Robust Classification of Histology Images Exploiting Adversarial Auto Encoders

Nikhil Cherian Kurian*, Gurparkash Singh*, Poorvi Hebbar*, Shreekanya Kodate*, Swapnil Rane†, Amit Sethi*
*Indian Institute of Technology Bombay, Mumbai, India
†Tata Memorial Centre, Mumbai, India

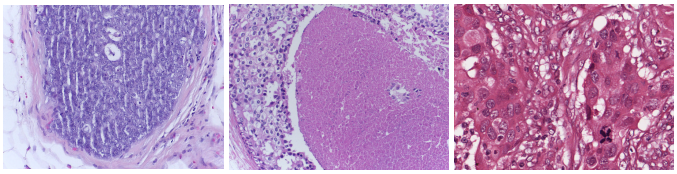
***Abstract*—Deep learning (DL) thrives on the availability of large numbers of high quality images with reliable labels. Due to the large size of whole slide images in digital pathology, patches of manageable size are often mined for use in DL models. These patches are often variable in quality, weakly supervised, individually less informative, and noisily labelled. To improve classification accuracy even with these noisy input and labels in histopathology, we propose a novel method for robust feature generation using an adversarial autoencoder (AAE) . We utilize the likelihood of the features in the latent space of AAE as a criterion to weigh the training samples. We propose different weighing schemes for our framework and test our methods on two publicly available histopathology datasets. We observe consistent improvement in AUC scores using our methods, and conclude that robust supervision strategies should be further explored for computational pathology.**

***Index Terms*—Deep learning, Histopathology, Robust-Supervision , Adversarial Autoencoder, Noisy label Classification**

I. Introduction

Supervised deep learning (DL) models have consistently shown promising results for automated image analysis in medical image analysis over the last eight years [1]–[3]. However, the success of DL models depends on the availability of large datasets of high quality and correctly labelled images for training. When the quality of the training images or the accuracy of their labels degrade, the accuracy of the DL models trained using them reduces drastically [4]. Consequently, for automated medical image analysis in general, and computational pathology in particular, medical experts on a research team need to carefully label, annotate, and curate whole slide images (WSIs) to prepare training and testing datasets. This process often involves precise annotations of regions of interest (ROIs) so that high quality and homogeneous patches (sub-images) of anatomical structures can be mined. These patches then inherit the same label for as the ROI from which these are mined. This data preparation process is time-consuming and expensive.

On the other hand, weakly supervised labels for WSIs (or large images, in general) are much easier to obtain by simply mining their associated electronic medical records (EMRs) for overall diagnosis. Such weak supervision disregards the heterogeneity in the quality and anatomical content of patches mined from a single slide. For example, image quality can vary with tissue preparation, staining, and slide preparation methods [5], as shown in Figure 1. Additionally, intra-tumoral heterogeneity is a natural phenomenon. For example, tumoral and benign structures occur side by side and there is spatial variations in disease grade or mutational landscape often in a single slide. Spatial heterogeneity in anatomy and quality combined with the gigapixel size of WSIs means that weakly supervised label propagation from the WSI to ts patches leads to mislabeling of a certain unknown proportion of the individual patches.



(a) (b) (c)

Fig. 1: Diversity in the quality of histopathology images: (a) Venetian blinds artifact due to improper cut, (b) tissue degradation due to improper fixation, and (c) over-staining of eosin dye.

Although weakly supervised learning techniques such as multiple instance learning (MIL) try to address these issues [6], the expressiveness of the generated features are sometimes not strong [7] from the classification perspective as these methods extracts aggregate bag level features consisting of multiple instances. Such bag formations from several instances can also reduce the number of training points available for supervision.

In this paper, we address the problem of learning robust models for image classification in the face of label noise and weak supervision. We use adversarial autoencoders to get sample-wise weights for each training image. We assume that the samples that can deteriorate the model training will fall into the lesser likelihood regions of the class-specific distribution priors. This assumption eliminates the need for additional optimization steps to calculate the sample-wise weights. We also explore different schemes that can be used to weigh variants of cross entropy loss function for robust supervision.

