BreastNet: Entropy-Regularized Transferable Multi-task Learning for Classification with Limited Breast Data

Jialin Shi[1*], Ji Wu[1], Ping Lv[1], Jiajia Guo[2]

[1] Department of Electrical Engineering, Tsinghua University, Beijing, China.
[2] The people's Hospital of Peking University, Beijing, China.

* Corresponding author. Tel.:+86-135 0442 6468; email: shi-jl16@mails.tsinghua.edu.cn
Manuscript submitted June 28, 2018; accepted July 20, 2018.
doi: 10.17706/ijbbb.2019.9.1.20-26

Abstract: We describe a framework to automatically separate malignant from benign breast lesions using limited breast ultrasound data. The main uniqueness of this framework includes: (1) in terms of the unique shape features of breast lesions, two types of image patches are designed to fine-tune pre-trained models, aiming to characterize the overall appearance and heterogeneity in shapes of breast lesions. (2) taking the BI-RADS regression task as an auxiliary task, a multi-task architecture is proposed to improve the accuracy of classification. (3) instead of prevalent cross-entropy loss, we introduce training with confusion by means of regularizing prediction entropy to prevent overfitting. Extensive experimental results on small-scale breast ultrasound dataset corroborate that the proposed framework is superior to the state-of-the-art approaches in breast lesions classification with limited data. Besides, we provide detailed analysis of the choice of regularizing parameter and visual evidence that introduction of confusion leads to increase in feature generalization.

Key words: Breast ultrasound classification, multi-task learning, regularizing prediction entropy, transfer learning.

1. Introduction

Breast cancer is the most common cancer in women worldwide and the second leading cause of female cancer deaths [1]. Early diagnosis and treatment are the most effective means to improve survival. Though mammography is the primary imaging modality for screening, medical breast ultrasound (BUS) screening has also been demonstrated to be an effective way especially for those with dense breasts [2]. While extensive breast cancer works focus on mammogram-based studies, an automatic robust BUS image analysis tool is highly demanded in clinical practice. Because of typically small amount of data in medical image domain, investigating the framework for benign-malignant lesions classification with limited BUS data is meaningful and inevitable.

Recently, deep learning methods have been applied for medical images analysis. However, works based on deep networks to perform BUS tasks are still scarce according to Table 6 from [3]. Seokmin Han et al. studied pre-training networks on gray natural images and margin augmentation for classification of breast lesions [4]. Mohamed Abdel-Nasser et al. used texture analysis and super-resolution methods for breast tumor classification [2]. Lu Bing et al. proposed a method based on sparse representation for breast image classification under the framework of multi-instance learning [5]. These studies either focus on hand-crafted features, or simply focus on predicting categorical variables in single classification task.

Integrating the unique characteristics of breast lesions to develop multi-task methods to promote classification accuracy has not been fully explored.

For large-scale classification tasks, strongly discriminative learning using the cross-entropy loss is successful in part due to the large amounts of data, which enables networks to learn generalized discriminatory features. Cross-entropy loss forces the network to learn features that distinguish two images with a high confidence to minimize training error. However, despite using large datasets, over confidence is prone to overfitting. This effect may be pronounced for small-scale medical datasets. Based on the hypothesis, cross-entropy loss formulation may not be ideal. We expect the introduction of confusion in output logit activations to enable the network to learn slightly less distinctive features. To the best of our knowledge, for the existing works of medical image analysis, the regularization techniques for preventing overfitting include early stopping, L1/L2 regularization, dropout, batch normalization, etc. However, these techniques act on either the hidden activations or weights of a neural network, regularizing the output distribution based on entropic confusion has not been explored in medical image domain.

