An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis

Xueying Shi[1]($\boxtimes$), Qi Dou[2], Cheng Xue[1], Jing Qin[4], Hao Chen[1,3], Pheng-Ann Heng[1,5]

[1] Department of Computer Science and Engineering
The Chinese University of Hong Kong, Hong Kong, China
xyshi@cse.cuhk.edu.hk
[2] Department of Computing Imperial College London, London SW7 2AZ, U.K.
[3] Imsight Medical Technology Co., Ltd., Shenzhen, China
[4] Centre for Smart Health, School of Nursing
The Hong Kong Polytechnic University, Hong Kong, China
[5]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality
Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of
Sciences, Shenzhen, China

Abstract. Automated skin lesion analysis is very crucial in clinical practice, as skin cancer is among the most common human malignancy. Existing approaches with deep learning have achieved remarkable performance on this challenging task, however, heavily relying on large-scale labelled datasets. In this paper, we present a novel active learning framework for cost-effective skin lesion analysis. The goal is to effectively select and utilize much fewer labelled samples, while the network can still achieve state-of-the-art performance. Our sample selection criteria complementarily consider both informativeness and representativeness, derived from decoupled aspects of measuring model certainty and covering sample diversity. To make wise use of the selected samples, we further design a simple yet effective strategy to aggregate intra-class images in pixel space, as a new form of data augmentation. We validate our proposed method on data of *ISIC 2017 Skin Lesion Classification Challenge* for two tasks. Using only up to 50% of samples, our approach can achieve state-of-the-art performances on both tasks, which are comparable or exceeding the accuracies with full-data training, and outperform other well-known active learning methods by a large margin.

1 Introduction

Skin cancer is among the most common cancers worldwide, and accurate analysis of dermoscopy images is crucial for reducing melanoma deaths [3]. Existing deep convolutional neural networks (CNNs) have demonstrated appealing efficacy for skin lesion analysis, even setting dermatologist-level performance. However, these achievements heavily rely on extensive labelled datasets, which is very expensive, time-consuming and skill-demanding. Recently, with increasing awareness of the impediment from unavailability of large-scale labeled data, researchers have been

frequently revisiting the concept of active learning to train CNNs in a more cost-effective fashion [7]. The goal is to learn CNNs with much fewer labelled images, while the model can still achieve the state-of-the-art performance.

Sample selection criteria usually use informativeness or representativeness [4]. Informative samples are the ones which the current model still cannot recognize well. For example, Mahapatra et al. [6] derived uncertainty metrics via a Bayesian Neural Network to select informative samples for chest X-ray segmentation. On the other hand, representativeness measures whether the set of selected samples are diverse enough to represent the underlying distributions of the entire data space. Zheng et al. [12] chose representative samples with unsupervised feature extraction and clusters in latent space. Moreover, rather than only relying on one single criterion, some works actively select samples by integrating both criteria. Yang et al. [9] selected samples which receive low prediction probabilities and have large distances in CNN feature space. Another state-of-the-art method is AIFT [13] (active, incremental fine-tuning), which employed the entropy of CNN predictions for a sample to compute its informativeness as well as representativeness, demonstrating effectiveness on three medical imaging tasks. However, these existing methods derive both criteria based on the same CNN model, which hardly avoid potential correlations within the selected samples. How to exploit such dual-criteria in a more disentangled manner still remains open.

With active sample selection, the data redundancy of unlabelled sample pool is effectively reduced. Meanwhile, we should note that the obtained set of images come with high intra-class variance in color, texture, shape and size [8,11]. Directly using such samples to fine-tune the model may fall into more-or-less hard example mining, and face the risk of over-fitting. Hence, we argue that it is also very critical to more wisely use the compact set of selected samples, for unleashing their value to a large extent. However, sample utilization strategies receive less attention in existing active learning literatures. One notable method is mix-up [10], that augments new training data as pixel-wise weighted addition of two images from different classes. However, mix-up is not suitable for situations where data have large intra-class variance while limited inter-class variance, which is exactly our case at skin lesion analysis.

