Ensemble Transductive Learning for Skin Lesion Segmentation

Zhiying Cui[1,2⋆], Longshi Wu[1,2⋆], Ruixuan Wang[1,2(✉)], and Wei-Shi Zheng[1,2]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China
wangruix5@mail.sysu.edu.cn

Abstract. Automated segmentation of skin lesions from dermoscopy images is helpful for the diagnosis and treatment of skin cancers. However, due to small annotated training set and the large visual difference in skins and lesions between subjects, the generalization performance of segmentation models are often limited. Inspired by the transductive learning for image classification, we propose a transductive segmentation approach for skin lesion segmentation, by choosing some of the pixels in test images to participate the training of any segmentation model together with the training set. In this way, visual features in the test images can be effectively learned during model training. Comprehensive evaluations with different model structures and transductive learning strategies showed that the proposed transductive segmentation approach always improve the performance of the corresponding state-of-the-art segmentation models in skin lesion segmentation.

Keywords: Transductive learning · Medical image segmentation · Skin lesions

1 Introduction

Skin cancer is one of the most common cancers, with over 5,000,000 new patients every year in the United States [5]. To effectively diagnose skin cancers and evaluate the effect of various treatments, it is necessary to record and measure the progression of skin lesion regions over time. However, it is time consuming for dermatologists to accurately delineate skin lesion regions. In this case, the state-of-the-art automated image segmentation techniques could potentially help clinicians to efficiently segment skin lesion regions from healthy parts.

Multiple deep learning models have recently been developed for image segmentation, including the first fully convolutional network (FCN) [8], the well-known U-Net [9] which was initially proposed for medical image segmentation by extending the original FCN model with skip-connections between the down-sample and the corresponding up-sample layers, and the state-of-the-art segmentation model DeepLab [2, 3]. For the segmentation task of skin lesions, the

⋆ The authors contribute equally to this paper.

method with top accuracy applied target detection on the skin lesion to reduce reverse effect from different size of skin lesion [4].This approach is cumbersome in training and requires a large number of pre-trained models, so we did not perform detection on our segmentation pipeline. But it should be noted that the baseline models we used is as same as baseline models in previous work.Due to the difficulty in annotating medical images for segmentation, deep learning models are often over-trained with limited annotated medical images. The over-training becomes exacerbated when there is large difference in images between subjects. In this case, images of certain subjects from the test set cannot be typically represented by any image from the training set, and therefore even more training data would not fundamentally solve the problem of limited generalization performance on the test data from new subjects. Unfortunately, such subject-level difference frequently appears in skin image analysis, where each subject may demonstrate distinctive visual features.

To reduce the effect of the subject-level difference, transfer learning is often applied in image classification and segmentation tasks [12, 13], by pretraining a model on another large (either natural or medical) dataset and then fine-tuning the pre-trained model on the task data. However, transfer learning based on a large set of natural images may be limited in improving the performance on medical images, while it is often difficult or infeasible to obtain a large set of medical images to pre-train a segmentation model for later-on use. Another solution is to employ ensemble models by combining multiple individual ones, including the well-known Bagging [1] and Boosting [6] methods. However, these learning strategies cannot fundamentally solve the issue caused by the subject difference.

Instead of exploring novel model architectures or knowledge from additional dataset to help improve segmentation performance on medical images, we propose to directly learn to extract knowledge from the test data during model training. Inspired by transductive learning for classification tasks [11], which tries to use both the annotated training data and the un-annotated test data during model training, we hypothesize that extraction of information from test data and then embedding to the process of model training would largely help the final model to effectively segment the test images. In transductive learning for classification tasks, considering that there are always incorrect prediction on test data by the (initially trained) model, often only those of the test data with high prediction confidence are selected to join the model training, where the predicted label for the selected test data were considered as the ground truth.