In this paper, we propose a deep shape-assisted entropy-regularized transferable multi-task (DSETM) framework for benign-malignant classification with limited BUS data. The main uniqueness of this framework includes: (1) in terms of unique shape features of breast lesions, two types of image patches are designed to fine-tune two pre-trained models, aiming to intergrade the overall appearance (OA) and heterogeneity in shapes of breast lesions (HS). (2) taking the Breast Imaging-Reporting and Data System (BI-RADS) regression task as an auxiliary task, a Convolutional Neural Network (CNN)-based multi-task architecture is proposed to improve the accuracy of classification. (3) instead of prevalent cross-entropy loss, we introduce regularizing prediction entropy to learn more generalizable feature representations to prevent overfitting. Extensive experimental results on a small-scale BUS dataset corroborated that our method achieved a superior accuracy, outperforming other methods by a significant margin and demonstrating competitive capability.

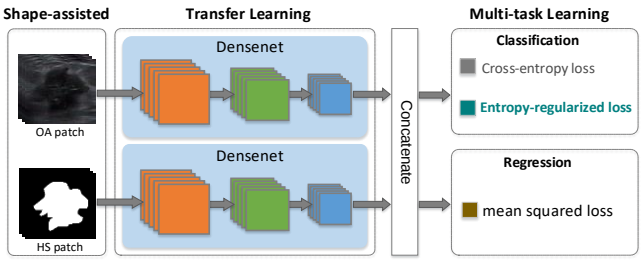


Fig. 1. The proposed DSETM framework. (1) Two stream characterize OA and HS; (2) Multi-task includes classification task (labels proved by biopsy) and regression task (labels proved by BI-RADS); (3) Entropy-regularized loss represents training with confusion.

2. Deep Shape-Assisted Entropy-Regularized Transferable Multi-task Framework

The framework for our proposed end-to-end deep shape-assisted entropy-regularized transferable multi-task network to differentiate between malignant and benign breast lesions is shown in Fig. 1. Check section 2.1 for more details.

2.1. Shape-Assisted Transferable Multi-task Learning Framework

Why shape-assisted? The most remarkable shape features of a benign mass image are oval or round shape, circumscribed margins. The most remarkable shape features of a malignant mass image are speculated margins, irregular shape [5]. The heterogeneity in shapes will gives us a useful guide for BUS

analysis. We extracting the region of interest (ROI) for preprocessing and data augmentation. To characterize the lesion's OA, a rectangle ROI encapsulating the lesion is identified. In order to describe the lesion's HS, non-lesion pixels are set to 0 and lesion pixels are set to 255. Using the OA and HS features means the combination of entire image features and unique shape features. In other words, the method will use more detailed information and improve the performance.

Why multi-task? Actually, the tasks of benign-malignant breast tumors classification and BI-RADS clinical score regression may be highly associated, since they aim to predict semantically similar targets. Jointly learning these two tasks can utilize the intrinsic useful correlation information among categorical and clinical variables to promote the learning performance. The use of multi-task where one architecture serves to address multiple problems can reduce the number of free parameters, further control overfitting to some extend [6]. The biopsy-proven data are used for primary classification task, while BI-RADS data are used for auxiliary regression task. From a clinical perspective, BI-RADS corresponds to radiologists' diagnoses depending on their clinical experiences.

Previous studies have evidenced the efficacy of transfer learning from natural image domain to medical image domain on image classification tasks [7]. We prefer the Densely Connected Convolutional Networks (DenseNet) model [8] that has been pre-trained on the ImageNet dataset. The DSETM architecture (see Fig. 1) adopts two-stream input data, where one stream is corresponding to OA and the other is used to characterize HS of breast lesions. DenseNet-121 which is removed the last fully connected layer and corresponding weights is used for each channel to perform transfer learning. Two types of image patches are designed to fine-tune two pre-trained DenseNet-121 models. To model the global structural information of breast lesions, we then concatenate the features from penultimate layer (the global average pooling layer) of each pre-trained DenseNet-121 network. Because there are no parameters to optimize in the global average pooling thus overfitting is avoided at this layer. According, concatenating the global average pooling layer instead of fully- connected layer is adopted. The output contains benign-malignant class labels based on biopsy-proven data for the primary binary classification task and clinical scores based on BI-RADS data for the auxiliary regression task. The proposed framework can also be mathematically described as follows.

