 1.0 1.0 1.0 1.0}	1.30	0.00	2816.66	0.00	2817.96	0.00
D	{0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.1 0.1 0.1}	1.30	0.00	189.54	0.00	190.84	0.00
E	{0.9 0.9 0.9 0.9 1.0 1.0 1.0 1.0 1.0 1.0 1.0}	1.30	0.00	189.54	0.00	190.84	0.00
F	{0.0 0.0 0.0 0.1 0.1 0.1 0.1 0.2 0.2 0.3 0.9 1.0}	3.24	0.33	1076.87	13.54	1080.11	13.86
G	{0.0 0.0 0.1 0.7 0.8 0.8 0.9 0.9 0.9 0.9 1.0 1.0}	3.24	0.33	1076.87	13.54	1080.11	13.86
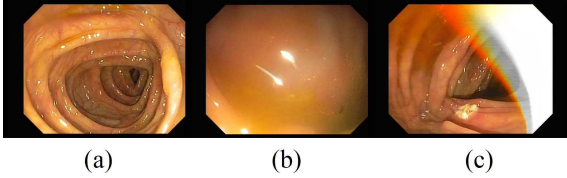
Figure 2: Three colonoscopy frames, (a) informative, (b) non-informative, and (c) ambiguous but labeled "informative" because it is mostly clear. The ambiguous frames contain both clear and blur parts, and generate noisy labels at the patch level via automatic data argumentation. Our AIFT method aims to automatically handle the label noise.

Figure 3: Comparing 8 methods in colonoscopy frame classification (see text for a detailed analysis).

sification task whereby an input image is labeled as either informative or non-informative.

For the experiments, 4,000 colonoscopy frames are selected from 6 complete colonoscopy videos. A trained expert then manually labeled the collected images as informative or non-informative. A gastroenterologist further reviewed the labeled images for corrections. The labeled frames at the video level are separated into training and test sets, each containing approximately 2,000 colonoscopy frames. For data augmentation, we extracted 21 patches from each frame.
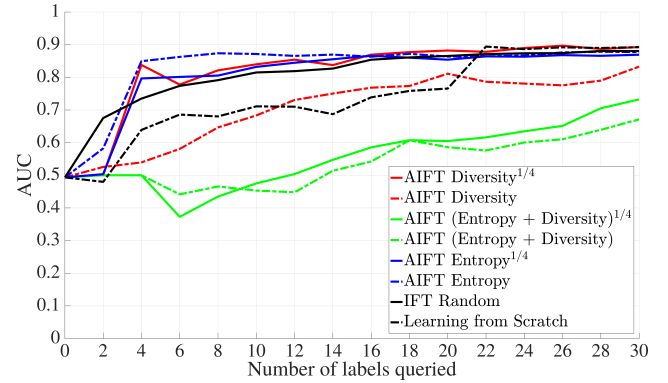
In all three applications, our AIFT begins with an empty training dataset and directly uses AlexNet pre-trained on ImageNet. Fig. 3 shows that at the first step (with 2 labels queried), IFT Random yields the best performance. There are two possible reasons: (1) random selection gives the samples with the positive/negative ratio compatible with the test dataset; (2) the pre-trained AlexNet gives poor predic-

tions on our dataset, as it was trained by natural images instead of biomedical images. Its output probabilities are mostly confused or even incorrect, yielding poor selection scores. However, AIFT Diversity$^{1/4}$, Entropy, Entropy$^{1/4}$ quickly surpass IFT Random after the first fine-tuning, as they select important samples for fine-tuning, making the training process more efficient than just randomly selecting from the remaining training dataset. AIFT Entropy and Diversity$^{1/4}$ with only 4 label queries can achieve the performance of IFT Random with 18 label queries, and that of Learning from Scratch with 22 randomly selected frames. Thereby, more than 75% labeling cost could be saved from IFT Random and 80% from Learning from Scratch.

