ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, СИСТЕМНИЙ АНАЛІЗ ТА КЕРУВАННЯ

DOI: 10.20535/kpisn.2024.1-4.301028
UDC 004.855.5

V.O. Nikitin*, V.Y. Danilov
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
*Corresponding author: nvo63911@gmail.com

TRANSFORMER VS. MAMBA AS SKIN CANCER CLASSIFIER: PRELIMINARY RESULTS

Background: Skin cancer is a deadly disease that claims tens of thousands of lives annually. Early diagnosis is crucial for successful treatment. Reliable diagnostic tools typically involve surgical methods, such as histological examination. However, issues arise when such methods are not feasible or desirable: for instance, the location of the lesion on the face, allergies to anesthesia, etc. This has led to active research in non-invasive methods, including those based on neural networks. This brings about the task of skin cancer image classification. Currently, and in our specific case, the models showing the best results in classification are Transformers. Nevertheless, these models have significant computational limitations due to quadratic scaling. In this study, a new machine learning architecture was explored proposed as an alternative to Transformers. Mamba is scaled linearly and demonstrates Transformer-like efficiency in various machine learning tasks.
Objective: The performance of two machine learning architectures was compared, Vision Transformer (ViT) and Mamba, for skin cancer classification using dermoscopic images. The goal is to investigate the classification results of both models.
Methods: A well-known benchmark in skin cancer classification, the HAM10000 dataset was used, which includes 10,015 dermoscopic images. The data to address issues such as class imbalance and normalized the images was prepared. Both models, ViT and Mamba, were pre-trained on the ImageNet dataset and fine-tuned for skin cancer classification. The models based on overall accuracy and F1-score for specific skin cancer classes were evaluated.
Results. The dataset for classification was processed. Using pre-trained weights of the two architectural variants, VMamba and ViT, they were fine-tuned on the proposed dataset. For quality assessment, accuracy and F1-score metrics were used. The results show that the ViT and Mamba models have similar overall accuracy, with Mamba models slightly better at classifying underrepresented classes such as Bowen's disease and dermatofibroma. Both models demonstrated high F1-scores in the case of melanoma, indicating their effectiveness in detecting this severe form of skin cancer.
Conclusions: The results indicate that Mamba is a viable alternative to ViT for skin cancer classification due to its similar accuracy. The application of VMamba could potentially make skin cancer diagnosis cheaper and more accessible due to its efficient scaling. Further research is needed to explore other variants of the Mamba model and to enhance its performance on larger datasets.
Keywords: machine learning; computer vision; skin cancer; transformers; spatial state models; VMamba.

Introduction

Skin cancer poses a significant challenge to healthcare systems worldwide. Over 1.4 million skin cancer cases were reported in 2020, of which 120,000 were fatal [1]. The key to successfully treating this disease lies in early detection – in most cases, a lesion can be removed with little to no consequences if caught in the early stages [2]. The gold standard in diagnostics is histology, which involves the surgical removal of the lesion. The problem is, there are cases when patients do not want the lesion removed for various reasons, ranging from aesthetic (removal leaves scars) to religious concerns. Recent advancements in image analysis have focused on using dermoscopic pictures for classification, which are easy and harmless to obtain. The integration of machine learning with these images has led to the development of effective diagnostic models. In particular, the Transformer architecture, originally designed for natural language processing, has shown promise in computer vision tasks, including skin cancer analysis in recent years. The Vision Transformer [3] has demonstrated significant performance

Пропозиція для цитування цієї статті: В.О. Нікітін, В.Я. Данилов, "Трансформер і Мамба для класифікації раку шкіри: попередні результати", *Наукові вісті КПІ*, № 1–4, с. 26–30, 2024. doi: 10.20535/kpisn.2024.1-4.301028

Offer a citation for this article: V.O. Nikitin, V.Y. Danilov, "Transformer vs. Mamba as skin cancer classifier: pre elementary results", *KPI Science News*, no. 1–4, pp. 26–30, 2024. doi: 10.20535/kpisn.2024.1-4.301028

in both benchmarks and real-world tasks [4, 5, 6], showing higher metric values than previously used Convolutional Neural Networks (CNNs). However, despite its advantages, the model has computational challenges, making it less accessible for processing high-quality images.