Let $\chi = \{x_n\}_{n=1}^{N}$ denote the training set, with the element x_n representing the n-th subject. For the primary classification task, denote the labels of $C (C = 2$ categories as $\boldsymbol{y}^c = \{y_n^c\}_{n=1}^{N} (c = 1, \cdots C)$. For the auxiliary regression task, denote the clinical scores as $\boldsymbol{z} = \{z_n\}_{n=1}^{N}$. The aim of the proposed framework is to learn a non-linear mapping $\chi \to (\{y\}, \{z\})$ from the input space to both spaces of the class labels and the clinical scores. For the multi-task, the loss function is as follows:

$$L_{loss} = -\frac{1}{N}\sum_{x_n \in \chi}\sum_{c=1}^{C} 1\{y_n^c = c\} \log\left(P\left(y_n^c = c \mid x_n; W\right)\right) + \frac{1}{N}\sum_{x_n \in \chi}\left(z_n - \overline{z_n}\right)^2 \tag{1}$$

where the first term is the prevalent cross-entropy loss and the second one is the mean squared loss for regression to evaluate the difference between the clinical score $\overline{z_n}$ and the ground truth z_n. Note that $1\{\cdot\}$ is an indicator function, with $1\{\cdot\} = 1$ if $\{\cdot\}$ is true; and 0, otherwise. $P(y_n^c = c|x_n; W)$ indicates the conditional probability of the subject x_n using the network coefficients W.

2.2. Regularizing Prediction Entropy

In this part, we investigate regularizing prediction entropy for limiting classifier overconfidence, which will introduce confusion at the fundamental level of neural network training. While we view the knowledge of a model as the conditional distribution, confident predictions correspond to output distributions that have low entropy. The confidence penalty constitutes a regularization term that prevents these peaked

distributions, leading to better generalization. In other words, limiting overconfidence will increase confusion of output conditional probability distribution, namely increasing the entropy.

If we assume the distribution of classes in training and prediction phases to be uniform, we can measure the deviation of the output probability from a random classifier as a measure of prediction certainty, and limit it in order to introduce confusion in output activations [9]. To measure this deviation, we consider the KL divergence $D_{KL}(\mathrm{p}_w(y|x)||\mathrm{p}_w(y|u))$ where u is the uniform vector with norm 1. We see that:

$$\begin{aligned} D_{KL}\left(p_w(y|x)\,\|\,p_w(y|u)\right) &= \sum\nolimits_c p_w(y_c|x)\times\log\left(\frac{p_w(y_c|x)}{N^{-1}}\right) \\ &= \log N + \sum\nolimits_c p_w(y_c|x)\times\log\left(p_w(y_c|x)\right) \\ &= \log N - H\left(p_w(y|x)\right) \end{aligned} \tag{2}$$

where $\mathrm{H}(p_w(y|x)$ is the Shannon entropy. Hence, minimizing certainty through$D_{KL}(p_w(y|x)||p_w(y|u))$is equivalent to maximizing the Shannon entropy. We formulate the final objective for a batch of b samples as:

$$L_{regularization\text{-}loss} = L_{loss}\left(p_w(y|x)\right) - \frac{a}{b}H\left(p_w(y|x)\right) \tag{3}$$

where L_{loss} denotes the prevalent cross-entropy loss for classification task or the multi-task loss. α is the regularization parameter. The new loss function promotes learning a classifier with a maximum entropy output distribution.

