Quantization of Fully Convolutional Networks for Accurate Biomedical Image Segmentation

Xiaowei Xu[1,2], Qing Lu[2], Lin Yang[2], Sharon Hu[2], Danny Chen[2], Yu Hu[1], Yiyu Shi[2]
[1] Huazhong University of Science and Technology
[2] Univerity of Notre Dame
{xuxiaowei,bryanhu}@hust.edu.cn; {xxu8, qlu2, lyang5, shu, dchen, yshi4}@nd.edu

Abstract

With pervasive applications of medical imaging in health-care, biomedical image segmentation plays a central role in quantitative analysis, clinical diagnosis, and medical intervention. Since manual annotation suffers limited reproducibility, arduous efforts, and excessive time, automatic segmentation is desired to process increasingly larger scale histopathological data. Recently, deep neural networks (DNNs), particularly fully convolutional networks (FCNs), have been widely applied to biomedical image segmentation, attaining much improved performance. At the same time, quantization of DNNs has become an active research topic, which aims to represent weights with less memory (precision) to considerably reduce memory and computation requirements of DNNs while maintaining acceptable accuracy. In this paper, we apply quantization techniques to FCNs for accurate biomedical image segmentation. Unlike existing literatures on quantization which primarily targets memory and computation complexity reduction, we apply quantization as a method to reduce overfitting in FCNs for better accuracy. Specifically, we focus on a state-of-the-art segmentation framework, suggestive annotation [26], which judiciously extracts representative annotation samples from the original training dataset, obtaining an effective small-sized balanced training dataset. We develop two new quantization processes for this framework: (1) suggestive annotation with quantization for highly representative training samples, and (2) network training with quantization for high accuracy. Extensive experiments on the MICCAI Gland dataset show that both quantization processes can improve the segmentation performance, and our proposed method exceeds the current state-of-the-art performance by up to 1%. In addition, our method has a reduction of up to 6.4x on memory usage.

1. Introduction

With pervasive applications of medical imaging in health-care, biomedical image segmentation has always been one of the most important tasks in biomedical imaging research. Biomedical image segmentation extracts different tissues, organs, pathologies, and biological structures, to support medical diagnosis, surgical planning and treatments. In common practice, segmentation is performed manually by pathologists, which is time-consuming and tedious. However, the ever-increasing quantity and variety of medical images make manual segmentation impracticable in terms of cost and reproducibility. Therefore, automatic biomedical image segmentation is highly desirable. But, this task is very challenging, because of high variability in medical images due to complex variations in biomedical objects and structures and because of low contrast, noise, and other imaging artifacts caused by various medical imaging modalities and techniques.

In the past years, substantial progress has been made on biomedical image segmentation with pixel based methods [8, 14, 21, 18] and structure based methods [1, 10, 9, 19]. These methods achieve promising results on nonmalignant objects using hand-crafted features and prior knowledge of structures. However, they suffer considerable degradation when applied to malignant objects with serious deformation. Recently, deep neural networks (DNNs), particularly fully convolutional networks (FCNs), have been highly effective for biomedical image segmentation, which require little hand-crafted features or prior knowledge. Ronneberger et al. [16] proposed U-Net, a U-shaped deep convolutional network that adds a symmetric expanding path to enable precise localization. With strong use of data augmentation, this segmentation model achieves significant improvement over previous methods. The DCAN model by Chen et al. [2, 3] added a unified multi-task object to the U-Net learning framework, which won the 2015 MICCAI Gland Segmentation Challenge [17]. Based on DCAN, Yang et al. [26] proposed suggestive annotation which extracts representative samples as a training dataset, by adopting active learning into their network design. With the refined training samples and optimized structure for DNNs, suggestive annotation achieves state-of-the-art performance on the MICCAI Gland segmentation dataset [17].

Corresponding author: bryanhu@hust.edu.cn; yshi4@nd.edu

At the same time, DNN quantization has become an active research topic [7], which aims to represent DNN weights with less memory (precision) while maintaining acceptable accuracy with efficient memory and computation costs. It has been observed in the literature, however, that sometimes quantization can improve accuracy which can be credited to the reduction of overfitting. Dynamic fixed point are adopted in [11][12], which achieves 4x less memory operation cost with only 0.4-0.6% Top-5 accuracy loss for ImageNet classification [6]. Ternary weight network [13] and binaryConnect [5] have further reduced the bit-width of weights to 2 bits or even 1 bit with a relatively larger accuracy loss. Recently, their enhanced version, trained ternary training [29] and binary weight network [15] have reduced the accuracy loss to only 0.6-0.8%. There also exists some works using non-linear quantization to represent the parameter distribution for better accuracy [11][27]. Unlike the above works, some studies aims to quantize not only the weights but also the activations. Quantized neural networks [12], binarized neural networks [4], and XNOR-net [15] reduced the weights to only 1 bit and the activations to 1-2 bits resulting in a large reduction on memory and computation cost yet with significant accuracy loss. In some of the above works, we notice that quantization can sometimes improve the performance [11][23][22][27], which can be credited to the reduction of overfitting.

In this paper, we adopt quantization as a method to reduce overfitting to FCNs for accurate biomedical image segmentation. Particularly, we focus on a recent effective biomedical image segmentation framework, suggestive annotation [26]. We develop two new quantization processes to incorporate into this state-of-the-art framework: (1) suggestive annotation with quantization for highly representative training samples, and (2) network training with quantization for high accuracy. Extensive experiments are presented on the widely-used MICCIA Gland dataset, and the results show that our proposed method exceeds the current state-of-the-art performance by up to 1%. In addition, our method has a reduction of up to 6.4x on memory usage.