While transductive learning has been applied for image classification tasks [10], there is little work particularly for medical image segmentation tasks. In this paper, we propose a transductive learning approach to the segmentation of skin lesion regions, aiming to improve the segmentation performance on the test images by directly learning subject-level visual features from test images during model training. Different from image classification tasks in which each image has a class label, image segmentation can be considered as a pixel-level classification task, in which different pixels in one image may have different class labels

and prediction confidence levels. Therefore, we propose choosing high-confidence pixels from test images during transductive learning rather than using all pixels of each image. Experiments showed the superior performance of the pixel-level test data selection for transductive learning. To further improve the segmentation performance, an ensemble strategy was combined with the transductive learning, in which multiple individual segmentation models are trained with transductive learning and then combined together to segment test images. Experiments with various deep learning models showed that the transductive learning with the ensemble strategy always performs better than corresponding baseline models in skin lesion segmentation.

2 Transductive skin lesion segmentation

The objective of the study is to alleviate the influence of subject-level difference between training and test set, such that the trained segmentation model can have better generalization ability. Instead of focusing on exploring information from training set or other seemingly irrelevant large dataset, here we focus on directly exploring information from the test set during model training, inspired by the transductive learning strategy.

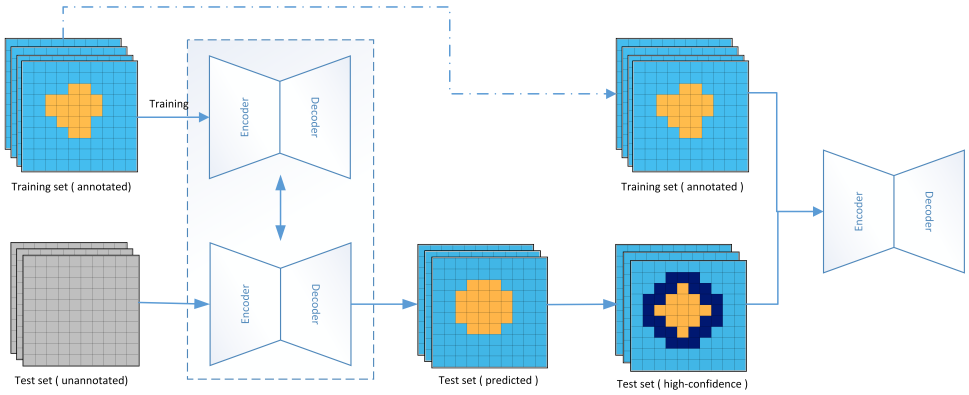

Fig. 1. The framework for transductive skin lesion segmentation. The line with double arrows indicates that the two segmentation models pointed by the arrows are identical.

2.1 Transductive learning

To use transductive learning strategy, an initial segmentation model based only on the training images need to be trained (Figure 1, upper left) and then used to predict the initial segmentation result for each test image (Figure 1, lower left). The initial predictions would be used as ground-truth annotations for the test images, and finally such 'annotated' test images are used together with the

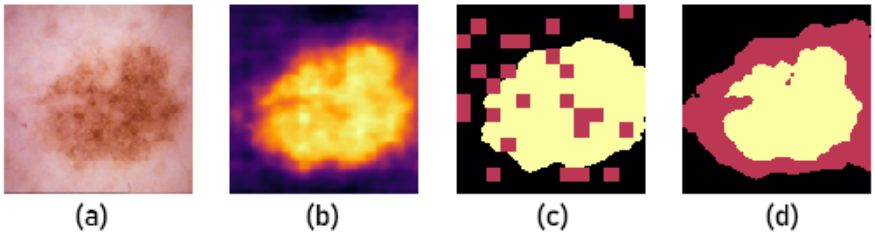

Fig. 2. Example of generating annotation of a test image for transductive learning. (a) a test skin image. (b) the probability output of the initial segmentation model for the test image, with brighter pixels indicating higher probability of belonging to lesion. (c) an example of random selection of pixels for transductive learning. Pixels within randomly generated purple regions are discarded and the other pixels are selected. Each selected pixel was assigned to either 'lesion' (yellow) or 'healthy' (black) class label by thresholding the probability output with 0.5. (d) High-confidence pixel selection for transductive learning. A pixel was considered high-confidence when the probability output is either close to 1.0 (highly likely being lesion) or to 0.0 (likely being healthy), therefore the pixels whose output probability is around 0.5 (purple regions) are discarded and all the other pixels are selected for tranductive learning.

training set to train a new segmentation model (Figure 1, right half), finishing the process of transductive learning. Considering the initial segmentation of test images are often noisy, for each *test* image, only the pixels with high-confidence predictions were selected for model training, where the confidence for each pixel can be directly obtained from the prediction (probability) output from the initial segmentation model. As supported by experiments (Section 3.2), such confidence-based pixel selection strategy is more effective than other candidate selections, including selecting pixels randomly from each test image or using all pixels from a subset of test images. This is also consistent with the strategy of selecting high-confidence test images for transductive image classification [10].