3. Experiments and Analysis

A biopsy-proven BUS dataset is used in this study, including 150 benign masses and 162 malignant masses. All the ultrasound images are annotated by experts with more than 15 years of clinical experience. All breast lesions are labeled by BI-RADS and proved by biopsy. We trained the network in a 5-fold cross-validation setting and repeated the experiments 10 times. The performance is reported as the average of testing results. In our BUS dataset, the BI-RADS have 7 categories (i.e. 2, 3, 4A, 4B, 4C, 5, 6). Suppose suspicious degree of malignancy of samples in each category follows a uniform distribution. We adopt the mean of distribution of every category as the regression scores. After the data has been dealt with appropriately, we will make it public.

To evaluate the contributions of the proposed three strategies (shape-assisted, multi-task and entropy-regularized), we compare the DSETM framework with its six variants. Six variants include (1) deep transferable single-task single-stream learning (DSSL) using OA information; (2) deep transferable single-task single-stream learning with regularizing output entropy (E-DSSL);(3) deep transferable single-task two-stream learning (DSTL) using both OA and HS details; (4) deep transferable single-task two-stream learning with regularizing output entropy (E-DSTL); (5) deep transferable multi-task two-stream learning (DMTL); (6) deep transferable multi-task two-stream learning with regularizing output entropy (E-DMTL). The networks are trained using stochastic gradient descent optimizer with a learning rate of 10^{-3}, the momentum of 0.9 and the learning rate decays by 10^{-6} every epoch. The code is implemented using the Keras library and will be publicly released upon acceptance.

We compare our DSETM framework to previous models validated on other breast ultrasound tasks and state-of-the-art powerful deep learning algorithms in Table 1. For primary classification task, accuracy, sensitivity and specificity are indicators of performance. Both root mean squared error (RMSE) and correlation coefficient (CC) are used to evaluated the performance of regression task. As hypothesized, our

proposed framework shows better metrics than fine-tuned Inception [10], fine-tuned ResNet-50 [11] and the model of Seokmin Han et al [4]. For shape-assisted strategy, compared DSSL (pre-trained DenseNet-121, OA information, classification task) with DSTL (pre-trained DenseNet-121, OA and HS information, classification task), we observe an increase in accuracy from 83.37% to 86.29%. That means the shape of breast lesions contains important detail information and incorporating shape information into the learning model can improve the performance.

To demonstrate the performance improvement that results from multi-task, we compare the performance of DSTL (pre-trained DenseNet-121, OA and HS information, classification task) to DMTL (pre-trained DenseNet-121, OA and HS information, classification task and regression task). In terms of accuracy (86.29%--87.63%), sensitivity (87.15%--88.21%) and specificity (84.84%--86.52%), all of the evaluation metrics of DMTL are better than DSTL. This result is consistent with our previous discussion that having an auxiliary regression task can help the primary classification task. Using multi-task methods can alleviate overfitting to boost the accuracy.

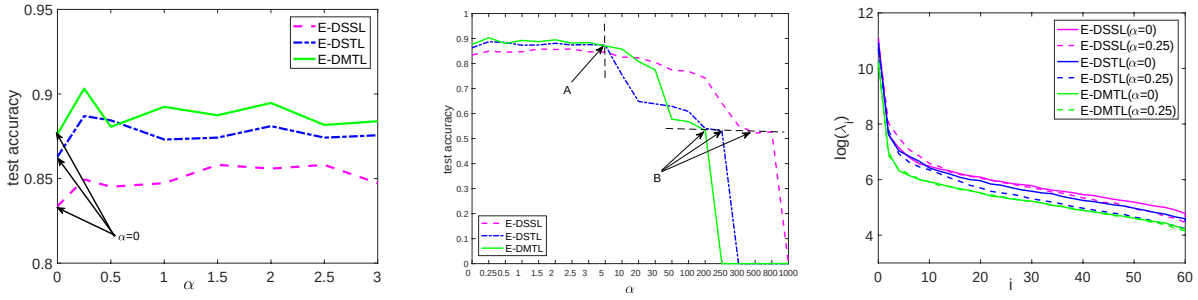
Table 1. Performance Comparisons of the Proposed DSETM Framework and Related Methods on Breast Ultrasound Dataset

