A Vision Transformer with Adaptive Cross-Image and Cross-Resolution Attention

Benjamin A. K. Murray[1][0000−0002−2413−923X], Wei R. Tan[2][0000−0002−5076−0206], Liane S. Canas[1][0000−0002−2553−1284], Catherine H. Smith[2][0000−0001−9918−1144], Satveer K. Mahil[2][0000−0003−4692−3794], Sebastien Ourselin[1][0000−0002−5694−5340], and Marc Modat[1][0000−0002−5277−8530]

[1] Biomedical Engineering & Image Sciences, King's College London, London, SE1 7EH, UK

[2] St John's Institute of Dermatology, Guy's and St Thomas' NHS Foundation Trust and King's College London, London, UK

Abstract. Vision Transformers (ViTs) are the current state-of-the-art in deep learning for computer vision tasks. They are trained on vast datasets and are capable of useful downstream tasks through clever use of the attention mechanism.

The biggest limiting factor for ViTs is the number of pixels and tokens that can be processed in a given pass. Memory constraints on both patch size and the number of patches mean that ViTs are most effective at processing relatively low-resolution images.

Whilst ViTs can attend very flexibly across an image, attending across images in a naive fashion requires memory proportional to the square of the number of images. This is a further limiting factor. Given the task of automated assessment of psoriasis severity, a chronic skin condition that can affect large portions of a person's skin, it is necessary to look across multiple images and at fine detail in large images.

We present a method that adapts ViTs to a two-stage design that allows for the regression of a patient's psoriasis score across multiple images and resolutions and shows its effectiveness relative to a baseline ViT.

The implementation of our method is available at `https://github.com/KCL-BMEIS/multivit.git`.

Keywords: Vision Transformer · Multi-image · Multi-resolution · Psoriasis

1 Introduction

Psoriasis is a common, incurable inflammatory skin disease associated with multimorbidity and reduced life expectancy. It affects 2 million individuals in the UK, and costs the NHS over £750 million (estimated) annually. Research has successfully delivered several new psoriasis therapies over the past 2 decades [14], but disease severity and evolution over time are still monitored through clinical evaluation of skin lesions, which has well-recognised limitations.

The Psoriasis Area Severity Index (PASI) [7] is the gold standard for disease severity assessments. It integrates area of psoriasis involvement with *erythema* (redness of skin), *induration* (thickening/hardening of skin) and *scaling* of psoriasis plaques. It is measured by clinicians and ranges from 0 (no disease) to 72. A PASI of 10+ is considered severe. Physicial Global Assessment (PGA) is another measure used to grade psoriasis. It is an overall assessment of all psoriasis lesions on the body according to a Likert scale, which is a 6 point score between 0 (no psoriasis) and 5 (severe psoriasis). PGA is correlated with PASI.

Both PASI and PGA are time-consuming, highly subjective and poorly reproducible (low intra- and inter-rater consistency). Assessments rely on face-to-face contact between clinician and patient, which may not be deliverable with increasingly limited healthcare resources.

Deep-learning-based computer vision has the potential to automate image-based assessment of psoriasis severity. Initially created for language processing, the state of the art transformer [12] architecture has been adapted to vision tasks [5] with great success. Vision Transformers (ViTs) use image patches converted to tokens rather than pixel values in isolation. This is done for two reasons. Firstly, a pixel conveys very little semantic information, but a patch of pixels can convey significant semantic information. Secondly, attention is $O(T^2)$ in memory usage for T tokens in a standard transformer architecture, as all pairwise patch interactions must be considered. A position encoding is added to each of the patches, and can be learned depending on the architecture, so that the network can understand the spatial relationship between patches.

Transformers use Scaled Dot-Product attention; a mechanism that allows ViTs to learn the relationship between image patches. ViTs typically use self-attention, meaning patches of the same image attend to each other and to a class token that captures image-level semantics and is typically used as the input to class-level task heads. The class token allows ViTs to be trained to pay attention to the parts of an image most important to the model task, or even to use the patch to class attentions to act as semantic segmentation with only image-level labels for training [2].

