DAT341 / DIT867 Programming Assignment 5
Dermatological image classification

First Author
Semir Sehic
semirs@chalmers.se

Second Author
Valentin Quoniam-Barré
valquo@chalmers.se

Abstract

This report outlines our investigation into machine learning architectures for classifying dermatological images, focusing on distinguishing birthmarks (nevus) from malignant melanomas. Beginning with a simple CNN architecture as a foundational model, we progressively explore more sophisticated and advanced architecture models. Including ResNet (He et al., 2015), VGG19 (Simonyan and Zisserman, 2015) and the innovative DINOv2 (Oquab et al., 2024). Through comparative analysis on a subset of the 2018 ISIC challange dataset we present the impact of architectural choices, normalization techniques, residual connections, data augmentation and transfer learning on model performance. Our findings accentuate the potential of advanced machine learning models in improving the accuracy of skin cancer detection.

1 Introduction

Skin cancer, particularly melanoma, is a significant health concern, and early detection is crucial for successful treatment outcomes. One potential approach to aid in early detection involves analyzing birthmarks, as patients with melanoma often exhibit larger and more irregular ones. Since this detection process relies on visual inspection, image classification techniques can be employed to automate and streamline the process.

In this study, various methods will be employed to train several classifier models for skin lesion classification. These methods may include batch normalization (Ioffe and Szegedy, 2015), transfer learning, data augmentation techniques and residual connections. Different models will be used as feature extractors, like ResNet (He et al., 2015) or state-of-the-art DINOv2 (Oquab et al., 2024).

The trained models will then be evaluated. The results obtained from different models will be compared and analyzed to identify the most effective approach for melanoma detection.

Furthermore, the study will delve into the limitations and potential challenges associated with the proposed methods, such as issues related to data quality, class imbalance, and model generalization. Additionally, ethical considerations surrounding the use of automated image classification systems in healthcare, including privacy concerns, bias mitigation, and the potential for misdiagnosis, will be discussed.

Overall, this study aims to contribute to the development of an effective and reliable automated system for melanoma detection, which could potentially aid in early diagnosis and improve patient outcomes.

2 Methods

Here, the process of training and evaluating the different models is explained in detail.

2.1 Data Augmentation

Data augmentation techniques are often applied to increase the diversity and size of the training dataset. Among these techniques, one could think of:

- Random rotations
- Flips
- Zooms
- Random crops

For this specific problem, zooming somewhere on the image might result in landing on a blank piece of skin. Therefore, the melanoma or the birthmark would not be visible. The same issue arises with random cropping. To save time, the decision was made to only apply random flipping.

The augmented dataset was then utilized to train the models. It's a small augmentation, but it's sufficient to improve the model's generalization ability and robustness to variations in the input data.

2.2 Batch Normalization

Batch normalization (Ioffe and Szegedy, 2015) is used to improve the training stability and convergence speed of deep neural networks. It normalizes the input to each layer by subtracting the batch mean and dividing by the batch standard deviation. This normalization can be represented by the following equation:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \tag{1}$$

where x_i is the input to the layer, μ_B is the batch mean, σ_B^2 is the batch variance, and ϵ is a small constant added for numerical stability.

This normalization helps reduce the internal covariate shift, a phenomenon where the distribution of inputs to a layer changes during training, leading to slower convergence and potential degradation of performance.

Batch normalization layers were incorporated into the model architectures, in the convolutional part, to enhance their training efficiency and overall performance.

2.3 Residual Connections

Residual connections, introduced in the ResNet (He et al., 2015) architecture, allow the flow of information through skip connections, bypassing certain layers in the network. This technique helps mitigate the vanishing gradient problem, which can occur in very deep neural networks, by providing an alternative path for gradients to flow during backpropagation.

Initially, a small residual connection was implemented in the model architecture. Subsequently, the ResNet architecture was employed as a feature extractor, with its pre-trained weights utilized to extract relevant features from the input images. These extracted features were then fed into fully connected layers for the final classification task. This approach, known as transfer learning, leverages the knowledge gained from the ResNet model, which was pre-trained on a large-scale dataset, and adapts it to the specific task of skin lesion classification.

2.4 Transfer Learning

Transfer learning is a technique that leverages knowledge gained from a model trained on a different but related task. In this study, pre-trained models such as VGG19 (Simonyan and Zisserman, 2015) and DINOv2 (Oquab et al., 2024) which were initially trained on large-scale datasets like ImageNet, were used as feature extractors on the skin lesion classification task. By leveraging the learned representations from these pre-trained models, the training process can be accelerated, and better performance can potentially be achieved, even with a relatively small dataset for the specific task at hand.

