Webly Supervised Learning for Skin Lesion Classification

Fernando Navarro[1(✉)], Sailesh Conjeti[1,2], Federico Tombari[1], and Nassir Navab[1,3]

[1] Computer Aided Medical Procedures, Technische Universität München, Munich, Germany
fernando.navarro@tum.de
[2] German Center for Neurodegenrative Diseases (DZNE), Bonn, Germany
[3] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. Within medical imaging, manual curation of sufficient well-labeled samples is cost, time and scale-prohibitive. To improve the representativeness of the training dataset, for the first time, we present an approach to utilize large amounts of freely available web data through web-crawling. To handle noise and weak nature of web annotations, we propose a two-step transfer learning based training process with a robust loss function, termed as Webly Supervised Learning (WSL) to train deep models for the task. We also leverage *search by image* to improve the search specificity of our web-crawling and reduce cross-domain noise. Within WSL, we explicitly model the noise structure between classes and incorporate it to selectively distill knowledge from the web data during model training. To demonstrate improved performance due to WSL, we benchmarked on a publicly available 10-class fine-grained skin lesion classification dataset and report a significant improvement of top-1 classification accuracy from 71.25% to 80.53% due to the incorporation of web-supervision.

1 Introduction

The success of deep learning in computer vision tasks such as image classification, object detection, segmentation *etc.* is owed to the availability of a large corpus of annotated training data [1–3]. However, translating these developments to medical imaging applications is often challenging as curating a representative dataset is cost-, time- and scale-prohibitive. On the other hand, excessive reliance on a small-sized, well-curated dataset offers limited guarantees on the generalizability to unseen scenarios and could lead to potential overfit on the training data due to excessive over-parameterization of deep networks. In this paper, we propose to leverage freely available data crawled from the web to offset the need for a large dataset and introduce the concept of *Webly Supervised Learning* (WSL) as a potential approach for training neural networks for medical imaging applications. We present a proof of concept for the task of fine-grained classification of skin lesions in dermatological images.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 398–406, 2018.
https://doi.org/10.1007/978-3-030-00934-2_45

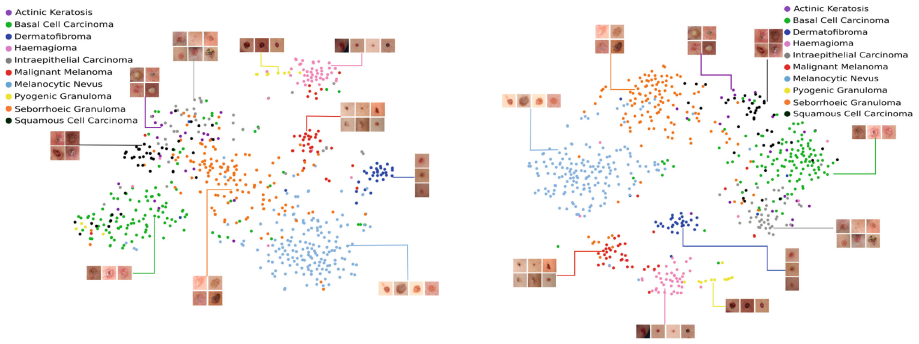

Fig. 1. Comparison of t-SNE embedding space generated from networks trained on limited clean data (Left) against network trained with Webly Supervised Learning (Right) generating compact class clusters with improved separability especially for under-represented classes.

The task of skin lesion classification is a representative example of a medical imaging application in which, annotated training data is limited in availability. However, there is an abundance of freely-available web data. We source our images from multiple publicly-accessible sites such as [4], where pictures of skin lesions are uploaded with the goal of getting feedback on the type of lesion with respect to visual features. Prior work on deep learning for skin lesion classification includes training networks that perform either a two or three-class classification (melanoma, non-cancerous and seborrheic keratosis) [5–7]. Authors in [8] propose a deeply learned network for nine-class categorization using a large dataset of 130000 images extensively curated from hospital archives and from dermatological websites. The data used in this work underwent extensive manual quality control with 23 human experts and filtering prior to fine-tuning InceptionV3 [2]. In contrast to [8], within this paper, we adopt a more unconstrained learning paradigm by focusing on learning in presence of extreme label noise by developing a dedicated robust loss function and employing transfer learning strategies to seamlessly leverage webly sourced data into training without employing any additional heuristics or expert knowledge.