2. Related Work

In this section, we briefly review suggestive annotation [26], on which our proposed method is based. Several representative quantization methods are discussed in detail, which will be adopted in our experiments. The readers are also referred to [11, 5, 15] for other quantization methods.

2.1. Suggestive Annotation for Biomedical Image Segmentation

We based our proposed framework on suggestive annotation [26], which achieves state-of-the-art performance on the Gland dataset. The key idea of the work is that better performance can be achieved with representative training samples instead of original training samples. As shown in Figure 1, the suggestive annotation framework [26] has two steps: suggestive annotation and network training. The first step extracts typical samples from the original training set with multiple suggestive FCNs, and the second step trains segmentation FCNs with the extracted samples. In the first step, multiple suggestive FCNs are trained in parallel. During the inference stage, multiple suggestive FCNs produce multiple predictions for the same input from the original training set, which can be used to calculate the representativeness of the samples. Note that each FCN generates two outputs: contour of the objects and the segmented object. The suggestive FCNs and segmentation FCNs have the same network structure which is based on DCAN [3] and active learning.

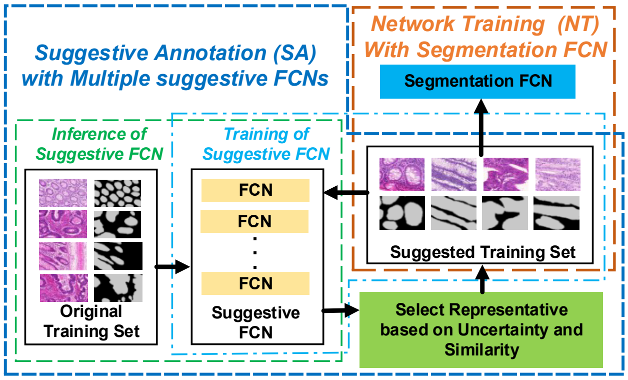


Figure 1. Illustration of the suggestive annotation framework [26]. With suggestive annotation, better samples (suggestive training set) can be extracted from the original training set for further training with better performance.

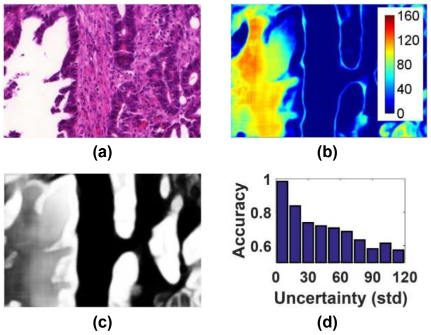


Figure 2. (a) An original image; (b) the probability map produced by multiple FCNs in suggestive annotation for (a); (c) uncertainty estimation of the results by the multiple FCNs; (d) relation between uncertainty estimation and pixel accuracy on the testing data. Obviously there is a strong correlation between the test accuracy and uncertainty (reprinted from [26]).

Two metrics are involved with representativeness: uncertainty and similarity. A representative training samples should be hard to predict as they are located on the ”boarder line” of the feature space, and have low similarity with each other as they can well describe the variety of the ”boarder line" with limited quantity. In suggestive annotation, the standard deviation of the multiple predictions from multiple suggestive FCNs are regarded as the uncertainty score. The averaged outputs of last convolutional layers of multiple suggestive FCNs are regarded as a domain-specific image descriptor, which can be used to evaluate the similarity of images with cosine similarity.

Selecting representative training samples with uncertainty and similarity is an NP-hard problem [26]. A simple heuristic method is adopted: extract K samples with the highest uncertainty scores first, and then select the final k ($k < K$) samples based on their similarity with each other. The reason to put uncertainty in the first step is that uncertainty is more important than similarity [26]. As shown in Figure 2, the test accuracy is highly correlated with the uncertainty score.

2.2. Quantization Techniques for DNNs

2.2.1 Incremental Quantization (INQ)

Incremental quantization [27] quantizes weights to powers of two in an iterative manner. In each iteration, a subset of weights is selected and quantized, and a fine-tuning process is then presented while the quantized weights are locked during both feed-forward and feed-back prorogation. The above process iterates until all weights are quantized. The quantization calculation is shown in Eq. (1), where w^q and w are quantized and original weights, respectively, and u and l are the upper and lower bounds of the quantized set, respectively. Note that how to choose the weights during each iteration is dependant on the magnitude of the weight. With incremental quantization, the weights can be represented with only 3-5 bits with almost no accuracy loss, and the multiplication can be simplified to shift operation.

$$w^q = \begin{cases} sign(w) \times 2^p & \text{if } 3 \times 2^{p-2} \leq |w| < 3 \times 2^{p-1}; \\ & \qquad l \leq p \leq u; \\ sign(w) \times 2^m & \text{if } |w| \geq 2^u; \\ 0 & \text{if } |w| < 2^{-l-1}. \end{cases} \tag{1}$$

2.2.2 DoReFa-Net

DoReFa-Net [28] trains DNNs with low bitwidth weights and activations represented using low bitwidth parameter gradients, and it enables training acceleration of low bitwidth neural network on general hardware. In the quantization process, weights and activations can be deterministically quantized, while gradients need to be stochastically quantized. DoReFa-Net adopts a simple quantization method to quantize 32 bits values to only 1 bits as shown in Eq. (2), where w_l and w_l^q are the original and quantized weights of the lth layer, respectively, and $E(|w_l|)$ calculates the mean of the absolute value of weights in the lth layer.

$$w_l^q = E(|w_l|) \times sign(w_l) \tag{2}$$

Thus, DoReFa-Net can achieve a 32x compression rate atmost with comparable accuracy compared with networks using floating-point representation, and the computation of multiplication is also simplified to addition and/or substraction. In the feed-back propagation, weights and gradients are maintained in floating point, and quantized weights are only used in the feed-forward propagation.

