Melanoma Detection with Convolutional Neural Networks: Evaluating Custom and Pre-Trained Models

Herman Olvik
Chalmers Student / Group 34
olvik@chalmers.se

Cornelia Swartling
Chalmers Student / Group 34
corswa@chalmers.se

Abstract

This report explores the use of convolutional neural networks (CNNs) for melanoma detection, focusing on the 2018 ISIC dataset. With melanoma incidence rising globally, early and accurate diagnosis is crucial for improving survival chances. Traditional diagnostic methods are primarily based on visual examinations, but machine learning, particularly CNNs, presents a promising alternative for faster and possibly more reliable diagnostics. Conducted as part of the DAT341 - Applied Machine Learning course at Chalmers University of Technology, this study evaluates various neural network architectures in PyTorch, including the effects of layer normalization, data augmentation, residual connections, and the application of a pre-trained ResNet-18 model. Despite facing constraints such as limited computational resources and a one-week deadline, the findings show that layer normalization combined with data augmentation led to a model accuracy of about 88.50%, while the pre-trained ResNet-18 model slightly outperformed custom models with an accuracy of 88.82%. The report concludes that machine learning techniques, especially layer normalization and data augmentation, significantly enhance melanoma detection accuracy, with pre-trained models showing promising results. However, limitations such as dataset bias towards fair-skinned individuals and restricted computational capacity were notable challenges.

1 Introduction

Melanoma, the deadliest skin cancer, is rising globally, mainly in people with frequent sunburns. Early detection and treatment greatly improve survival chances, emphasizing the need for quick, precise diagnosis. Key signs are irregularly shaped and colored lesions. Diagnosis typically requires a professional's visual examination (Gyllencreutz, 2023).

Recent advancements in machine learning (ML) have shown potential for enhancing melanoma detection, offering greater accessibility and efficiency. Recent reviews indicate ML, particularly convolutional neural networks (CNNs), can classify skin lesions with high accuracy, presenting a faster, possibly more reliable alternative to traditional diagnostics (Patel, 2023; Thomsen, 2019; Sharma et al., 2021).

This report examines different neural network architectures in PyTorch, aiming to develop a model that accurately identifies melanoma using the 2018 ISIC dataset from Austria and Australia. It will assess the performance-impact of layer normalization, data augmentation, residual connections, as well as transfer learning (using a pre-trained model).

This report, a compulsory task for the DAT341 - Applied Machine Learning course at Chalmers University of Technology, faces its main constraint in the one-week deadline. The limited computational capabilities of student-owned computers and GPUs, being the sole resources for model training within this report's scope, further exacerbate this challenge.

2 Technical solutions

2.1 Method

A baseline neural network model without special configurations such as layer normalization

or residual connections was implemented as a reference model. Then, specialized models for each configuration (layer normalization and residual connections) were implemented. All mentioned models utilize PyTorch's "nn.Module" interface. In addition, the pre-trained ResNet18 from PyTorch's torchvision package (PyTorch) was imported and configured.

The process by which these different architectures were later compared consisted of training each model over a pre-defined number of epochs on the 2018 ISIC dataset (International Skin Imaging Collaboration, 2018), and evaluating their accuracy scores on a provided validation set (provided as part of the assignment). Different data augmentation alternatives were then evaluated for all models. The dataset consisted of 3213 images each of melanoma and nevus lesions.

2.1.1 Custom Architectures

In the baseline CNN model, the architecture consisted of three convolutional layers ('conv1', 'conv2', 'conv3') followed by max-pooling ('pool'). ReLU activation functions were applied after each convolutional layer. The output of the last convolutional layer was flattened and fed into two fully connected layers ('fc1', 'fc1') for classification. The final layer ('fc1') outputted the predicted probabilities for each class (melanoma or nevus). In the Layer Normalization CNN, the architecture had the same structure, but with the addition of normalization layers ('ln1', 'ln2', 'ln3') being applied after each convolutional layer. In the Residual Connection model, the same architecture as the baseline was used as well, but with the addition of residual connections between the convolutional layers.

A stochastic gradient descent optimizer was applied to all custom models, with a learning rate of 0.001 as well as a momentum value of 0.9. This was implemented using the SGD class in PyTorch.

2.1.2 ResNet-18 Configuration

To tailor the pre-trained ResNet-18 model for binary classification, its final layer was modified to output for two classes instead of the original 1000, aligning it with the binary task. The training setup included the CrossEntropyLoss function, and the Adam optimizer, which is known for its efficient learning rate adjustments.

2.1.3 Image Resolutions

Inspired by outputs from the ChatGPT (GPT-4) language model, the project initially utilized images at a resolution of 224x224 for training and validation. However, after consulting recent studies, notably a publication on NCBI (Kaur, 2022), the decision was made to reduce the image resolution to 128x128. This change, based on empirical evidence suggesting comparable performance at lower resolutions, facilitated faster training cycles and allowed for a greater number of epochs to be executed, enhancing model refinement over time.