ViTs require images to be of a given resolution and trained at or fine-tuned to a desired resolution. ViTs are configured to a given number of patches that is typically constrained by memory usage, given that attention is $O(T^2)$ in memory for T patches. A typical patch size in pixels is around 16^2 to 32^2 as larger patches require larger token memory to properly capture the information in a patch. As such, ViTs generally handle image sizes between 256^2 and 512^2 pixels. Commodity cameras have resolutions in the thousands of pixels squared, so fine-grained detail may not be readily detectable by a ViT.

Given the nature of psoriasis, standard ViT architectures cannot perform at their full potential. Psoriasis tends to occur across much of a patient's body rather than being limited to a single area, and is a heterogeneous condition with the pattern of presentation varying widely between individuals. A patient's full presentation of psoriasis is not provided by a single image in the general case, and many of the images are whole body images of a high resolution that must be downsampled to a much lower resolution to fit in a ViT architecture, necessitating a dramatic drop in detail.

Skin type is a confounding factor automated psoriasis assessment. The Fitzpatrick Skin Type (FST) [6] puts skin tones into six categories. Automated assessment for people with FST V and VI is a challenge as psoriasis does not present as distinctly in images as it does for FST I to IV, and datasets tend to have few people with FST V and VI.

In order for ViTs to pay attention across multiple images, images can be packed into a single composite image. For N images having T patches each, the memory requirement is $O((NT)^2)$ as every image patch must attend to every other image patch across the images. This typically necessitates reducing the resolution of the images or using fewer patches per image. For biomedical imaging, in which fine-grained texture is generally important in the assessment of a condition, this might negate any advantage of being able to attend between images.

We present a novel method, MultiViT, with a mechanism that adaptively focuses on regions across multiple related images in order to focus on important features and thus capture a person's whole psoriasis presentation. Mechanisms to increase resolution for the most clinically relevant subsections of the images are also presented. We demonstrate its effectiveness in the prediction of PASI scores given images.

2 Related Work

Multiple Instance Learning is a technique that attempts to learn across multiple related images, and approaches such as [9] use earlier forms of attention to look across multiple images.

SparseViT [3] is a shifted window (SWIN) [10] transformer-based model developed to efficiently parse high resolution images with low latency. SparseViT is designed to skip less important regions of a high-resolution image via the $L2$ norm of each window activation to focus on key image areas (for example the detection of pedestrians for self-driving cars). SparseViT is focused on mitigating latency with minimal performance loss, whereas the goal for MultiViT is to maximise auto-diagnostic power, as latency is not a significant factor in our clinical pipeline.

NaViT [4] is a method to improve training time for large scale networks with large scale datasets during pre-training. It allows ViTs to be trained on multiple images of differing aspect ratios. During training, a set of images is packed into a set of tokens and masked to prevent attention between images. They use a novel position encoding to avoid deforming images of different aspect ratios that factorizes position in to x, y components, meaning that image patch grids can be adapted to any aspect ratio. In contrast to NaVit, MultiViT allows attention between images to exploit relationships between them. MultiViT also avoids windowed attention as we want to maximise the use of the attention mechanism on each image.

3 Method

MultiViT is a novel adaptation of the ViT architecture that enables memory-efficient adaptive attention across multiple images. It does so via two mechanisms. Firstly, MultiViT implements adaptive attention with a fixed patch budget, meaning that for N images, each having T patches, it uses $O(NT^2)$ memory, rather than $O((NT)^2)$ memory for a vanilla ViT. Secondly, MultiViT can optionally focus on critical tokens at an increased resolution, narrowing its field of focus but enabling it to better exploit fine-grained resolution where required.