2.2.3 Ternary Weight Networks

TWN [13] trains DNNs with weights constrained to only three values $\pm\alpha_l$ and 0. Compared with DoReFa-Net, TWN has an extra zero, which requires 2 bits to represent weights while also improving the performance. Note that TWN is also applied in a layer-wise manner, which is the same with DoReFa-Net. For each layer, the quantization of TWN is shown in Eq. (3).

$$w_l^q = \begin{cases} \alpha_l & \text{if } |w_l| > \delta_l; \\ 0 & \text{if } -\delta_l \leq |w| \leq \delta_l; \\ -\alpha_l & \text{if } |w_l| < -\delta_l; . \end{cases} \tag{3}$$

As there is no deterministic solution for δ_l and α_l, an approximated optimal solution is presented as shown in Eq. (4) and Eq. (5). Note that the feed-back propagation is the same as that for DoReFa-Net.

$$\delta_l = 0.7 \times E(|w_l|) \tag{4}$$

$$\alpha_l = \underset{i \in \{i|w_l(i))|\} > \delta_l}{E} (|w_l(i))|) \tag{5}$$

3. Motivation

Usually quantization of DNNs are used to reduce the bit length of weights in DNNs. In fact, quantization can not only reduce memory consumption, but also can improve the performance sometimes [11][23][27]. For example, Han et al. [11] has improved the Top-1 error by 0.01% for ImageNet classification. Zhou et al. [27] has quantized DNNs to only 4 and 5 bits for ImageNet classification, and the Top-1 and Top-5 error for the two configurations are all improved with a reduction of 0.2%-1.47%. One interesting phenomenon is that the Top-5 error with quantization of 3 bits is lower than that with quantization of 4 bits. A possible explanation is that lower bits representation is a more strict constraint to reduce overfitting. We would like to apply the above idea to suggestive annotation [26] to reduce overfitting and improve performance.

Two quantization processes for the two steps in the suggestive annotation framework have different purposes. For suggestive annotation, the purpose is to obtain representative samples, and therefore, uncertainty is more critical

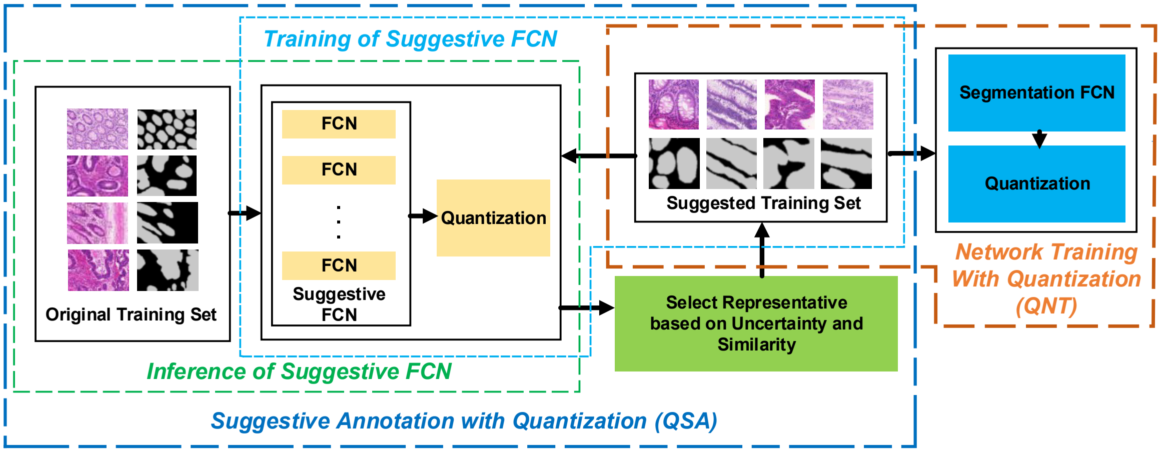

Training of Suggestive FCN
Original Training Set
FCN
FCN
FCN
Suggestive FCN
Quantization
Suggested Training Set
Select Representative based on Uncertainty and Similarity
Segmentation FCN
Quantization
Network Training With Quantization (QNT)
Inference of Suggestive FCN
Suggestive Annotation with Quantization (QSA)

Figure 3. Illustration of quantization framework based on the suggestive annotation framework. In suggestive annotation with quantization, better training samples (suggestive training set) can be extracted from the original training set. In network training with quantization, better performance can be achieved by reduce overfitting.

than accuracy. For network training, the purpose is to increase accuracy, and several characteristics of FCNs need to be considered. First, unlike general DNNs with multiple fully connected layers, all layers in FCNs are convolutional or deconvolutional layers, which is an extreme case of weight sharing. Second, unlike general classification tasks with only several outputs, FCNs generate the same number of outputs as that of the inputs. This makes quantization of FCNs much harder, which has less space for quantization compared with general DNNs. We would like to explore suitable quantization method for FCNs in network training.

