FUZZY-LABEL WEIGHTED DEEP LEARNING CLASSIFICATION FOR CT IMAGE QUALITY EVALUATION

Ee Ping Ong[1], Ruchir Srivastava[1], and Wenbo Chen[2]
[1]Institute For Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR), Singapore, {epong, srivastavar}@i2r.a-star.edu.sg
[2]GE Healthcare (China), wenbo.chen@ge.com

***Abstract*— This paper proposes a fuzzy-label weighted deep learning-based image classification approach for assessing computed tomography (CT) image quality. More specifically, we want to determine if a captured CT image passes Quality Assessment (QA) with a certain radiation dose. Our contributions here include proposing a fuzzy-label weighting method and introduces the concept of a "fuzzy-label" (to reflect the confidence of the ground-truth annotation by annotator) to aid in the training of deep learning-based image classification. We proposed an ensemble/assimilation method to determine the image quality at entire CT image-level using CT windowing (i.e. clipping of the CT image to 8-bit grayscale with respect to the various window-width (WW) and window-length (WL)), similar to what a human would do manually to assess the CT image quality in the factory setting. Experimental results showed that our proposed fuzzy-label weighted deep learning-based image classification approach (trained using annotations provided by one single annotator) significantly outperforms that of its traditional baseline image classification approach.**

I. INTRODUCTION

Image classification is the process of attempting to categorize or label an image based on specific rules. The categorization rules can be devised using one or more spectral or textural characteristics or specific "object" of interest (e.g. cat, dog) in the image. The two general approaches of image classification are "supervised" and "unsupervised". Supervised classification is the process of learning from labeled data and this is the dominant paradigm in machine learning and artificial intelligence. More recently, deep learning-based image classification algorithms, particularly convolutional neural networks, are becoming more and more popular for solving image classification problems [4, 8]. Deep learning algorithms have also rapidly become a methodology of choice for analyzing medical images for the purpose of image classification, object detection, segmentation, registration, and other tasks [6].

In computed tomography (CT) scanning for medical imaging applications, it would be ideal to use the lowest possible radiation dose to reduce the harmful effect of radiation to the patients [2]. However, reducing radiation dose will result in increases in CT noise and deteriorating image quality and it will come to a point where the image quality is not acceptable for assessment by the doctors. As such, CT scanning devices' manufacturers attach great deal of focus and attention to manufacturing CT scanners that are able to provide reasonably good CT image quality at the "lowest possible radiation dose" for their CT scanners [10]. In the factory setting, the Quality Control (QC) engineers would use captured CT images of phantoms to evaluate the CT image quality. In one of such tests, the captured CT image of a particular phantom would look like the one shown in Figure 1, and they would want to determine if a captured CT image passes Quality Assessment (QA) or not. As the captured CT images are originally acquired in 16-bit format to present measures of tissue density, and since human eye can only distinguish roughly 100 shades of grey simultaneously (according to perception literature), there is a limit to direct presentation of these values as an image on a display screen to the QC engineers and the radiologists [7]. Thus, for them to be able to parse the complex images both faster and easier, one can use image windowing to increase the contrast across a region of interest. As such, one approach the QC engineer uses (in the factory setting) is to visually look at the CT image at various pre-determined window-width (WW) and window-length (WL) 8-bit grayscale image [7].

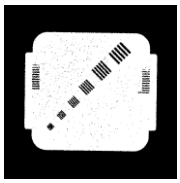

Figure 1: Sample CT image.

In our case here, the QC engineer would be most focused on the set of stripes, particularly the set labelled as "B" group (see Figure 2). As long as one or more of the B group of stripes is visible in these set of pre-determined WW-WL 8-bit grayscale images, the CT image will be considered to have passed quality assessment.

The goal of our classification application is to simulate the manual QC process performed by the QC engineer and to design a software algorithm to automatically determine if a captured CT image (captured from a phantom using a CT scanner) passes Quality Assessment (QA) or not. This is a 2-class classification problem ("Fail" or "Pass") for the purpose of quality control of manufactured CT scanners in a factory setting with varying radiation doses. The input data or CT image to be used for our classification task is stored as DICOM files (one of which is shown in Figure 1).

In order to determine the image quality at entire CT image-level, we proposed an ensemble/assimilation method to predict image quality of the entire CT image using CT windowing, i.e. clipping of the CT image to 8-bit grayscale with respect to the various window-width (WW) and window-length (WL) of the CT

image. This is similar to what a quality control engineer would do manually to assess the CT image quality in the factory setting.