AIFT Diversity works even poorer than IFT Random because of noisy labels generated through data augmentation. AIFT Diversity strongly favors frames whose prediction pattern resembles Pattern C (see Tab. 1). Naturally, it will most likely select an ambiguous frame such as Fig. 1 and Fig. 2 (c), because predictions of its patches are highly diverse. All patches generated from the same frame inherit the same label as the frame; therefore, at the patch level, the labels are very noisy for the ambiguous frames. AIFT Entropy, Entropy$^{1/4}$, and Diversity$^{1/4}$ can automatically exclude the noisy label, naturally yielding outstanding performance. Given the outstanding performance of AIFT Entropy, Entropy$^{1/4}$, and Diversity$^{1/4}$, one may consider combining entropy and diversity, but unfortunately, combinations do not always give better performance, because finding a nice balance between entropy and diversity is tricky as shown in our example analysis in Tab. 1 and supplementary material.

4.2. Polyp Detection

Colonoscopy is the preferred technique for colon cancer screening and prevention. The goal of colonoscopy is to find and remove colonic polyps—precursors to colon cancer—as shown in Fig. 4. For polyp detection, our database contains 38 short colonoscopy videos from 38 different patients, and they are separated into the training dataset (21 videos; 11 with polyps and 10 without polyps) and the testing dataset (17 videos; 8 videos with polyps and 9 videos without polyps). There are no overlaps between the training dataset and testing dataset at the patient level. Each colonoscopy frame in the data set comes with a binary ground truth image. 16300 candidates and 11950 candidates were generated from the training dataset and testing dataset, respectively.

At each polyp candidate location with the given bounding box, we perform a data augmentation by a factor $f \in \{1.0, 1.2, 1.5\}$. At each scale, we extract patches after the candidate is translated by 10 percent of the resized bounding box in vertical and horizontal directions. We further rotate each resulting patch 8 times by mirroring and flipping. The
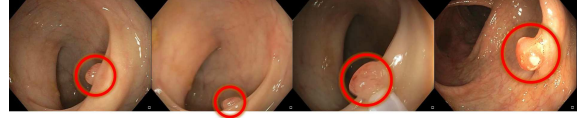

Figure 4: Polyps in colonoscopy videos with different shape and appearance.
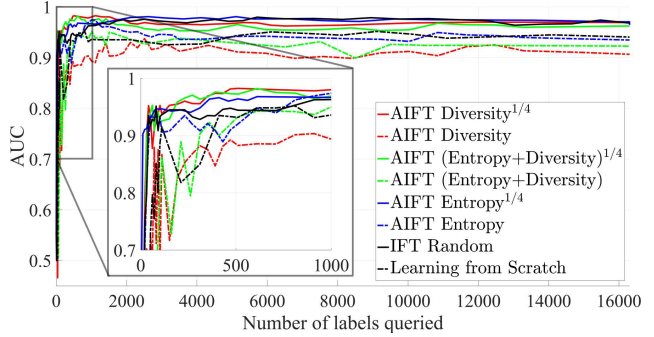

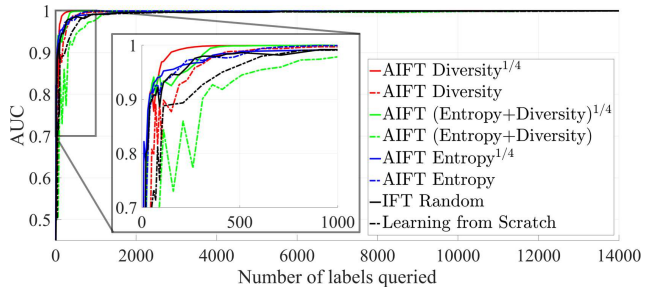Figure 5: Comparing 8 methods in polyp detection (see text for a detailed analysis).

Figure 6: Monitor the performance of the proposed method on the remaining training dataset. Using 5% of the whole training dataset (800/16300), the CNN can predict almost perfectly on the remaining 95% dataset.

patches generated by data augmentation belong to the same candidate.