4. Method

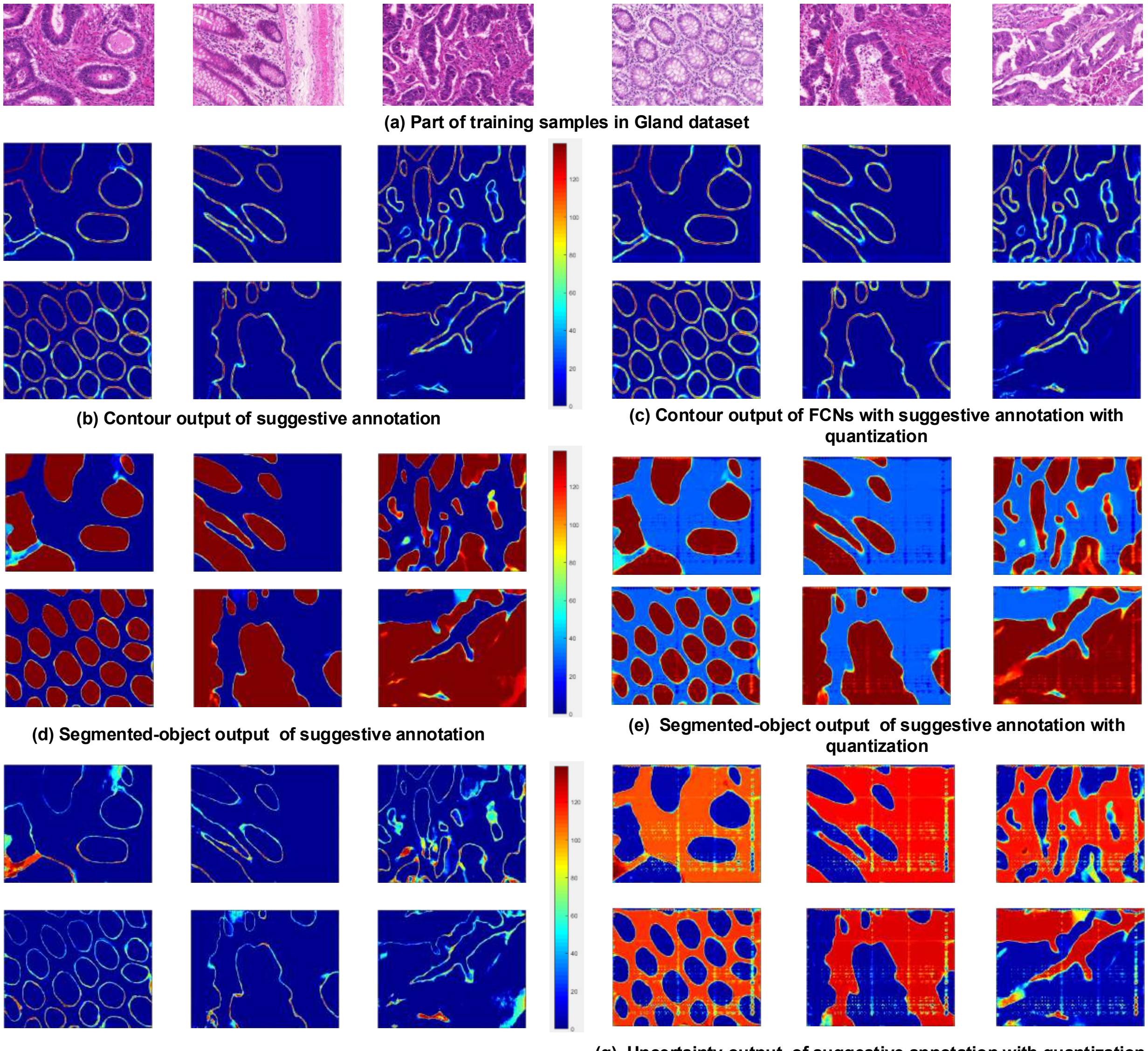
In this section, we focus on suggestive annotation with quantization as network training with quantization is relatively simple. Additionally, uncertainty and similarity of the enhanced suggestive annotation are also analysed in details.

4.1. Suggestive Annotation with Quantization

As shown in Figure 3, the proposed quantization framework has two steps: suggestive annotation with quantization and network training with quantization. In the first step, we add a quantization module to suggestive FCNs for high uncertainty. In the second step, quantization of segmentation FCNs are performed with the suggestive training samples for higher accuracy. In order to obtain high representativeness, each FCN in suggestive FCNs should be diverse for high uncertainty with acceptable accuracy. However, usually DNNs including FCNs are over-parameterized, and a large portion of the parameters is redundant. Thus, multiple suggestive FCNs will have very small variance of the final prediction though with different weight initialization. The adopted regularization techniques including weight decay and dropout scheme [20] will further make the multiple suggestive FCNs to be almost the same. By adding quantization to suggestive annotation, the above requirement can be satisfied. Though it may be a little offensive since most of the time it will degrade the accuracy, it is particularly appreciated by suggestive FCNs that focus on uncertainty. Particularly quantization transforms the originally continuous weight space, where the weights of several networks can be arbitrarily close, into a sparse and discrete one, and thus increasing the distances between the trained networks and accordingly the diversity of the outputs. Note that accuracy should be also considered and too offensive quantization methods should be avoided.