Three pre-trained models were used as feature extractors :

- ResNet (He et al., 2015), where the last layer is modified into an Identity layer. The output is then given to a 3-depth linear classifier.
- VGG19 (Simonyan and Zisserman, 2015), where only the feature part is used and frozen. Then, its output is passed to a more complex classifier than for ResNet. Using a deeper classifier because we thought it would be too simple to have only 3 linear layers with something as complicated as vgg19.
- DINOv2 (Oquab et al., 2024): DINOv2, a state-of-the-art vision transformer model, was also explored as a feature extractor for the skin lesion classification task. Similar to the other pre-trained models, DINOv2 was utilized to extract relevant features from the input images, which were then passed to a custom classifier head for the final classification.

By leveraging these pre-trained models as feature extractors and combining them with custom classifier architectures, the study aimed to benefit from the powerful representations learned on large-scale datasets while adapting the models to the specific task of skin lesion classification through transfer learning.

2.5 Training Procedure

The models were trained using a gradient descent (SGD) optimizer with a learning rate of 0.01, which is fairly common. The loss function used was the cross-entropy loss. It transforms the two outputs of the classifier to probabilities, and compare them to the real label. This loss penalizes the model when it assigns a low probability to the correct class and encourages it to assign high probabilities to the true class labels, thereby minimizing the discrepancy between the predicted and true distributions.

The training was conducted over 10 epochs, with a batch size of 32. This partitioned the training set into 201 batches, allowing each epoch to be computed within a minute. Given the various tests and modifications performed on the code, a relatively short computation time was desirable to facilitate efficient experimentation and iteration.

For the validation, the output with the highest probability was chosen as the predicted label. By comparing it to the real label, the validation accuracy was calculated.

3 Results

In this section we present our results from the different ML techniques and applications used.

3.1 Data Augmentation

Applying basic data augmentation techniques enhance our models performance by introducing a broader variety of training data. By implementing flips, zooms and crops results in a improvement in the models ability to generalize across unseen images. Using these basic data augmentation techniques on CNNs resulted in a accuracy of 78%. By adding random rotations to the data augmentation resulted in a increase of accuracy to 80%. Importance of presenting diverse scenarios to the model during training. Especially for tasks like dermatological image classification where variance in visual features is high.

3.2 Batch Normalization

By incorporating batch normalization led to faster convergence rate and stable training across the models. The normalization of inputs within each layer reduced internal covariate shift. This adjustment resulted in a reduction of epoch count needed to reach similar results as before. We were able to achieve a small level of improvement of the performance, reaching a accuracy of 82%. Enhancing the models capability to learn even more than before.

3.3 Residual Connections

The introduction of residual connections had a impact on the ability to train even deeper networks without hindrance of vanishing gradients. Models equipped with residual connections not only exhibited faster convergence but also achieved higher accuracy levels. The improvement was quantified as a 84% increase in overall accuracy with a small enhancement in the model's sensitivity to detect melanomas. But an overall improvement from the baseline accuracy.

3.4 Transfer Learning

Utilizing pre-trained models through transfer learning accelerated the training process and improved the accuracy of our dermatological image classification models. By adapting the ResNet, VGG19 and state of the art DINOv2 architectures pre-trained on large image datasets, we leveraged their learned features to enhance our models. The transfer learning approach yielded an immediate improvement, with models reaching a accuracy of 87% (ResNet), 86% (VGG19) and 88% (DINOv2) on the test set compared to those trained from scratch. This demonstrates the value of transfer learning in effectively applying the knowledge acquired from large datasets to specific domain-focused tasks like melanoma detection.

3.5 Summary

Our results through various experiments show improvements in classification accuracy through the application of advanced machine learning techniques. Data augmentation increased the model robustness to variations in input images. While batch normalization and residual connections contributes to faster convergence and the ability to train deeper networks more effectively. Transfer learning proved invaluable in leveraging already existing knowledge. Enhancing our models performance with minimal additional training. Final evaluation accuracy:

CNN	ResNet	VGG19	DINOv2
84%	87%	86%	88%

4 Discussion

4.1 Analysis

The results obtained from this study demonstrate a progressive improvement in performance across the employed methods, with the combination of DINOv2 as a feature extractor and a simple fully connected classifier emerging as the most effective approach. This model exhibited superior performance compared to the other configurations evaluated.

