Improving active learning by data balance to reduce annotation efforts

Han Lei\textsuperscript{1,2}, Shuai Wang\textsuperscript{1,2}, Dezhi Zheng\textsuperscript{1,2}, Xiaolei Qu\textsuperscript{1,2}, Shangchun Fan\textsuperscript{1,2}, Chongyang Cui\textsuperscript{1,2}

\textsuperscript{1}School of Instrumentation Science and Opto-electronics Engineering, Beihang University, 100191, Beijing, People’s Republic of China
\textsuperscript{2}Beijing Advanced Innovation Center for Big Data-based Precision Medicine, Beihang University, 100191, Beijing, People’s Republic of China
\textsuperscript{3}School of Computer Science and Engineering, Beihang University, 100191, Beijing, People’s Republic of China
\textsuperscript{\textcopyright}E-mail: wangshuai@buaa.edu.cn

Abstract: Image classification is a fundamental task in image analysis. Recent advances in deep learning have achieved promising results on many image classification benchmarks. However, in some particular tasks, especially in biomedical image analysis, such as natural image classification. Our algorithms, we had experiments on the binary natural dataset and achieved promising performance, which could help save more than a half of efforts of annotation of dataset. More importantly, our method has found that the ATF* algorithm would lead to a worse station as the candidates selected by active learning are not in the same distribution as original dataset has, and the algorithm was limited to binary classification, which should broaden to multi classification.

1 Introduction
Deep learning’s rapid development has empowered many breakthroughs in computer vision, such as large scale visual recognition [1], instance segmentation in complex scenes [2] etc. Those deep learning models were elaborately designed and trained carefully on large scale labelled dataset, making them perform beyond humankind in some particular field. The technology’s revolution has also spread in biomedical image analysis, such as convolutional neural networks (CNNs). However, deep learning models’ spread could be restrained by the lack of such large annotated datasets for its tedious and costly annotation process, and demands of specialty in biomedicine. Therefore, how to choose the most valuable candidates from unlabelled dataset to train models more effectively and robustly in limited efforts for labelling is our task to resolve in this paper.

Researches have realised that even all data onto the same dataset belongs to an independent and identical distribution, the individuals have different values for model’s training in different stages [3]. Bengio et al. [4] inspired by humans and animals’ learning procedure, presented that examples in a meaningful order could speed up convergence of the training process, which is called curriculum learning. More specifically, a curriculum starts with a focus on the easier examples, rather than a uniform distribution over the training set which helps model learning faster. However, in nowadays training, we often use, transfer learning based on some classic and efficient pre-trained models (e.g. AlexNet [5], Inception v3 [6]), which could count as a curriculum learning’s first stage somewhat. Therefore, we can combine active learning [7] with transfer learning, using examples of maximal information as a certain input, to help model learn fastest in second stage after learning sufficient easier examples.

Yang et al. [8] combines fully convolutional network (FCN) and active learning to significantly reduce annotation effort by making judicious suggestions on the most effective annotation areas, utilising uncertainty and similarity information provided by FCN. Moreover, in binary classification tasks, Zongwei Zhou utilised multiple patches generated automatically from each candidate through data augmentation, to boost the performance of CNNs in biomedical imaging, and proposed an ATF* algorithm which combines transfer learning and active learning, and start model’s training procedure from no labelled data. However, we found that the ATF* algorithm would lead to a worse station as the candidates selected by active learning are not in the same distribution as original dataset has, and the algorithm was limited to binary classification, which should broaden to multi classification.

For those shortcomings, we developed a new algorithm based on the original thought of ATF*, where we added an active data balancing method to take precautions against models’ getting worse in some situations caused by unbalanced candidates, developed the denoising algorithm by using median to predict the candidates’ classes, and broadened active learning based on patches from binary classification to multi classification. Based on our algorithms, we had experiments on the binary natural dataset and three classes biomedical image dataset, and achieved promising performance, which could help save more than a half of efforts of annotation of dataset. More importantly, our method has shown capacity to train model in fields besides biomedical image analysis, such as natural image classification.

