Synthetic Augmentation and Feature-based Filtering for Improved Cervical Histopathology Image Classification

Yuan Xue[1], Qianying Zhou[1], Jiarong Ye[2], L. Rodney Long[3], Sameer Antani[3], Carl Cornwell[3], Zhiyun Xue[3], and Xiaolei Huang[1]

[1] College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, USA
[2] Department of Statistics, Pennsylvania State University, University Park, PA, USA
[3] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract. Cervical intraepithelial neoplasia (CIN) grade of histopathology images is a crucial indicator in cervical biopsy results. Accurate CIN grading of epithelium regions helps pathologists with precancerous lesion diagnosis and treatment planning. Although an automated CIN grading system has been desired, supervised training of such a system would require a large amount of expert annotations, which are expensive and time-consuming to collect. In this paper, we investigate the CIN grade classification problem on segmented epithelium patches. We propose to use conditional Generative Adversarial Networks (cGANs) to expand the limited training dataset, by synthesizing realistic cervical histopathology images. While the synthetic images are visually appealing, they are not guaranteed to contain meaningful features for data augmentation. To tackle this issue, we propose a synthetic-image filtering mechanism based on the divergence in feature space between generated images and class centroids in order to control the feature quality of selected synthetic images for data augmentation. Our models are evaluated on a cervical histopathology image dataset with limited number of patch-level CIN grade annotations. Extensive experimental results show a significant improvement of classification accuracy from 66.3% to 71.7% using the same ResNet18 baseline classifier after leveraging our cGAN generated images with feature based filtering, which demonstrates the effectiveness of our models.

1 Introduction

Cervical cancer is the fourth-most frequently diagnosed cancer among women all over the world [1]. The diagnosis of cervical cancer and its precancerous stages can be accomplished through assessment of histopathology slides of cervical tissue by pathologists. An important outcome of the assessment is the cervical intraepithelial neoplasia (CIN) grade, an important indicator for abnormality assessment identified by the abnormal growth of cells on the surface of the cervix. CIN grade can be divided into CIN1, CIN2, and CIN3 with increased severity from mild to severe. Thus, accurate assessment of CIN grade is crucial for diagnosis and treatment planning of cervical cancer.

Considering the shortage of pathologists, an automatic cervical histopathology image classification system has great potential in under developed regions for its low cost and accessibility. Moreover, such a system can help pathologists with diagnosis and

potentially mitigate the inter- and intra- pathologist variation. Existing literature [2, 6] have studied various supervised learning methods for nuclei-based cervical cancer classification. Chankong *et al.* [2] proposed automatic cervical cancer cell segmentation and classification using fuzzy C-means (FCM) clustering and various types of classifiers. Guo *et al.* [6] designed hand-crafted nuclei-based features for fusion-based classification on digitized epithelium histopathology slides with linear discriminant analysis (LDA) and support vector machines (SVM) classifier. While accomplishments have been achieved with fully-supervised learning methods, they require large amounts of expert annotations of cervical whole slide images. Since the annotation process can be tedious and time-consuming, it often results in limited number of labeled data available for supervised learning models.

Recently, several works have leveraged unsupervised learning methods, more specifically, Generative Adversarial Networks (GANs) [4] in medical image analysis to mitigate the small dataset sizes and limited annotations [10, 14, 3]. Frid-Adar *et al.* [3] investigated conditional GANs (cGANs) [11] to generate synthetic CT images and improved the performance of CNN in liver lesion classification, by adding generated images into the training data as data augmentation. Similarly, Madani *et al.* [10] found GAN based data augmentation achieved higher accuracy than traditional augmentation in Chest X-ray classification. Ren *et al.* [14] explored a method to classify the prostate histopathology images by domain adaptation so that knowledge learned in the source domain can be transferred to the target dataset without annotation. Although the GANs in previous applications can generate visually appealing synthetic images, the feature quality of generated images varies among examples and not all of them are guaranteed to contain meaningful features to improve the model performance in the original task.

In this paper, we study the 4-class (normal, CIN 1-3) cervical histopathology image classification problem based on a ResNet18 [7] baseline classifier. We run and evaluate our models on a heterogeneous epithelium image dataset with limited number of patch-level annotations. Images in the dataset have various color, shapes and texture which makes the classification very challenging. While the capability of the baseline model is limited by the number of training data, we propose a cGAN based image synthesis model to generate high-fidelity synthetic epithelium histopathology patches to expand the training data. To improve the diversity of synthetic images, we incorporate the minibatch discrimination [15] to reduce the closeness between examples inside a minibatch. Moreover, unlike previous works which directly added the generated data into the training set, we apply a feature based filtering mechanism to further improve the feature quality of the synthetic images added. We first pre-train a feature extractor using baseline ResNet18 model and calculate feature centroids for each class as the average feature of all training images. The generated images are then filtered based on the distance to the corresponding centriod in the feature space. Experimental results show that our proposed cGAN model along with the feature based filtering significantly outperform the baseline ResNet18 model and the traditional augmentation methods.