Fig. 5 shows that AIFT (Entropy+Diversity)$^{1/4}$ and Diversity$^{1/4}$ reach the peak performance with 610 label queries, while IFT Random needs 5711 queries, indicating that AIFT can cut nearly 90% of the annotation cost required by IFT Random. The fast convergence of AIFT (Entropy+Diversity)$^{1/4}$ and Diversity$^{1/4}$ is attributed to the majority selection method, which can efficiently select the informative and representative candidates while excluding those with noisy labels. When the queried number is about 5000, the AIFT Entropy$^{1/4}$ reaches its peak performance. The reason is that the entropy can only measure the informativeness so the queried sample is very likely to be similar to each other. It needs more queries to select most

of the informative candidates. AIFT Diversity and (Entropy+Diversity) cannot perform as well as the counterparts with the majority selection due to noisy labels. Learning from Scratch never achieves the performance of fine-tuning even if all training samples are used, which is in agreement with [24].

To gain further insights, we also monitor the performance of the 8 methods on the remaining training dataset. Each time after we have fine-tuned the previous CNN, we test it on the remaining training dataset. We have observed that only 800 candidates are needed to reach the maximum performance. As is shown in Fig. 6, the candidates selected by our method, which are only 5% (800/16300) of all the candidates, can represent the remaining dataset, because in colonoscopy videos consecutive frames are usually similar to each other.

4.3. Pulmonary Embolism Detection

Our experiments are based on the PE candidates generated by the method proposed in [16] and the image representation introduced in [23] as shown in Fig. 7. We adopt the 2-channel representation because it consistently captures PEs in cross-sectional and longitudinal views of vessels, achieving greater classification accuracy and accelerating CNN training process. In order to feed the RGB-like patches into CNN, the 2-channel patches are converted to 3-channel RGB-like patches by duplicating the second channel. For experiments, we use a database consisting of 121 CTPA datasets with a total number of 326 PEs. The tobogganing algorithm [16] is applied to obtain a crude set of PE candidates. 6255 PE candidates are generated, of which 5568 are false positives and 687 are true positives. To train CNN, we extract patches of 3 different physical sizes, i.e.,10 mm-, 15 mm-, and 20 mm-wide. Then, we translate each candidate location along the direction of the affected vessel 3 times, up to 20% of the physical size of each patch. Then, data augmentation for training dataset is performed by rotating the longitudinal and cross-sectional vessel planes around the vessel axis, resulting in 5 additional variations for each scale and translation.

Finally, a stratified training dataset with 434 true positive PE candidates and 3406 false positive PE candidates would be generated for training and incrementally fine-tuning the CNN and a testing dataset with 253 true positive PE candidates and 2162 false positive PE candidates. The overall PE probability is calculated by averaging the probabilistic prediction generated for the patches within PE candidate after data augmentation.

Fig. 8 compares the 8 methods on the testing dataset. The performance of each method becomes saturated after 2000 labels queried. AIFT (Entropy+Diversity)$^{1/4}$ and Diversity$^{1/4}$ converge the fastest among the 8 methods and yields the best overall performance, attributed to
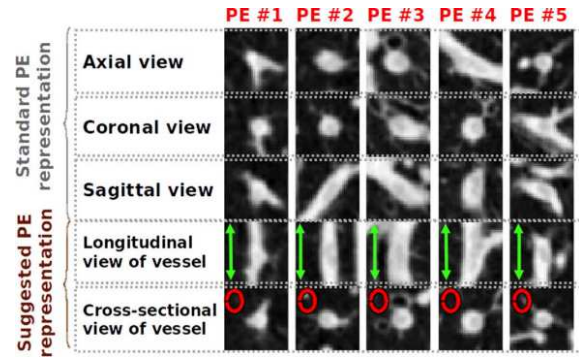

Figure 7: Five different PEs in the standard 3-channel representation, as well as in the 2-channel representation [23] , which was adopted in this work because it achieves greater classification accuracy and accelerates CNN training convergence. The figure is used with permission.
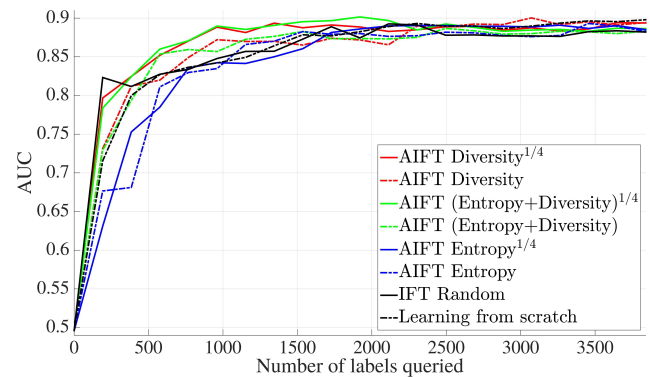