In this work, we propose a novel active learning method for skin lesion analysis to improve annotation efficiency. Our framework consists of two components, i.e., sample selection and sample aggregation. Specifically, we design dual-criteria to select informative as well as representative samples, so that the selected samples are highly complementary. Furthermore, for effective utilization of the selected samples, we design an aggregation strategy by augmenting intra-class images in pixel space, in order to capture richer and more distinguishable features from these valuable yet ambiguous selected samples. We validate our approach on two tasks with the dataset of *ISIC 2017 Skin Lesion Classification Challenge*. We achieve state-of-the-art performance by using 50% data for task 1 and 40% for task 2 of skin lesion classification tasks. In both tasks, our proposed method consistently outperforms existing state-of-the-art active learning methods by a large margin.
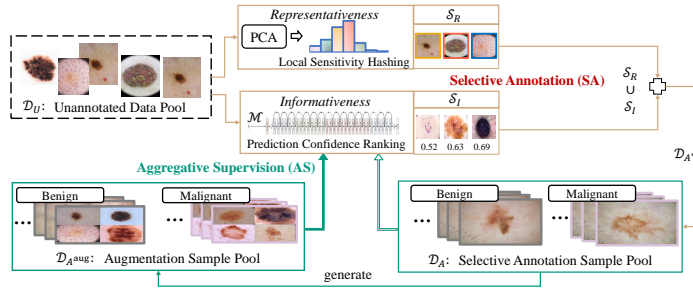
Fig. 1. Overview of our proposed active learning framework for skin lesion analysis. In each iteration, from unannotated data pool $\mathcal{D}_U$, we select a worthy-annotation set $\mathcal{D}_{A^*}$ composing representative samples $\mathcal{S}_R$ and informative samples $\mathcal{S}_I$. Moreover, we generate augmentations $\mathcal{D}_{A^{\mathrm{aug}}}$ of all the gathered annotated data pool $\mathcal{D}_A$. Finally, the model is updated by supervised learning with $\mathcal{D}_A \cup \mathcal{D}_{A^{\mathrm{aug}}}$.

2   Methods

Our framework is illustrated in Fig. 1. We first train the ResNet-101 model $\mathcal{M}$ with the annotated set of $\mathcal{D}_A = \{(x_j, y_j)\}_{j=1}^Q$, which is initialized with randomly selected 10% data from the unlabelled sample pool $\mathcal{D}_U = \{x_i\}_{i=1}^T$. Next, we iteratively selecting samples, aggregating samples, and updating the model.

2.1   Selective Annotation (SA) with Dual-Criteria

We select samples considering both criteria of informativeness and representativeness. The informativeness is calculated based on the prediction of the trained model. The representativeness is obtained by PCA features and hashing method. In our framework, we call this procedure as *selective annotation (SA)*.

Firstly, we test the unlabelled samples with the current trained model. The images with low prediction confidences computed from the model are selected as informative ones, since they are nearby the decision boundary. The model would present relatively lower confidence when encountering those new "hard" unlabelled samples, which usually have either ambiguous pattern or rare appearance. For each sample, the highest prediction probability across all classes, is regarded as its model certainty. With ranking $\mathcal{D}_U$ according to certainties, the lower certainty indicates stronger informativeness. The selected samples following this aspect of criteria are represented as $\mathcal{S}_I$:

$$\mathcal{S}_I \leftarrow \operatorname*{Rank}_{x_i}(\{\mathcal{M}(x_i)\}, N_I), \tag{1}$$

where $\mathcal{M}(x_i)$ is certainty of current model $\mathcal{M}$ for each sample x_i in $\mathcal{D}_U$, ranking is in ascending order, and the first N_I samples are selected. We set $N_I = 10\% \times N \times \gamma$ where N is the total number of available samples, and γ is the sample selection ratio of informativeness criterion. 10% is the hyper-parameter which controls the scale of newly selected samples during each round of sample selection.
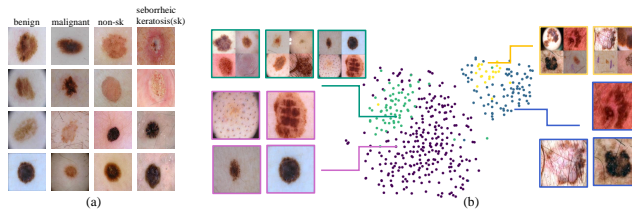
Fig. 2. (a) Skin lesion images with limited inter-class variance while large intra-class variance. (b) Embedding of high-level CNN features of the original and augmented samples using t-SNE. The purple and green dots are original and augmented benign data. The blue and yellow dots are original and augmented malignant data. Augmented samples are natural clustering with original ones in the high-level semantic space.