2   Methodology

An overall illustration of our proposed data augmentation pipeline can be found in Fig. 1. In traditional fully-supervised training, the model is trained on training images
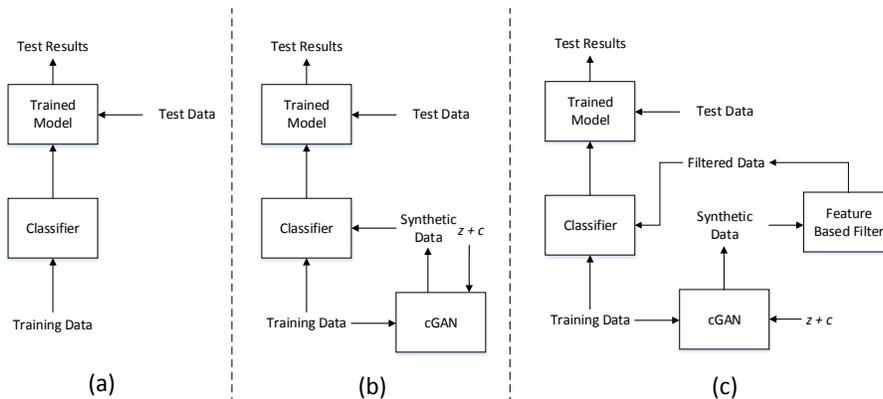
Fig. 1: Illustration and comparison between different training processes. (a) Traditional training pipeline; (b) Conditional GAN augmented training pipeline; (c) Our proposed cGAN augmentation with feature based filtering. The input to the cGAN are noise vector z and label condition vector c.

and the inference is done by feeding the test data to the trained model. In previous GAN-based augmentation works [3, 10], a GAN model is first trained to generate some synthetic images based on the training data, then the generated images are added to the original training data as a data augmentation strategy. However, since the discriminator in GAN only outputs a high level judgement (0 or 1) of the fidelity of generated images, such pipeline cannot guarantee that the generated data have similar features to the real images which contribute to the classification task. To tackle this issue, we propose a feature based filtering mechanism to further improve the feature quality and fidelity of the synthetic images. We first introduce the cGAN model used in our framework.

2.1 Theoretical Preliminaries

The conventional cGANs [11] have an objective function defined as:

$$\min_{\theta_G} \max_{\theta_D} \mathcal{L}_{\text{cGAN}} = \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x \mid c)] + \mathbb{E}_{z \sim \mathcal{N}}[\log(1 - D(G(z \mid c)))] \ . \tag{1}$$

In the objective function above, θ_G and θ_D represent the parameters for the generator G and discriminator D in cGAN, respectively. x represents the real data from an unknown distribution p_{data} and c is the conditional label (*e.g.*, CIN grades). z is a random vector for the generator G, drawn from a normal distribution $\mathcal{N}(0, 1)$. G is trained to fool the discriminator with synthetic data by minimizing the objective. Meanwhile, D that takes both z and c as input is trained to maximize the objective, aiming to distinguish real data and synthetic images generated by G.

During the experiments, we observe that the intra-class diversity of the generated images greatly affects the data augmentation performance. Consider a mode-collapse

GAN model which only generates very limited number of modes, all generated images will look similar. In such case, no matter how many examples are added to the training data, only few of them contribute to the data augmentation. To this end, we incorporate the minibatch discrimination module [15] into our discriminator to reduce the homogeneity between generated examples inside a minibatch. Let $f(x_i) \in R^A$ denote a feature vector for input x_i from an intermediate layer in the discriminator, where A = out channels $\times$ weight $\times$ height. We multiply $f(x_i)$ by a transformation matrix $T \in R^{A \times B \times C}$ to compute a matrix $M_i \in R^{B \times C}$, where B represents the number of output features of the minibatch discriminator, and C refers to the number of kernel dimension which is set to 3 by default in our experiments. The similarity of the image x_i with the rest of images x_j in the same batch is computed as

$$o(x_i)_b = \sum_{j=1}^{n} \exp(-||M_{i,b} - M_{j,b}||_{L_1}) \ , \tag{2}$$

and $o(x_i) = [o(x_i)_1, o(x_i)_2, ..., o(x_i)_B] \in R^B$. Then the similarity $o(x_i)$ is concatenated with $f(x_i)$ and fed into the next layer of the discriminator. Minibatch discrimination penalizes the generator if mode is collapsing and encourages the model to generate diverse images. With minibatch discrimination, our cGAN models can generate more diverse images which can be effectively added to the training data for augmentation.

