Efficient Active Learning for Image Classification and Segmentation Using a Sample Selection and Conditional Generative Adversarial Network

Dwarikanath Mahapatra[1(✉)], Behzad Bozorgtabar[2], Jean-Philippe Thiran[2], and Mauricio Reyes[3]

[1] IBM Research Australia, Melbourne, Australia
dwarim@au1.ibm.com
[2] Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
{behzad.bozorgtabar,jean-philippe.thiran}@epfl.ch
[3] University of Bern, Bern, Switzerland
mauricio.reyes@istb.unibe.ch

Abstract. Training robust deep learning (DL) systems for medical image classification or segmentation is challenging due to limited images covering different disease types and severity. We propose an active learning (AL) framework to select most informative samples and add to the training data. We use conditional generative adversarial networks (cGANs) to generate realistic chest xray images with different disease characteristics by conditioning its generation on a real image sample. Informative samples to add to the training set are identified using a Bayesian neural network. Experiments show our proposed AL framework is able to achieve state of the art performance by using about 35% of the full dataset, thus saving significant time and effort over conventional methods.

1 Introduction

Medical image classification and segmentation are essential building blocks of computer aided diagnosis systems where deep learning (DL) approaches have led to state of the art performance [13]. Robust DL approaches need large labeled datasets which is difficult for medical images because of: (1) limited expert availability; and (2) intensive manual effort required for curation. Active learning (AL) approaches overcome data scarcity with existing models by incrementally selecting the most informative unlabeled samples, querying their labels and adding them to the labeled set [7]. AL in a DL framework poses the following challenges: (1) labeled samples generated by current AL approaches are too few to train or finetune convolution neural networks (CNNs); (2) AL methods select informative samples using hand crafted features [9], while feature learning and model training are jointly optimized in CNNs.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 580–588, 2018.
https://doi.org/10.1007/978-3-030-00934-2_65

Recent approaches to using AL in a DL setting include Bayesian deep neural networks [1], leveraging separate unlabeled data with high classification uncertainty and high confidence for computer vision applications [14], and fully convolution networks (FCN) for segmenting histopathology images [16]. We propose to generate synthetic data by training a conditional generative adversarial network (cGAN) that learns to generate realistic images by taking input masks of a specific anatomy. Our model is used with chest xray images to generate realistic images from input lung masks. This approach has the advantage of overcoming limitations of small training datasets by generating truly informative samples. We test the proposed AL approach for the key tasks of image classification and segmentation, demonstrating its ability to yield models with high accuracy while reducing the number of training samples.

2 Methods

Our proposed AL approach identifies informative unlabeled samples to improve model performance. Most conventional AL approaches identify informative samples using uncertainty which could lead to bias as uncertainty values depend on the model. We propose a novel approach to generate diverse samples that can contribute meaningful information in training the model. Our framework has three components for: (1) sample generation; (2) classification/segmentation model; and (3) sample informativeness calculation. An initial small labeled set is used to finetune a pre-trained $VGG16$ [12] (or any other classification/segmentation model) using standard data augmentation (DA) through rotation and translation. The sample generator takes a test image and a manually segmented mask (and its variations) as input and generates realistic looking images (details in Sect. 2.1). A Bayesian neural network (BNN) [6] calculates generated images' informativeness and highly informative samples are added to the labeled image set. The new training images are used to fine-tune the previously trained classifier. The above steps are repeated till there is no change in classifier performance.

2.1 Conditional Generative Adversarial Networks

GANs [3] learn a mapping from random noise vector z to output image y: $G : z \rightarrow y$. In contrast, conditional GANs (cGANs) [5] learn a mapping from observed image x and random noise vector z, to y: $G : \{x, z\} \rightarrow y$. The generator G is trained to produce outputs that cannot be distinguished from"real" images by an adversarially trained discriminator, D. The cGAN objective function is:

$$L_{cGAN} = E_{x,y\ p_{data}(x,y)} [\log D(x, y)] + E_{x\ p_{data}(x),z\ p_z(z)} [\log (1 - D(x, G(x, z)))], \quad (1)$$

where G tries to minimize this objective against D, that tries to maximize it, i.e. $G^* = \arg\min_G \max_D L_{cGAN}(G, D)$. Previous approaches have used an additional $L2$ loss [10] to encourage the generator output to be close to ground truth in an $L2$ sense. We use $L1$ loss as it encourages less blurring [8], and defined as:

$$L_{L1}(G) = E_{(x,y)\ p_{data}(x,y),z\ p_z(z)} [\|y - G(x, z)\|_1]. \quad (2)$$

Thus the final objective function is :

$$G^* = \arg\min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G), \tag{3}$$

where $\lambda = 10$, set empirically, balances the two components' contributions.