Next, considering sample diversity, we desire the added samples present dissimilar appearances, and hence, are representative for the entire dataset. Specifically, we regard feature-level difference as an indicator of sample diversity. To avoid using features from the same CNN as used for informative sample selection, we compute the first principal component of the image as features for data diversity. With the PCA features, we map similar unlabelled items into the same buckets using local sensitivity hashing (LSH), which is for efficient approximate nearest neighbor search. Next, we uniformly fetch samples from each bucket and obtain the set of $\mathcal{S}_R$ as representative samples. This process is formulated as:

$$\mathcal{S}_R \leftarrow \underset{x_i}{\mathrm{UniSample}}(LSH(\{PCA(x_i)\}, K), N_R), \qquad (2)$$

where $K = 10$ is our number of buckets in LSH. We set $N_R = 10\% \times N \times (1 - \gamma)$ with $(1 - \gamma)$ being the sample selection ratio of representativeness criterion. As the PCA features are independent of the learned CNN, our obtained $\mathcal{S}_R$ and $\mathcal{S}_I$ are decoupled and highly complementary. With one round of SA, we get the additional labelled set of samples as $\mathcal{D}_{A^*} = \mathcal{S}_I \cup \mathcal{S}_R$ and update $\mathcal{D}_A \leftarrow \mathcal{D}_A \cup \mathcal{D}_{A^*}$.

2.2 Aggregative Supervision (AS) with Intra-class Augmentation

In active learning, majority previous efforts have focused on how to select samples, but somehow neglected how to effectively harness them to produce more distinguishable features. As usually the selected training samples are very challenging and ambiguous, it is important to design strategies which can sufficiently unleash the potential values of these newly labelled samples. If just directly using such samples to fine-tune the model, we may encounter high risks of over-fitting, since the updated decision boundary would be curly to fit the ambiguous images. To enhance the model's capability to deal with those ambiguous samples, we propose to aggregate the images into new form of augmented samples to update the model. In our framework, we call this procedure as *aggregative supervision (AS)*.

Specifically, we aggregate images from the same class in pixel space, by stitching four intra-class images in a 2×2 pattern, as presented in Fig. 2 (b). Such a

concatenation of samples from the same class can provide richer yet more subtle clues for the model to learn more robust features to reduce intra-class variance, especially given the highly ambiguous and limited number of samples obtained from SA process. In a sense that the model aims to discriminate between distributions of benign and malignant images, the proposed sample aggregation scheme can be beneficial to reduce the influence of individual complicated sample on the model, and percolate the underlying pattern inherent in each category. Finally, the aggregated image is resized to the same size as the original resolution, and its label is the same class of those composed images. Generally, our strategy shares the pixel-level augmentation spirit as mix-up [10], while we can avoid overlapping the ambiguous contents of inter-class images with limited appearance difference.

To demonstrate the effectiveness of the proposed aggregation scheme at feature level, we embed the CNN features of the original images and the aggregated images with our intra-class stitching onto a 2D plane using t-SNE, see Fig. 2. We employ the features obtained from the last fully connected layer (before softmax), as these features have strong semantic meanings. Note that these aggregated samples haven't yet been used to train the model. We observe that the aggregated samples naturally group together with the ordinary images within the class, when mapped into the higher-level space with a pre-learned feature extractor (i.e., the CNN model). This demonstrates that our aggregation scheme can provide a new and informative form of training images, offering apparently different view in raw pixel space while maintaining the essential patterns of its category in the highly-abstracted semantic space.

3 Experimental Results

Dataset. We extensively validate our proposed active learning framework on two different tasks using the public data of *ISIC 2017 Skin Lesion Classification Challenge* [1]. These two tasks hold different aspects of challenges and sample ambiguity characteristics.

Same as the state-of-the-art methods on the leaderboard [2,5], in addition to the 2,000 training images provided by the challenge, we acquired 1,582 more images (histology or expert confirmed studies) from the ISIC archive [1] to build up our available database. In total, we got 3,582 labelled images (2,733 benign and 849 malignant) as our training data pool. We directly utilized the validation set (150 images) and test set (600 images) of the ISIC challenge.

Implementations. The luminance and color balance of input images are normalized exploiting color constancy by gray world. Images are resized to $224{\times}224$ to match input size of pre-trained ResNet-101 model. The images are augmented with rotating by up to $90°$, shearing by up to $20°$, scaling within [0.8, 1.2], and random flipping horizontally and/or vertically. We use weighted cross-entropy loss with Adam optimizer and initial learning rate as 1e-4. Code will be released.