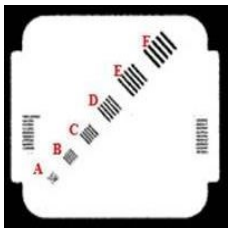

Figure 2: Sample CT image with "B" group of stripes labelled.

II. PROPOSED CLASSIFICATION APPROACH

The block diagram of our proposed fuzzy-label weighted deep learning-based image classification approach for assessing computed tomography (CT) image quality is shown in Figure 3 below.

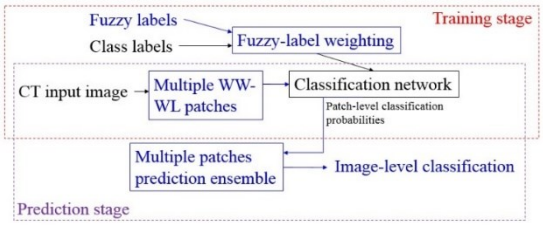

Figure 3: Block diagram of proposed fuzzy-label weighted deep learning-based image classification approach.

The CT images captured from the phantoms using the CT scanners in our case has a spatial resolution of 512x512 pixels and an intensity resolution of 16 bits (2^16 discrete intensity values) and they are stored in dicom image format. For each CT image, we only extract the 32x32 pixels' sub-region (or patch of image) enclosed by (*x*,*y*) coordinates (153, 308) at top left corner and (185, 340) at bottom right corner for training and validation. This corresponds to the second smallest group of bars from the bottom of the image (labelled as "B" group of stripes as shown in Figure 2). Note that because the phantom is placed in roughly fixed position when being scanned, thus the coordinates of the various group of stripes do not deviate much. For each CT image, a certain fixed set of window-width (WW) and window-length (WL) 8-bit range clipped from the 16-bit CT image are used, where the (WW, WL) are as follow (as pre-determined by the Quality Control engineers): [(1,60), (1,70), (1,80), (40,60), (40,70), (40,80), (80,60), (80,70), (80,80)]. The class labels are 1 for passed QA (i.e. the group of stripes are clearly "Visible") and 0 for failed QA (i.e. the group of stripes are "Not clearly Visible").

We proposed the introduction of a "fuzzy-label" into our deep learning training process, where the fuzzy-label indicates whether this image is fuzzy or not, or in another words, the confidence of the assigned ground-truth annotation by annotators in our case since there are large inter-annotators' variances and also large intra-annotator's variances [5]. If the fuzzy-label is assigned to be "1", it means that the annotator has a low confidence in the class label he/she assigned, otherwise the fuzzy-label is assigned to be "0". Several sample patches of images annotated with class label = 1 (i.e. passed QA) are shown in Figure 4, while several sample patches of images annotated with class label = 0 (i.e. failed QA) are shown in Figure 5. Note that some of the image patches annotated with class label = 0 (i.e. failed QA) and some annotated with class label = 1 (i.e. passed QA) look very similar and there are large inter-annotators' subjectivity as well as significantly large intra-annotator's subjectivity when performing annotations. This makes it difficult to train a traditional image classification model with accurate classification because of the large annotators' subjectivity.

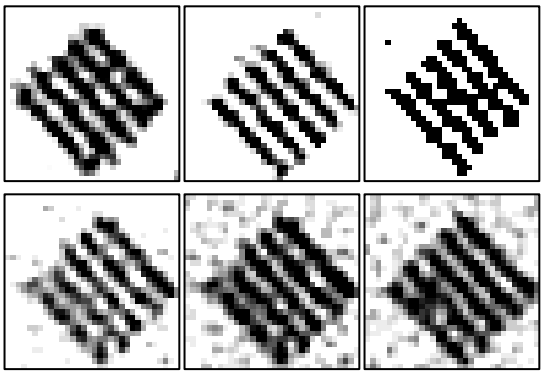

Figure 4: Sample image patches annotated with class label = 1 (i.e. passed QA).

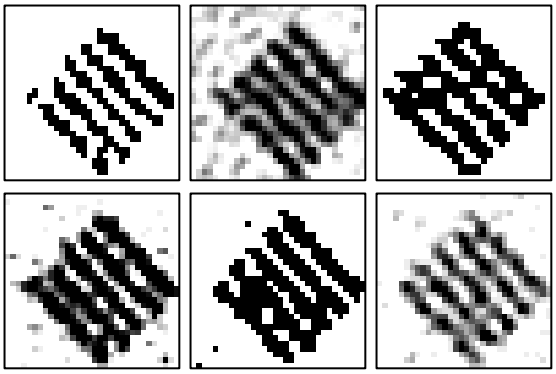

Figure 5: Sample image patches annotated with class label = 0 (i.e. failed QA).