Figure 8: Comparing 8 methods in pulmonary embolism detection (see text for a detailed analysis).

majority selection method proposed in this work. AIFT (Entropy+Diversity)$^{1/4}$ and Diversity$^{1/4}$ with only 1000 labels required can achieve the performance of random selecting 2200 labels fine-tune from AlexNet (IFT Random). Note that even AIFT Diversity reach its peak performance when about 3100 samples queried because PE data set injected little noisy labels. Since entropy favors the uncertain ambiguous samples, both AIFT Entropy$^{1/4}$ and Entropy perform bad at the beginning. IFT Random outperforms at the first few steps as analysed in Sec. 4.1, but increase slowly overall. Based on this analysis, the cost of annotation can be cut at least half by the our method.

4.4. Observations on selected patterns

We meticulously monitored the active selection process and examined the selected candidates, as an example, we include the top 10 candidates selected by the six AIFT methods at Iteration 3 in colonoscopy frame classification in the

supplementary material (see Fig. 10). From this process, we have observed the following:

- Patterns A and B are dominant in the earlier stages of AIFT as the CNN has not been fine-tuned properly to the target domain.
- Patterns C, D and E are dominant in the later stages of AIFT as the CNN has been largely fine-tuned on the target dataset.
- The majority selection—AIFT Entropy$^{1/4}$, Diversity$^{1/4}$, or (Entropy+Diversity)$^{1/4}$—is effective in excluding Patterns C, D, and E, while AIFT Entropy (without the majority selection) can handle Patterns C, D, and E reasonably well.
- Patterns B, F, and G generally make good contributions to elevating the current CNN's performance.
- AIFT Entropy and Entropy$^{1/4}$ favor Pattern A because of its higher degree of uncertainty as shown in Fig. 10.
- AIFT Diversity$^{1/4}$ prefers Pattern B while AIFT Diversity prefers Pattern C (Fig. 10). This is why AIFT Diversity may cause sudden disturbances in the CNN's performance and why AIFT Diversity$^{1/4}$ should be preferred in general.
- Combing entropy and diversity would be highly desirable, but striking a balance between them is not trivial, because it demands application-specific λ_1 and λ_2 (see Eq. 3) and requires further research.

5. Conclusion, discussion and future work

We have developed an active, incremental fine-tuning method, integrating active learning with transfer learning, offering several advantages: It starts with a completely empty labeled dataset, and incrementally improves the CNN's performance through continuous fine-tuning by actively selecting the most informative and representative samples. It also can automatically handle noisy labels via majority selection and it computes entropy and diversity locally on a small number of patches within each candidate, saving computation time considerably. We have evaluated our method in three different biomedical imaging applications, demonstrating that the cost of annotation can be cut by at least half. This performance is attributed to the advanced active and incremental capability of our AIFT method.

We based our experiments on the AlexNet architecture because a pre-trained AlexNet model is available in the Caffe library and its architecture strikes a nice balance in depth: it is deep enough that we can investigate the impact of AIFT on the performance of pre-trained CNNs, and it is also shallow enough that we can conduct experiments quickly. Alternatively, deeper architectures such as VGG, GoogleNet, and Residual network could have been used and have shown relatively high performance for challenging computer vision tasks. However, the purpose of this work is
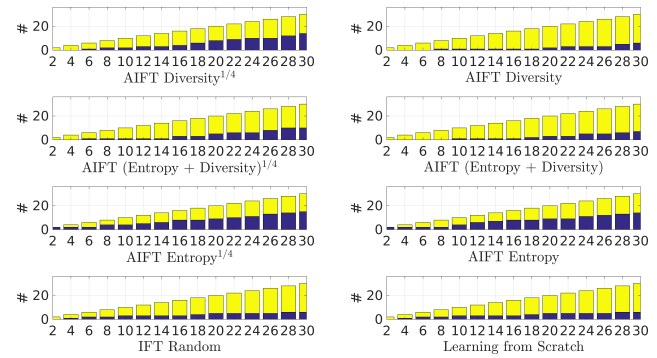

Figure 9: Positive/negative ratio in the samples selected by six methods. Yellow bar represents the negatives and blue bar represents the positives.

not to achieve the highest performance for different biomedical image tasks but to answer the critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging.* The architecture and learning parameters are reported in the supplementary material.