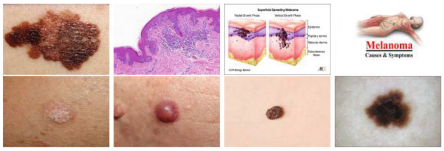

Fig. 2. Type of noise in WSL for Melanoma class as keyword. The images in the first row represent examples of cross-domain noise. The second row represents the cross-category noise.

Harvesting images from the web presents opportunities for abundant availability and the ability to encompass sufficient heterogeneity during training. However, learning from them is challenging due to the presence of different types of noise. These include *cross-domain noise*: retrieved web images may include non-dermatoscopic images such as histology, artistic illustrations, icons *etc.*

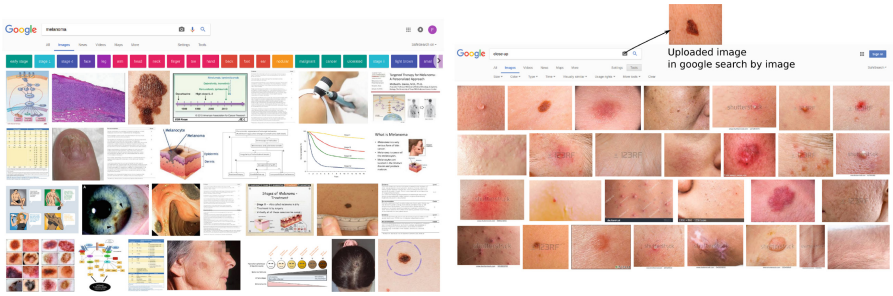

Fig. 3. Comparison of crawled results. The left image shows an example of a search by keyword "melanoma": the resulting images contain high cross-domain noise. The right image shows the results of a search by image, where the cross-domain noise is significantly reduced sharing strong visual similarity to the query image.

and *cross-category noise*: images that are visually similar to the query yet belong to a different class. Cross-domain noise is introduced by bias due to a specific search engine and associated search criterion (such as user tags). Additionally, image-search engines are biased as they often operate in high-precision low-recall regimes and preferentially present objects centered with clean background and a canonical view-point. Figure 2 illustrates different types of noise present in retrieved images upon web-crawling with "melanoma" as the search tag. Learning from web data is one approach for learning under extreme label noise. Methods within the computer vision community that leverage web supervision for training can be broadly categorized as: (1) *Filtering*: approaches that aim to clean or filter the collected web images prior to training [9,10]; (2) *Modeling Relationships*: approaches that model the relationship between web images and noisy labels with a small subset of clean images and utilize the discovered relationships to improve training [11,12] and (3) *Robust Loss Functions*: approaches that learn in the presence of label noise by introducing robustness within their loss-function design [13,14]. Our proposed approach encompasses the best of the aforementioned approaches with the following contributions:

1. **Reduction in Cross-domain Noise:** This is the first work to leverage *search by image* to improve search specificity and reduce cross-domain noise by fetching images that share close visual features to the query image.
2. **Noise Modeling:** We model the noise as a class-transition matrix which is estimated from the bag of retrieved images. This noise modeling approach allows for distillation of knowledge from noisy web-images to train very-deep networks.
3. To the best of our knowledge, this is the first work within the medical image computing that leverages web-supervision to train deep neural networks and specifically targeted at fine-grained ten-class categorization of skin lesions.