2 Methodology
If a complicated measurement is required to find the value of the target function for a certain input. Therefore, it is desirable to only use examples of maximal information about the function. Methods, where the learner points out good examples, are often called active learning [7]. In this paper, we developed an active learning algorithm based on Zongwei Zhou’s ATF* algorithm, modifying the training process with data balancing and broadened the original binary classification to multi classification.

Fig. 1 outlines the main ideas and steps of our deep active learning framework based on CNNs, starting with a huge data pool containing all unlabelled data. We first select a small number of samples (e.g. 1000) from the original data pool as a mini data pool to reduce the computation of sorting. Then, we generate patches from each image of mini data pool, and get a series of predictions of patches via deep learning model. We extract useful information (such as uncertainty estimation, similarity estimation, Kullback–Leibler KL distance and entropy) from those predictions of each
data, and select the most useful part of it (e.g. top ten) as a batch of candidates for fine-tuning our deep learning model. After that, those chosen data will be removed from original data pool, and the training and selecting loop will break until classification performance is satisfactory.

2.1 Incremental fine-tuning based on pre-trained model

In order to reduce the dependence of model's training process on labelled data, we use the pre-trained model (e.g. Inception v3) as base model to train on our task. At the beginning, the labelled dataset is empty; we take the pre-trained model to get the information of each candidate in mini data pool, and select a batch of top candidates for labelling. The newly labelled candidates will be incorporated into labelled dataset to continuously fine-tune the deep learning model incrementally until the performance reaches satisfaction. In our experiments, the chosen candidates would be discarded after use, and did not put them back to the original data pool.

2.2 Active data balancing

Although incremental fine-tuning could help the model to converge faster in the training process, the model's performance could get worse in some condition, which depends on the base model's meta-parameter adjustments [9]. In the experiments, we found that base models’ initialisation has a significant impact on the speed of convergence, and the performance of final models on test dataset using active learning. The reason for it may include two parts:

i. The base model used for transfer learning is pre-trained on large scale natural dataset, which has a gap of knowledge between natural images and biomedical images, or other special datasets. Further fine-tuning is demanded for filling the gap.

ii. When transfer pre-trained model to our particular task, at least output layers of the base model should be modified to tasks’ demand (e.g. predictions’ classes). However, the modified layers are always initialised randomly, giving uncertainty to start of active learning.

After fine-tuned with a small number of labelled data, base model would train and perform better with active learning, but what if we have no initial labelled data? Or we want to fine-tune base model with no initial labelled data. To address this problem, we proposed an active data balance method, and make the active leaning process start with base model directly. Take binary classification as an example: at the present moment, we could obtain a series prediction according patches generated randomly by one image via present model, then we can compute the median of those predictions. If the median is bigger than 0.5, we regard the candidate's label as positive. If else, we regard the candidate's label as negative. Based on that simple method, we select candidates actively for further annotation from the rank of candidates’ information from mini data pool, while keeping the data balance as the original dataset's (e.g. 1:1). (More details of the active data balance algorithm for multi classification is illustrated in Algorithm 1 (see Fig. 2.).)

2.3 Multi classification and metric based on median

To broaden binary classification to multi classification, we modified the denoising algorithm and proposed a new metric utilising median number. Take three subtypes classification as an example: when patches generated by one image are input to the present model, we could get a series of lists, each one has three predictions (e.g. (p1, p2, p3)), and calculate the median number along the series of lists. We regard the maximum median's class as the candidate's label. If else, we regard the candidate's label as negative. Based on that simple method, we select candidates actively for further annotation from the rank of candidates’ information from mini data pool, while keeping the data balance as the original dataset's (e.g. 1:1). (More details of the active data balance algorithm for multi classification is illustrated in Algorithm 1 (see Fig. 2.).)

3 Results

In this section, we apply our active learning with data balance to two different applications, including dogs and cats classification, and lymphoma subtype classification, whose dataset contains three subtypes.