In the real world, datasets are usually unbalanced. In order to achieve good classification performance, both classes of samples should be used in training. Fig. 9 shows the positive/negative label ratio of the samples selected by the six methods in each iteration in colonoscopy quality application. For random selection, the ratio is nearly the same as whole training dataset, a reason that IFT Random has stable performance at the cold-start. AIFT Diversity$^{1/4}$, Entropy$^{1/4}$ and Entropy seem capable of keeping the dataset balanced automatically, a new observation that deserves more investigation in the future.

We choose to select, classify and label samples at the candidate level. Labeling at the patient level would certainly reduce the cost of annotation more but introduce more severe label noise; labeling at the patch level would cope with the label noise but impose a much heavier burden on experts for annotation. We believe that labeling at the candidate level offers a sensible balance in our three applications.

Finally, in this paper, we use only entropy and diversity as the criteria. In theory, a large number of active selection methods may be designed, but we have found that there are only seven fundamental patterns as summarized in the Sec. 3.4. As a result, we could conveniently focus on comparing the seven patterns rather than the many methods. Multiple methods may be used to select a particular pattern: for example, entropy, Gaussian distance, and standard deviation would seek Pattern A, while diversity, variance, and divergence look for Pattern C. We would not expect significant performance differences among the methods within each group, resulting in six major selction methods for deep comparisons based on real-world clinical applications.

References

[1] M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016. 2

[2] G. Carneiro, J. Nascimento, and A. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pages 652–660. Springer International Publishing, 2015. 2

[3] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1945–1958, 2015. 4

[4] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 507–514. Springer, 2015. 2

[5] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng. Standard plane localization in fetal ultrasound via domain transferred deep neural networks. *Biomedical and Health Informatics, IEEE Journal of*, 19(5):1627–1636, Sept 2015. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[7] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H.-C. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. In *the 1st Workshop on Deep Learning in Medical Image Analysis, International Conference on Medical Image Computing and Computer Assisted Intervention, at MICCAI-DLMIA'15*, 2015. 2

[8] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016. 1, 2

[9] I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov. *JMLR Workshop and Conference Proceedings (Volume 16): Active Learning Challenge*. Microtome Publishing, 2011. 2, 4

[10] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008. 2

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 4

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[13] M. Kukar. Transductive reliability estimation for medical diagnosis. *Artificial Intelligence in Medicine*, 29(1):81–106, 2003. 3

[14] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 1

[15] J. Li. Active learning for hyperspectral image classification with a stacked autoencoders based neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1062–1065, Sept 2016. 2

[16] J. Liang and J. Bi. Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in ct pulmonary angiography. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 630–641. Springer, 2007. 7

[17] L. Lu, Y. Zheng, G. Carneiro, and L. Yang. *Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Large-Scale Datasets*. Springer, 2016. 1, 2

[18] J. Margeta, A. Criminisi, R. Cabrera Lozoya, D. C. Lee, and N. Ayache. Fine-tuned convolutional neural nets for cardiac mri acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–11, 2015. 2

[19] T. Schlegl, J. Ofner, and G. Langs. Unsupervised pre-training across image domains improves lung tissue classification. In *Medical Computer Vision: Algorithms for Big Data*, pages 82–93. Springer, 2014. 2

[20] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11. 2, 4

[21] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. M. Summers. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2015. 2

[22] F. Stark, C. Hazırbaş, R. Triebel, and D. Cremers. Captcha recognition with active deep learning. In *Workshop New Challenges in Neural Computation 2015*, page 94. Citeseer, 2015. 2