2 Methodology

Given a small representative training dataset $\mathcal{C} = \{(\mathbf{x}_c^n, \mathbf{y}_c^n)\}_{n=1}^N$ of dermatological images with expert annotations, we source web-images for WSL by utilizing the *Search by Image* option within standard search engines (here, https://images.google.com/) by submitting each of the *clean* images independently and crawling the top retrieved results. Let the bag of images (of size M) crawled from the web for a query image of $\mathbf{x}_c^n$ be represented as $\mathcal{W}^n = \{\mathbf{x}_w^m \mid \text{Query: } \mathbf{x}_c^n\}_{m=1}^M$. The semantic label associated with the query clean image $\mathbf{y}_c^n$ is given to the corresponding web-crawled bag $\mathcal{W}_n$. Let the complete web-crawled images with their corresponding annotations be denoted as $\mathcal{W} = \{(\mathcal{W}^n, \mathbf{y}_c^n)\}_{n=1}^N$. Contrasting with prior WSL approaches [15,16] that use search by keyword (such as *melanoma*, *keratosis etc.*), we observe that our approach significantly reduces the cross-domain noise by fetching only images that share strong visual features with the query. Figure 3 contrasts the proposed search by image approach against the search by keyword methodology for the construction of a web-dataset for WSL. It must be noted that the labels transferred from $\mathcal{C}$ to $\mathcal{W}$ are extremely noisy as the web-search relies on non-task specific solely visual features for ranking and carries no guarantees on the fetched results sharing the same semantic class as the query. Additionally, in such an uncontrolled setting, multiple queries could fetch the same images thus the web-images could carry potential cross-category noise. With per-image web-crawling, our training dataset is significantly augmented (with at most $M \times N$ unique images) and the resultant dataset is rich in representativeness and heterogeneity but fraught with extreme label noise which needs to be factored out in the subsequent training steps.

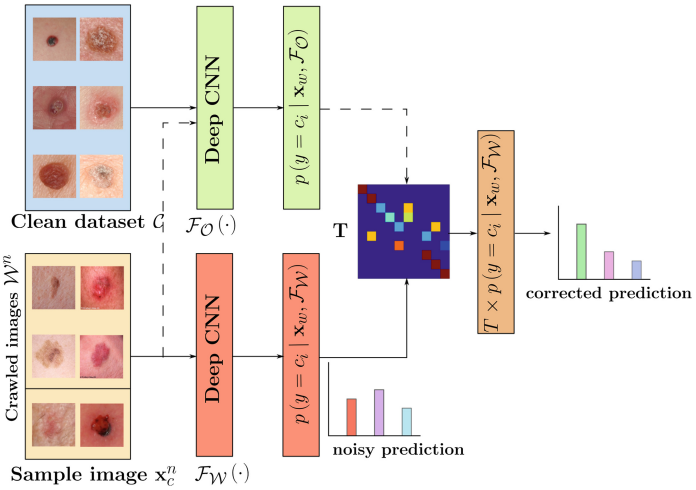

Fig. 4. Overview of the proposed WSL approach consisting of a two-step training. First, training a network on web data, follow by fine-tuning a second network utilizing the latter as strong prior initialization. Noise correction is performed when training on web data.

2.1 Model Learning

Noise Correction: Assuming that we have access to a perfect oracle network $\mathcal{F}_{\mathcal{O}}(\cdot)$, we model the noise within the web images as a class-transition matrix T that can be used to diffuse the predictions on web data across confusing classes. In naïve terms, T models the probability of each class being confused into one another. Considering three classes (c_1, c_2 and c_3), if c_1 and c_2 are visually more similar than c_3, there is a higher probability for cross-category noise across c_1 and c_2 in comparison to c_1 and c_3 (or c_2 and c_3) and this reflects back in the estimated class-transition matrix as $T(c_1, c_2) > T(c_1, c_3)$. From within the web crawled images $\mathcal{W}$, we use the predictions of oracle network $\mathcal{F}_{\mathcal{O}}$ to mine the most representative sample $\hat{\mathbf{x}}_{w,c_i}$ of class c_i from within $\mathcal{W}$ as:

$$\hat{\mathbf{x}}_{w,c_i} = \text{argmax}_{\mathbf{x}_w \in \mathcal{W}}\ p\left(y = c_i \mid \mathbf{x}_w, \mathcal{F}_O\right), \tag{1}$$

where $p(\cdot)$ is the class posterior probability. This is repeated for all target classes and the class transition matrix for the web-data is estimated as:

$$T_{ij} = p\left(y = c_j \mid \hat{\mathbf{x}}_{w,c_i}, \mathcal{F}_O\right) \tag{2}$$

The aforementioned approach globally estimates the noise transition matrix across the web-crawled images and allows for selective diffusion across confounding classes associated with that bag. As the availability of a perfect oracle network is highly unlikely in real-world, we use a deep network trained on the limited clean dataset $\mathcal{C}$ as a potential surrogate for $\mathcal{F}_{\mathcal{O}}$.

