TRAINING A NEURAL NETWORK BASED ON UNRELIABLE HUMAN ANNOTATION OF MEDICAL IMAGES

Yair Dgani[1] *Hayit Greenspan*[2] *Jacob Goldberger*[1]

[1]Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel
[2]Department of Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel.

ABSTRACT

Building classification models from clinical data often requires human experts for example labeling. However, it is difficult to obtain a perfect set of labels due to the complexity of the medical data and the large variability between experts. In this study we present a neural-network training strategy that is more robust to unreliable labeling by explicitly modeling the label noise as part of the network architecture. Our method is demonstrated on breast microcalcifications classification into benign and malignant categories, given multi-view mammograms. We show that the proposed training procedure outperforms standard training methods that ignore the existence of label noise.

Index Terms— Mammography, Microcalcifications, deep-learning, noisy-labels, robust training.

1. INTRODUCTION

Manual annotation of objects in scenery images (cat, dog, bicycle etc.) is usually a very simple task. In medical imaging tasks, annotations are made by radiologists with expert knowledge of the data and the task. However, it is difficult to obtain a perfect set of labels to serve as a gold standard, due to the complexity of the medical data and the large interrater variability. Two expert physicians may disagree on a complex patient case due to differences in clinical training, prior experience, and understanding of the disease. Annotation differences can also arise due to the limited amount of time available in annotating large number of cases. One of the main challenges facing the medical imaging domain is therefore the need to cope with unreliable annotated samples. This is especially important when employing supervised machine learning algorithms that require labeled data as training examples.

One strategy to cope with unreliable annotation is to collect labeling from multiple experts. In this case we can assess the level of expertise of each annotator as part of the training procedure [1]. However, collecting medical data annotation is a time consuming and expensive procedure that requires the collaboration of researchers and radiologists - thus collecting annotation form several experts is not always feasible. In this study we explore deep learning training strategies, based on a single annotation for each object, which are robust to unreliable labeling.

There are not many studies that have attempted to address the problem of training deep neural networks (DNN) algorithms with unreliable labels [1][2][3][4]. Sukhbaatar et al. [4] suggested adding a regularized linear layer on the top of the softmax layer, and made strong assumptions in order to prove that the proposed noisy layer can be viewed as the transition matrix between the true and observed data labels. Larsen et al. [3] simplified the noise model by assuming a single noise parameter that can be estimated by performing a cross validation procedure. Goldberger and Ben-Reuven [5] recently suggested a training procedure based on adding another softmax layer to the network, that uses the output of the last hidden layer of the network to predict the probability of the label being flipped.

All previous methods for training based on noisy labels have been demonstrated either on toy datasets (e.g. MNIST) or simple object classification datasets (e.g. CIFAR-100). Medical imaging tasks are difficult even if the labels are perfect since there is not always a clear indication of the pathology from the image. Hence, it is not clear whether robust training methods are relevant in medical diagnosis tasks. We are not aware of any study that investigated the applicability of label-noise robust training in medical imaging. In this study we follow the line of research [5] that was found to be

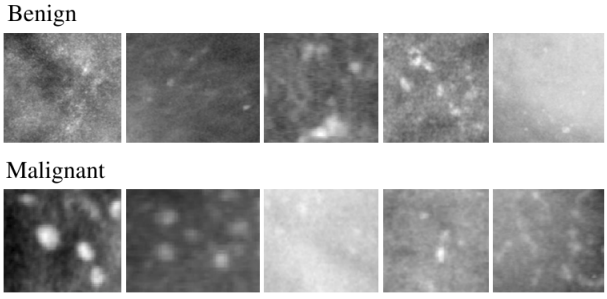

Fig. 1: Examples of benign and malignant MC clusters (from the DDSM dataset). ROI sizes range from $2mm \times 2mm$ to $6mm \times 6mm$.

superior to alternative strategies [6] and check its applicability for medical data.

In the current study we demonstrate the applicability of the proposed training method to the classification of breast microcalcifications. We address the problem of multi-view mammography analysis in the presence of microcalcifications (MC). In this case of computer-aided diagnosis (CAD) the task is to build a classifier to predict whether a suspicious region on a medical image is malignant or benign. In order to train such a classifier, a set of images is collected from hospitals. The actual gold standard (whether it is cancer or not) can be obtained from biopsies, but since it is an expensive and an invasive process, often CAD systems are built from labels assigned by a radiologist that visually examines the medical images and provides a subjective (possibly noisy) version of the gold standard. In this study we investigate the possibility of training a benign vs. malignant classifier based only on manual annotation without having gold standard biopsy results.