MultiViT is composed of two ViT stages that are connected together by our *Attention Filter* module. Each stage of the MultiViT architecture consists of a ViT model backbone that has been pre-trained on a large, general-purpose dataset. One or more attention heads are added to the class token output of the ViT model and used to fine-tune the model on image-level labels.

The MultiViT architecture is depicted in Figure 1 and consists of the following stages:

Stage 1 is executed on a set of N related images (from a given patient). Each image execution generates tokens and class-to-patch attentions on a per-image basis based on the task head losses for *Stage 1*.

The *Attention Filter* module ranks, selects and composes the most influential tokens from the *Stage 1* executions. It creates a composite hidden state and composite image from the image patches corresponding to the tokens kept from *Stage 1*.

Stage 2 is a slightly modified ViT that takes the composite hidden state and composite image as its inputs, and calculates a final task score using its own task heads. It has a *Token Concatenator* module that can merge the composite tokens with tokens embedded from the composite image.

3.1 Attention Filter

The *Attention Filter*, shown in figure 1 (a) creates the composite hidden state and composite high-resolution image from the tokens output by the executions of *Stage 1* as follows:

The *Attention Scorer* ranks tokens for all N input images output by *Stage 1* by their activation strength relative to their corresponding class token. The T most highly ranked tokens are selected and the rest discarded.

The *Mapping Generator* generates a mapping for the selected tokens from their source hidden states to the composite hidden states and a mapping for the corresponding image patches that the tokens map to in the high resolution images.

The *Token Mapper* takes the *Stage 1* hidden states and the token mapping and creates a composite hidden state. Given a resolution multiplier of ratio R, each patch token from *Stage 1* maps to $R \times R$ input patch tokens for *Stage 2*. Class tokens from *Stage 1* are discarded.

The *Patch Mapper* takes the image patch mapping and uses it to select image patches from the high-resolution version of the images. The resulting composite

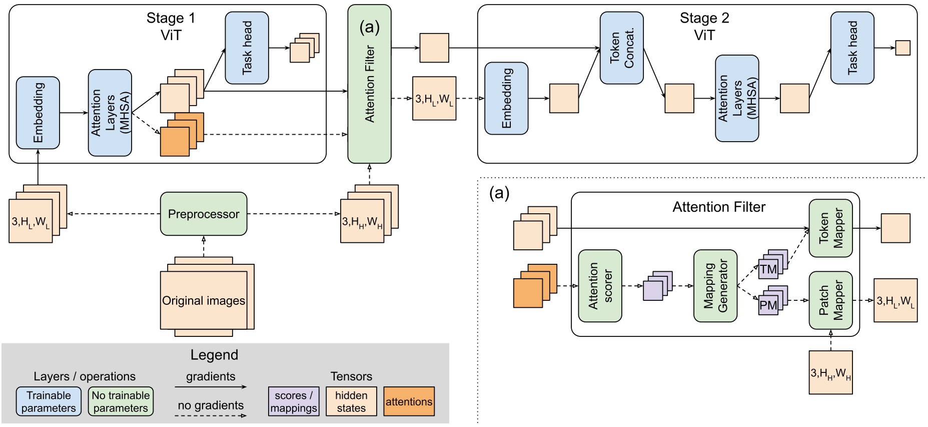

Fig. 1. The MultiVit architecture. MultiViT consists of three stages. *Stage 1* is a standard ViT that is executed on multiple images, sequentially. The *Attention Filter* (a.) takes the resulting patch tokens and attentions from *Stage 1*, ranks those tokens by importance to the *Stage 1* class token and then generates mappings for the T most important tokens. The *Attention Filter* generates a composite hidden state of surviving tokens and the corresponding compound image. These are then input to a modified *Stage 2* ViT.

image is passed through the *Stage 2* patch embedder to generate a set of high-resolution tokens.

Critically, because tokens selected by the *Attention Filter* are passed to *Stage 2*, gradients from *Stage 2* also inform *Stage 1*, and the overall network learns from the task heads of both stages concurrently.