We then introduce the feature based filtering method after generating synthetic images from our trained cGAN model. One of the advantages of GANs is that the input is drawn from a distribution and one can generate an infinite number of images with a trained GAN. Thus, any filtering methods will not affect the number of images available to be added to the training data. We first pre-train a feature extractor using the baseline classifier on the original training set to extract visual features of the input images. The features consist of activations produced by different layers in the feature extractor. The feature distance between image x and centroid c is then defined as

$$D_f(x, c) = \sum_l \frac{1}{H_l W_l} ||\hat{\phi}_l(x) - \hat{\phi}_l(c)||_2^2 \ , \tag{3}$$

where $\hat{\phi}_l$ is the unit-normalized activation in the channel dimension A_l of the lth layer of a feature extraction network with shape $H_l \times W_l$. This ℓ_2 distance between unit-normalized activation can be regarded as a cosine distance in the feature space.

The centroid c is calculated as the average feature of all training images in the same class. For class i, its centroid c_i is represented by

$$c_i = [\frac{1}{N_i} \sum_{j=1}^{N_i} \phi_1(x_j), \frac{1}{N_i} \sum_{j=1}^{N_i} \phi_2(x_j), ..., \frac{1}{N_i} \sum_{j=1}^{N_i} \phi_L(x_j)] \ , \tag{4}$$

where N_i denotes the number of training samples in ith class and x_j is the jth training sample. Similar to Eq. 3, ϕ_l is the activation extracted from the lth layer of the feature extraction network. L is the total number of layers utilized in the feature based filtering.

2.2   Implementation Details

Our GAN model is built upon the DCGAN [13] and WGAN-GP [5] with several alterations. The conditional generator consists of 8 convolutional blocks, where the first

block consists of two separate transpose convolutional blocks for conditional labels c and random vector z. After the first block, the activations of c and z are concatenated along the channel dimension and fed into the next layer of the generator. Different from DCGAN which uses transpose convolution for upsampling, our convolutional block consists of a bilinear upsampling with factor 2 except for the first shared block which has factor 2×1, a 3×3 convolution with stride 1, a batch normalization layer and a ReLU activation. The final block is a 7×7 convolution followed by a tanh activation.

The conditional discriminator consists of 7 convolutional blocks, where the first block includes two separate convolutional blocks for conditional labels c and random vector z with no batch normalization as in DCGAN. The activations of c and z are concatenated and fed into the rest of convolutional blocks. Each of the rest of the convolutional blocks contains a 4×4 convolutional layer, a batch normalization layer and a LeakyRelu activation. Each convolution has stride 2 except for the 6th and 7th block which has stride 2×1 and 1, respectively. Activations of the last convolutional block are then fed into the minibatch discrimination layer as described in Eq. 2. After the minibatch discrimination, a fully connected layer outputs the final logits of discriminator.

Our cGAN model is trained with WGAN-GP loss with batchsize 100, fixed learning rate $2e-4$ and 700 training epochs. The baseline classifier is the widely used ResNet18 [7] which has shown promising performance in various vision tasks. All classification models are trained using the same baseline classification model with batchsize 64 by Adam optimizer [8] with weight decay $2e-5$ and the cross-entropy loss. The initial learning rate of all classification models is $1e-3$ and is reduced by factor 0.2 when the validation accuracy has stopped being improved for 5 epochs.

3   Experiments

The experimental dataset is a cervical histopathology image dataset collected from a collaborating health sciences center. It contains multiple data sources, all of which are annotated by the same pathologist. The data processing follows [6] by dividing an annotated image into patches according to the medial axis of the epithelium tissue in that image. We first divide the medial axis into several straight segments, and crop the patches with each containing one straight segment. The cropped patches are then rotated such that all their medial axes align horizontally after rotation. The cropped patches are resized to a unified size of 256×128. In total, there are $1,112$ normal, 181 CIN1, 463 CIN2, 454 CIN3 patches. Examples of the images can be found in the first row of Fig. 2. All evaluations are done based on the patch-level ground truth annotations. We randomly split the dataset, by patients, into training, validation, and test sets, with ratio $7 : 1 : 2$ and keep the ratio of image classes almost the same among different sets. All evaluations and comparisons reported in this section are done on the test set.