Recently, a new spatial state architecture known as Mamba [7] has been introduced. It is believed that Mamba could potentially address these issues. In this paper, a basic comparison of the effectiveness of ViT and a modified Mamba model for computer vision in diagnosing skin cancer is provided. This comparison aims to determine whether Mamba could be an efficient skin cancer classifier while requiring fewer computational resources.

Problem statement

The goal of this paper is to evaluate the effectiveness of the Mamba model for skin cancer diagnosis, using the ViT as a recognized and effective model for comparison. By comparing the results of these two models, we aim to determine whether the Mamba approach offers a solution of comparable effectiveness for the early detection of skin cancer through image analysis. Ultimately, this research seeks to contribute to the development of advanced diagnostic systems that can improve patient outcomes by facilitating timely and less invasive interventions.

Motivation

The ViT, introduced by Alex Dosovitskiy et al. in 2021, represents a significant advancement in adapting the Transformer architecture, originally designed for Natural Language Processing (NLP), to the realm of computer vision. At the core of the Transformer is the attention mechanism [8], which selectively focuses on the most salient parts of input data, enabling the model to capture critical features with high precision.

However, the attention mechanism also introduces a challenge: the computational complexity increases quadratically with the size of the input, as the attention scores must be calculated between all pairs of input vectors. This complexity becomes a bottleneck when dealing with high-resolution images, leading to increased resource requirements and reduced accessibility. Despite numerous attempts to optimize the algorithm, maintaining the balance between efficiency and performance remains a challenge. According to recent studies, such as the one by Tay et al. (2022) [9], achieving significant reductions in computational complexity often results in a trade-off with the model's performance, which is a key factor in the popularity of the Transformer architecture.

Natively, ongoing research seeks models that can solve the Transformer's computational problems while retaining high performance. One of the competitive architectures is State Space Models (SSMs). These models can generally be described by the next system of equations, where $x(t) \in R^n$ – variables, $u(t) \in R^m$ – inputs, $y(t) \in R^p$ – outputs, $A \in R^{m \times n}$ – state matrix, $B \in R^{n \times m}$ – control matrix, $C \in R^{p \times n}$ – output matrix:

$$\begin{cases} x'(t) = Ax(t) + Bu(t); \\ y(t) = Cx(t). \end{cases}$$

After a data discretization [10] step SSMs can achieve considerable performance in machine learning tasks. We find the latest advancements in this field, published in Mamba paper [7] very promising. While output items are still dependant on each other in a similar way to Recurrent Neural Networks (RNNs), on training step these dependencies (matrix A) can be precomputed allowing Mamba to train parallelly. Same time, with adding new block (token), the complexity of additional computation is not dependant on a size of a context that that was already processed (so not the way it works in Transformers). Summarising, authors were able to achieve Transformer's performance with model that scales linearly.

If the same approach works for the task of skin cancer classification that could potentially decrease the resources that are needed and therefore make final diagnostics cost more affordable. In this work we would like to do initial empirical validation of this assumption.

In order to apply Mamba to Computer Vision task we utilize its modification – VMamba [11]. In order to fit the data to the task it uses mechanism called 2D selective scan. The approach involves unfolding image patches into sequences along rows and columns and then scanning these sequences in four different directions: from top-left to bottom-right, from bottom-right to top-left, from top-right to bottom-left, and from bottom-left to top-right. Authors state that this process ensures that each pixel gathers information from all other pixels in various directions. It is important to note that the architecture used in this study, VMamba, is not the only Mamba variant for computer vision tasks. Currently, there is at least one other model known as Vision Mamba [12]. However, for this study VMamba was spe-

cifically chosen over Vision Mamba for a couple of reasons. Firstly, it demonstrated slightly better performance in classification tasks reported to the model's authors [11, 12]. Secondly, the additional feature of bidirectional layers in Vision Mamba was not proven to be effective, making this model seem unnecessary for now. Of course, this could change in future versions of the model, and we will be eagerly anticipating new improvements proposed by the community.

Methods

The HAM10000 [13] dataset is a well-established benchmark in the field of skin cancer classification, consisting of 10,015 dermoscopic images that include both malignant and benign lesions, with each image being cleaned and centered. It was decided to utilise this renowned dataset in order for our results to be easily reproducible for further research. In our study, a data preprocessing pipeline was implemented that integrated methods from previous research [6, 7] along with some novel techniques. The primary steps in our data preparation included:

· Deduplication: Duplicate images were eliminated from the dataset to ensure that each sample was unique. Each record of the dataset is marked with lesion id field by which duplicates can be identified.