Webly Supervised Learning: We adopt a transfer-learning like paradigm to train our fine-grained classification network as shown in Fig. 4. From an overall perspective, the web-crawled dataset $\mathcal{W}$ is used to train an initial model $\mathcal{F}_{\mathcal{W}}(\cdot)$ with weighted-cross entropy loss. Noise correction is modulated by changing the network predictions with the estimated noise transition matrix T from the retrieved web-images $\mathcal{W}$, the modulated cross-entropy loss for training $\mathcal{F}_{\mathcal{W}}$ is estimated as:

$$\mathcal{L} = -\sum_{\mathbf{x}_c^n \in \mathcal{W}} w(\mathbf{x}_c^n)\ y(\mathbf{x}_c^n)\ log(T\ \times\ p(\mathbf{x}_c^n)) \tag{3}$$

where $w(\mathbf{x}_c^n)$ is the weight associated with the class, estimated using median-frequency balancing, $y(\mathbf{x}_c^n)$ is the ground truth of sample $\mathbf{x}_c^n$ and $p(\mathbf{x}_c^n)$ provides the estimated probability of sample $\mathbf{x}_c^n$ to belong to class c. The trained network $\mathcal{F}_{\mathcal{W}}$ is used an initialization for subsequent fine-tuning with clean data $\mathcal{C}$ to obtain the final target model $\mathcal{F}_{\mathcal{C}}(\cdot)$. Such a training strategy ensures that all available rich information from web-supervision is transferred as a strong prior to $\mathcal{F}_{\mathcal{C}}$ and that only expert annotated data is used to train the final network.

3 Experiments

Dataset: The limited manually annotated dataset was sourced from the Dermofit Image Library [17], which consists of 1300 high quality skin lesion images annotated by expert dermatologists. The lesions are categorized into ten fine-grained classes including melanomas, seborrhoeic keratosis, basal cell carcinomas, *etc.* The dataset has an extreme class imbalance (*e.g.* the melanocytic nevus (25.4%) *vs.* pyogenic granuloma (1.8%)). The under-representation of these classes further motivates the need for augmentation with web-crawled data. For our experiments, we performed a patient-level split and used 50% of the data for training and the rest for testing. It must be noted that due to the proprietary nature of the Dermofit library, we do not expect any of our test-data to be freely accessible via the web and hence would not be duplicated within the web-data while training the networks.

Networks: To demonstrate the contributions in terms of both the effectiveness of the proposed search by image as well as the introduction of noise correction while model learning, we established three baselines as presented in Table 1. Specifically, BL1 is the *vanilla* version of training exclusively with the clean training dataset, while contrasting with BL2 we can test the hypothesis that creating a web-dataset through *search by image* induces higher search specificity and significantly reduces the cross-domain noise compared to the web data mined with keywords or user tags. We chose to use the Inception V3 deep architecture [2] as the base model for this work. All the aforementioned models were trained with stochastic gradient descent with a decaying learning rate initialized at 0.01, momentum of 0.9 and dropout of 0.8 for regularization and the code was developed in TensorFlow [18].

Table 1. Design parameters and average performance observed for incremental baselines designed to validate WSL for skin lesion classification.

#	Name	Model learning			Performance	
		Training data	Initialization	Noise correction	Average accuracy	Cohen's Kappa
1	BL1	Clean	ImageNet	–	0.713	0.650
2	BL2	Web data	ImageNet	×		
3		Clean	#4		0.799	0.760
4	Proposed	Web data	ImageNet	✓		
5		Clean	#6		**0.805**	**0.768**

4 Results and Discussion

To evaluate the effect of inclusion of WSL and the proposed noise correction, we report the average accuracy across all classes and the Cohens Kappa coefficient in Table 1. The latter metric is particularly motivated due to the presence of significant class imbalance within our dataset. We also report the class-wise area under the curve (ROC) in Table 2. The confusion matrices are visualized in Fig. 5. To contrast the learned intermediate features, we embed them into a two-dimensional subspace using t-Stochastic Neighbor Embedding (t-SNE) illustrated in Fig. 1.

Analyzing the Embedding Space: Comparing the t-SNE embeddings of the test data generated by BL1 and the proposed approach in Fig. 1, we observe that the embeddings in WSL approach cluster examples from the same semantic class more compactly and maintain better class-separability. Within the embedding of BL1, we notice poor separability between the less-frequently occurring classes (represented with ●, ● and ● bullets), which is significantly improved in the embedding of WSL. We also observe that the misclassification of Pyogenic Granuloma(benign) ● class into Basal Cell Carcinoma (malignant) ● in case of BL1 is not observed for WSL. This is quite critical as these classes are mutually exclusive. Meaning that, a vast malignant samples can be classified as benign, leading to a wrong diagnosis.