[23] N. Tajbakhsh, M. B. Gotway, and J. Liang. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 62–69. Springer, 2015. 7

[24] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016. 2, 7

[25] D. Wang and Y. Shang. A new active labeling method for deep learning. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 112–119, July 2014. 2

[26] H. Wang, Z. Zhou, Y. Li, Z. Chen, P. Lu, W. Wang, W. Liu, and L. Yu. Comparison of machine learning methods for

classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 f-fdg pet/ct images. *EJNMMI research*, 7(1):11, 2017. 2

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 1

[28] K. Zhou, H. Greenspan, and D. Shen. *Deep Learning for Medical Image Analysis*. Academic Press, 2016. 1, 2

Supplementary material

The AlexNet architecture and learning parameters used in our experiments

As discussed in Sec. 5, the purpose of this work is not to achieve the highest performance for different biomedical image tasks but to answer the critical question: *How to dramatically reduce the cost of annotation when applying CNNs in biomedical imaging.* For this purpose, we base our experiments on AlexNet, whose architecture is shown in Table 2, as it is deep enough that we can investigate the impact of AIFT on the performance of pre-trained CNNs, and also small enough that we can conduct experiments quickly. Learning parameters used for the training and fine-tuning of AlexNet in our experiments are summarized in Table 3.

Table 2: The AlexNet architecture used in our experiments. Of note, C is 2 as all our three applications are binary classifications by nature.

layer	type	input	kernel	stride	pad	output
data	input	3x227x227	N/A	N/A	N/A	3x227x227
conv1	convolution	3x227x227	11x11	4	0	96x55x55
pool1	max pooling	96x55x55	3x3	2	0	96x27x27
conv2	convolution	96x27x27	5x5	1	2	256x27x27
pool2	max pooling	256x27x27	3x3	2	0	256x13x13
conv3	convolution	256x13x13	3x3	1	1	384x13x13
conv4	convolution	384x13x13	3x3	1	1	384x13x13
conv5	convolution	384x13x13	3x3	1	1	256x13x13
pool5	max pooling	256x13x13	3x3	2	0	256x6x6
fc6	fully connected	256x6x6	6x6	1	0	4096x1
fc7	fully connected	4096x1	1x1	1	0	4096x1
fc8	fully connected	4096x1	1x1	1	0	Cx1

Table 3: Learning parameters used for the training and fine-tuning of AlexNet in our experiments. μ is the momentum, α_{fc8} is the learning rate of the weights in the last layer, α is the learning rate of the weights in the rest layers, and γ determines how α decreases over epochs. The learning rate for the bias term is always set twice as large as the learning rate of the corresponding weights. "Epochs" indicates the number of epochs used in each AIFT iteration. $AIFT_1$ indicates the first iteration of AIFT while $AIFT_+$ indicates all the following iterations of AIFT.

Application	Method	μ	α	α_{fc8}	γ	epochs
	$AIFT_1$	0.9	0.0001	0.001	0.95	20
Colonoscopy Frame Classification	$AIFT_+$	0.9	0.0001	0.0001	0.95	15
	Learning from Scratch	0.9	0.0001	0.001	0.95	20
	$AIFT_1$	0.9	0.001	0.01	0.95	5
Polyp Detection	$AIFT_+$	0.9	0.0001	0.001	0.10	3
	Learning from Scratch	0.9	0.001	0.01	0.95	10
	$AIFT_1$	0.9	0.001	0.01	0.95	10
Pulmonary Embolism Detection	$AIFT_+$	0.9	0.001	0.01	0.10	5
	Learning from Scratch	0.9	0.001	0.01	0.95	20

