ON THE IMPACT OF SELF-SUPERVISED LEARNING IN SKIN CANCER DIAGNOSIS

Maria Rita Verdelho and Catarina Barata

Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal

ABSTRACT

Deep neural networks (DNNs) are the standard approach for image classification. However, they require a large amount of data and corresponding annotations. Collecting medical data is a difficult task, due to privacy restrictions. Moreover, it is even harder to obtain the clinical labels, since these must be provided by specialists. Self-supervised learning (SSL) has emerged as a possibility to overcome this issue, since it uses non-annotated data to pre-train the DNN. Recently SSL has been applied in the context of skin cancer. However, the results were not conclusive. Moreover, a proper analysis of the impact of different SSL approaches is still missing. In this paper we investigate two SSL approaches: **Rotation** and **SimCLR**. Our results highlight the benefits of applying self-supervised learning to the classification of dermoscopy images. Additionally, we demonstrate that these approaches learn different and complementary features.

Index Terms— Skin Cancer, Deep Learning, Self-Supervised Learning, Dermoscopy

1. INTRODUCTION

Skin cancer is one of the most common types of cancer worldwide [1]. In the past decade, the number of melanoma cases diagnosed has increased by 47% and in non-melanoma cancer about 5.400 people worldwide die every month due to this disease [2]. Skin cancer is also one of the most treatable forms of cancer when detected in an early stage. However, late detection can have a significant impact in mortality rates. Therefore, there is a need to develop a convenient and precise method to perform early diagnosis [3].

Over the past decade, deep neural networks (DNNs) have been developed to assist human experts and accelerate the process of skin cancer diagnosis [3]. However these methods require a huge amount of annotated data to obtain satisfactory results. Collecting medical data is a difficult task, due to privacy and law restriction, and it is even harder to obtain clinical annotations, since these must be provided by specialists [4]. To overcome this issue, the research community has been relying on transfer learning. This method consists of first training a model for a task using a large data base and then "recycle" it for a new target task [5]. These pre-trained models usually have deeper architectures than what is needed in medical image analysis [6]. Additionally, the color distribution of natural images is also very different from the medical ones [7], which can result in models that have difficulties in generalizing to the other data [6].

Self-supervised learning (SSL) has emerged as a strategy to avoid the annotation process. This technique takes advantage of unlabeled data to perform a pre-training of the DNN [8] [9], allowing the model to learn relevant image features that can later be applied to a specific task. Recently, SSL has been used in the skin image context. Both Li *et al.* [9] and Tajbakhsh *et al.* [6] applied SSL techniques with color-based pretext tasks to the segmentation of skin lesions. Kwasigroch *et al.* [4] applied two SSL techniques based on geometric distortion to the skin cancer classification task. The closest work to ours is that of Chaves *et al.* [10], in which they asses five SSL contrastive techniques against a competitive supervised baseline and conclude that SSL is competitive both in reducing variability and improving model accuracy. Despite the promising results, it is still unclear which is the best SSL strategy for skin images. Additionally, all works focus solely on a quantitative analysis, disregarding the impact of SSL on the features learned by the model.

This work aims to shed a new light on the application of SSL in the skin cancer context. Towards this goal we have develop a robust experimental framework to:

(i) investigate the impact of SSL on the training and generalization of a DNN for skin lesion diagnosis into 8 different classes, and demonstrate that even with a small dataset there are benefits in using SSL.
(ii) compare two different SSL approaches, one based on geometric distortion and another on contrastive learning.
(iii) for the first time provide a qualitative assessment of the impact of the different pre-training strategies, using explainability approaches.
(iv) demonstrate the complementarity of the features learned by the SSL strategies and the benefits of combining them.

To the best of our knowledge, this is the first work to perform a robust quantitative and qualitative validation of the impact of SSL, and to demonstrate the importance of combining different SSL techniques.

The remaining of the paper is organized as follows. Section 2 introduces the used methodologies, and Section 3 describes the experimental setup. Section 4 presents the results and Section 5 concludes the paper.

2. METHODOLOGIES

This section gives a brief overview of SSL and the two strategies adopted in this work, as well as the experimental setup adopted in the skin cancer problem.