The *Token Concatenator* added to *Stage 2* takes the hidden state from *Stage 1* and concatenates it with the hidden state from the embedded high-resolution image patches. This is done by concatenating the two sets of tokens along the hidden axis and then passing them through a *Linear* layer to reduce them back to the standard hidden size. These resulting tokens are then passed to the rest of the *Stage 2* ViT.

3.2 MultiViT Configurations for Ablation

MultiViT has three configuration parameters, *Mode*, *N*, and *R*.

Mode controls whether patch tokens are passed between *Stage 1* and *Stage 2*. In *Multi* mode, this is enabled, but in *Partial* mode it is disabled. *N* is a positive integer value that controls how many images are being attended across for each sample. *R* is a positive integer value that sets the resolution multiplier for *Stage 2*.

Mode and *R* determine in combination whether patch tokens and image patches are output from the *Attention Filter*. For *Multi-N-1* the *Attention Filter* outputs only composite patch tokens. For *Multi-N-2*, it outputs both composite patch tokens and a composite (high-resolution) image. For *Partial-N-1* and *Partial-N-2*, it only outputs a composite image to *Stage 2*.

The *Token Concatenator* is only used if both composite tokens and a composite image are required by the configuration.

4 Experiments

4.1 Dataset

The dataset used here is an internal clinical dataset gathered from consenting psoriasis patients. It is comprised of 1109 color images of 152 patients with demographics and clinical information. 763 images are professional medical photographs and 346 are self-taken photographs. Images are de-identified by digitally removing heads and features such as tattoos. Demographic and clinical data includes FST, PASI, and PGA.

Images range from 480 to 6016 pixels in width and 480 to 4640 pixels in height, with aspect ratios between $1 : 0.45$ and $2.41 : 1$. The dataset is augmented by randomly cropping up to 10% of the borders of each image. This is done before resampling the data to $(384, 384)$ for low resolution images and $(384 \times R, 384 \times R)$ for high resolution images so that the augmentation is consistent for low and high resolution images. 80% of the patients are used for the training fold and 20% kept as the test fold and never trained upon.

The dataset is heavily biased toward FSTs of I-IV. In the training fold, there are 104 patients with FSTs I-IV and 18 with FSTs V-VI. In the test fold, there are 24 patients with FSTs I-IV and 6 with FSTs V-VI.

4.2 Model

The model uses the HuggingFace [13] *ViTModel* as a donor architecture, with configuration and starting weights from *'google-vit-large-patch32-384'*. This model accepts images of 384×384 and has a patch size of 32, meaning there are 12×12 patches per image. Each model has 23 multi-head attention blocks with 16 channels and a hidden size of 1024. The model uses absolute positions that are learned during training. The model weights have no task heads; these are trained during our fine-tuning step.

We make use of two values from the dataset when training the model PASI and clinical severity. Two scores are provided for each patient, we use the score from the first rater throughout. The task heads are configured to be Multi Layer Perceptrons [11] with a hidden layer of size 128 or 256. The two configurations used during the experiments are (PASI(128) & PGA(128)) and (PASI(256) & PGA(128)).

The model is trained for 300 epochs. Fine-tuning is employed, with the embedding layers, the last 5 attention layers, and task heads unfrozen. Weight decay is set at 0.01 and the hidden layer dropout is set to 0.1. Learning rate is initialised to $1e - 4$ and linearly decays over 300 epochs to $1.25e - 7$.

5 Results and Discussion

Tables 1 and 2 show the performance of the different network configurations in terms of Mean Absolute Error (MAE) and Mean Signed Error (ME) for the PASI regression task.

Table 2 shows that *Baseline* is outperformed by all four configurations of MultiViT, coming last or second to last in all evaluation metrics.

Multi-3-1 has the best MAE for PASI and is within clinically acceptable inter-rater variance over the dataset as a whole, taken to be an MAE of 4.67 from 43 dermatologists reported by [8]. It has the second best MAE over FSTs I to IV.