Effect of Web Supervision: Contrasting BL1 against the proposed method in Table 1 and Fig. 5, we clearly observe a significant improvement in the model performance across all classes, with a more pronounced diagonal in its confusion matrix. This is clearly attributed to a better network initialization derived through transfer learning with web-supervision. This also demonstrates that we are effective in factoring out the cross-domain and cross-category noise within the web-dataset and effectively use it for supervising deep models in the presence of limited manual annotations. In Table 2, contrasting the class-wise performance, we observe that the performance on under-represented classes is significantly improved upon WSL. This is clearly evident in Intraepithelial Carcinoma ● (5.99% Clean data) and Pyogenic Granuloma ● (1.17% Clean data) where the performance improves by 3.6% and 7.3% respectively. Contrasting BL2 with the proposed approach in Table 2, we observe an overall improvement when performing noise correction. The AUC has a slight improvement across the majority of classes. With this observation, it can be concluded that the web-crawled images retrieved in a search by image proposed methodology are so rich in terms of visual features that the effect of noise correction is only marginal when comparing BL2 and the proposed approach.

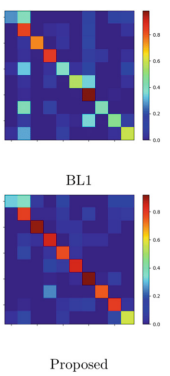

Fig. 5. Confusion matrices showing the effect of the proposed WSL approach compared to BL1.

Table 2. Results showing the AUC for each class and the average overall accuracy for every model.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Avg AUC
Class % in train	3.42	18.49	4.96	7.53	5.99	5.82	25.51	1.17	19.86	6.67	–
BL1	0.898	0.943	0.976	**0.983**	0.936	0.955	0.979	0.927	0.933	0.872	0.940
BL2	0.873	0.955	0.995	0.966	0.967	0.984	**0.987**	0.991	**0.975**	0.935	0.963
Proposed	**0.920**	**0.966**	**0.995**	0.968	**0.972**	**0.985**	0.983	**0.991**	0.961	**0.957**	**0.970**

5 Conclusions

In this work, we have demonstrated for the first time the effectiveness of webly supervised learning for the task of skin lesion classification. We demonstrate that WSL can be very effective for training in limited data regimens with high-class imbalance as web data can augment under-represented classes and boost the model performance. By crawling the web through search by image to generate the web-dataset, we induce high search specificity and effectively minimize the influence of cross-domain noise. The proposed noise correction approach by modeling cross-category noise helps in learning an effective network initialization from web data.

Acknowledgements. The authors gratefully acknowledge CONACYT for the financial support.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2017)
3. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
4. DermQuest. https://www.dermquest.com/
5. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE TMI **36**, 994–1004 (2017)
6. Matsunaga, K., Hamada, A., Minagawa, A., Koga, H.: Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. arXiv:1703.03108 (2017)
7. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection. arXiv:1710.05006 (2017)
8. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115 (2017)
9. Krause, J., et al.: The unreasonable effectiveness of noisy data for fine-grained recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 301–320. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_19

10. Massouh, N., Babiloni, F., Tommasi, T., Young, J., Hawes, N., Caputo, B.: Learning deep visual object models from noisy web data: how to make it work. arXiv:1702.08513 (2017)
11. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: CVPR (2015)
12. Vahdat, A.: Toward robustness against label noise in training deep discriminative neural networks. In: NIPS (2017)
13. Sukhbaatar, S., Fergus, R.: Learning from noisy labels with deep neural networks. arXiv:1406.2080 (2014). **2**(3), 4
14. Patrini, G., Rozza, A., Menon, A., Nock, R., Qu, L.: Making neural networks robust to label noise: a loss correction approach. arXiv:1609.03683 (2016)
15. Chen, X., Gupta, A.: Webly supervised learning of convolutional networks. In: ICCV (2015)
16. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
17. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi, M., Schaefer, G. (eds.) Color Medical Image Analysis, pp. 63–86. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-5389-1_4
18. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous distributed systems. CoRR abs/1603.04467 (2016)