Algorithms	Accuracy	Sensitivity	Specificity	RMSE	CC
Seokmin Han et al [4]	0.8148	0.8456	0.7731	----	----
Pretrained Inception [10]	0.7021	0.7389	0.6853	----	----
Pretrained ResNet-50[11]	0.8073	0.7369	0.8806	----	----
DSSL	0.8337	0.8323	0.8347	----	----
E-DSSL ($\alpha = 0.25$)	0.8495	0.8549	0.8433	----	----
DSTL	0.8629	0.8715	0.8484	----	----
E-DSTL ($\alpha = 0.25$)	0.8870	0.8675	0.9024	----	----
DMTL	0.8765	0.8821	0.8652	0.2973	0.7985
E-DMTL ($\alpha = 0.25$)	0.9032	0.9175	0.8881	0.2668	0.8566

In this experiment, we also demonstrate the impact of regularizing prediction entropy. We observe that all of E-DSSL, E-DSTL and E-DMTL models with entropy-regularized strategy obtain performance gains. For example, specifying α as 0.25, E-DSSL gives an increase of accuracy from 83.37% to 84.95%, while E-DSTL provides an increase from 86.29% to 88.70%. Besides accuracy, regularizing prediction entropy strategy also demonstrates the utilities for metrics of sensitivity and specificity. It is crucial to note that E-DMTL obtains better performance for both primary classification task and auxiliary regression task (in terms of RMSE and CC) than DMTL. The results confirm the effectiveness of regularizing prediction entropy on limited BUS dataset.

An integral component of regularization is the choice of regularization parameter α. For details on the choice of α used, we experiment with grid-searching over a small or large spectrum of α values in Fig. 2(a) and Fig. 2(b) respectively. Firstly, for a small spectrum in Fig. 2(a), we observe that entropy-regularized improves the performance on all α. $\alpha = 0$ means there are no entropic confusion in loss function. With a nonzero α, all the models obtain better results than those α equals zero for both classification task and the auxiliary regression task (see more details in Table 1 in the supplement). We also find that performance is insensitive to the choice of α. Secondly, for a large spectrum of α in Fig. 2(b), we observe that performance begins to decrease starting from the first inflection point A. The possible reason is that the regularization loss becomes smaller as α gradually increases and the difference (between L_{loss} and the entropy multiplying by α/b) demonstrates as negative after the point A. Further, we observe that there are significant performance declines when curves beyond the point B. The test accuracy of point B is about 0.5,

which exactly corresponds to statistics of our BUS dataset. The main reasons for these inferior results maybe that the loss is too small and the training is stopped due to the absence of further improvement in loss function. Therefore, the performance can be improved when parameter α varies over a small range.



(a) (b) (c)

Fig. 2. Variation of test performance with regularization parameter varies (a) over a small spectrum; (b) over a large spectrum; (c) Eigenvalue decomposition of covariance (unnormalized PCA) for concatenate layer or penultimate layer.

We hypothesize that the introduction of confusion can reduce the specificity of features and improve generalization. To evaluate it, we provide visual evidence with eigenvalue decomposition of the covariance matrix (unnormalized PCA) on the concatenate layer or penultimate layer features of E-DSSL, E-DSTL, E-DMTL in Fig. 2(c). For a feature matrix with large covariance between the features of different classes, we would expect the first few eigenvalues to be large, and the rest to diminish quickly. We plot the value $\log(\lambda_i)$ for the ith eigenvalue λ_i obtained after decomposition and observe that there are fatter tails for models with $\alpha = 0$ than with $\alpha = 0.25$. Compared to the three models with $\alpha = 0$ or $\alpha = 0.25$, we observe a reduction in the tail of the curves respectively, implying that some generality in features has been introduced through shape-assisted, entropy-regularized and multi-task. The recently published work of [12], [13] explore the applicability of regularizing low-entropy outputs, which is similar to our entropy-regularized formulation. However, they achieve no gains in the context of image classification on CIFAR-10 and CIFAR-100 datasets. We extend this work, provide more detailed analysis and suggestions, demonstrating the effectiveness of entropy-regularized for small-scale medical image dataset classification.