2.1.4 Data Augmentation Techniques

Image transformation techniques from the "torchvision.transforms" package were tested and evaluated through iterative experimentation. This included normalization (using standard ImageNet dataset values, with means of '[0.485, 0.456, 0.406]' and standard deviations of '[0.229, 0.224, 0.225]'), random cropping, horizontal flips with a probability of 0.5, random rotations of 30°, and color jittering (brightness, contrast, and saturation) were evaluated based on their effectiveness in increasing the final accuracy scores of the models.

3 Results

3.1 Custom Models

As for the accuracy assessment of different model configurations, the model that acquired the best accuracy was the one utilizing the layer normalization architecture, with random transformations applied to the training data only, and leaving the validation data untouched, apart from the resolution adjustment to 128 x 128. The transformations applied were random cropping, horizontal flipping and random rotations.

This model achieved an accuracy of about 0.8850, or 88,50%, for training epoch 158. Below is the plot of the model's loss and accuracy over the 200 epochs it was trained for, as well as the corresponding confusion matrix of the evaluation

part.

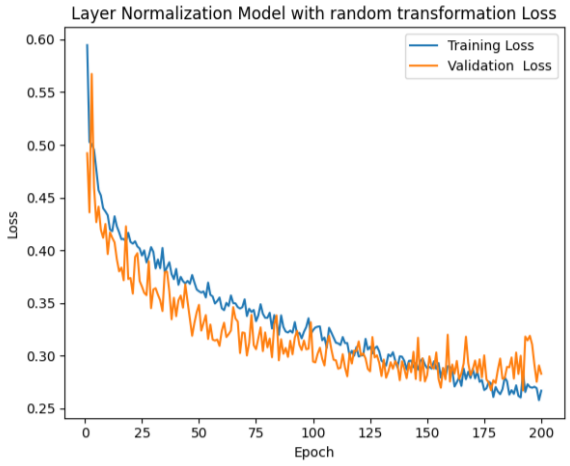

Figure 1: Figure 1. Loss Over 200 Epochs For The Best Custom Model.

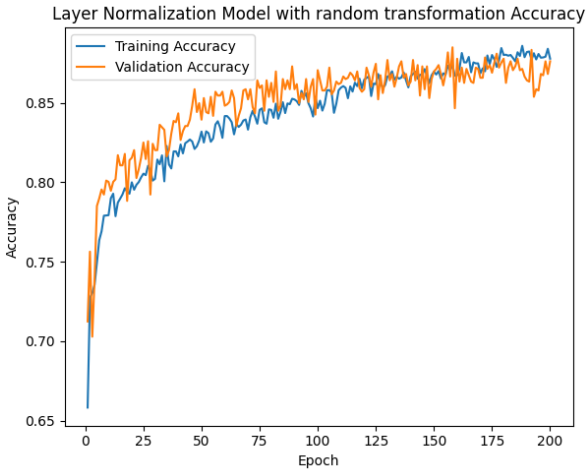

Figure 2: Figure 2. Accuracy Over 200 Epochs For The Best Custom Model.

According to the confusion matrix in Figure 3, the model is observed to be slightly more inclined to classify nevus, the harmless lesions, where about 13,3% of melanoma images were incorrectly classified as nevus. In comparison, only about 9,7% of nevus lesions were incorrectly classified as melanoma.

3.2 ResNet18 (Pre-trained Model)

Allowing the configured pre-trained ResNet18 model with its optional weights included to run for the same amount of epochs, resulted in an accuracy score of 88,82% on epoch 166. Its loss

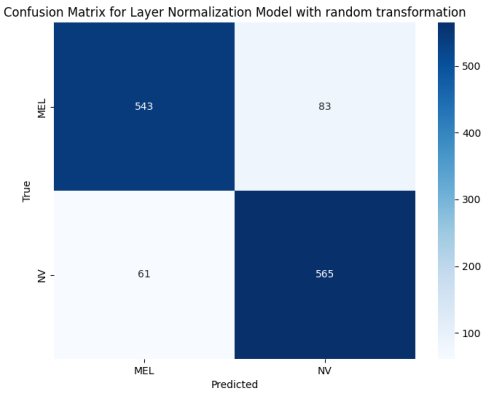

Figure 3.

Figure 3: Confusion Matrix For The Custom Model.

and accuracy is visualized in Figure 4.