We report quantitative evaluation scores between all baseline models and our models including the accuracy, area under the ROC curve (AUC), sensitivity and specificity to provide a comprehensive comparison. All models are run for 5 rounds with random initialization for fair comparison. The mean and standard deviation results of 5 runs are illustrated in Table 1. We use the same baseline classifier with same hyperparameters setting in all experiments to ensure differences only come from the augmentation mechanisms.
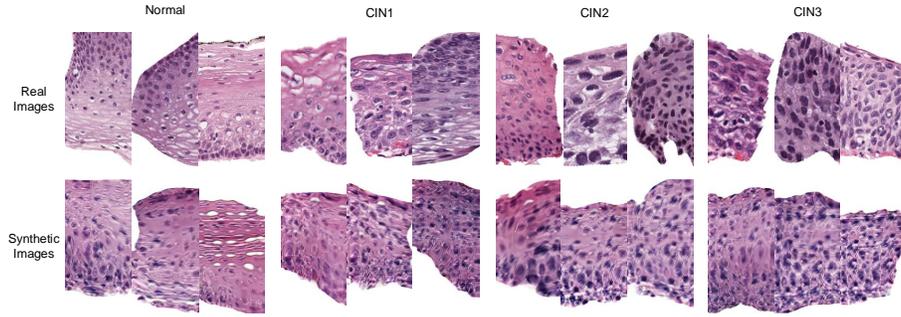
Fig. 2: Examples of real and synthetic images for all CIN grades.

We first conduct experiments with the baseline ResNet18 [7] classifier trained on the original training data and the data with traditional data augmentation. The traditional data augmentation used in our experiments includes random horizontal and vertical flipping, random adjustment of brightness, contrast and saturation to serve as a color based augmentation. From the first two rows of Table 1, one can see that there are no obvious improvements from leveraging traditional data augmentation, which further indicates that traditional augmentation methods do not perform well on such classification problem. Since we cannot ensure the feature quality of the traditional augmentation, we only train all our GAN models using the original training data.

Examples of the synthetic images generated by our cGAN models compared to the real images in the training set are shown in Fig. 2. As one can observe from Fig. 2, our generated images are realistic and keep important features such as color, shape and location of nuclei, and texture information so that the generated images can be used to extend the original training set.

While the synthetic images are visually appealing, high realism is not equivalent to meaningful features for improving classification results. We further explore the distribution of generated images in the feature space in Fig. 3 to ensure they are separable. We use the t-SNE [9] dimension reduction algorithm which can convert the embedding of high-dimensional data into a two-dimensional space for better visualization. After training a baseline ResNet18 classifier with the original training data, we use the pre-trained ResNet18 model as the feature extractor to extract features from the last convolutional layer in the ResNet18 model. We use the same feature extractor for both the expanded training data without and with feature based filtering in Fig. 3 (a) and (b), respectively. Although overall there are overlapping between classes in the feature space which indicates the difficulty of such classification problem, the new training data with feature based filtering clearly have more distinguishable features than the ones without the feature filtering, which supports our claim and is in accord with the improvements in classification performance.

In all GAN-based augmentation models, the generated synthetic images are added into the original training set in equal proportion and we use ratio R to represent the number of synthetic images to the number of real images. We keep the same proportion

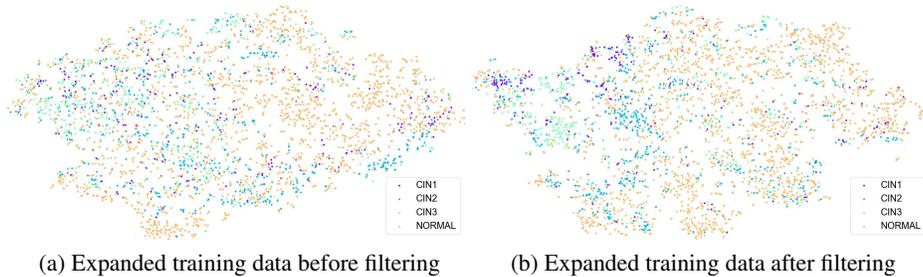(a) Expanded training data before filtering   (b) Expanded training data after filtering

Fig. 3: t-SNE visualization of extracted image features of expanded training data. The ratio of synthetic images to the original training images is set to 2 for both before and after filtering (R=2).

between classes as in the original set to make sure synthetic images generated from minority classes are meaningful for augmentation. We start with a cGAN model with an auxiliary classifier (AC) [12] which is similar to [3]. The auxiliary classifier is added to the last layer of the discriminator network to output the class labels for both real and synthetic images. While cGAN with AC works fine when the ratio is 0.5, the same model with ratio being 2 significantly degrades the classification performance, resulting in even worse scores than the baseline model with no augmentation. One possible reason is that the auxiliary classifier is trained on both real and synthetic images (mainly trained on synthetic images when $R > 1$), the features learned in the training process may not represent the original real data very well. Rather, if a feature extractor is trained only on the real data, the learned features could be more meaningful to the original data.