II. Related work

Previous attempts to prevent overfitting on noisy outliers includes curriculum learning, self-paced learning, and robust loss functions. In curriculum learning the model is trained gradually using easier samples first, similar to how human are taught [8]. Unsupervised measures, such as entropy of classification output, are used to calculate the hardness of the

training samples. Self-paced learning schemes incorporate label information by including the loss of a sample as a measure of hardness [9]. These two ideas have been extended in self-paced curriculum learning [10], self-paced boosting [11] and diversity-based self-paced learning [12]. Meta-learning based sample weighting has also been explored [13]. Loss functions that are not over-eager to fit (have high gradient) on the outliers have been demonstrated to be robust to sample noise, such as the L1 loss [14] and the generalized cross-entropy loss [15].

Adversarial autoencoders (AAE), which we use in this work, is an extension of regular autoencoder by inducing a prior distribution in the bottleneck layer [16]. Sampling from the prior distribution leads to a generative model that can be used for feature extraction [17] and anomaly detection [18].

Robust learning for histopathology image classification has been explored only in a few works. the importance of robust training schemes are explored in. A novel loss function and graph-based ensemble boosters to enhance the strength of training samples have been proposed [19]. Self-similarity between multiple patches has also been used to counter label noise [20].

III. Methodology

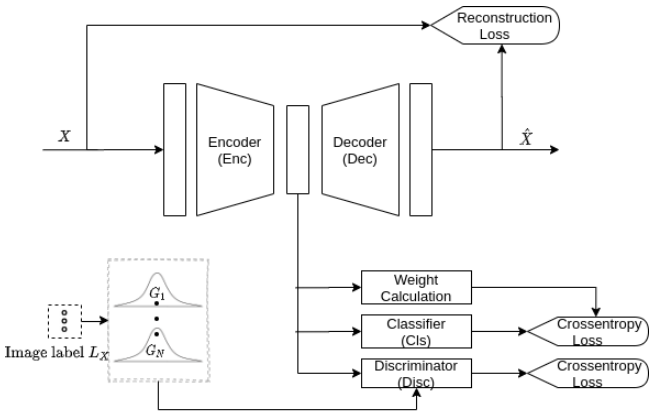


Fig. 2: Adversarial auto-encoder based architecture used for robust classification.

Our proposed robust supervision technique falls in the class of sample weighting schemes. Unlike other techniques that involve additional optimization steps or predefined curricula, we utilize the likelihood of the features of a sample in the latent space of AAE to derive its dynamic weight. We hypothesized that the less informative or the noisy samples will fall into the low likelihood regions of the regularized latent space.

As shown in figure 2, our model has an encoder block that acts as a feature generator for the task-specific classifier as well as for the AAE. The adversarial training for the generated features in the d-dimensional (d=32 in our experiments) latent space is performed with the aid of the discriminator block. The discriminator compares the features generated by the encoder with a random vector sampled from its corresponding class specific prior distribution, which we assume to be a d-dimensional Gaussian distribution. Since the task-specific classifier is also needs to be optimized, the encoder block in this adversarial task has to generate samples that are optimized both for the classifier as well as to fool the discriminator. Here the role of decoder block in our architecture is to ensure that all images belonging to a particular class are pulled towards a mean feature vector of its Gaussian prior. During the training phase, feature generator and discriminator will perform the following min-max game to generate the samples:

$$\arg\max_{Disc} \arg\min_{Enc,Dec,Cls} [C(X, L_X) + \lambda_1 R(X) - \lambda_2 D(X, P)]$$

where C is the classification loss, R is the reconstruction loss, and D is the discriminator loss. These losses are sample-wise weighted cross-entropy, mean square error, and cross entropy respectively in our scheme. Further, X is a training sample and L_X is its label, and P is a sample from the prior distribution. Hyperparameters λ_1 and λ_2 were decided based on validation.