2.1. Self-Supervised Learning (SSL)

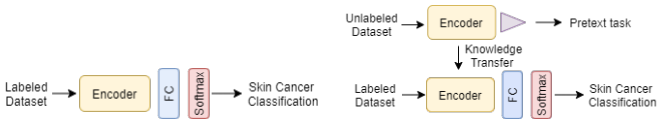
SSL is a technique used to extract visual features from unlabeled data [4]. The main goal is to use the learned weights to initialize a DNN for a specific target task, which is, in the skin cancer image analysis, the classification of the different skin lesions. To achieve this goal, the model is trained to execute a pretext task, for which labels can be easily generated without human supervision. Pretext tasks aims to extract different feature representations from the images. Therefore, it is important to select a SSL technique that is adequate to the wanted target supervised task. In this paper we will use two SSL techniques, which we believe to have a good performance on the skin image classification problem: Rotation [11] and the SimCLR [12].

2.1.1. Rotation technique

Rotation [11] is a classification-based technique, where the network is trained to predict which rotation (0º, 90º, 180º or 270º) has been applied to the image. Therefore, by predicting which rotation was applied to the input, the model is capable of extracting useful information from each image. The training pipeline starts with a small set of geometric transformations, which will be applied to the dataset. Secondly, the transformed images are fed to the model and the DNN is trained to identify which rotation was applied to the original image. As mentioned before, the set of geometric transformations defines the classification task, meaning that if there are four rotations then it is a 4-class classification problem.

2.1.2. SimCLR technique

SimCLR [12] is a SSL approach that applies the concept of contrastive learning to infer feature representations from the unlabeled dataset. Feature representations are learned by maximizing the agreement between differently augmented views of the same image via a contrastive loss, which will also accentuate the dissimilarity among different images. The key idea is when comparing the multiple images using the contrastive objective, the representations of corresponding views are 'attracted' to one another and the others are 'repelled'. SimCLR can be divided into four main steps: 1) Random transformations are applied to the input, in order to obtain a pair of two augmented images. 2) Each augmented image within the pair is sent to an encoder. 3) The output representations of the encoder are then sent to a multi-layer perceptron (MLP). 4) The contrastive loss is applied in the feature space given by the MLP.



(a) Supervised Learning. (b) Self-supervised Learning.

Fig. 1. Proposed framework using different initialization techniques applied to the skin cancer diagnoses. In both models the last layer is fully-connected one with 8 units. The triangle represents the last layers of the DNN specific of the pretext-task.

2.2. Experimental Framework

This paper aims to perform a robust assessment of the impact of SSL as a pre-training technique, to initialize the weights of a DNN for skin cancer diagnosis. To better understand the impact of SSL, we perform a systematic assessment, adopting the following pipeline:

(i) **Baselines** - two standard supervised learning strategies, where the weights of the DNN are initialized either at random (trained from scratch) or using a pre-trained model on ImageNet (fine-tuning).

(ii) **Scratch + SSL** - standard SSL methodology, where the weights of the DNN are initialized at random and refined using either the Rotation or the SimCLR technique.

(iii) **ImageNet + SSL** - a variant of the SSL approach, that aims to leverage the information from a model pre-trained on the ImageNet dataset. Here, we initialize the weights of the model used in the SSL phase from ImageNet and refine them using either the Rotation or SimCLR approach.

(iv) **Fusion** - fusion of the DNNs pre-trained using the Rotation and SimCLR techniques both at the feature (early fusion) and classification (late fusion) level [1].

Fig. 1 (a) describes the proposed generic approach for the application of supervised learning (baselines) and Fig. 1 (b) describes the proposed approach for the application of self-supervised learning. For the latter, the first step consists of pre-training the DNN using the chosen pretext task and, secondly, fine-tuning the parameters of model to the classification task (this time using labels), by recycling the encoder and adding a fully connected layer to output the 8 classes presented in our skin cancer dataset. In all our experiments, the encoder is a ResNet-50 [13].