Following this idea, we provide an ablation study of our cGAN models with and without the feature filtering. For feature based filtering, we first generate $5,000$ images for each class, then the images with lowest distance to the corresponding centroids will be kept. Retained quantity is calculated based on the ratio R. As shown in Table 1, our no filtering cGAN model with different ratios all show improvements in classification performance. Meanwhile, the feature filtering brings obvious benefits to all evaluation metrics, and our full models with different number of synthetic images added to the training data achieved best performance in all metrics. More importantly, the quality levels of generated images from our full models are very stable. Classification performances are similar among different ratios and during different runs, with consistently high mean and low std, which demonstrates that the feature qualities of generated images after filtering are superior to images generated by other models or without filtering.

4 Conclusions

In this paper, we investigate a novel GAN based aumentation pipeline for cervical histopathology image classification problem. We mainly focus on one of the major limitations of using GAN for augmentation: one cannot measure and control the quality of the synthetic images. While traditional GANs try to improve the fidelity of synthetic im-

Method	Accuracy	AUC	Sensitivity	Specificity
ResNet18 [7]	0.663 ± 0.032	0.775 ± 0.021	0.553 ± 0.019	0.866 ± 0.010
Traditional Augmentation	0.670 ± 0.015	0.780 ± 0.010	0.587 ± 0.016	0.868 ± 0.006
cGAN + AC R=0.5	0.687 ± 0.009	0.792 ± 0.006	0.596 ± 0.009	0.874 ± 0.003
cGAN + AC R=2	0.660 ± 0.020	0.773 ± 0.014	0.551 ± 0.032	0.862 ± 0.011
Ours w/o Filtering R=0.5	0.695 ± 0.009	0.796 ± 0.006	0.559 ± 0.032	0.874 ± 0.006
Ours w/o Filtering R=2	0.705 ± 0.010	0.804 ± 0.007	0.555 ± 0.019	0.872 ± 0.005
Ours w/ Filtering R=0.5	0.716 ± 0.009	0.810 ± 0.005	$\mathbf{0.611} \pm 0.012$	$\mathbf{0.886} \pm 0.006$
Ours w/ Filtering R=2	$\mathbf{0.717} \pm 0.008$	$\mathbf{0.811} \pm 0.006$	0.608 ± 0.009	0.882 ± 0.003

Table 1: Quantitative comparisons of baseline classification models and different augmentation models. All evaluation metrics are averaged over four classes. AC stands for the auxiliary classifier in the discriminator and R represents the data augmentation ratio.

ages, an augmentation system should try to generate images which have better feature quality rather than visual realism. By introducing a feature based filtering mechanism, our model boosts the performance of baseline classifier significantly on a challenging cervical histopathology dataset. As an attempt to make better use of GAN based augmentation models in medical imaging, our proposed pipeline has great potentials in other medical imaging applications where the number of labeled data is limited.

5   Acknowledgements

References

1. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A.D.L., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68 6**, 394–424 (2018)
2. Chankong, T., Theera-Umpon, N., Auephanwiriyakul, S.: Automatic cervical cell segmentation and classification in pap smears. Computer methods and programs in biomedicine **113**(2), 539–556 (2014)
3. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)

5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: NeurIPS. pp. 5767–5777 (2017)
6. Guo, P., Banerjee, K., Stanley, R.J., Long, R., Antani, S., Thoma, G., Zuna, R., Frazier, S.R., Moss, R.H., Stoecker, W.V.: Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. IEEE journal of biomedical and health informatics **20**(6), 1595–1607 (2016)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
10. Madani, A., Moradi, M., Karargyris, A., Syeda-Mahmood, T.: Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: Medical Imaging 2018: Image Processing. vol. 10574, p. 105741M. International Society for Optics and Photonics (2018)
11. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
12. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML. pp. 2642–2651 (2017)
13. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
14. Ren, J., Hacihaliloglu, I., Singer, E.A., Foran, D.J., Qi, X.: Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In: MICCAI. pp. 201–209. Springer (2018)
15. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS. pp. 2234–2242 (2016)