2. TRAINING A DEEP NEURAL NETWORK WITH A NOISY CHANNEL

Our goal is to train a standard neural-network classifier using training data with unreliable manual annotation. We consider the correct unknown labels as hidden random variables. We add another component to the network, denoted by noisy channel, which models the stochastic relation between the correct label and the observed noisy label. The parameters of the noisy channel are learned as part of the neural network training. In this section we first describe the probabilistic framework we use to model the label noise and then derive a training algorithm that simultaneously finds a classifier and learns the parameters of the noise channel. At test time we want to predict the true labels. Hence, we remove the noisy-channel component.

Assume we are given a k-class classification problem with labels denoted by $1, ..., k$. In a neural net model with parameter-set w, the probability of input x being labeled as i is:

$$p(y = i|x; w) = \frac{\exp(w_{oi}^{\top} h(x) + b_{oi})}{\sum_j \exp(w_{oj}^{\top} h(x) + b_{oj})} \qquad (1)$$

where we denote the non-linear function applied to the input x by $h = h(x)$, and w_o, b_o are the parameters of the output soft-max layer.

We further assume that in the training process we cannot directly observe label y. Instead we can only observe a noisy version of it, denoted by z. In our case y is the correct medical diagnosis and the label z is the manual annotation provided by an expert. We assume here a simplified noise model where the noisy label is a stochastic function of the true label. Formally, the noise model is defined by a parameter-set θ such that $\theta(i, j) = p(z = j|y = i)$ is the probability of observing label j given that the true label is i. In our case the task is to build a classifier to predict whether a suspicious region in a medical image is malignant or benign. The actual gold standard whether it is cancer or not (denoted by y), is obtained from biopsies. The label assigned by a radiologist that visually examines the medical images is denoted by z.

The combined neural-net model with noisy labels, therefore, is:

$$p(z = j|x; w, \theta) = \sum_{i=1}^{k} \theta(i, j) p(y = i|x, w). \qquad (2)$$

The model is illustrated in Figure 2.

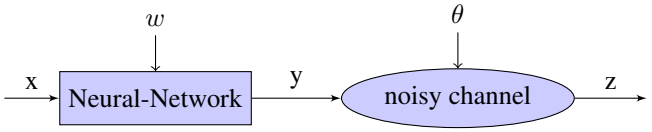

Fig. 2: The network combined with a noisy channel for producing estimation of the observed noisy labels.

Assume we are given n images $x_1, ..., x_n$ with corresponding noisy labels $z_1, ..., z_n$ which are viewed as noisy versions of the hidden gold standard labels $y_1, ..., y_n$. The log-likelihood function of the model parameters is:

$$L(w, \theta) = \sum_t \log p(z_t|x_t; \theta, w) = \qquad (3)$$

$$\sum_t \log \sum_i (p(z_t|y_t = i; \theta) p(y_t = i|x_t; w)).$$

The goal of the training procedure is to find the noise parameter θ and the neural-net set of parameters w that maximizes the likelihood function. Since the random variables $y_1, ..., y_n$ are hidden, to solve this maximization problem a standard approach is to apply the Expectation-Maximization (EM) algorithm [7]. The EM algorithm, however, is a greedy optimization procedure that is notorious for getting stuck in local optima. In most EM applications there is a closed-form solution for the optimization performed at the M-step. Here, since we need to retrain the DNN at each M-step, even a monotonic improvement of the likelihood is not guaranteed. Another problem of iterating between EM-steps and neural network training is that it requires training a neural network in each iteration of the EM and this procedure does not scale well.

The approach we suggest here is to directly use the likelihood function (3) as an objective function for training a DNN. We can view the parameters of the noisy channel, θ, as a linear function of the soft-max output vector. Let P_y and P_z be the network soft-decision distribution of the correct and noisy label respectively. Then Eq. (2) can be written in the following vectorial form:

$$P_z = \theta^{\top} \cdot P_y.$$

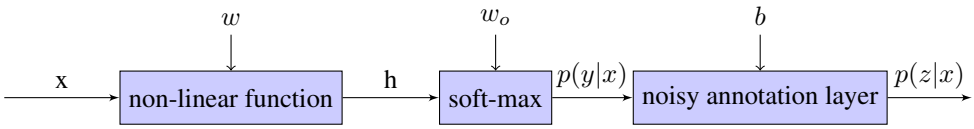

Fig. 3: An illustration of the noisy-label neural network architecture for the training phase.