3.1 Dogs and cats classification

Fig. 1 Illustrating the detailed workflow of our active learning method

Algorithm 1: Multi classification based on active fine-tuning method

\[ C = \{x_i\}, i \in [1, m] \quad [C \ has \ m \ patches] \]
\[ M: \ pre-trained \ model \]
\[ \text{Output:} \]
\[ K: \ metrics \ of \ C \]
\[ T: \ class \ of \ C \]
\[ \text{Functions:} \]
\[ P \leftarrow F(C, M) \quad \text{[outputs of } M \ \text{given } V \times C \text{]} \]
\[ P = \{p_i\}, i \in [1, m] \quad [C \ has \ m \ patches] \]
\[ p_i = \{p'_i\}, j \in [1, 3] \quad [3 \ classes] \]
\[ m \leftarrow R(P') \quad \text{[outputs median of a series of } p'] \]
\[ \text{Initialization:} \]
\[ C \]
\[ \text{compute } P \leftarrow F(C, M) \]
\[ \text{for each } p'_i \in P', j \in [1, 3] \text{ do} \]
\[ \text{compute } m' \leftarrow R(p'_i) \]
\[ \text{for } k \in [1, 3] \text{ do} \]
\[ \text{if } m_k = \max(m') \ k \in [1, 3] \]
\[ T \leftarrow k \]
\[ S \leftarrow \text{top 25\% of } p'_i \quad i \in [1, m] \]
\[ \text{for each } s \in S \]
\[ X = \sum s \times \log s \]

Fig. 2 Algorithm 1: Multi classification based on active fine-tuning method

This is an open access article published by the IET under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)

J. Eng., 2019, Vol. 2019 Iss. 23, pp. 8650-8653

8651
In this part, we use dogs and cats dataset as a binary classification task, demonstrating our algorithms including active learning combining with transfer learning and data balance. Fig. 3 shows dog and cat samples and corresponding patches generated by them randomly. The dataset of dogs and cats contain 25,000 images with labels 'cat' and 'dog', and we generate ten patches randomly from each image. We trained the classification modal based on Inception v3 as Keras has the pre-trained weights.

Fig. 4 shows the test accuracy of Inception v3 in the precondition of active select with data balance method, active select without data balance method and random select as a comparison. Utilising active select with data balance method, the steep accuracy curve reaches the black dotted line immediately, while yellow curve (dashed line) by random select increases slowly. More specifically, Fig. 4 shows that transfer learning combining with active select and data balance with only 80 candidate queries (0.32%) can achieve the performance of accuracy over 0.93, while transfer learning queried over 10,000 images (4.0%) from dataset, whose model was fine-tuned from Inception v3 by random select. Yellow dashed line could be seen in Fig. 4, where test accuracy is 0.93. 92% cost for labelling candidates could be saved to reach an accuracy of 0.93. Moreover, the training accuracy using active learning combined with data balance could be improved with careful adjustment of learning rates and the number of repeats when training the model.

Besides that, the blue dotted curve shows that with only active learning, test accuracy may reach the black dotted line, but immediately descends near the random select curve, and creeps over the accuracy of 0.96 with a low speed compared with yellow dashed curve by random select. Fig. 4 shows that the active select with data balance has faster and more robust learning process until convergence. Further, we extracted the samples of each class, and calculated the proportion of positive data along the percentage of label queried. In Fig. 5 the red real line of active select with data balance starts at 0.1, and has a steep descend at the beginning of training process, but it recovers immediately and rise to 0.4 before 0.05% of label queried. The increasing tendency shows red real line would reach the black dotted line, which is also the ideal proportion of data balancing. Compared with the red real line, the blue dashed line almost descends to zero at the beginning, and creep near 0.3 in the whole training procedure, which caused the gap between the test accuracy with data balance and without data balance as shown in Fig. 4.

### 3.2 Lymphoma subtypes classification

Lymphoma subtype classification is the hardest task for pathologists in histopathological image analysis, which also equates to the hardest annotation works (Fig. 6). The more specialists of pathology we need, the more cost we have to assume. To dramatically reduce the cost of annotation, we used multi classification based on active fine-tuning method showed in Algorithm 1 (Fig. 2). The dataset has three subtypes of lymphoma, which looks similar to each other. A subtype of chronic lymphocytic leukaemia has 113 images, follicular lymphoma has 139 images, and mantle cell lymphoma has 122 images. All images’ size is 1388 pixels×1040 pixels. Using active learning combining data balance could reach top-1 accuracy of 0.91 with only 65% of whole data, while training model by randomly selecting should have training all train dataset's images for five epochs. All experiments about lymphoma subtype classification were based on VGG16 model, and demonstrates the efficiency of the broadened active leaning algorithm.