· Normalization: The pixel intensities of the images were normalized to standardize the input data, facilitating better convergence of the models. As a result, all pixels were transformed to 0–1 range.

· Oversampling: To mitigate the issue of class imbalance, we applied augmentation techniques such as rotations, flips, and zooms to artificially enhance the presence of underrepresented classes. We aimed to have at least 7,000 samples of each class, addressing the class balancing problem with that strategy.

· Test-Validation Split: Since the HAM10000 dataset provides separate training and test data, further the train data was divided into training and validation sets using a consistent random seed to ensure reproducibility and enable a fair evaluation of model performance. The training was stopped when the model showed no improvements in the validation loss metric for 3 consecutive epochs.

To maintain the integrity of our results, we ensured uniformity in the use of augmented images and the train-test-validation split across all experiments. Each model underwent training for up to 20 epochs, with early stopping implemented to terminate training if there was no improvement in the model's performance beyond a threshold of 10^{-2} over three consecutive epochs. This strategy was employed to prevent overfitting and reduce the usage of the use of computational resources. Additionally, different learning rates were explored to enhance model convergence and prevent overlooking the optimal solution. On average, each model was trained for approximately 15 epochs, all with learning rates from 10^{-6} to 10^{-8}.

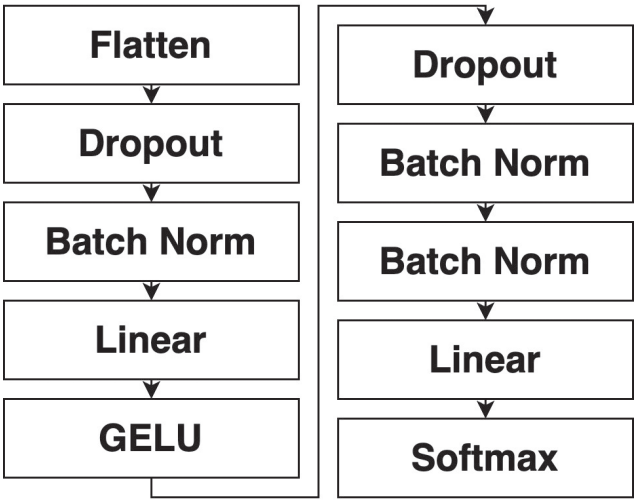

Fig. 1. Classification block used for the task

Both models were pretrained on the ImageNet dataset by other research teams [5, 6] for ImageNet classification and fine-tuned by us for the specific task at hand. Our custom classifier comprised a sequence of layers – Flatten, Dropout, Norm, Linear, GELU, Dropout, Norm, Linear, Softmax – which can be seen in Fig. 1. Three distinct learning rates were tested for each model and here presented only the ones with top learning rates. The classification block is also a result of experiments. The key metrics we focused on were overall accuracy and F1 scores for three classes: Actinic Keratosis, Intraepithelial Carcinoma, and Dermatofibroma (as the most underrepresented classes), as well as Melanoma (as the most severe and dangerous type of the disease).

Results

Analyzing the data presented in Table 1, it is apparent that both the larger ViT models and VMamba models exhibit comparable performance in terms of overall accuracy. The ViT_L_32 model, utilizing the imagenet1_kv1 weights, shows a training

***Table* 1.** Fine Tuning Results

Model	Weights	Train Acc	Top-1 Acc	F1 AKEIC	F1 Df	F1 Mel
vit_l_32	imagenet1_v1	0.98	0.76	0.49	0.64	0.88
vit_l_16	imagenet1k_swag_linear_v1	0.85	0.8	0.44	0.75	0.9
VMamba-T	vssmtiny_dp02_ckpt_epoch_258	0.95	0.8	0.6	0.76	0.9
VMamba-B	vssm_base_0229_ckpt_epoch_237	0.91	0.77	0.65	0.65	0.87

accuracy of 0.98 and a commendable top-1 accuracy of 0.76. This suggests the lowest generalizability of the model to new data if compared to every other model trained in this research. As for F1 scores for the targeted classes, VMamba models show a slightly superior performance, particularly in classifying Actinic Keratoses and Intraepithelial carcinoma / Bowen's Disease (AKIEC), with the VMamba-T and VMamba-B achieving an F1 score of 0.6 and 0.65 for AKIEC, which is notable considering the underrepresented nature of this class.