Synthetic Image Generation: The parameters of G, θ_G, are given by,

$$\widehat{\theta} = \arg\min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} l\left(G_{\theta_G}(x, z), x, z\right), \tag{4}$$

N is the number of images. Loss function l combines content loss and adversarial loss (Eqn. 1), and $G(x, z) = y$. Content loss ($l_{content}$) encourages output image y to have different appearance to x. z is the latent vector encoding (obtained from a pre-trained autoencoder) of the segmentation mask. $l_{content}$ is,

$$l_{content} = NMI(x, y) - VGG(x, y) - MSE(x, y). \tag{5}$$

NMI denotes the normalized mutual information (NMI) between x and y, and is used to determine similarity of multimodal images. VGG is the $L2$ distance between two images using all 512 feature maps of Relu $4 - 1$ layer of a pre-trained $VGG16$ network [12]. The VGG loss improves robustness by capturing information at different scales from multiple feature maps. MSE is the intensity mean square error. For similar images, NMI gives higher value while VGG and MSE give lower values. In practice $l_{content}$ is measuring the similarity (instead of dissimilarity in traditional loss functions) between two images, and takes higher values for similar images. Since we are minimizing the total loss function, $l_{content}$ encourages the generated image y to be different from input x.

The generator G (Fig. 1(a)) employs residual blocks having two convolution layers with 3×3 filters and 64 feature maps, followed by batch normalization and ReLU activation. It takes as input the test Xray image and the latent vector encoding of a mask (either original or altered) and outputs a realistic Xray image whose label class is the same as the original image. The discriminator D (Fig. 1(b)) has eight convolution layers with the kernels increasing by a factor of 2 from 64 to 512. Leaky ReLU is used and strided convolutions reduce the image dimension when the number of features is doubled. The resulting 512 feature maps are followed by two dense layers and a final sigmoid activation to obtain a probability map. D evaluates similarity between x and y. To generate images with a wide variety of information we modify the segmentation masks of the test images by adopting one or more of the following steps:

1. **Boundary Displacement**: The boundary contours of the mask are displaced to change its shape. We select 25 continuous points at multiple boundary locations, randomly displace each one of them by $\pm[1, 15]$ pixels and fit a b-spline to change the boundary shape. The intensity of pixels outside the original mask are assigned by linear interpolation, or by generating intensity values from a distribution identical to that of the original mask.

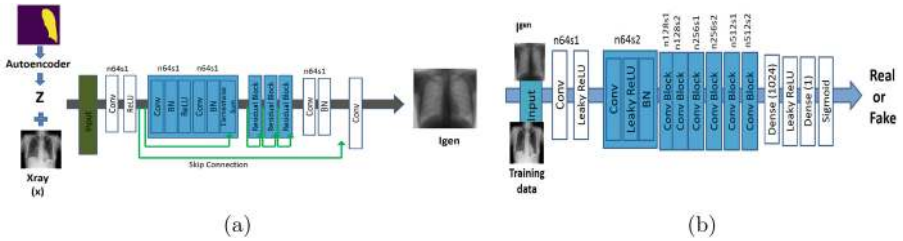

Fig. 1. (a) Generator Network; (b) Discriminator network. $n64s1$ denotes 64 feature maps (n) and stride (s) 1 for each convolutional layer.

2. The intensity values of the lung region are changed by generating values from a uniform distribution modeled as $\alpha\mu + \beta\sigma$, where μ is the original distribution's mean, σ is its standard deviation, and $\alpha = [1, 5], \beta = [2, 10]$ (varied in steps of 0.2).
3. Other conventional augmentation techniques like flipping, rotation and translation are also used.

For every test image we obtain up to 200 synthetic images with their modified masks. Figure 2 (a) shows an original normal image (bottom row) and its mask (top row), and Figs. 2 (b, c) show generated 'normal' images. Figure 2 (d) shows the corresponding image mask for an image with nodules, and Figs. 2 (e, f) show generated 'nodule' images. Although the nodules are very difficult to observe with the naked eye, we highlight its position using yellow boxes. It is quite obvious that the generated images are realistic and suitable for training.

2.2 Sample Informativeness Using Uncertainty Form Bayesian Neural Networks

Each generated image's uncertainty is calculated using the method described in [6]. Two types of uncertainty measures can be calculated from a Bayesian neural network (BNN). Aleotaric uncertainty models the noise in the observation while epistemic uncertainty models the uncertainty of model parameters. We adopt [6] to calculate uncertainty by combining the above two types. A brief description is given below and refer the reader to [6] for details. For a BNN model f mapping an input image x, to a unary output $\widehat{y} \in R$, the predictive uncertainty for pixel y is approximated using:

$$Var(y) \approx \frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t^2 - \left(\frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t\right)^2 + \frac{1}{T}\sum_{t=1}^{T}\widehat{\sigma}_t^2 \tag{6}$$

$\widehat{\sigma}_t^2$ is the BNN output for the predicted variance for pixel y_t, and $\widehat{y}_t, \widehat{\sigma}_t^{2\,T}_{t=1}$ being a set of T sampled outputs.