### 4 Conclusion and future work

In this paper, we have developed an incremental active learning method combing with active data balance and broaden the original method from binary classification to multi classification. Using our method, the active training process based on transfer learning could be more robust and efficient, no matter how bad initialisation the base model is, while active learning may not perform well without data balance. Besides that, we have also made an adjustment on reducing noisy labels by choosing median from a series of predictions of patches, which performs better compared with choosing mean number of predictions. Based on that, we simply
broaden active learning classification method to multi classification tasks, and achieve ideal result.

All the experiments and results demonstrated in the last part are based on Inception v3 and VGG16 [10], two classic convolution neural networks. However, the methods on AlexNet and MobileNet [11] have also experimented, and obtain a similar result. In reality, we may confront many situations where we have enough unlabelled data and few labelled data or limited effort for labelling, including both biomedical images and other natural images, and the active learning with data balancing could be a potential method of training in such a situation of limited labelled data.

In this paper, all models are training on batch selected by active learning’s choose engine incrementally at one time, but more batches could be generated at one updating process, and train more than one time, which may help model to develop a better performance and save more computation on selecting candidates from a large unlabelled data pool. Those different training methods may help the model to reach convergence faster and reduce computation in some tasks. As for convergence, it is also vital to choose a proper optimiser, and those optimisers with adaptive learning rates, such as RMSprop, Adagrad, Adam and Nadam etc., did not perform well in training of active learning in our experiments mentioned in the last section, as the training process is more unstable and discontinuous than classic models’ training. Mini-batch stochastic gradient decent was adopted in our experiments with careful set of learning rates to achieve our results. Finally, as human’s learning course, selecting data onto different information on distinct periods of training based on active learning may have a promising performance on training from scratch.

5 Acknowledgments

This research was supported by the National Youth Talent Support Program held by Beihang University. For this research, the authors thank the innovation team of the Ministry of Education for Advanced Sensing Technology in Aerospace. More importantly, they thank their supervisors Dr. Zheng and Dr. Wang for their trust and generous support. They also thank their classmates for their self-giving help in coding and valuable suggestions.

6 References

[1] Russakovsky, O., Deng, J., Su, H., et al.: ‘Imagenet large scale visual recognition challenge’, Int. J. Comput. Vis., 2015, 115, (3), pp. 211–252
[2] Lin, T., Maire, M., Belongie, S.J., et al.: ‘Microsoft COCO: common objects in context’. European Conf. on Computer Vision, Zurich, Switzerland, 2014, 8693, pp. 740–755
[3] Elman, J.L.: ‘Learning and development in neural networks: the importance of starting small’, Cognition, 1993, 48, (1), pp. 71–99
[4] Bengio, Y., Louradour, J., Collobert, R., et al.: ‘Curriculum learning’. Int. Conf. on Machine Learning, Montreal, Canada, 2009, pp. 41–48
[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ‘Imagenet classification with deep convolutional neural networks’. Neural Information Processing Systems, Lake Tahoe, NV, USA, 2012, pp. 1097–1105
[6] Szegedy, C., Vanhoucke, V., Ioffe, S., et al.: ‘Rethinking the inception architecture for computer vision’. Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 2818–2826
[7] Krogh, A., Vedelsby, J.: ‘Neural network ensembles, cross validation and active learning’. Int. Conf. on Neural Information Processing Systems, Denver, CO, USA, 1994, pp. 231–238
[8] Yang, L., Zhang, Y., Chen, J., et al.: ‘Suggestive annotation: a deep active learning framework for biomedical image segmentation’. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Quebec City, Canada, 2017, pp. 399–407
[9] Zhou, Z., Shin, J., Zhang, L., et al.: ‘Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally’. IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 4761–4772
[10] Simonyan, K., Zisserman, A.: ‘Very deep convolutional networks for large-scale image recognition’, arXiv preprint arXiv:1409.1556, 2014
[11] Howard, A.G., Zhu, M., Chen, B., et al.: ‘Mobilenets: efficient convolutional neural networks for mobile vision applications’, arXiv preprint arXiv:1704.04861, 2017