4.2. Impact on Uncertainty and Similarity

In suggestive annotation with quantization, high uncertainty can be obtained without sacrificing much accuracy. As shown in Figure 4, accuracy including contour and segmented object and uncertainty are compared. Note that the suggestive FCNs output both contour and segmented object for high segmentation performance. Comparing Figure 4(b) and Figure 4(c), we can notice that the contour for both approaches are almost the same, and they can both obtain clear contours. However, for segmented object in Figure 4(d) and Figure 4(e), suggestive annotation identifies a very clear segmented object, while the quantized version is relatively vague. This is mainly due to the fact that suggestive annotation with quantization has a larger uncertainty of the background data, and this is verified in Figure 4(f) and Figure 4(g). The uncertainty scores of suggestive annotation with quantization are much higher than that of suggestive annotation. Therefore, suggestive training set with higher uncertainty can be obtained with quantization at the same time with little accuracy loss. Note that as shown in Figure 4(b,c,d,e), the magnitudes of the artefacts (all in blue) are much smaller compared with those of contours and objects and thus having little impacts on the results.



(a) Part of training samples in Gland dataset

(b) Contour output of suggestive annotation

(c) Contour output of FCNs with suggestive annotation with quantization

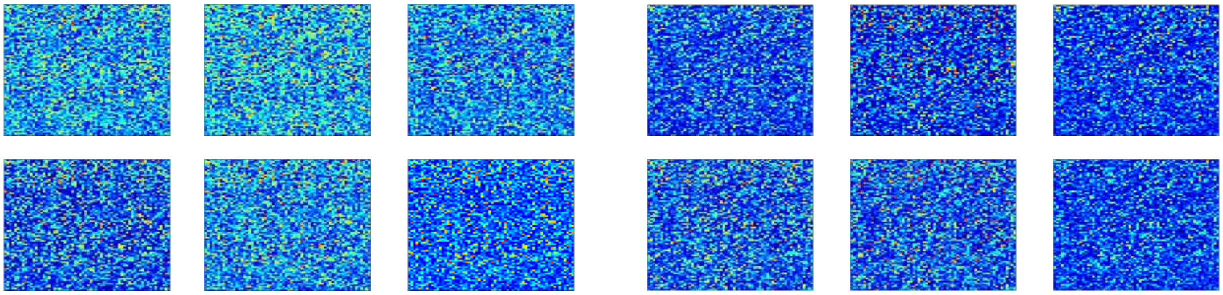
(d) Segmented-object output of suggestive annotation

(e) Segmented-object output of suggestive annotation with quantization

(f) Uncertainty output of suggestive annotation

(g) Uncertainty output of suggestive annotation with quantization

Figure 4. Uncertainty comparison between suggestive annotation and suggestive annotation with quantization. The accuracy of contour and segmented object and uncertainty are compared, respectively. There is almost no accuracy loss. However, suggestive annotation with quantization has higher uncertainty scores.



(a) Suggestive annotation

(b) Suggestive annotation with quantization

Figure 5. Similarity (the outputs of the final convolutional layer of FCNs, which can be regarded as an image descriptor) comparison between suggestive annotation and suggestive annotation with quantization.

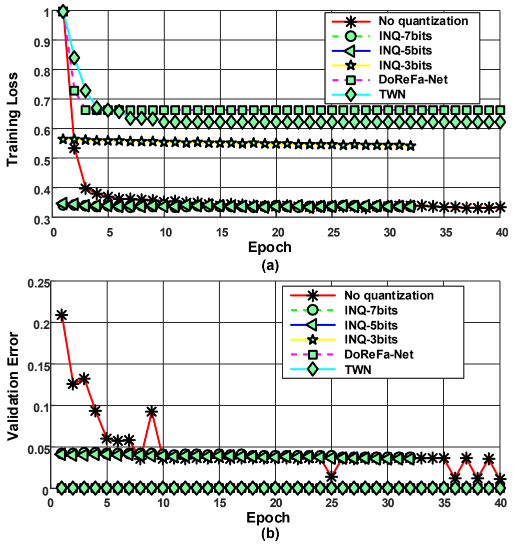


Figure 6. (a) Learning loss and (b) validation error on the Gland dataset with various quantization methods.

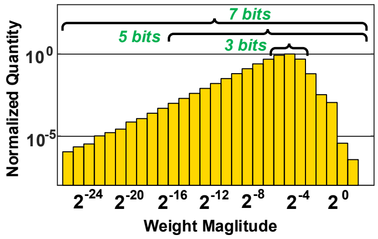


Figure 7. Magnitude distribution of weights in FCNs.

As shown in Figure 5, the similarity (the output of the last convolutional layer) comparison between suggestive annotation and suggestive annotation with quantization is discussed. As the dimension of the output image matrix is relatively large (64×80), its details are not clear. However, we can still notice that the distributions of the image of the two approaches have some differences. For each approach, there also exists variance among the outputs.

5. Experiment and Results

5.1. Experiment Setup

We adopt the 2015 MICCAI Gland Challenge dataset [17] which have 85 training images (Part A: 37 normal glands, and Part B: 48 abnormal glands) and 80 testing images (Part A: 60 normal glands, and Part B: 20 abnormal glands). In suggestive annotation, 16 images with the highest uncertainty scores are extracted first, and then 8 images are collected based on their representativeness using similarity, which are added to the suggested training set in each iteration. Note that the training samples in the first iteration are selected randomly. Totally there are 120 iterations in suggestive annotation, and totally 960 suggested training samples are produced. 5 FCNs are used in suggestive annotation, and the waiting time between two annotation suggestion stages is about 10 minutes on a workstation with 4 NVIDIA Tesla P100 GPUs. We adopt a simple learning rate scaling strategy: set learning rate to 0.0005 in the initial stage, and to 0.00005 when the iteration times reaches a threshold. As the training time is long, all the configurations are repeated 4 times and the best ones are selected for comparison.