All configurations tend to under evaluate PASI, although MultiViT improves this tendency over the *Baseline* configuration. Figure 2 presents this visually.

Interestingly, the MultiViT configurations with R set to 2 do not demonstrate a consistent improvement across the metrics evaluated. We note however that *Partial-3-2* significantly outperforms other configurations on FST V-VI for MAE and ME, and *Multi-3-2* has the best overall ME performance, indicating less tendency to under evaluate PASI.

Performance of all networks on FSTs V to VI indicates that more patients with these skin types are needed.

A direct comparison with the literature is complicated by lack of publicly available datasets and implementations. Current studies employ convolutional neural networks trained on professional medical photographs of clinical poses, demonstrating an MAE of 3.12 with a dataset of 14096 images from 2367 patients [8], and an MAE of 3.3 on a dataset of 2700 images of 60 patients [14] taken over treatment course.

Task heads	*Baseline*	*Multi-3-1*	*Multi-3-2*	*Partial-3-1*	*Partial-3-2*
PASI(128)+PGA(128)	(5.361) 5.210	4.024	4.860	4.836	5.242
PASI(256)+PGA(128)	(5.281) 5.127	**3.877**	4.220	4.674	4.516

Table 1. Mean Absolute Error (MAE) for various MultiViT configurations against the baseline ViT model. Lower scores are better. **Bold** entry indicates the best score. Entry in parentheses indicates the per-image MAE.

6 Conclusion and Further Work

Our proof of concept method highlights that ViTs can be adapted to effectively handle the challenges posed by psoriasis. Our MAE of 3.877 is within a reported inter-rater MAE of 4.67, despite the limited size of our dataset. Critically, the ability to learn from a mixture of self-taken photos and clinical photos may open new avenues for remote app-based psoriasis monitoring.

The following are identified as limitations of this work:

Metrics	*Baseline*	*Multi-3-1*	*Multi-3-2*	*Partial-3-1*	*Partial-3-2*
MAE (per-image)	5.281	—	—	—	—
MAE	5.127	**3.877**	*4.220*	4.674	4.516
MAE (I-IV)	3.968	*2.804*	**2.736**	3.402	4.283
MAE (V-VI)	9.763	*8.167*	10.158	9.761	**5.451**
ME	-3.192	*-2.402*	**-1.741**	-2.644	-2.721
ME (I-IV)	-1.717	-1.178	**0.117**	*-0.865*	-2.081
ME (V-VI)	-9.093	*-7.298*	-9.172	-9.761	**-5.281**

Table 2. MAE and Mean signed Error (ME), for the dataset as a whole and for FST I-IV vs. V-VI for Baseline and MultiViT configurations with PASI(256)+PGA(128) task heads. **Bold** indicates the best score for a given metric and *italics* indicates the second best score.

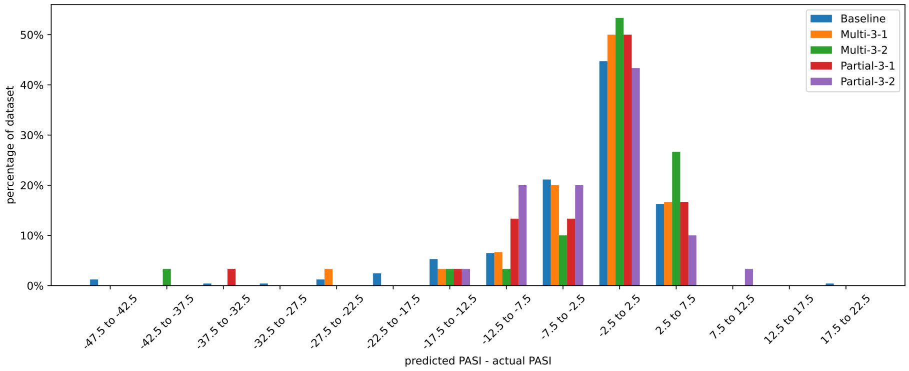

Fig. 2. Histograms of predicted PASI - actual PASI for the test fold of our psoriasis dataset. This figure highlights the tendency of all the different architectures to underestimate PASI.