2.2 Ensemble transductive learning

To further improve the generalization of segmentation, we propose combining the ensemble strategy with the transductive learning, i.e., multiple segmentation models were respectively trained by transductive learning and then combined together when predicting the segmentation result for any new image. The variations between these models can be obtained by either training from differential initialized model parameters or by randomly selecting a subset of test images for tranductive learning. To make full use of test images, the former option was selected here, i.e., using high-confidence pixels of all test images, but training models from different initialized parameters.

3 Experimental evaluations

3.1 Experimental setup

The proposed transductive segmentation approach was evaluated on ISIC dataset which was released for the MICCAI'2018 grand challenge "ISIC task1: Lesion Boundary Segmentation" [5]. The training set consists of 2594 dermoscopic images and corresponding ground-truth annotations for lesion regions. The validation set and test set contain 100 images and 1000 images respectively, with ground-truth annotations kept by the organizer. The predicted segmentation result by any model was submitted online to obtain the prediction result via the live leaderboard.

During training of a segmentation model, unless mentioned otherwise, SGD optimizer was used with initial learning rate 0.007 and the momentum value 0.9. Learning rate was updated with a poly scheduler. For the evaluation metric, besides the general measurements (accuracy, dice score, Jaccard index or intersection over union, sensitivity, specificity), the organizer particularly chose the Threshold Jaccard Index (TJI) as the essential metric. In this metric, Jaccard index for the predicted segmentation of any test image was set to zero when the index is lower than a pre-defined threshold (0.65 here), while the index was kept unchanged when it is higher than the threshold. TJI was calculated by averaging the thresholded index values over all test images. TJI can more accurately reflect the number of images in which automated segmentation fails or falls outside expert inter-observer variability. Note that the number of images in which automated segmentation fails is a direct measure of the amount of labor required to correct an algorithm.

In all the subsequent experimental results, if there is no special explanation, we use the same settings to train three identical models at the same time, and then simply use voting strategy and average strategy for ensemble learning. The three models differ only in the random process of parameter initialization and the randomness of the selected samples during the training process. The purpose of the integration is to make the experimental results more reliable and stable.

3.2 Effectiveness of transductive segmentation

To evaluate the effectiveness of transductive segmentation, we first compared the proposed ensemble transductive model with several alternative strategies. One baseline is the traditional ensemble of three segmentation models without using transductive learning ('No transductive' in Table 1). Another strategy is the ensemble of three transductive segmentaion models, with each model trained with all pixels of randomly selected 80% test images ('Random test images' in Table 1). The third strategy is the ensemble of three transductive segmentaion models, with each model trained with all pixels of all test images ('All test images' in Table 1), the fourth strategy is the ensemble of three transductive segmentaion models, with each model trained with randomly selected 80% pixels from each test images ('Random pixels' in Table 1). The last row ('High-confidence

pixels' in Table 1) shows the segmentation performance of the proposed semble of three transductive segmentaion models, with each model trained with high-confidence pixels from each test images. High-confidence pixels were selected by discarding those pixels whose prediction probability values is within the range $[0.25, 0.75]$. Table 1 clearly shows that the proposed ensemble transductive segmentation model outperforms all the other strategies, with 3.5% improvement in TJI compared to the traditional ensemble model (78.1% vs. 74.6%), and more than 1% improvement compared to the ensemble transductive model based on all test images (78.1% vs. 76.8%). Another observation is that the all the ensemble transductive segmentations (the last four rows in Table 1) outperform the ensmble model without using transductive learning, supporting that transductive learning is effective in improving the segmentation of skin lesions, no matter which strategy was used to select pixels or images from the test set.