A. Pre-processing the input training data

We performed simple data pre-processing on the input data before feeding them to the training network. Here, we performed data centering and normalization by using z-score, as follows:

$$z = (X - \mu)/\sigma \quad (1)$$

where X is the input data of each channel, μ is the mean of X, and σ is the standard deviation of X. The normalized training images (instead of the original raw images) are then fetched as input to the CNN-based classification network.

B. CNN-based Classification network

To demonstrate our proposed fuzzy-label weighted deep learning-based image classification approach, we employ a commonly used deep learning classification network, namely a Residual Networks with 50 layers (ResNet50) based on Convolutional Neural Networks (CNN) [3] in our experiments. We changed the last layer to use a "softmax" activation with 2 classes to suit our 2-class classification problem. We train this model from scratch based on the input data (which are the image patches) together with the image patches' class-labels and the fuzzy-labels (annotated by only one annotator) using a single NVIDIA GeForce

RTX 2080 Ti GPU. The model is trained with the following parameters: epochs of 200, batch size of 32, cross entropy loss function, Adam optimizer with initial learning rate of 1e-3 and this learning rate is progressively dropped as epoch number increases.

C. Fuzzy-label weighting

We proposed a fuzzy-labels' weighting technique to train the input CT image data (or more specifically in our case is the CT-windowed image patches containing the "B" group of stripes) to perform a 2-class classification. Here, our aim is to tell our classification network that if the human annotator providing the ground-truth annotation (or the true class-label) is uncertain of his/her own annotation (because in our case here the size of the image patch containing the "B" group of stripes is very small and the stripes may be very blur as well), then we want the human annotator to provide this information to the classification network as well. To this end, we introduce a new label information that we called "fuzzy-label" information, $\boldsymbol{f}$. In this "fuzzy-label" information, we assign a "1" if the annotator is unsure (and has low confidence) of his/her own class-label annotation, and otherwise assign a "0" if he/she is very confident of the class-label annotation assigned. We also define a constant called "ambiguous label factor" (or a) corresponding to the probability assigned to the one-hot representation of true labels when "fuzzy-label" is equal to 1. For example, given the one-hot representation of true class-labels with m elements, $\boldsymbol{Y}$, for a 2-class classification (as in our case) where:

$$\boldsymbol{Y} = \begin{bmatrix} y_{1,1} & y_{1,2} \\ \vdots & \vdots \\ y_{m,1} & y_{m,2} \end{bmatrix} \tag{2}$$

Then, our fuzzy-label weighted labels, $\boldsymbol{Z}$, used for training can be written as:

$$\boldsymbol{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} \\ \vdots & \vdots \\ z_{m,1} & z_{m,2} \end{bmatrix} \tag{3}$$

where:

$$z_{i,j}|_{i=1,\dots,m;\ j=1,2} = \begin{cases} y_{i,j}, & if\ f_i = 0 \\ a, & if\ f_i = 1\ and\ y_{i,j} = 1 \\ (1-a), & if\ f_i = 1\ and\ y_{i,j} = 0 \end{cases} \tag{4}$$

In all our experiments here, we use a = 0.6 in our 2-class classification problem.

It should be noted that our above fuzzy-label weighting approach can be extended to multi-class classification problem as well. In the multi-class classification case, we could write the one-hot representation of the true class labels with m elements, $\boldsymbol{Y}$, for a c-class classification scenario as follows:

$$\boldsymbol{Y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,c} \\ \vdots & \cdots & \vdots \\ y_{m,1} & \vdots & y_{m,c} \end{bmatrix} \tag{5}$$

Then, our fuzzy-label weighted labels, $\boldsymbol{Z}$, used for training will be given by:

$$\boldsymbol{Z} = \begin{bmatrix} z_{1,1} & \cdots & z_{1,c} \\ \vdots & \cdots & \vdots \\ z_{m,1} & \vdots & z_{m,c} \end{bmatrix} \tag{6}$$

$$z_{i,j}|_{i=1,\dots,m;\ j=1,\dots,c} = \begin{cases} y_{i,j}, & if\ f_i = 0 \\ a, & if\ f_i = 1\ and\ y_{i,j} = 1 \\ (1-a)/(c-1), & if\ f_i = 1\ and\ y_{i,j} = 0 \end{cases} \tag{7}$$