We will discuss three aspects in the experiment regarding quantization of suggestive annotation (SA), number of parallel FCNs in suggestive annotation, and quantization of network training (NT). Note that without explicit specifications, one FCN is used in training for segmentation. All the experiments are evaluated considering detection (F1 score), segmentation (dice score) and shape similarity (object Hausdorff distance) [17]. Several widely-used quantization methods are discussed: incremental quantization, DoReFa-Net, and TWN. We first perform a simple FCN training with the above quantization methods. For incremental quantization, we first analyzed the distribution of the weights as shown in Figure 7, and select three configurations: 7 bits, 5 bits, and 3 bits. As shown in Figure 6, it can be noticed that only incremental quantizations with 7 bits and 5 bits have low training loss and achieve comparable performance on the validation dataset with unquantized networks. Incremental quantization with 3 bits, DoReFa-Net, and TWN obtain a large training loss, and their validation accuracy is almost zero. Though this is common in network quantization, the accuracy degradation of FCNs is much larger compared with general DNNs, which is possibly due to the following two reasons. First, unlike general DNNs, FCNs has no fully connected layers resulting in less redundance. Second, the performance of segmentation is determined in the object level, which means successful segmentation requires correct classification of a doze of pixels in an object. This is much harder than general classifications using DNNs. Considering the above discussions, we adopt incremental quantization with 7 bits and 5 bits in the rest of the experiments.

5.2. Impact of Number of Parallel FCNs

We first discuss the impact of number of parallel FCNs in suggestive annotation. As shown in Figure 8, six configurations are discussed. We find that the same trend exists in all configurations: a moderate accuracy is obtained with number of two, and then the accuracy decreases, and a local minimum occurs with the number of around four; then the accuracy will increase to a local maximum and decrease afterwards. It seems that there exists much redundance in FCNs for suggestive annotation, and proper number of parallel FCNs will contribute to the performance. We will adopts 5 parallel FCNs in suggestive annotation in the experiments afterwards. In Figure 8(e), we can find that network training can achieve higher accuracy in most of the configurations

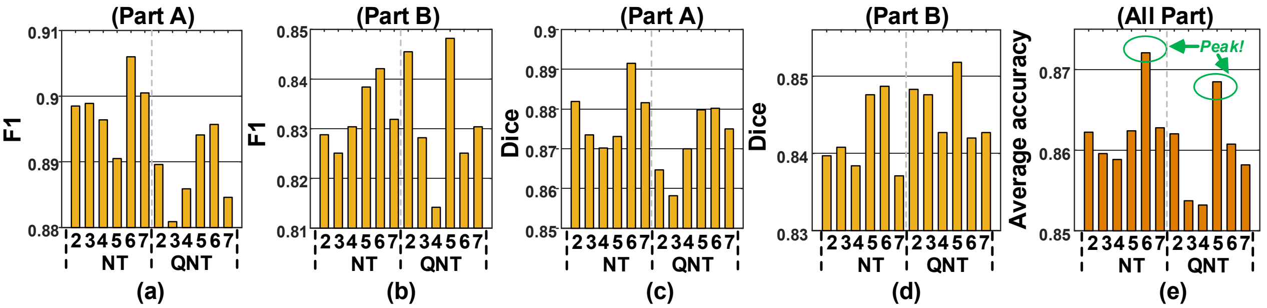


Figure 8. Performance comparison of **different number (2,3,4,5,6,7) of suggestive FCNs** with **network training (NT)** and **network training with quantization (QNT)**. The QNT is quantized using INQ with 7 bits.

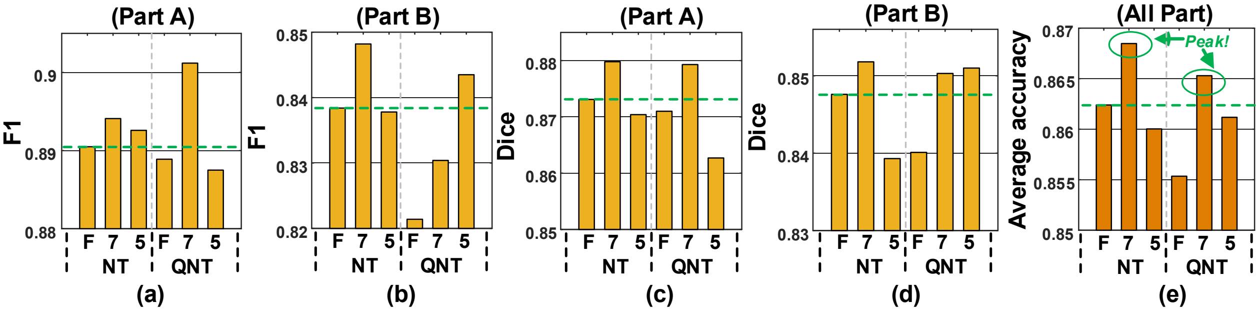


Figure 9. Performance comparison of quantization methods (**F: float, 7: INQ with 7 bits, 5: INQ with 5 bits**) in **suggestive annotation with quantization** with **network training (NT)** and **network training with quantization (QNT)**. The QNT is quantized using INQ with 7 bits. The green dash-line corresponds to the performance of the work [26] with the same configuration.