Table 1. Comparison of transductive segmentations with different strategies. DeepLab v3+ was used as the backbone segmentation model.

Transductive strategy	accuracy	dice	jaccard	sensitivity	specificity	TJI
No transductive	0.935	0.883	0.808	0.95	0.928	0.746
Random test images	0.934	0.887	0.816	0.935	0.936	0.756
All test images	0.933	0.888	0.82	0.921	0.938	0.768
Random pixels	0.933	0.889	0.82	0.929	0.933	0.765
High-confidence pixels	**0.941**	**0.896**	**0.83**	0.919	**0.957**	**0.781**

3.3 Robustness of transductive segmentation over model structures

To evaluate the robustness of the ensemble transductive segmentation, we compared its performance with the 'No transductive' ensemble model and the ensemble tranductive segmentation with 'All test images' under three different segmentation model structures, the well-known U-Net, the DeepLab V3+, and the recently proposed Dual Attention Network(DAN) [7]. Table 2 shows that, while different backbone segmentation models performed differently, all the three models with the proposed high-confidence pixel transductive learning (last row) performed better than the two strong baselines (first and second rows) in skin lesion segmentation. This confirms that the transductive segmentation is independent of segmentation model structures.

3.4 Effectiveness of single transductive segmentation

So far, the evaluation was based on ensemble of multiple single segmentation models. To show that the proposed transductive segmentation approach works not just on ensemble models, here we compared single transductive segmentation model ('High-confidence' in Figure 3) with the baseline single segmentation

Table 2. Performance of ensemble transductive learning with different segmentation model structures. Threshold Jaccard Index (TJI) was used as the metric.

Method	UNet	DAN	DeepLab
No transductive	0.702	0.756	0.746
All test images	0.728	0.762	0.768
High-confidence pixels	**0.730**	**0.779**	**0.781**

without using transductive learning ('Baseline' in Figure 3). In each case, three single models were trained and averaged. Consistent with the previous results with ensemble models, transductive segmentation works better than the one without transductive learning on single models as well, no matter which model structure is used.

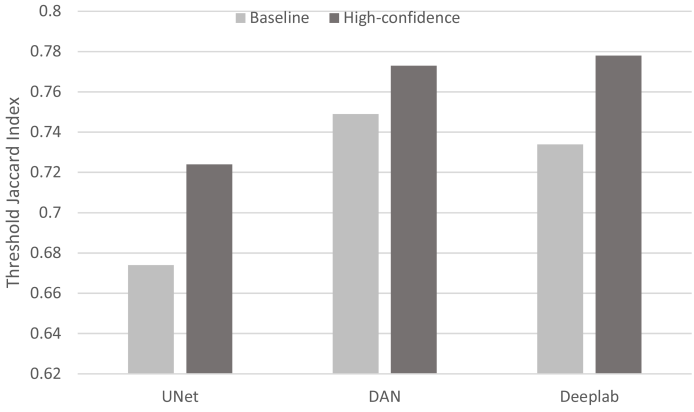

Fig. 3. Comparison between single transductive segmentation and the traditional segmentation without transductive learning on three different model structures.

3.5 Influence of hyper-parameters

One key hyper-parameter in the proposed approach is the threshold value to select high-confidence pixels. All the reported performance above was based on the threshold 0.75, i.e., selecting pixels in each test image whose prediction probability is either larger than 0.75 (likely 'lesion') or less than 0.25 (likely 'healthy'). Here we evaluated the performance of the proposed transductive segmentation with different threshold values 0.65 (i.e., selecting pixels whose prediction probability is either larger than 0.65 or less than 0.35), 0.75, 0.85, and 0.90 (i.e., selecting pixels whose prediction probability is either larger than 0.90 or less than 0.10). Figure 4 demonstrates the selected pixels at different thresholds,

with higher threshold leading to fewer selected pixels, and lower threshold leading to more selected pixels. It is not surprising that higher (e.g., 0.85 or 0.90) or lower threshold value (0.65) would cause relatively worse performance compared to the threshold 0.75 (Table 3), because higher threshold values would make the transductive learning discard too many pixels from test images and therefore cannot extract enough information from test images, while lower threshold value would make the transductive learning select too many pixels from test images which would increase the likelihood of incorrect prediction labels for model training.