Primarily, we are limited by the small size of our dataset and the lack of public psoriasis datasets with gold-standard annotations due in part to the identifiable and sensitive nature of skin photographs. We are in the process of greatly expanding our internal dataset, especially for FST $V - VI$.

We will further demonstrate this architecture on non-psoriasis datasets with public benchmarks. We are looking for other medical and non-medical tasks that benefit from efficient attention over multiple images. Additionally, many classification / regression tasks involve high resolution single images in which only small areas of the image are of importance, such as digital histopathology and dermoscopic images, and our method can be readily adapted to these.

There are a number of architectural improvements that we intend to make to MultiViT. These include but are not limited to adaptive patch geometry, attention scoring mechanism [1], a more general adaptive attention mechanism and the use of attention-based semantic segmentation to enhance token selection and model interpretability.

7 Acknowledgements

This research was supported by the Wellcome Engineering and Physical Sciences Research Council Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z) and the UK Research and Innovation London Medical Imaging and the Artificial Intelligence Centre for Value-Based Healthcare. The Jade Consortium, funded by the Engineering and Physical Sciences Research Council (grant EP/T022205/1) provided compute for model training. The dataset was collected as part of Psoriasis Association (grant BSTOP50/5 [to CHS]), the National Institute for Health and Care Research (NIHR Advanced Fellowship NIHR302258 [to SKM] and NIHR Senior Investigator Award [to CHS]), under 20/YH/0135 ethics. Ben Murray is an affiliate of the AI Institute at KCL.

Disclosure of Interests. C. H. Smith reports departmental research funding as an investigator in the European Union's Innovative Medicines Initiative consortia involving multiple industry partners (see biomap-imi.eu and hippocrates-imi.eu for details). S. K. Mahil reports departmental income from AbbVie, Almirall, Eli Lilly, Janssen, Leo, Novartis, Pfizer, Sanofi, and UCB, outside the submitted work.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4190–4197. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.385
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9650–9660 (October 2021)
3. Chen, X., Liu, Z., Tang, H., Yi, L., Zhao, H., Han, S.: Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2061–2070 (June 2023)
4. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., Oliver, A., Padlewski, P., Gritsenko, A., Lucic, M., Houlsby, N.: Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 2252–2274. Curran Associates, Inc. (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
6. Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types I through VI. Arch. Dermatol. **124**(6), 869 (Jun 1988)

7. Fredriksson, T., Pettersson, U.: Severe psoriasis–oral therapy with a new retinoid. Dermatologica **157**(4), 238–244 (1978)
8. Huang, K., Wu, X., Li, Y., Lv, C., Yan, Y., Wu, Z., Zhang, M., Huang, W., Jiang, Z., Hu, K., Li, M., Su, J., Zhu, W., Li, F., Chen, M., Chen, J., Li, Y., Zeng, M., Zhu, J., Cao, D., Huang, X., Huang, L., Hu, X., Chen, Z., Kang, J., Yuan, L., Huang, C., Guo, R., Navarini, A., Kuang, Y., Chen, X., Zhao, S.: Artificial intelligence–based psoriasis severity assessment: Real-world study and application. J Med Internet Res **25**, e44932 (Mar 2023). https://doi.org/10.2196/44932
9. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018)
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (October 2021)
11. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
13. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6, `https://aclanthology.org/2020.emnlp-demos.6`
14. Xing, Y., Zhong, S., Aronson, S.L., Rausa, F.M., Webster, D.E., Crouthamel, M.H., Wang, L.: Deep Learning-Based Psoriasis Assessment: Harnessing Clinical Trial Imaging for Accurate Psoriasis Area Severity Index Prediction. Digital Biomarkers **8**(1), 13–21 (03 2024). https://doi.org/10.1159/000536499