Evaluation metrics. For quantitative comparisons, our evaluations followed

Table 1. Quantitative evaluations of our proposed active learning framework for skin lesion analysis on two different classification tasks.

Methods		Data Amount	Extra Label	Task1					Task2				
				ACC	AUC	AP	SE	SP	ACC	AUC	AP	SE	SP
Leaderboard	Monty [2]	100%	√	0.823	0.856	0.654	0.103	**0.998**	0.875	**0.965**	**0.839**	0.178	**0.998**
	Popleyi [5]	100%	×	0.858	**0.870**	**0.694**	0.427	0.963	**0.918**	0.921	0.770	0.589	0.976
	Full-data (ResNet-101)	100%	×	**0.863**	0.821	0.590	**0.496**	0.952	0.903	0.941	0.773	**0.856**	0.912
Selection	Random (Rand)	50%/40%	×	0.825	0.795	0.520	0.359	0.934	0.878	0.923	0.731	0.722	0.906
	AIFT [13]	50%/40%	×	0.810	0.754	0.447	**0.385**	0.913	0.885	0.907	0.677	0.711	**0.916**
	SA (Ours)	50%/40%	×	**0.847**	**0.800**	**0.575**	0.368	**0.963**	**0.903**	**0.938**	**0.784**	**0.844**	0.914
Aggregation	SA (Ours)+Mix-up [10]	50%/40%	×	0.467	0.572	0.273	**0.615**	0.431	0.720	0.638	0.361	0.124	0.824
	SA+AS (**Ours**)	50%/40%	×	**0.860**	**0.831**	**0.600**	0.479	**0.952**	**0.908**	**0.934**	**0.755**	**0.756**	**0.935**

the challenge given metrics, which consist of accuracy (ACC), area under ROC curve (AUC), average precision (AP), sensitivity (SE) and specificity (SP).

3.1   Results of Cost-effective Skin Lesion Analysis

In our active learning process, based on the initially randomly selected 10% data, we iteratively added training samples until obtaining predictions which cannot be significantly improved ($p > 0.05$) over the accuracy of last round. It turns out that we only need 50% of the data for Task-1 and 40% of the data for Task-2.

The overall performance for Task-1 and Task-2 are representatively presented in Fig. 3(a) and Fig. 4(a). In Fig. 3(a), we present the baseline of active learning which is random sample selection (purple). By using our proposed dual-criteria sample selection (green), the accuracy gradually increases and keeps higher than the baseline through different query ratios. Further using our aggregative supervision (red), the accuracy achieves 86.0% when using only 50% samples, which is very close to the accuracy of 86.3% with full-data training (yellow). In Fig. 4(a), by actively querying worth-labelling images, our proposed method can finally exceed the performance of full-data training only using 40% samples. In addition, when comparing with the state-of-the-art method of AIFT (blue) [13], our proposed method can outperform it consistently across all sample query ratios on both tasks. This validates that our deriving dual-criteria in a decoupled way is better than only relying on currently learned network.

In Table 1, we categorize the different comparison methods into three groups, i.e., the leading methods in challenge, active learning only with sample selection strategy, and further adding the sample augmentation strategy. The amount of employed annotated data is indicated in *data amount* column. For leaderboard, only rank-2 [2] and rank-4 [5] methods are included, as rank-1 method used non-image information (e.g., sex and age) and rank-3 method used much more extra data besides the ISIC archive ones. Nevertheless, we present the challenge results for demonstrating the state-of-the-art performance of this dataset. We focus on active learning part, with our implemented full-data training as standard bound. From the Table 1, we see that our SA can outperform AIFT [13], and AS can outperform mix-up [10], across almost all evaluation metrics on both tasks. Overall, our proposed method achieves highly competitive results against full-data training and challenge leaderboard, with significantly cost-effective labellings.
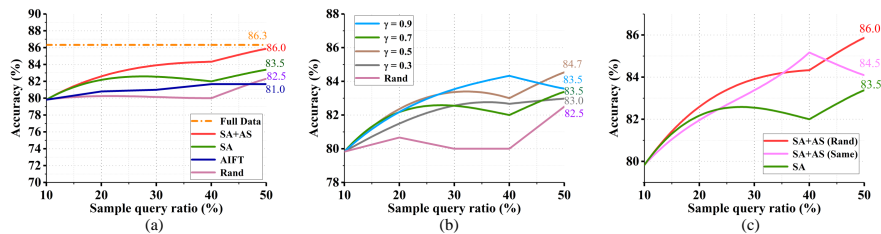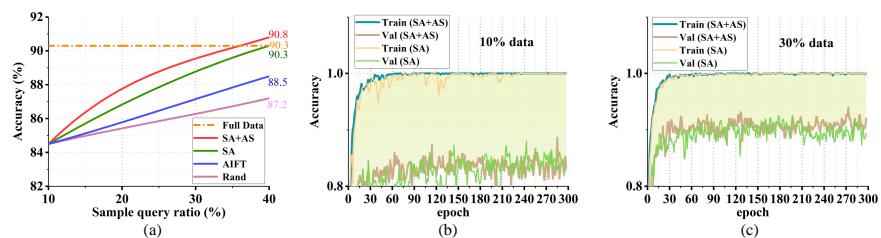
Fig. 3. Experimental results of our proposed active learning framework on Task 1. (a) Overall accuracy of different methods at sample query ratios. (b) Ablation study of SA, by adjusting γ. (c) Ablation study of AS, by changing the choice of stitched images.