When the discriminator reports low confidence in distinguishing a real Gaussian sample against the feature vector, the adversarial training is declared successful. Training the classifier separately on top of a well-trained AAE generator gave poor classifier performance because such a feature generator was agnostic to the classification task a priori.

We explored the use of following schemes for sample-wise weighting in the loss C to train the adversarial autoencoder based classifier (AAEC):

- **Binary weighting (BW):** The sample's class-specific likelihood is compared to a global threshold, which is a tunable hyperparameter, to decide on its inclusion or exclusion (binary weight).
- **Binary normalized weighting (BNW):** Binary weighting is computed separately within each training mini-batch by normalizing the likelihood within the batch and comparing that to a threshold.
- **Normalised weighting (NW):** Continupis weights are by normalizing the likelihood within each batch.

The binary weighting schemes described above are similar to curriculum and self paced learning frameworks that allows only easy samples to appear in the training phase based on the age (state of learning iterations) of the model. On the other hand, the NW scheme ensures that all samples are represented in training, albeit with different weights. We further extended BW and BNW schemes, choosing the best of the models as the initial weights and continue training on these without any explicit weighting. We call these set of schemes that continue their training from BW and BNW without any weights as **Binary Weighting-No Weighting (BWNW)** scheme and **Binary Normalised Weighting-No Weighting (BNWNW)** scheme respectively.

IV. Experiments

We used a single model architecture for an experiment where we artificially added label noise, and another where we worked with an unknown level of noise.

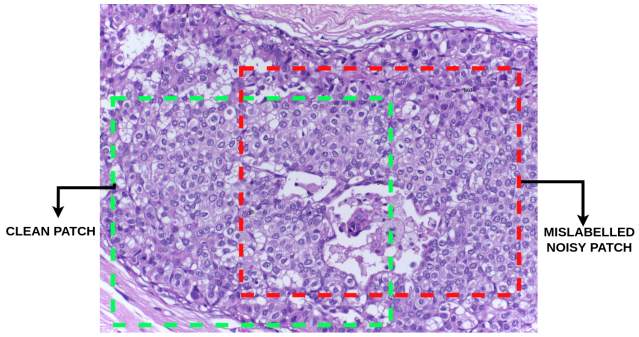


Fig. 3: Patch extraction from DCIS region can result in noisy samples (red box) where the basement membrane, which is its tell-tale feature, is not visible, or a good sample (green box) where it is visible.

A. Model Architecture

Table I shows the model architecture used in our experiments. In both of our experiments we used the value of $N = 2$ (number of classes), with the centres of the two prior multivariate Gaussian distributions located at the opposite ends of a d-dimensional hypersphere in the AAE latent space.

Encoder			
Layer Type	Kernels	Dropout	Activation
Convolutional	4	0.5	Leaky-ReLU
Convolutional	8	0.5	Leaky-ReLU
Convolutional	16	None	None

Decoder			
Layer Type	Kernels	Dropout	Activation
Input	16	–	–
Trans.-Conv	8	0.3	Leaky-ReLU
Trans.-Conv	4	0.3	Leaky-ReLU
Trans.-Conv	36	None	None

Classifier and Discriminator			
Layer Type	Nodes	Dropout	Activation
Input	–	–	–
Fully-Connected	32	0.5	Leaky-ReLU
Fully-Connected	16	0.5	Leaky-ReLU
Fully-Connected	2	None	Softmax

TABLE I: Model architecture used in our experiments

For convolution and transpose-convolution (upsampling) layers, we used kernels of size 5×5 with a stride of 3. Batch-normalization was used after each layer of all four segments of the model. The activation function used throughout the network is leaky-ReLU with a slope of 0.2 for the negative inputs. We used Adam optimizer with a learn rate of 0.001. Data imbalance in our experiments was accounted using a proportionate weighted sampling for each mini-batch updates.