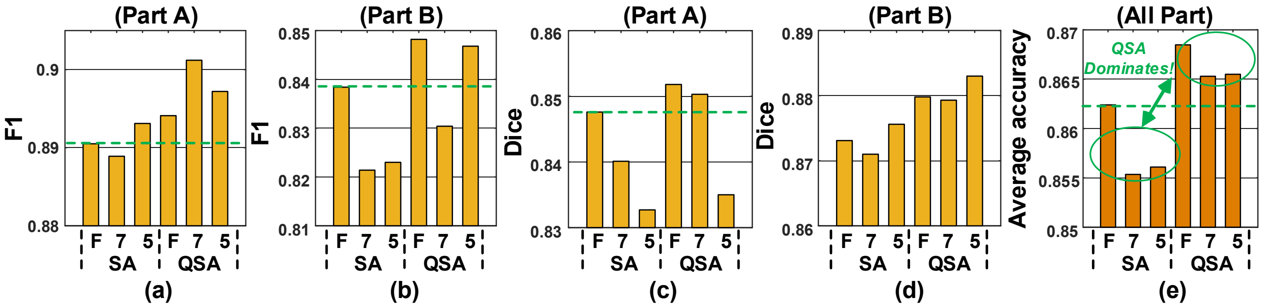


Figure 10. Performance comparison of quantization methods (**F: float, 7: INQ with 7 bits, 5: INQ with 5 bits**) in **network training with quantization** with training samples from suggestive annotation (SA) and suggestive annotation with quantization (QSA). The QSA is quantized using INQ with 7 bits. The green dash-line corresponds to the performance of the work [26] with the same configuration.

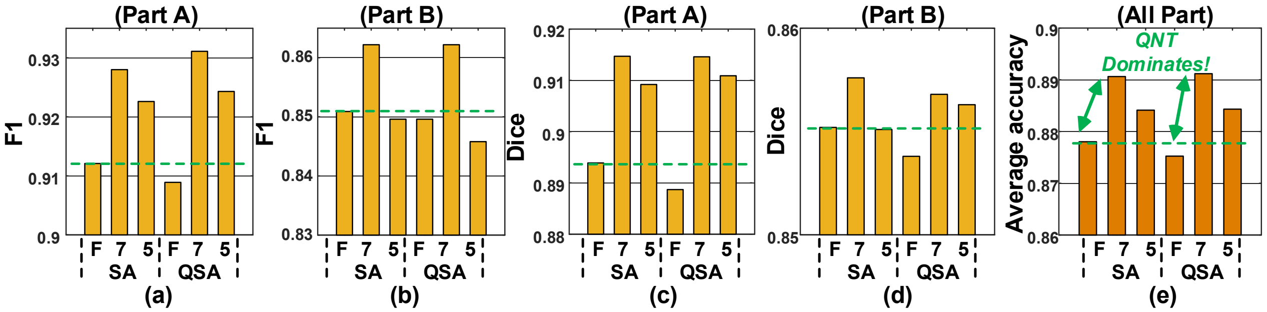


Figure 11. Performance comparison of quantization methods (**F: float, 7: INQ with 7 bits, 5: INQ with 5 bits**) in **network training with quantization** using training samples from suggestive annotation (SA) and suggestive annotation with quantization (QSA). The QSA is quantized using INQ with 7 bits. Note that **ensemble technique with 5 FCNs** are adopted here. The green dash-line corresponds to the performance of the work [26] with the same configuration.

Table 1. Performance comparison with existing works using **five** FCNs on the MICCAI Gland dataset. The work [26] achieves state-of-the-art performance on the dataset.

Configuration	F1 Score		Object Dice		Object Hausdorff	
	Part A	Part B	Part A	Part B	Part A	Part B
SA (5 FCNs + INQ-7bits) + NT (5 FCNs + INQ-7bits)	**0.930**	**0.862**	**0.914**	**0.859**	**41.783**	97.390
Suggestive annotation [26]	0.921	0.855	0.904	0.858	44.736	**96.976**
Multichannel [25]	0.893	0.843	0.908	0.833	44.129	116.821
Multichannel [24]	0.858	0.771	0.888	0.815	54.202	129.930
CUMedVision [3]	0.912	0.716	0.897	0.781	45.418	160.347

compared with network training with quantization. That is to say quantization of network training will hurt the accuracy for some configurations.

5.3. Discussion on Suggestive Annotation Quantization

As shown in Figure 9, suggestive annotation with INQ with 7 bits can always obtain higher accuracy compared with that with the other two. This reversed U-shape trend indicates that suggestive annotation with INQ with 7 bits may be close to the best fitting point, and loose quantization (no quantization or floating-point representation) and tight quantization (INQ with 5 bits) both degrade the fitting and accuracy loss arises. By comparing network training and network training with quantization, we can find that network training with quantization will not always improve the accuracy.

5.4. Discussion on Network Training

As shown in Figure 10, unlike suggestive annotation with quantization, the highest accuracy of network training with quantization is achieved with floating-point representation in most of the configurations. This means network training with quantization will degrade the performance. By comparing network training and network training with quantization, we can notice that suggestive annotation with quantization has a great contribution to performance improvement, and the average improvement is 0.9%.

5.5. Comparison with Existing Works

In order to make fair comparison with existing works, we adopts ensemble methods and set the number of FCNs in network training to five, which is the same as [26]. Several configurations are evaluated as shown in Figure 11. Suggestive annotation with quantization shows the same trend as network training with quantization. In Figure 9 and Figure 10, suggestive annotation with quantization has a great impact on the performance with one FCN, while network training with quantization has a significant influence on the performance with five FCNs. This is due to the fact that the network behaviour of multiple networks with ensemble methods differs from that of only one network.