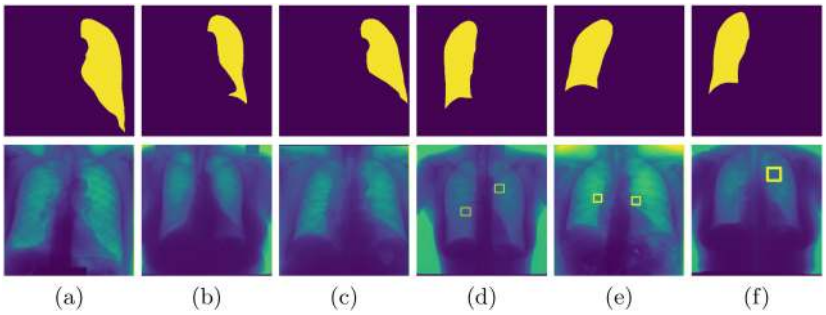

Fig. 2. Mask (Top Row 1) and corresponding informative xray image (Bottom Row); (a)–(c) non-diseased cases; (d)–(f) images with nodules of different severity at the center of yellow box. (a), (d) are the original images while others are synthetic images generated by altering the mask characteristics.

2.3 Implementation Details

Our initial network is a $VGG16$ network [12] or $ResNet18$ [4] pre-trained on the Imagenet dataset. Our entire dataset had 93 normal images and 154 nodule images. We chose an initially labeled dataset of 16 (chosen empirically) images from each class, augment it 200 times using standard data augmentation like rotation and translation, and use them to fine tune the last classification layer of the $VGG16$. The remaining test images and their masks were used to generate multiple images using our proposed cGAN approach (200 synthetic images for every test image as described earlier), and each generated image's uncertainty was calculated as described in Sect. 2.2. We ranked the images with highest uncertainty score and the top 16 images from each class were augmented 200 times (rotation and translation) and used to further fine-tune the classifier. This ensures equal representation of normal and diseased samples in the samples to add to the training data. This sequence of steps is repeated till there is no further improvement of classifier accuracy when tested on a separate test set of 400 images (200 images each of nodule and normal class). Our knowledge of image label allows quantitative analysis of model performance.

3 Experiments

Dataset Description: Our algorithm is trained on the SCR chest XRay database [2] which has Xrays of 247 (93 normal and 154 nodule images, resized to 512×512 pixels) patients along with manual segmentations of the clavicles, lungs and heart. The dataset is augmented 500 times using rotation, translation, scaling and flipping. We take a separate test set of 400 images from the NIH dataset [15] with 200 normal images and 200 images with nodules.

3.1 Classification Results

Here we show results for classifying different images using different amounts of labeled data and demonstrate our method's ability to optimize the amount of labeled data necessary to attain a given performance, as compared to conventional approaches where no sample selection is performed. In one set of experiments we used the entire training set of 247 images and augmentation to fine tune the $VGG16$ classifier, and test it on the separate set of 400 images. We call this the fully supervised learning (FSL) setting. Subsequently, in other experiments for AL we used different number of initial training samples in each update of the training data. The batch size is the same as the initial number of samples.

The results are summarized in Table 1 where the classification performance in terms of sensitivity ($Sens$), specificity ($Spec$) and area under the curve (AUC) are reported for different settings using $VGG16$ and $ResNet18$ [4] classifiers. Under FSL, 5–fold indicates normal 5 fold cross validation; and 35% indicates the scenario when 35% of training data was randomly chosen to train the classifier and measure performance on test data (the average of 10 such runs). We ensure that all samples were part of the training and test set atleast once. In all cases AL classification performance reaches almost the same level as FSL when the number of training samples is approximately 35% of the dataset. Subsequently increasing the number of samples does not lead to significant performance gain. This trend is observed for both classifiers, indicating it is not dependent upon classifier choice.