3. EXPERIMENTAL SETUP

3.1. Dataset and Evaluation Metrics

All experiments were performed using the ISIC 2019 [14] [15] [16]. This dataset comprises 25,331 dermoscopy images, divided into 8 lesions classes: Actinic keratosis (AKIEC),

[1] More details will be presented in the Support Material document

Basal cell carcinoma (BCC), Benign keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Nevus (NV), Squamous cell carcinoma (SCC) and Vascular (VASC). These labels are only used to train the classification models (recall Fig. 1). The images were collected at different medical centers (each center generated images with different sizes, color and aspect ratio). Therefore, it was necessary to pre-process them. This process compensated the color and allowed all the images to have the same size, while maintaining their aspect ratio. After having resized all the images to the desired size (224x224), we applied the color constancy algorithm Shades of Gray as it is proposed in [17]. In order to compare the different initialization approaches and assess their robustness, we adopted a 5-time Monte Carlo sampling strategy, where the ISIC 2019 dataset was partitioned five times into training (70%) and validation (30%) sets. Based on this, we report the median and standard deviation of the following metrics: Balanced Accuracy (BACC), Precision, F1-Score, and Specificity.

3.2. Network Training and Computational Environment

The experimental framework was implemented using Tensorflow/Keras and one NVIDIA Tesla K80 GPU [2]. All models were trained for 60 epochs, using early stopping and the Adam optimizer [18]. The batch size was set to 32. For SSL, the losses are the categorial cross-entropy for the rotation task and for the SimCLR it was used the NT-Xent loss (with $\tau = 0.1$). For this task, we transformed the input image using horizontal flips, central crops and rotations of 0, 90, 180 or 270 degrees. We also studied the impacts of random color distribution and random Gaussian blur, however these experiments resulted in a lower performance of the model. Both tasks had a initial learning rate of $\eta = 10^{-4}$, however the rotation had a reduction factor of 0.75 and the SimCLR a exponential decay of 0.96. To train the classifier, we adopted the weighted categorical cross-entropy loss, where the weights are set to the relative frequency of each class, in order to account for the unbalance. Here the learning rate was set to $\eta = 10^{-5}$, with a reduction factor of 0.75. In order to prevent over-fitting, we also used online data-augmentation (random flips and rotations of multiples of 90 degrees).

4. RESULTS

The results section is divided into three parts: i) a quantitative analysis, where we compare the different approaches taking into consideration the selected evaluation metrics; ii) a qualitative analysis that used the Grad-CAM technique [19] to convey a more interpretable analysis of the impact of the various initialization strategies in the features learned by the model; and iii) a quantitative analysis of the fusion of SSL strategies.

4.1. Quantitative Analysis

Table 1 summarizes the median and standard deviation of the scores obtained for the different initialization techniques. By looking at Table 1 it is possible to see that there are some benefits in using SSL when compared to the baseline supervised training. By looking at the baseline trained from scratch (row 1) and to both rows trained from scratch with self supervised learning techniques (row 3 and 4) it is visible that both SSL techniques presented higher median and lower standard deviations. This proves that when comparing models trained from scratch there is a tendency to have **higher accuracy** and **more stability** (the standard deviation has a lower value) in the **models** that use **SSL**. By looking at the models trained using the ImageNet weights - the baseline (row 2) and to both models that used the SSL techniques (row 5 and 6)- it is visible that the latter two have a higher stability (lower standard deviation) even though both had smaller or similar accuracy to the baseline. This proves that when comparing models trained with the ImageNet weights there is a tendency to **have more stability** in the models that use SSL. Finally, looking at the SSL pre-trained models (row 2, 3, 5 and 6) it is possible to see that the **rotation technique has a higher accuracy** when compared to the model initialized with the SimCLR technique. More results are presented in the support material document. A shared conclusion between our work, [4], and [10] is that, when using SSL pre-trained models, there is an out-performance in general terms, especially in variability.