For Dermatofibroma (Df), a rare category, both models perform well, but the ViT_L_16 model, with the imagenet1k_swag_linear_v1 weights, lags slightly behind with an F1 score of 0.75 compared to the VMamba-T's 0.76. This indicates VMamba's potential in handling classes with fewer training samples more effectively. In the crucial case of Melanoma (Mel), which is of the highest concern due to its severity, both models deliver high F1 scores, with VMamba-B and ViT_L_16 both achieving score 0.9, demonstrating their robustness in identifying this dangerous form of skin cancer.

Summarizing, while both ViT and VMamba architectures demonstrate high efficacy in skin cancer classification tasks, VMamba models, in particular, show promise in their handling of specific, less represented classes without compromising on the detection of more severe conditions such as Melanoma. These findings underscore the potential of using these models in clinical settings, where they should be less expensive while being as effective as ViT based ones.

Conclusions

There is a problem in non invasive skin cancer classification where top performance modules happen to be computationally expensive. Because of quadratic scaling, implementing ViT in diagnostics could cause a drastic increase in costs, as higher image resolutions lead to higher diagnostic prices. To make early diagnostics an option, it must be financially affordable and therefore ViT is not a perfect fit for the task. In this research, we aimed to provide initial evidence of the effectiveness of a novel Transformer competitor, Mamba, in the area of skin cancer classification.

Our research lays a solid base for further exploration of VMamba for medical diagnostics, and oncology in particular. Our results suggest that VMamba can be just as effective for the task as Transformer is, which in combination of cheaper computation makes VMamba a good fit for medical diagnostics. We hope further researchers will explore the Mamba model more extensively. We assume the next vector should be: reconsidering classification head and different learning strategies, exploring other Mamba's Computer Vision modifications and fine tuning model on higher volumes of data. More advanced researched may also consider results of VMamba's cross scan module in order to better fit this mechanism to dermoscopy images which should, theoretically lead to better model's performance. Hardware aspect of task has some space for exploration as well. The exact resource ratios of the models were not assessed, as such a comparison is complex and beyond the scope of this paper, especially since Mamba is already established as more resource-efficient than Transformer-based models.

References

[1] World Cancer Research Fund International, "Skin cancer statistics", 2022. Available: https://www.wcrf.org/cancer-trends/skin-cancer-statistics/. [Accessed: 28-Mar-2024].

[2] A.F. Jerant, J.T. Johnson, C.D. Sheridan, and T.J. Caffrey, "Early detection and treatment of skin cancer", *Am. Fam. Physician*, vol. 62, no. 2, pp. 357–368, Jul. 2000. Available: https://www.aafp.org/pubs/afp/issues/2000/0715/p357.html. [Accessed: 28-Mar-2024].

[3] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929v2, 2020. Available: https://doi.org/10.48550/arXiv.2010.11929. [Accessed: 28-Mar-2024].

[4] V. Nikitin, and N. Shapoval, “Vision Transformer for Skin Cancer Classification”, *Scientific Collection “InterConf+”*, no. 33 (155), May 2023, pp. 449–60. Available: https://doi.org/10.51582/interconf.19-20.05.2023.039. [Accessed: 28-Mar-2024].

[5] G. Yang, S. Luo, and P.A. Greer, “A Novel Vision Transformer Model for Skin Cancer Classification”, *Neural Process. Lett.*, vol. 55, pp. 9335-9351, 2023. Available: https://doi.org/10.1007/s11063-023-11204-5. [Accessed: 28-Mar-2024].

[6] C. Xin et al., “An improved transformer network for skin cancer classification”, *Comput. Biol. Med.*, vol. 149, p. 105939, 2022. Available: https://doi.org/10.1016/j.compbiomed.2022.105939. [Accessed: 28-Mar-2024].

[7] A. Gu, and T. Dao, “Mamba: Linear-time sequence modeling with selective state spaces”, arXiv preprint arXiv:2312.00752, 2023. Available: https://doi.org/10.48550/arXiv.2312.00752. [Accessed: 28-Mar-2024].

[8] A. Vaswani et al., “Attention is all you need” arXiv preprint arXiv:1706.03762, 2017. Available: https://doi.org/10.48550/arXiv.1706.03762. [Accessed: 28-Mar-2024].