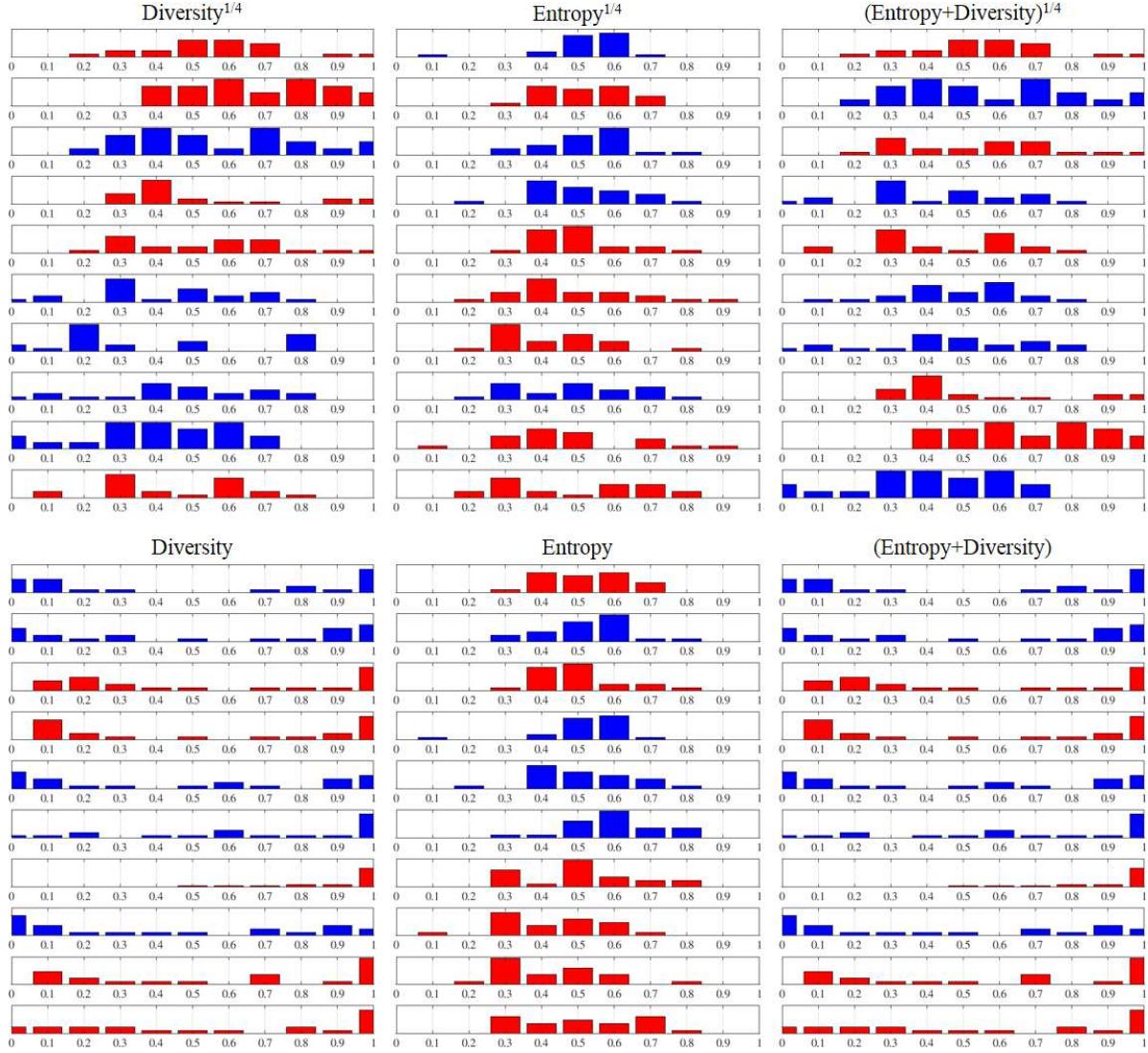[1] Polyp Detection AIFT Diversity$_+$: 0.9 | 0.001 | 0.01 | 0.50 | 3

Figure 10: Top 10 candidates selected by the six AIFT methods at Iteration 3 in colonoscopy frame classification. Positive candidates are in red and negative candidates are in blue. Both AIFT Entropy and AIFT Entropy$^{1/4}$ favor Pattern A because of its higher degrees of uncertainty. AIFT Diversity$^{1/4}$ prefers Pattern B while AIFT Diversity suggests Pattern C. With $\lambda_1 = \lambda_2 = 1$ (Eq. 3), diversity is dominant in AIFT (Entropy+Diversity) and (Entropy+Diversity)$^{1/4}$.