Fig. 4. High-confidence pixel selection from each test image with a different threshold. (a) a test image, (b-e) selected pixels (black and yellow) when threshold is 0.65, 0.75, 0.85, and 0.90 respectively. Yellow regions correspond to high-confidence lesions and black regions correspond to high-confidence healthy regions, while purple regions correspond to excluded pixels.

Table 3. The performance of transductive segmentation with different threshold to select high-confidence pixels from each test image.

Threshold	accuracy	dice	jaccard	sensitivity	specificity	TJI
0.65	0.935	0.889	0.818	**0.943**	0.933	0.767
0.75	**0.941**	**0.896**	**0.83**	0.919	**0.957**	**0.781**
0.85	0.937	0.89	0.819	0.936	0.945	0.767
0.90	0.937	0.887	0.813	0.942	0.943	0.757

One may doubt that the reported comparison results above is based on the optimal selection of the hyperparameter for the proposed model, but not for the baseline models. Here we also performed experiments by varying relevant hyperparameters within the baseline models. Specifically, we varied the percent of test images under the 'Random test images' condition (see Table 1 and Section 3.2) for transductive learning, and also varied the percent of randomly selected pixels under the 'Random pixels'. Table 4 showed that even with the optimal hyperparameters, the transductive segmentations under these two conditions were outperformed by the the proposed transductive segmentation with high-confidence pixel selection.

Table 4. The performance of transductive models with varying relevant hyperparameters for alternative pixel/image selection. Note that the 'Random test images' with 100% image selection is equivalent to the 'Random pixels' with 100% pixel selection.

Method	accuracy	dice	jaccard	sensitivity	specificity	TJI
Random test images (60%)	0.934	0.887	0.815	**0.939**	0.936	0.759
Random test images (80%)	0.934	0.887	0.816	0.935	0.936	0.756
Random test images (100%)	0.933	0.888	0.82	0.921	0.938	0.768
Random pixels (60%)	0.933	0.888	0.819	0.923	0.934	0.767
Random pixels (80%)	0.933	0.889	0.82	0.929	0.933	0.765
Random pixels (100%)	0.933	0.888	0.82	0.921	0.938	0.768
High-confidence pixels (0.75)	**0.941**	**0.896**	**0.83**	0.919	**0.957**	**0.781**

4 Conclusion

This paper proposed an ensemble transductive learning strategy for automatically segmenting lesion regions from skin images. By learning directly from both training and test set, the proposed approach can effectively reduce the subject-level difference between training and test set, thus improving the generalization performance of segmentation models. The superior performance of transductive segmentation has been consistently confirmed with varying model structures and strategies to select pixels from test images. Considering that the number of annotated training images are often very limited in medical image segmentation tasks, the transductive segmentation approach may provides an alternative effective way to improve the performance of any segmenatation model, besides the widely adopted transfer learning and ensemble modeling.

References

1. Breiman, L.: Bagging predictors. Machine Learning. pp. 123–140 (1996)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062 (2014)
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
4. Chengyao Qian, Ting Liu, H.J.Z.W.P.W.M.G.B.S.: A detection and segmentation architecture for skin lesion segmentation on dermoscopy images. arXiv preprint arXiv:1809.03917 (2018)
5. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection. In: IEEE International Symposium on Biomedical Imaging. pp. 168–172 (2018)
6. Freund, Y., Schapire, R.E., et al.: Experiments with a new boosting algorithm. In: International Conference on Machine Learning. pp. 148–156 (1996)
7. Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H.: Dual attention network for scene segmentation. arXiv preprint arXiv:1809.02983 (2018)

8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241 (2015)
10. Song, J., Shen, C., Yang, Y., Liu, Y., Song, M.: Transductive unbiased embedding for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1024–1033 (2018)
11. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
12. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3712–3722 (2018)
13. Zhuang, N., Yan, Y., Chen, S., Wang, H., Shen, C.: Multi-label learning based deep transfer neural network for facial attribute classification. Pattern Recognition. **80**, 225–240 (2018)