Comparison with existing works are shown in Table 1. With proper quantization techniques, our proposed method can achieve the best performance on all aspects except object Hausdorff distance on part B. For part A with nonmalignant subjects, out methods can achieve a 0.9%-1% improvement with the current state-of-the-art method. For part B with malignant subjects, it is much harder to segment, and our method gets a 0.1%-0.7% improvement. We achieve comparable performance on object Hausdorff distance on part B, which is only 0.4% worse than suggestion annotation. In addition, our method can also obtain 4.6x and 6.4x reduction on memory usage for INQ with 7 bits and 5 bits, respectively. As activations are in floating point representation, the runtime are not affected.

6. Conclusion

Usually quantization is used to reduce the bit length of parameters with some accuracy loss. In this paper, we apply quantization to FCNs for accurate biomedical image segmentation, and quantization is used to reduce overfitting in FCNs. Particularly we base our work on the current state-of-the-art work [26], and it has two steps: suggestive annotation and network training. We add two quantization processes to the two steps, respectively: one to suggestive annotation for high-representative training samples, and the other to general training for high accuracy. Extensive experiments are presented on the widely-used MICCIA Gland dataset. Results show that both quantization processes can improve the segmentation performance by around 1% for some configurations. However, for specific networks, usually there is only one process dominates in the performance. For network training with only one FCN, suggestive annotation with quantization dominates, while network training with quantization dominates for network training with five FCNs. The number of parallel FCNs in suggestive annotation will also affect the performance. Our proposed method exceeds the current state-of-the-art performance by up to 1%. In addition, our method has a up to 6.4x reduction on memory usage. Our future work will focus on a general quantization principle that should also work for other DNN frameworks.

References

[1] D. Altunbay, C. Cigir, C. Sokmensuer, and C. Gunduz-Demir. Color graphs for automated cancer diagnosis and grading. *IEEE Transactions on Biomedical Engineering*, 57(3):665–674, 2010. 1

[2] H. Chen, X. Qi, J.-Z. Cheng, P.-A. Heng, et al. Deep contextual networks for neuronal structure segmentation. In *AAAI*, pages 1167–1173, 2016. 1

[3] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: Deep contour-aware networks for accurate gland segmentation. In *CVPR*, pages 2487–2496, 2016. 1, 2, 8

[4] M. Courbariaux and Y. Bengio. Binarynet: Training deep neural networks with weights and activations constrained to+ 1 or-1. corr abs/1602.02830 (2016). 2

[5] M. Courbariaux, Y. Bengio, and J.-P. David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, pages 3123–3131, 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 *CVPR*, pages 248–255. IEEE, 2009. 2

[7] Y. Ding, J. Liu, and Y. Shi. On the universal approximability of quantized relu neural networks. *arXiv preprint arXiv:1802.03646*, 2018. 2

[8] S. Doyle, A. Madabhushi, M. Feldman, and J. Tomaszewski. A boosting cascade for automated detection of prostate cancer from digitized histology. *MICCAI*, pages 504–511, 2006. 1

[9] H. Fu, G. Qiu, J. Shu, and M. Ilyas. A novel polar space random field model for the detection of glandular structures. *IEEE transactions on medical imaging*, 33(3):764–776, 2014. 1

[10] C. Gunduz-Demir, M. Kandemir, A. B. Tosun, and C. Sokmensuer. Automatic segmentation of colon glands using object-graphs. *Medical image analysis*, 14(1):1–12, 2010. 1

[11] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015. 2, 3

[12] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *arXiv preprint arXiv:1609.07061*, 2016. 2

[13] F. Li, B. Zhang, and B. Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016. 2, 3

[14] K. Nguyen, A. Sarkar, and A. K. Jain. Structure and context in prostatic gland segmentation and classification. In *MICCAI*, pages 115–123. Springer, 2012. 1

[15] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016. 2

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[17] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017. 1, 6

[18] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot. A novel texture descriptor for detection of glandular structures in colon histology images. In *SPIE Medical Imaging*, pages 94200S–94200S. International Society for Optics and Photonics, 2015. 1

[19] K. Sirinukunwattana, D. R. Snead, and N. M. Rajpoot. A stochastic polygons model for glandular structures in colon histology images. *IEEE transactions on medical imaging*, 34(11):2366–2378, 2015. 1

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4

[21] A. Tabesh, M. Teverovskiy, H.-Y. Pang, V. P. Kumar, D. Verbel, A. Kotsianti, and O. Saidi. Multifeature prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 26(10):1366–1378, 2007. 1

[22] X. Xu, Q. Lu, T. Wang, J. Liu, H. Yu, and Y. Shi. Efficient hardware implementation of cellular neural networks with powers-of-two based incremental quantization. In *Neuromorphic Computing Symposium*, 2017. 2

[23] X. Xu, Q. Lu, T. Wang, J. Liu, C. Zhuo, X. S. Hu, and Y. Shi. Edge segmentation: Empowering mobile telemedicine with compressed cellular neural networks. In *ICCAD*, pages 880–887. IEEE, 2017. 2, 3

[24] Y. Xu, Y. Li, M. Liu, Y. Wang, M. Lai, I. Eric, and C. Chang. Gland instance segmentation by deep multichannel side supervision. In *MICCAI*, pages 496–504. Springer, 2016. 8

[25] Y. Xu, Y. Li, Y. Wang, M. Liu, Y. Fan, M. Lai, and E. Chang. Gland instance segmentation using deep multichannel neural networks. *IEEE Transactions on Biomedical Engineering*, 2017. 8

[26] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. *arXiv preprint arXiv:1706.04737*, 2017. 1, 2, 3, 7, 8

[27] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017. 2, 3

[28] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016. 3

[29] C. Zhu, S. Han, H. Mao, and W. J. Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016. 2