To enforce the constraint that θ is a stochastic matrix, i.e. $\theta(i,j) > 0$ and $\sum_j \theta(i,j) = 1$, we use the following standard reparametrization:

$$\theta(i,j) = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}.$$

The network architecture model is illustrated in Figure 3.

There is a degree of freedom in the proposed model. We can apply a permutation on the softmax distribution of the correct label and the inverse permutation on the transition from the correct labels to the observed labels. To avoid this problem we first train the network without the additional noise layer. We then use the parameters of the original network to initialize the parameters of the extended network that contains the noise layer.

3. EXPERIMENTAL SETUP

3.1. Dataset and features

A screening mammographic examination usually consists of four images, corresponding to each breast scanned in two views: the mediolateral oblique (MLO) view and the craniocaudal (CC) view. The MLO projection is taken in a 45° angle and shows part of the pectoral muscle. The CC projection is a top-down view of the breast. Both views are included in the diagnostic procedure. When reading mammograms, radiologists judge whether or not a malignant lesion is present by examining both views and breasts. In an expert diagnosis procedure, the expert looks at each of the views separately, and delivers one final assessment. When a radiologist does not observe a lesion in both views this can influence interpretation and decision making. This study is based on the DDSM dataset [8] which provides large number of annotated mammograms with a biopsy-proven diagnosis.

The contours of the lesions are provided by a chain code which we used to extract irregular shaped ROIs. We extracted ROIs that contained clusters of MCs for which a proven pathology was found. We chose patients in the DDSM dataset that had both CC and MLO views in order to test our model. Our dataset was comprised of 1410 clusters (705 of CC, and 705 of MLO), of which 372 were benign and 333 were malignant. Feature vectors were extracted from the CC and MLO views, respectively. Following [9], the features were extracted from the Curvelet coefficients at intermediate scales (in our study, two scales), and included the four features mentioned in [10] for each scale, with three additional features: entropy, skewness and kurtosis. Overall, each extracted ROI was represented by 14 features. Many other texture features that can be used for mammography analysis have been reported in the literature, e.g. GLCM [11], GLRLM [12], Gabor filters [13] and features that are based on the wavelet transform. Using the Curvelet features we obtained the best results. Due to lack of space and since this is not the focus of this work, we do not describe here classification results based on the alternative features.

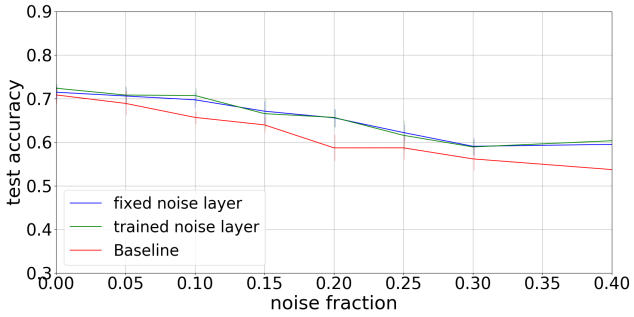

Fig. 4: Test classification accuracy results on the fatty part of the DDSM dataset as a function of the noise level.

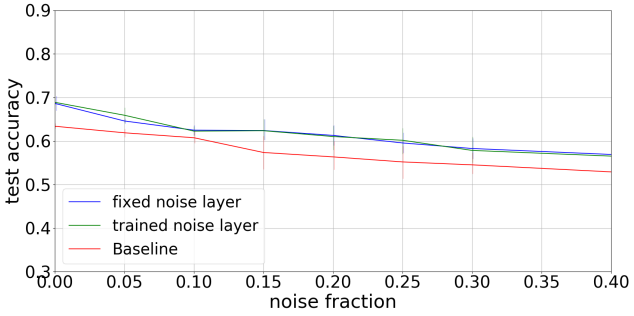

Fig. 5: Test classification accuracy results on the dense part of the DDSM dataset as a function of the noise level.