Table 1. Classification and Segmentation results for active learning framework of Xray images. DM-Dice metric and HD- Hausdorff distance

	Active learning (% labeled + Classifier)										FSL			
	10%		15%		25%		30%		35%		5-fold		35%	
	VGG16 [12]	ResNet18 [4]	[12]	[4]	[12]	[4]	[12]	[4]	[12]	[4]	[12]	[4]	[12]	[4]
Sens	70.8	71.3	75.3	76.2	89.2	89.7	91.5	91.8	91.7	91.9	92.1	92.4	78.1	78.5
Spec	71.1	71.9	76.0	76.8	89.9	90.5	92.1	92.4	92.4	92.5	92.9	93.1	78.4	78.7
AUC	74.3	75.0	78.7	79.4	92.5	93.0	94.9	95.1	95.2	95.3	95.7	95.9	80.6	81.0
DM	68.2		74.1		86.4		90.4		91.0		91.3		79.3	
HD	18.7		14.3		9.3		8.1		7.9		7.5		15.1	

3.2 Segmentation Performance

Using the labeled datasets at different stages we train a UNet [11] for segmenting both lungs. The trained model is then evaluated on the separate set of 400 images from the NIH database on which we manually segment both lungs. The segmentation performance for FSL and different AL settings is summarized in Table 1 in terms of Dice Metric (DM) and Hausdorff Distance (HD). We observe that the segmentation performance reaches the same level as FSL at a fraction of the full dataset - in this case between $30-35\%$, which is similar for classification.

Figure 3 shows the segmentation results for different training models. When the number of training samples are less than 10% the segmentation performance is quite bad in the most challenging cases. However the performance improves steadily till it stabilizes at the 35% threshold.

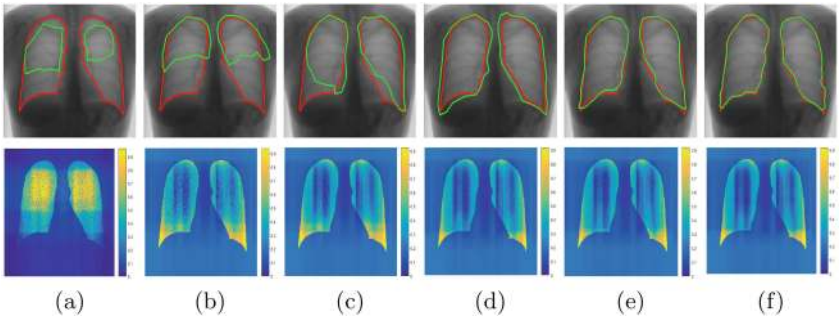

Fig. 3. Segmentation (top row) and uncertainty map (bottom row) results for different numbers of labeled examples in the training data (a) 5%; (b) 10%; (c) 20%; (d) 30%; (e) 35%; (f) 40%. Red contour is manual segmentation and green contour is the UNet generated segmentation.

3.3 Savings in Annotation Effort

Segmentation and classification results demonstrate that with the most informative samples, optimum performance can be achieved using a fraction of the dataset. This translates into significant savings in annotation cost as it reduces the number of images and pixels that need to be annotated by an expert. We calculate the number of pixels in the images that were part of the AL based training set at different stages for both classification and segmentation. At the point of optimum performance the number of annotated pixels in the training images is 33%. These numbers clearly suggest that using our AL framework can lead to savings of nearly 67% in terms of time and effort put in by the experts.

4 Conclusion

We have proposed a method to generate chest Xray images for active learning based model training by modifying the original masks of associated images. A generated image's informativeness is calculated using a bayesian neural network, and the most informative samples are added to the training set. These sequence of steps are continued till there is no additional information provided by the labeled samples. Our experiments demonstrate that, with about $33-35\%$ labeled samples we can achieve almost equal classification and segmentation performance as obtained when using the full dataset. This is made possible by selecting the most informative samples for training. Thus the model sees all the

informative samples first, and achieves optimal performance in fewer iterations. The performance of the proposed AL based model translates into significant savings in annotation effort and clinicians' time. In future work we aim to further investigate the realism of the generated images.

Acknowledgement. The authors acknowledge the support from SNSF project grant number 169607.

References

1. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: Proceedings of the International Conference on Machine Learning (2017)
2. van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal. **10**(1), 19–40 (2006)
3. Goodfellow, I., et al.: Generative adversarial nets. In: Proceedings of the NIPS, pp. 2672–2680 (2014)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the CVPR (2016)
5. Isola, P., Zhu, J., Zhou, T., Efros, A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
6. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (2017)
7. Li, X., Guo, Y.: Adaptive active learning for image classification. In: Proceedings of the CVPR (2013)
8. Mahapatra, D., Bozorgtabar, B., Hewavitharanage, S., Garnavi, R.: Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 382–390. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_44
9. Mahapatra, D., Schüffler, P.J., Tielbeek, J.A.W., Vos, F.M., Buhmann, J.M.: Semi-supervised and active learning for automatic segmentation of Crohn's disease. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 214–221. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_27
10. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.: Context encoders: feature learning by inpainting. In: Proceedings of the CVPR, pp. 2536–2544 (2016)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
13. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
14. Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L.: Cost-effective active learning for deep image classification. IEEE Trans. CSVT **27**(12), 2591–2600 (2017)

15. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the CVPR (2017)
16. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46