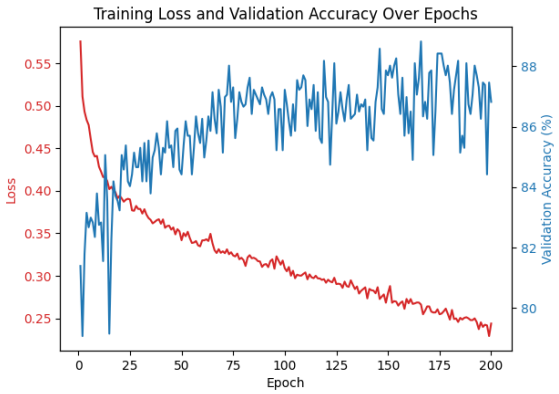

Figure 4.

Figure 4: Loss and Validation Accuracy over 200 Epochs for ResNet18.

Unfortunately, because of the limited time constraints of the project, the pretrained model was not analyzed further.

3.3 A demonstration of the effects of applying random transformations

An early version of a custom model that implemented residual connections was used to evaluate the impact of certain random transformations to the training data during the training process. This example in particular, highlighted the effects of applying randomized cropping for that specific model. For the model trained on the data without cropping, a notable decrease in performance

on the validation set is observed as the epochs progress.

Figure 5.

Figure 5: Loss and Validation Accuracy With Random Transformations Enabled

Figure 6.

Figure 6: Loss and Validation Accuracy With Random Transformations Disabled

3.4 Blind-test Evaluation

As part of the assignment, an informal classification competition was arranged by the course staff. Still within the initial experimentation phase of the project, an early model (trained for 70 epochs), was used to classify a blind test set. The generated labels were submitted to the course examiner, and received a reported score of 86,02%.

4 Conclusion

Based on the results of this report, it is concluded that augmentation through random transformations generally had a positive impact on model performance. So did layer normalization, which was selected as the best performing normalization architecture for this task, given that combinations or hybrids of such architectures were not considered. Residual connections surprisingly was not observed to have a big impact on the results of the custom architecture models, in contrast to the unbeaten performance of ResNet18.

The pre-trained model ResNet18 achieved the best overall accuracy score. Given that it is already incorporating a combination of residual connections and batch normalization, it possesses properties that were outside the scope of our custom architecture evaluation for this report. For our custom models approach, the one implementing residual connections did not match the one incorporating layer normalization which is considered a notable observation, given that residual connections is a defining feature of the ResNet18.

5 Limitations

5.1 Time Frame and Computing Power

Lowering image resolution to 128x128 pixels did cut processing time, but didn't fully mitigate the lengthy 13-hour training procedure. With the limited computational power of the students' own personal laptop computers, and the limited project timeline of one week, this combination hindered extensive experimentation. Hence, combinations of the custom model architectures (like residual connections and normalization techniques) were not evaluated.

5.2 Dataset Limitations

The dataset used (ISIC 2018) is heavily biased towards fair-skinned individuals. It had 6,426 training and 1,252 validation images. Whether this affected the models ability to classify skin lesions of darker skin tones was not explored.

5.3 Prior Experience and Knowledge

The students' limited prior experience with the tools used presented notable drawbacks, as it led to continuous recalibrations and shift of direction in the project's approach.

6 Ethical considerations

In the development of machine learning models for classifying skin lesions, ethical considerations play a crucial role. It's important to maintain informed consent, data privacy, and clarity about the model's functionality, biases, and associated risks.

Efforts to eliminate biases are essential to ensure fairness across diverse skin types and demographics. Applications built on these models should prioritize safety, avoiding discrimination. Respect for user dignity and cultural differences must underpin these technological solutions, aiming for inclusivity and preventing harm.

Finally, evaluating societal impacts and potential misuse is critical before deployment, necessitating dialogue with stakeholders to guide responsible use in healthcare applications.

References

Johan Dahlén Gyllencreutz. 2023. https://www.1177.se/Vastra-Gotaland/sjukdomar–besvar/cancer/cancerformer/malignt-melanom–hudcancer/ Melanom – hudcancer.

International Skin Imaging Collaboration. 2018. https://challenge.isic-archive.com/data/2018 ISIC Challenge. Accessed: 2024-03-17.

GholamHosseini H. Sinha R. Lindén M. Kaur, R. 2022. https://doi.org/10.3390/s22031134 Melanoma classification using a novel deep convolutional neural network with dermoscopic images.

Foltz E. Witkowski A. Ludzik J. Patel, R. 2023. https://doi.org/10.3390/cancers15194694 Analysis of artificial intelligence-based approaches applied to non-invasive imaging for early detection of melanoma: A systematic review.

PyTorch. https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html torchvision.models.resnet18. Accessed: 2024-03-17.

Ajay Nair Sharma, Samantha Shwe, and Natasha Atanaskova Mesinkovska. 2021. https://doi.org/10.1007/s00403-021-02236-9 Current state of machine learning for non-melanoma skin cancer. *Archives of Dermatological Research*.

Iversen L. Titlestad T. Winther O. Thomsen, K. 2019. https://doi.org/10.1080/09546634.2019.1682500 Systematic review of machine learning for diagnosis and prognosis in dermatology.