B. Tumor versus non-tumor

For our first set of experiments, we use the BreakHis dataset [21], which is available at different magnifications. We took 400x magnification consisting of 1,450 images divided between tumor and non-tumor images. To test the robustness of our method, we synthetically added label noise by randomly flipping the labels of a pre-determined percent of training samples. We varied this percentage from 0 to 20 in steps of 5 and we show the AUC values on a clean (without label noise) held-out dataset in Table II and plot the ROC curves for 20% noise in Figure 4. We further added color jitter based data-augmentation on the fly to simulate the staining variations found in the real world. We compared our models with the proposed weighting schemes by training a ResNet18 network (Transfer learning on pretrained model) that does not use any sample-wise weighting.

C. DCIS versus IDC

In this experiment we used ICIAR 2018 Grand Challenge dataset called Breast Cancer Histology (BACH) [22]. We took on the challenging problem of classifying between ductal carcinoma in situ (DCIS) and invasive ductal carcinoma (IDC). The distinction between these classes is the presence (in DCIS) or the absence (in IDC) of a basement membrane around tumorous cells that otherwise look quite similar.

To generate the training data, we sampled patches of size 512x512 from the original images of the size of 2048x1536. The sampling resulted in the dataset that contained around 1540 patches. Such sub-sampling presents a more realistic scenario than the previous experiment for medical image classification, although we do not have much control over the percent of labels that are noisy. For instance, a patch sampled from a DCIS image may not include a basement membrane, which makes it indistinguishable from an IDC sample. Further, some samples may contain no tumor region at all - neither DCIS nor IDC. Additionally, some patches may have other artifacts as shown in Figure 1.

Once again, we evaluated the robustness of our models on clean held-out samples by manually curating the test cases before training the models. We show the AUC values in Table III and the ROC curves are shown in Figure 4.

V. Results and Conclusion

The AUC values for both of our experiments are shown in table II and III. The worst case ROCs curves for two experiments are shown in figures4. From the results of the tumor versus non-tumor experiment, we observe that the performance of a regular CNN network worsens drastically with increasing noise. Further the DCIS versus IDC experiment shows that the robust weighting strategies perform much better than a conventional CNNs that implicitly overfits on noisy samples. The weighing strategies we found to be most robust in both the experiments. The experiments support the direction of further developing strategies for more robust histopathology, especially when the quality in image or strong supervision cannot be ensured.

Acknowledgment

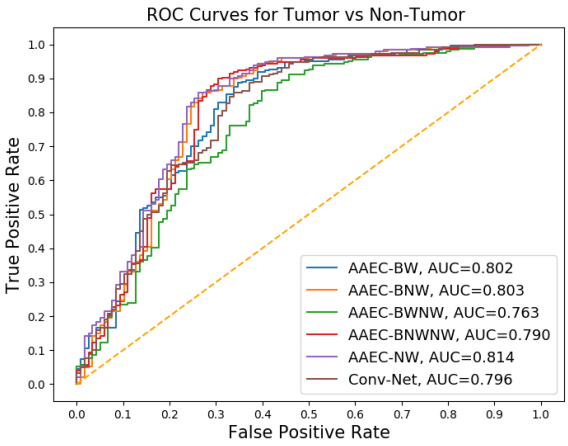
Authors would like to thank Nvidia Corporation for donation of GPUs used for this research.

Training Scheme	ROC-AUC scores for various label noise levels				
	0%	5%	10%	15%	20%
AAEC-BW	0.823	0.819	0.819	0.805	0.802
AAEC-BNW	0.836	0.828	**0.827**	0.808	0.803
AAEC-BWNW	0.815	0.802	0.802	0.806	0.763
AAEC-BNWNW	0.810	0.808	0.808	0.797	0.790
AAEC-NW	0.838	0.829	0.826	**0.821**	**0.814**
Conv-Net	**0.903**	**0.851**	0.802	0.790	0.796