D. Data augmentation and training

Data augmentation technique has been applied on-the-fly when training our model to prevent over-fitting. These include random horizontal shift and vertical shift as well as random rotations by up to about 15%. Here, we perform training using a cross-entropy loss function CEL_w with the fuzzy-label weighted one-hot representation of true labels as follows:

$$CEL_w = -\sum_{i=1}^{m}\sum_{k=1}^{c}\left[Z_i(k).log\left(P_i(k)\right)\right] \tag{8}$$

where c is the number of classes, m is the number of data points, $Z_i(k)$ is the fuzzy-label weighted "true" probability of the class k for data point i as provided by the ground-truth annotation. $P_i(k)$ is the probability of the class k as predicted by classification model for data point i (and in our case here this is the output of our CNN network after Softmax activation).

E. Image-level validation

For each dicom CT image sample, 9 WW-WL patch of images would be generated using (WW, WL) = [(1,60), (1,70), (1,80), (40,60), (40,70), (40,80), (80,60), (80,70), (80,80)] and a prediction would be applied to each WW-WL image. Each WW-WL image would be passed through the trained model to obtain a predicted probability, and its predicted label would be regarded as "1" (passed QA) if the predicted probability exceeds a prediction threshold (default = 0.5). In addition, each WW-WL image would also be flipped at 45 degrees axis (i.e., each WW-WL image would first be flipped horizontally (so that left becomes right and right becomes left) and then rotated by 90 degrees anticlockwise) before being passed through the trained model again. In this way, we would obtain another predicted probability, and similarly its predicted label would be regarded as "1" (passed QA) if the predicted probability exceeds a prediction threshold (default = 0.5).

The final prediction of a CT image sample is based on the assimilation of the predictions of the 18 WW-WL patches (i.e., the 9 WW-WL patch of images (of the "B" group of stripes) and its corresponding flipped versions). The CT image would be considered to have passed QA if one or more of the image patches passed the prediction test, otherwise it would be considered to have failed QA (this is to simulate what the QC engineer currently manually performs in the factory setting). The assimilated predicted label would be compared to the annotated class label to compute the evaluation metrics (such as precision, recall, F1-score, and accuracy) and these are used to gauge how good the trained model is at the CT image-level.

F. Evaluation Metrics

We will report our classification results using the following evaluation metrics [9]:

(1) Accuracy:

Accuracy is the total number of correct predictions divided by the total number of predictions made for a dataset:

$$ACC = (TP + TN)/(TP + TN + FP + FN) \tag{9}$$

where TP and TN denote the number of positive and negative instances that are correctly classified. Meanwhile, FP and FN denote the number of misclassified negative and positive instances, respectively.

(2) Sensitivity:

Sensitivity or true positive rate (TPR) is the proportion of samples that are genuinely positive that give a positive result:

$$TPR = TP/(TP + FN) \tag{10}$$

(3) Precision:

Precision or positive predictive value (PPV) is the probability that a subject/sample that returns a positive result really is positive:

$$PPV = TP/(TP + FP) \quad (11)$$

(4) F1-score:

The F1 score is the harmonic mean of the precision and recall:

$$F1 = (2 * PPV * TPR)/(PPV + TPR) \quad (12)$$

G. Dataset

Our dataset consists of dicom files containing CT images captured from phantoms in the factory production line. In general, these CT images are considered to have "Passed" quality assessment if 1 or more of WW-WL patches of the "B" group of stripes in the captured CT image is visible. The complete dataset contains 619 CT image samples.

III. RESULTS

For the purpose of training the classification model, we use 4178 WW-WL image patches for training and 1393 WW-WL image patches for validation. Note that 9 WW-WL image patches were generated per CT image sample, and there were 2072 Failed image patches (i.e. with class-label=0) and 3499 Passed image patches (i.e. with class-label=1). More specifically, out of the 4178 WW-WL training image patches, there are 2624 training patches annotated with class-label=1 and 1554 training patches annotated with class-label=0. Among the training patches with class-label=1, 2529 patches has a fuzzy-label=1 and 95 patches has a fuzzy-label=0. Also, among the training patches with class-label=0, 784 patches has a fuzzy-label=1 and 770 patches has a fuzzy-label=0. Out of the 1393 WW-WL validation image patches, there are 875 validation patches annotated with class-label=1 and 518 validation patches annotated with class-label=0. Among the validation patches with class-label=1, 844 patches has a fuzzy-label=1 and 31 patches has a fuzzy-label=0. Also, among the validation patches with class-label=0, 261 patches has a fuzzy-label=1 and 257 patches has a fuzzy-label=0. All the annotations were provided by one single annotator (who is the QC Engineer).