One of the notable strengths of the DINOv2 + fully connected classifier model is its ability to achieve high accuracy on the training set, approaching 90%. This suggests that the model effectively learned the underlying patterns and representations within the training data, enabling it to make accurate predictions on the samples it was trained on.

Importantly, there were no indications of overfitting, a common issue in machine learning models where the model performs exceptionally well on the training data but fails to generalize to unseen data. In this study, the loss function continued to decrease steadily, and the training accuracy remained consistent with the validation accuracy, indicating that the model was able to effectively generalize and make reliable predictions on both the training and validation sets.

4.2 Pre-calculating features

One missing analysis was that of the computation time. Training the classifier with DINOv2 as a feature extractor takes more that 4 minutes for 10 epochs and a fairly small training dataset (there could be millions of images). This is running on a fairly good GPU (Nvidia Geforce 4060), so the computation time is too long.

To solve this problem, the idea is to apply the feature extractor once on each of the datasets, and save the results. Then, the features can be reloaded and fed to the classifier. It was implemented, but no pertinent results could be obtained as the loss was constant, indicating it had reached a local minima. Several attempts to correct this were made but showed no improvement.

5 Limitations

While the methods explored in this study demonstrate promising results in improving skin lesion classification, there are several limitations that should be acknowledged.

Dataset Limitations: The performance of the models is heavily dependent on the quality and diversity of the training data. The dataset used in this study, a subset of the 2018 ISIC challenge dataset, may not be representative of the full range of skin lesion variations encountered in real-world scenarios. Factors such as skin tone, lighting conditions, and image acquisition methods can introduce biases and affect the model's generalization capabilities. For example, the model could have a lot of trouble classifying lesions on people with darker skin tones, or if the photographs are taken from afar or very close, the lesion may appear smaller or larger than in other images.

Lack of Interpretability: While deep learning models can achieve high accuracy, they often lack interpretability, making it challenging to understand the decision-making process and the specific features the model is using for classification. This limitation raises concerns about model transparency and trust, particularly in medical applications where explainability is crucial. This will be further discussed in the Ethical section.

Potential for Overfitting: Despite the use of techniques like data augmentation and transfer learning, there is still a risk of overfitting, where the model becomes too specialized to the training data and fails to generalize well to unseen samples. With a small dataset like the one used here, very deep models cannot be used as they would be too strong for the small amount of images provided. This is why a ”light” version of DINOv2 was used.

Computational Resources: Training and deploying advanced deep learning models, particularly those with large architectures like DINOv2, can be computationally expensive and may require specialized hardware (e.g., GPUs). This limitation can make it challenging to deploy these models in hospitals or laboratories with low budgets and older equipment.

6 Ethical Considerations

The use of automated image classification systems in healthcare applications, such as melanoma detection, raises several ethical concerns that must be carefully considered.

Privacy and Data Protection : Dermatological images contain sensitive personal information, including identifiable features and medical details. Ensuring the privacy and protection of patient data is paramount. In this specific case, it might be unharmful, but one should always think of data leaking in such a context.

Transparency and Interpretability : As mentioned earlier, deep learning models often lack interpretability, making it difficult to understand the decision-making process and the specific features driving the classification. In medical applications, where human experts need to scrutinize and validate the model's decisions, explainable AI techniques should be explored to enhance transparency and build trust in the system. However, using "black box" systems like deep neural networks, there will always be unknown parts about the model's logic. Appropriate training and education should be provided to ensure effective collaboration between human experts and AI systems.

Accountability and Liability : In the event of misdiagnosis or incorrect classification by the automated system, questions arise regarding accountability and liability. Who should be responsible? Is there even someone accountable for that? One could think of the A-level grading fiasco in England 2020, which had huge consequences and raised this issue. Clear guidelines and regulatory frameworks should be established to determine responsibility and ensure appropriate measures are taken to address potential harm or adverse outcomes. The final evaluation should always be made by a human, while the AI is used as a tool.

7 Conclusion

In summary, this study evaluated advanced machine learning techniques like data augmentation, batch normalization, residual connections, and transfer learning using models like DINOv2 for automated dermatological image classification. The DINOv2 model with a simple classifier achieved high accuracy in distinguishing benign birthmarks from melanomas.

However, limitations such as dataset biases, lack of interpretability, and potential overfitting should be addressed. Ethical concerns around privacy, transparency, and accountability must also be carefully considered before deploying such systems clinically.

Further research is needed to develop reliable and responsible automated melanoma detection tools to improve patient outcomes.

References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.