[9] F.D. Keles, P.M. Wijewardena, and C. Hegde, “On the computational complexity of self-attention” arXiv.org, 2022. Available: https://doi.org/10.48550/arXiv.2209.04881. [Accessed: 28-Mar-2024].

[10] A. Gu, K. Goel, and C. Rй, “Efficiently modeling long sequences with structured state spaces”, arXiv.org, 2022. Available: https://doi.org/10.48550/arXiv.2111.00396. [Accessed: 28-Mar-2024].

[11] Y. Liu et al., “Vmamba: Visual state space model”, arXiv preprint arXiv:2401.10166, 2024. Available: https://doi.org/10.48550/arXiv.2401.10166. [Accessed: 28-Mar-2024].

[12] L. Zhu et al., “Vision mamba: Efficient visual representation learning with bidirectional state space model”, arXiv preprint arXiv:2401.09417, 2024. Available: https://doi.org/10.48550/arXiv.2401.09417. [Accessed: 28-Mar-2024].

[13] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”, Sci. Data, vol. 5, 180161, 2018. Available: https://doi.org/10.1038/sdata.2018.161. [Accessed: 28-Mar-2024].

В.О. Нікітін, В.Я. Данилов

ТРАНСФОРМЕР І МАМБА ДЛЯ КЛАСИФІКАЦІЇ РАКУ ШКІРИ: ПОПЕРЕДНІ РЕЗУЛЬТАТИ

Проблематика. Рак шкіри – це смертельне захворювання, яке щороку забирає життя десятків тисяч людей. Ключовим елементом успішного лікування є рання діагностика. Типовим інструментом надійної діагностики є методи, що потребують хірургічного втручання, як, наприклад, гістологічне дослідження. Проблема постає, коли застосування таких методів не є можливим чи бажаним: розміщення утворення на обличчі, алергія на анестезію тощо. Через це активно ведеться дослідження неінвазійних методів, зокрема на основі нейронних мереж. Так постає завдання класифікації зображень раку шкіри. Моделі, які показують найкращі результати у класифікації наразі – трансформери. Тим не менше цей тип моделей має значні обчислювальні обмеження – квадратичне машстабування. У цій роботі ми досліджуємо нову архітектуру машинного навчання, що була запропонована як альтернативна трансфорерам. Mamba масштабується лінійно і демонструє схожу до трансформерів ефективність у низці задач машинного навчання.

Мета дослідження. Ми порівняли ефективність двох архітектур машинного навчання – Vision Transformer (ViT) та Mamba – для класифікації раку шкіри за допомогою дермоскопічних зображень. Метою є дослідження результатів класифікації двома моделями.

Методика реалізації. Ми використали набір даних HAM10000, відомий бенчмарк у класифікації раку шкіри, що включає 10 015 дермоскопічних зображень. Ми підготували дані для вирішення проблем, таких як дисбаланс класів, і нормалізували зображення. Обидві моделі, ViT та Mamba, були попередньо навчені на наборі даних ImageNet та допрацьовані для класифікації раку шкіри. Ми оцінили моделі на основі загальної точності та F1-score для конкретних класів раку шкіри.

Результати дослідження. Ми обробили набір даних для класифікації. Взявши заздалегідь натреновані ваги двох варіантів архітектур VMamba та ViT, ми донавчили їх на запропонованому наборі. Для оцінювання якості ми використовували значення accuracy та F1-score. Результати показують, що моделі ViT та Mamba мають схожу загальну точність, при цьому модель Mamba трохи краще класифікує менш представлені класи, такі як хвороба Боуена та дерматофіброма. Обидві моделі продемонстрували високі значення F1-score у випадку меланоми, що свідчить про їхню ефективність у виявленні цієї важкої форми раку шкіри.

Висновки. Результати свідчать, що Mamba є справжньою альтернативою ViT для класифікації раку шкіри через схожу точність. Застосування VMamba у перспективі могло б зробити діагностику раку шкіри дешевшою та більш доступною через ефективне машстабування. Надалі дослідження потрібні для вивчення інших варіантів моделі Mamba та для доопрацювання її продуктивності на більших наборах даних.

Ключові слова: машинне навчання; комп’ютерний зір; рак шкіри; трансформери; просторові моделі стану; VMamba.

Рекомендована Радою
НН інституту прикладного системного аналізу
КПІ ім. Ігоря Сікорського

Надійшла до редакції
1 квітня 2024 року

Прийнята до публікації
10 вересня 2024 року