4.2. Qualitative Analysis

We opted to execute a qualitative analysis, since we wanted to understand what each model saw differently and what it learned in order to make the diagnostic decisions. Therefore, to analyze the differences between the learned representations for each initialization technique the Grad-CAM [19] was used. This is a technique used for visualizing the features learned by the DNN and the regions of an image that activate a certain label. Figure 2 shows the Grad-CAM results for the different initialization techniques (fine-tuned with ImageNet weights). We present more examples in the Support Material document. Figure 2 proves that for the same input image all three models look at different parts of each lesion. Therefore, apart from having different performances each model seems to learn different information about each class of lesion. The SimCLR pre-trained model tended to focus more in the parts of the lesion that presented higher contrast, while the Rotation looked more at the structure of each lesion. The ImageNet pre-trained model, was the least intuitive to interpret, since its focus varied between lesion and skin. After, analyzing a set of different images it was possible to confirm that each method

[2]The source code and the Support Material document will be released in: `https://github.com/mrverdelho/IMPACT-OF-SELF-SUPERVISED-LEARNING-IN-SKIN-CANCER-DIAGNOSIS.git`

Table 1. Application of the Monte Carlo Sampling with different initialization techniques: training the model from scratch or fine-tuning with ImageNet weights; application of two self-supervised learning (SSL) techniques -Rotation and SimCLR - and fusion of both techniques.

Initialization	Technique	BACC (%)	Precision (%)	F1-Score (%)	SP(%)
Baseline	Scratch	46,82 ± 2,00	35,37 ± 3,84	37,24 ± 4,64	92,89 ± 0,55
	ImageNet	71,48 ± 1,82	65,14 ± 2,78	67,93 ± 1,75	96,04 ± 0,12
Scratch + SSL	Rotation	54,92 ± 1,15	40,54 ± 1,84	43,19 ± 2,04	93,39 ± 0,18
	SimCLR	52,54 ± 0,86	44,62 ± 1,39	47,53 ± 0,96	93,94 ± 0,18
ImageNet + SSL	Rotation	71,47 ± 0,30	62,37 ± 0,74	65,70 ± 0,47	95,77 ± 0,05
	SimCLR	65,37 ± 0,55	54,47 ± 2,71	58,28 ± 1,95	95,17 ± 0,18
Fusion	Early Fusion	**73,78 ± 0,24**	**68,41 ± 2,07**	**70,99** ± 2,61	**96,40** ± 0,36
	Late Fusion (mean)	57,09 ± 2,19	50,28 ± 1,41	52,02 ± 1,08	94,24 ± 0,19

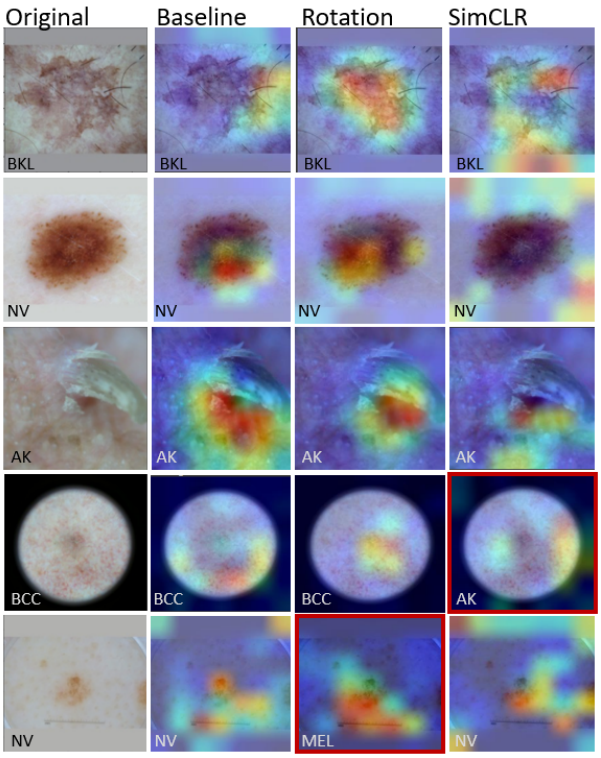


Fig. 2. Example of different lesion visualizations using the Grad-CAM algorithm (Baseline, Rotation and SimCLR). Each image contains the predicted lesion class and we present the incorrect classification with a red square.