Fig. 4. Experimental results of our proposed active learning framework on Task 2. (a) Overall accuracy of different methods at sample query ratios. (b)-(c) Observation of narrowing generalization gap and alleviating over-fitting by our active learning method.

3.2   Analysis of Components in Active Learning Framework

Firstly, we investigate the impact of hyper-parameter setting in sample selection. We adjust the ratio between $\mathcal{S}_I$ and $\mathcal{S}_R$ by changing the γ, as shown in Fig. 3(b). Varying the ratio γ would bring fluctuation on performance, but even the worst case is much better than random selection. We choose to use $\gamma = 0.7$ as the basis AS process, since it reflects the average-level performance of the SA step.

Secondly, we investigate the practically effective manner to stitch the intra-class samples. As shown in Fig. 3(c), we compare stitching four randomly selected intra-class images and replicating the same image by four times. For aggregative supervision, stitching different intra-class images can outperform replicating the same image, which exactly reflects that our designed augmentation strategy can help to improve performance by suppressing intra-class variance and sample ambiguity.

Finally, the results in Fig. 4(b) and Fig. 4(c) show that our proposed augmentation strategy can alleviate overfitting during model training. The shadow area indicates the generalization gap between the training and validation sets. It unsurprisingly decreases with increasing the data amount from 10% to 30%. With more careful observation, we find that the SA+AS can generally surpass pure SA on validation set, which demonstrates the effectiveness of alleviating over-fitting using our augmented new-style samples.

4   Conclusion

This paper presents a novel active learning method for annotation cost-effective skin lesion analysis. We propose a dual-criteria to select samples, and an intra-class sample aggregation scheme to enhance the model. Experimental results demonstrate that using only up to 50% of the labelled samples, we can achieve the state-of-the-art performance on two different skin lesion analysis tasks.

References

1. Codella et al., N.C.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: ISBI. pp. 168–172 (2018)
2. Diaz, I.G.: Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis. IEEE Journal of Biomedical and Health Informatics pp. 547–559 (2018)
3. Esteva, Andre et al., K.B.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
4. Huang, S.J., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. In: NIPS. pp. 892–900 (2010)
5. Lei, B., Jinman, K., Euijoon, A., Dagan, F.: Automatic skin lesion analysis using large-scale dermoscopy image and deep residual networks. arXiv:1703.04197 (2017)
6. Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M.: Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In: MICCAI. pp. 580–588 (2018)
7. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
8. Xue, C., Dou, Q., Shi, X., Chen, H., Heng, P.A.: Robust learning at noisy labeled medical images: Applied to skin lesion classification. ISBI (2019)
9. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: MICCAI. pp. 399–407 (2017)
10. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. ICLR (2017)
11. Zhang, J., Xie, Y., Wu, Q., Xia, Y.: Skin lesion classification in dermoscopy images using synergic deep learning. In: MICCAI. pp. 12–20 (2018)
12. Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Chen, D.Z., et al.: Biomedical image segmentation via representative annotation. In: AAAI (2019)
13. Zhou, Z., Shin, J.Y., Zhang, L., Gurudu, S.R., Gotway, M.B., Liang, J.: Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally. In: CVPR. pp. 4761–4772 (2017)