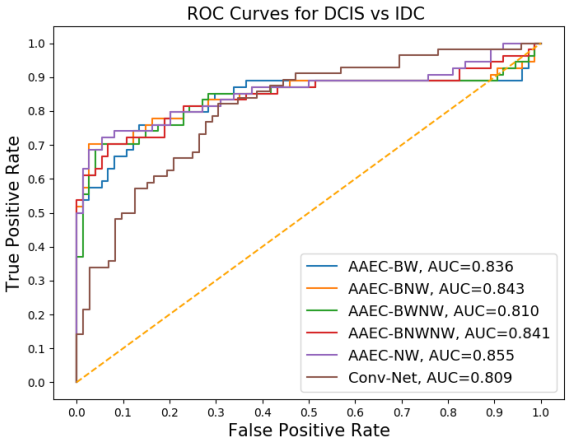
TABLE II: ROC-AUC scores for held out data for various label noise levels for tumor versus non-tumor

Training Scheme	AUC-ROC scores
AAEC-BW	0.836
AAEC-BNW	0.843
AAEC-BWNW	0.810
AAEC-BNWNW	0.841
AAEC-NW	**0.855**
Conv-Net	0.809

TABLE III: ROC-AUC scores for held out data for DCIS versus IDC.



(a) Tumor v. Non-tumor (20% noise)



(b) DCIS v. IDC

Fig. 4: ROC curves of the experiments

REFERENCES

[1] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," 2016.

[2] S. Jha and E. J. Topol, "Adapting to artificial intelligence: radiologists and pathologists as information specialists," *Jama*, vol. 316, no. 22, pp. 2353–2354, 2016.

[3] K. Yasaka, H. Akai, A. Kunimatsu, S. Kiryu, and O. Abe, "Deep learning with convolutional neural network in radiology," *Japanese journal of radiology*, vol. 36, no. 4, pp. 257–272, 2018.

[4] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: exploring techniques and remedies in medical image analysis," *Medical Image Analysis*, vol. 65, p. 101759, 2020.

[5] S. A. Taqi, S. A. Sami, L. B. Sami, and S. A. Zaki, "A review of artifacts in histopathology," *Journal of oral and maxillofacial pathology: JOMFP*, vol. 22, no. 2, p. 279, 2018.

[6] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

[7] M. Li, L. Wu, A. Wiliem, K. Zhao, T. Zhang, and B. Lovell, "Deep instance-level hard negative mining model for histopathology images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 514–522.

[8] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[9] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in neural information processing systems*, 2010, pp. 1189–1197.

[10] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning." in *AAAI*, vol. 2, no. 5.4, 2015, p. 6.

[11] T. Pi, X. Li, Z. Zhang, D. Meng, F. Wu, J. Xiao, and Y. Zhuang, "Self-paced boost learning for classification." in *IJCAI*, 2016, pp. 1932–1938.

[12] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Advances in Neural Information Processing Systems*, 2014, pp. 2078–2086.

[13] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," *arXiv preprint arXiv:1803.09050*, 2018.

[14] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," *arXiv preprint arXiv:1712.09482*, 2017.

[15] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in neural information processing systems*, 2018, pp. 8778–8788.

[16] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[17] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *arXiv preprint arXiv:1806.02146*, 2018.

[18] N. Li and F. Chang, "Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder," *Neurocomputing*, vol. 369, pp. 92–105, 2019.

[19] X. Shi, H. Su, F. Xing, Y. Liang, G. Qu, and L. Yang, "Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis," *Medical Image Analysis*, vol. 60, p. 101624, 2020.

[20] H.-T. Cheng, C.-F. Yeh, P.-C. Kuo, A. Wei, K.-C. Liu, M.-C. Ko, K.-H. Chao, Y.-C. Peng, and T.-L. Liu, "Self-similarity student for partial label histopathology image segmentation," *arXiv preprint arXiv:2007.09610*, 2020.

[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[22] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan *et al.*, "Bach: Grand challenge on breast cancer histology images," *Medical image analysis*, vol. 56, pp. 122–139, 2019.