also had some limitations. The rotation had difficulties in detecting centered and symmetrical lesions, since each rotation of 90 degrees is similar, then the model does not learn useful information about this lesion. This limitation is visible in the fifth row and second column of fig. 2. However, the SimCLR as some images contained margins with high contrast (black borders), this method tended to focus more on the margins than the lesion (exemplified in the fourth row and third column of fig. 2). Based on the qualitative results, the question that arose next was: **Is the information learned by both SSL techniques complementary?**

4.3. Fusion of SSL Approaches

As a consequence of the previous interrogation, we performed two tests that fused the models pre-trained with SSL. First, we used early fusion, which fuses the different methods in the feature space. Secondly, we used late fusion that fuses the models in the classification scores level (we applied the mean strategy). The fusion results appear in the last two rows of table 1. It is possible to conclude that the **early fusion (row 7) had better results than any other model both in stability and accuracy**, proving that in fact the features of both models have complementary information. However, the late fusion (row 8) proved to have worse results, meaning that the features are complementary, but not learned classification models.

5. CONCLUSIONS

This paper performed a robust assessment of the impact of SSL as a pre-training technique for skin cancer diagnosis. In particular, we performed a quantitative and qualitative analysis of the different pipelines. During this assessment we compared two SSL techniques: Rotation and SimCLR. Our experimental results show that there are benefits while using SSL. We observed that when applying this technique, the classification DNN appeared to have less variability in its performance. To the best of our knowledge this is the first work that provides a qualitative analysis of the features learned by the SSL strategies. This study led us to conclude that each model learned different information from the data. Therefore, we also studied the combination of the two SSL techniques which resulted in the highest performance. SSL is known to benefit from using more unlabeled data. Therefore, we plan to repeat both experiments using more unlabeled data in future work

Acknowledgements

This work was supported by the FCT project and multi-year funding [CEECIND/ 00326/2017] and LARSyS - FCT Plurianual funding 2020-2023; and by a Google Research Award'21 (Project DeepMutation).

6. REFERENCES

[1] *Euromelanoma: Cancro da Pele, 2020.*

[2] "Skincancer foundation, 2020," .

[3] Yasuhiro Fujisawa, Sae Inoue, and Yoshiyuki Nakamura, "The possibility of deep learning-based, computer-aided skin tumor classifiers," *Frontiers in medicine*, vol. 6, pp. 191, 2019.

[4] Arkadiusz Kwasigroch, Michał Grochowski, and Agnieszka Mikołajczyk, "Self-supervised learning to increase the performance of skin lesion classification," *Electronics*, vol. 9, no. 11, pp. 1930, 2020.

[5] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flavia Vasques Bittencourt, Sandra Avila, and Eduardo Valle, "Knowledge transfer for melanoma screening with deep learning," *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 297–300, Apr 2017.

[6] Nima Tajbakhsh, Yufei Hu, Junli Cao, Xingjian Yan, Yi Xiao, Yong Lu, Jianming Liang, Demetri Terzopoulos, and Xiaowei Ding, "Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1251–1255.

[7] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, pp. 101539, 2019.

[8] Carl Doersch and Andrew Zisserman, "Multi-task self-supervised visual learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2051–2060.

[9] Yuexiang Li, Jiawei Chen, and Yefeng Zheng, "A multi-task self-supervised learning framework for scopy images," *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 2005–2009, 04 2020.

[10] Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila, "An evaluation of self-supervised pre-training for skin-lesion analysis," *CoRR*, vol. abs/2106.09229, 2021.

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations," in *International Conference on Learning Representations (ICLR)*, Vancouver, Canada, Apr. 2018.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 08 2018.

[15] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 168–172.

[16] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Verónica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, and Josep Malvehy, "Bcn20000: Dermoscopic lesions in the wild," *ArXiv*, vol. abs/1908.02288, 2019.

[17] Graham D Finlayson and Elisabetta Trezzi, "Shades of gray and colour constancy," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2004, vol. 2004, pp. 37–41.

[18] Mohammed Alom, "Adam optimization algorithm," 06 2021.

[19] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct 2019.