3.2. Training procedure

Using the feature described above, the size of the input feature set is 28 (14 features for each view). We used a two hidden layer DNN comprised of 24 neurons each. Overall, the number of parameters (linear coefficients and bias terms) is $29 \times 24 + 25 \times 24 + 25 \times 2$. The non-linear activation we used was ReLU and we used dropout with parameter 0.5.

We trained the network using the Adam optimizer with default parameters, which we found to converge more quickly and effectively than stochastic gradient descent.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate the robustness of deep learning to training data with noisy labels with and without explicit noise modeling. We show results on the DDSM dataset described above. The biopsy results that define if the abnormities were benign or malignant, are viewed as the correct labels. To imitate a manual annotation by a human expert we generated noisy data from clean data by randomly changing some of the labels. We flipped each binary label with probability p. The labels of the test data remained, of course, unchanged to validate and compare our method to the regular approach. The results were computed using 10-fold cross validation. In this experiment set-up there is a complete isolation of the test set from the train set. Each fold was only used for testing and never for training. The noise-free classification results are similar to the ones obtained on the same dataset in previous studies [14]. We implemented three methods: The first is a baseline that is based on standard architecture (Sec. 3.1). The second approach is the one proposed in this study which includes another linear layer that models the noise. We also applied a variant where the noise level is known and kept fixed. We separated the mammograms into two different tissue-density categories and studied them individually: fatty tissues (ratings 1 and 2), and dense tissues (ratings 3 and 4). Figure 4 depicts the comparative test error results as a function of the fractions of noise for fatty tissues and Figure 5 depicts the results for dense tissues. The results show that methods that are explicitly aware of the noise in the labels are better than the baseline which is the standard training approach. We also show that the results in case the noise parameters are learned as part of training are comparable to the situation where the correct noise level is used. It is interesting to observe that in the case of dense tissue our method outperformed baseline even without injecting noise. The noise probability that was learned in that case was 0.05. This can be explained by the gap between the biopsy and the indication of the pathology from the image.

5. CONCLUSION

In this paper we addressed the problem of training a DNN classifier based on medical images with unreliable manual annotation. We proposed a method that is aware of the fact that the training labels are noisy. The learning of the noise level is done as part of the DNN training. We demonstrated the performance of the proposed method on the task of classification breast microcalcifications into benign and malignant given multi-view mammograms. We showed that the proposed training yields improved performance. In the future we plan to investigate the applicability of proposed method to other medical imaging tasks where the ground truth labeling is provided by manual annotation.

6. REFERENCES

[1] A. J. Bekker and J. Goldberger, "Training deep neural-networks based on unreliable labels," in *Int Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] V. Minh and G. Hinton, "Learning to label aerial images from noisy data," in *Int. Conf. on Machine Learning (ICML)*, 2012.

[3] J. Larsen, L. Nonboe, M. Hintz-Madsen, and K. L. Hansen, "Design of robust neural network classifiers," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1998.

[4] S. Sukhbaatar and R. Fergus, "Learning from noisy labels with deep neural networks," in *arXiv preprint arXiv:1406.2080*, 2014.

[5] J. Goldberger and E. Ben-Reuven, "Training deep neural networks using a noise adaptation layer," in *Int. Conference on Learning Representations (ICLR)*, 2017.

[6] E. Malach and S. Shalev-Shwartz, "Decoupling when to update from how to update," in *Neural Information Processing Systems (NIPS)*, 2017.

[7] A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[8] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," *Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, ed*, pp. 212–218, Medical Physics Publishing, 2001.

[9] A. J. Bekker, M. Shalhon, H. Greenspan, and J. Goldberger, "Multi-view probabilistic classification of breast microcalcifications," *IEEE Trans. Medical Imaging*, vol. 35:2, pp. 645–653, 2016.

[10] Y. Shang, Y. Diao, and C. Li, "Rotation invariant texture classification algorithm based on curvelet transform and SVM," in *Int. Conf. on Machine Learning and Cybernetics*, 2008, vol. 5, pp. 3032–3036.

[11] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[12] X. Tang, "Texture information in run-length matrices," *IEEE Transactions on Image Processing*, vol. 7, no. 11, pp. 1602–1609, 1998.

[13] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[14] I. I. Andreadis, G. M. Spyrou, and K. S. Nikita, "A comparative study of image features for classification of breast microcalcifications," *Meas Sci Technol*, vol. 22, no. 11, pp. 114005–114013, 2011.