The results obtained from validation data using our proposed fuzzy-label weighted deep learning-based image classification approach is shown in Table 1 while the results obtained from the same validation data using the traditional approach without fuzzy-label weighting is shown in Table 2. As our paper's contribution is not another "new CNN classification network", and there isn't any work directly comparable to ours, hence we only compare our approach to the baseline. The closest work to ours ever published to date is [1], who propose a solution to combine multi-raters' subjectivity information (65 raters in their experiments) as latent embeddings while using ResNet as baseline image classification network. For situations like ours where we only have one rater's subjectivity information for our dataset (since human annotations are time consuming and expensive), we are unable to use their approach on our dataset, and hence alternative such as our fuzzy-label weighting approach is necessary. Compared to their approach, our approach requires only one rater, is simpler, and does not add any additional network layers to the baseline network.

Comparing Table 1 and Table 2, it can be seen that our proposed fuzzy-label weighted classification approach provides significantly better results for both patch-level classification and image-level classification. Thus, this confirmed the effectiveness of our proposed fuzzy-label weighted classification approach for automatic CT image quality evaluation, where we are dealing with classification of dataset that may have many fuzzy annotations (due to large inter-annotators' variances and also large intra-annotator's variances). Note that for practical use (in the factory QC setting), the QC engineer's main interest and concern is the CT image-level prediction accuracy.

	Accuracy	Precision	Recall	F1-score
Patch-level	91.67%	0.9069	0.9207	0.9125
Image-level	98.14%	0.9834	0.9732	0.9781

Table 1: Classification results with proposed fuzzy-label weighting.

	Accuracy	Precision	Recall	F1-score
Patch-level	85.78%	0.8556	0.8805	0.8551
Image-level	93.55%	0.9571	0.8972	0.9203

Table 2: Classification results without fuzzy-label weighting.

IV. CONCLUSIONS

The major contribution of our paper is in introducing the fuzzy-label weighting method and the concept of "fuzzy-label" (provided by a single rater), and showing how we can utilize these on top of a baseline image classification network to achieve better classification results. The baseline image classification network we used here (ResNet) could be substituted with any other state-of-the-art model to be used together with our approach to improve their results, just like how we used it to solve our real-life industrial image Quality Assessment problem. Our experimental results showed that our proposed fuzzy-label weighted deep learning-based image classification approach significantly outperforms that of its traditional baseline image classification approach. Our proposed approach has been shown to work well with annotations provided by one single annotator and there is no requirement for many multiple annotators in order to use our proposed approach.

V. REFERENCES

[1] Li, B. et al., Learning from subjective ratings using auto-decoded deep latent embeddings." *Int. Conf. Medical Image Computing and Computer Assisted Intervention*, 2021.

[2] Goldman, L.W., "Principles of CT: Radiation Dose and Image Quality", *J. Nuclear Med Technology*, 2007.

[3] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Krizhevsky, A., Sutskever I., and Hinton G. E. "Imagenet classification with deep convolutional neural networks", *Conference on Neural Information Processing Systems* (NIPS), 2012.

[5] Lampert, T.A., A. Stumpf, P. Gancarski, "An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation", *IEEE T. Image Processing*, 25(6):2557-2572, Jun 2016.

[6] Litjens, G. et al., "A survey on deep learning in medical image analysis." *Medical Image Analysis*, 42:60-88, 2017.

[7] Masoudi, S., S.A.A. Harmon, S. Mehralivand, S.M. Walker, H. Raviprakash, U. Bagci, P.L. Choyke, B. Turkbey, "Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research", *J. of Medical Imaging*, 8(1), 2021.

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

[9] Sokolova, M., and Lapalme, G., "A systematic analysis of performance measures for classification tasks", *Int. J. Information Processing and Management*, Volume 45, Issue 4, pp 427-437, 2009.

[10] Zarb, F., L. Rainford, M.F. McEnteeb, "Image quality assessment tools for optimization of CT images", *Radiography*, Volume 16, Issue 2, pp. 147-153, May 2010.