4. Conclusion and Future Work

In this paper, we propose a deep shape-assisted entropy-regularized transferable multi-task framework for benign-malignant classification with limited BUS data. First, in terms of unique shape features of breast lesions, two types of image patches are designed to fine-tune two pre-trained models, aiming to characterize the OA and HS. Second, taking the BI-RADS regression task as an auxiliary task, the CNN-based multi-task architecture is proposed. Third, because prevalent cross-entropy loss formulation may not be ideal for small-scale medical image dataset, we introduce training with confusion by means of regularizing prediction entropy to prevent overfitting. We compare the DSETM framework with its six variants. The results show that our proposed framework demonstrates more robust performance than previous work, which has validated the effectiveness of shape-assisted, multi-task and entropy-regularized. Moreover, we provide detailed analysis of the choice of regularizing parameter and provide visual evidence that introduction of confusion leads to increase in feature generalization. In future work, it is promising to extend the current work by (1) incorporating multi-scale information to further improve classification performance; (2) together with segmentation task to build automatic classification system for breast ultrasound. Our method should be generally applicable to other bio-image analysis problems where the

datasets are small-scale.

References

[1] Centers for disease control and prevention (CDC) 2017 cancer among women. (2017). Retrieved from the website: https://www.cdc.gov/cancer/dcpc/data/women.html

[2] Abdel-Nasser, M., Melendez, J., Moreno, A., & Omer, O. A. (2017). Breast tumor classification in ultrasound images using texture analysis and super-resolution methods. *Engineering Applications of Artificial Intelligence*, *59,* 84-92.

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., & Ciompi, F. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42,* 60-88.

[4] Han, S., Kang, H. K., Jeong, J. Y., & Park, M. H. (2017). A deep learning framework for supporting the classification of breast lesions in ultrasound images. *Physics in Medicine & Biology, 62(19),* 7714.

[5] Bing, L., & Wang, W. (2017). Sparse representation based Multi-Instance learning for breast ultrasound image classification. *Computational and Mathematical Methods in Medicine.*

[6] Jamaludin, A., Kadir, T., & Zisserman, A. (2017). SpineNet: Automated classification and evidence visualization in spinal MRIs. *Medical Image Analysis. 41,* 63-73.

[7] Chen, H., Zheng, Y., Park, J. H., Heng, P. A., & Zhou, S. K. (2016). Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images. *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 487-495).

[8] Huang, G., Liu, Z., Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *CVPR*, *1(2)*, 3.

[9] Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., & Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrapping.

[10] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*, *4*, 12.

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[12] Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., & Hinton, G. (2017). Regularizing neural networks by penalizing confident output distributions.

[13] Abhimanyu D., Otkrist G., Ramesh R., & Iyad R. (2017). Regularizing prediction entropy enhances deep learning with limited data. *NIPS.*

Ji Wu is an associate professor and deputy director of the Department of Electronic Engineering, Tsinghua University, Beijing, China. He received his B.S and Ph.D degrees from Tsinghua University, in 1996 and 2001, respectively, both in electronic engineering. He is heading the Multimedia Signal and Intelligence Information Processing Lab at Tsinghua University. Since 2006, he has been the director of Tsinghua-iFlyTek Joint Lab for Speech Technologies. He is also the leader of SIAC-TWG (Technical Work Group of Speech Industry Alliance of China).

He won the second prize of National Science and Technology Progress Award in 2011 and the first prize of Beijing Science and Technology Award in 2014. His research interests include speech recognition, natural language processing, pattern recognition, machine learning data mining and medical image analysis. He has been elected a Senior Member of the Institute of Electrical and Electronic Engineers in 2015.

He has more 100 publications on Scientific Reports, IEEE Trans on ASLP, ICASSP *et al*.