MedMLP: An Efficient MLP-like Network for Zero-shot Retinal Image Classification

Menghan Zhou[1], Yanyu Xu* [4], Zhi Da Soh[2], Huazhu Fu[1], Rick Siow Mong GOH[1], Ching-Yu Cheng[2,3], Yong Liu[1], Liangli Zhen[1]

[1] The Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.
[2] Singapore Eye Research Institute
[3] Singapore National Eye Centre
[4] The Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, 250100, P. R. China.

Abstract. Deep neural networks (DNNs) have demonstrated superior performance compared to humans across various tasks. However, DNNs often face the challenge of domain shift, where their performance notably deteriorates when applied to medical images with distributions differing from those seen during training. To address this issue and achieve high performance in new target domains under zero-shot settings, we leverage the ability of self-attention mechanisms to capture global dependencies. We introduce a novel MLP-like model designed for superior efficiency and zero-shot robustness. Specifically, we propose an adaptive fully-connected (AdaFC) layer to overcome the fundamental limitation of traditional fully-connected layers in adapting to inputs of various sizes while maintaining GPU efficiency. Building upon AdaFC, we present a new MLP-based network architecture named MedMLP. Through our proposed training pipeline, we achieve a significant 20.1% increase in model testing accuracy on an out-of-distribution dataset, surpassing the widely used ResNet-50 model.

Keywords: Zero-shot Setting · MLP-like Network.

1 Introduction

Deep learning has achieved remarkable success in various domains, often matching or even surpassing human performance in certain tasks [12,21,6,2,22]. However, the distribution gaps caused by differences in imaging devices and patients in medical images [18,4,19] can significantly degrade the performance of deep learning models trained on large-scale image datasets. Hence, there is an urgent need to explore methods to achieve high performance in new target domains.

This challenge has been extensively explored by researchers, leading to several promising solutions, including adversarial training, domain adaptation, and data

* Corresponding author

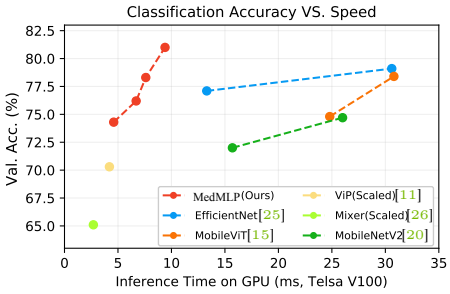

Fig. 1. Top-1 classification accuracy comparisons between the proposed MedMLP and state-of-the-art mobile models. The x-axis denotes the inference latency and the y-axis denotes the classification accuracy on the ImageNet validation dataset. We use different width multipliers (including 0.5, 0.75, 1.0, and 1.4) to trade-off between the model complexity and accuracy. MLP-like models are significantly faster than CNN-based models. The proposed MedMLP surpasses all these models with higher accuracy. The inference time in GPU is measured with Pytorch built-in profiler functions. More details can be found in Sec. 3.2.

augmentation [9,16,24,8,17]. The adversarial training and domain adaptation algorithms present challenges in real-world scenarios, particularly as predicting target domain data is difficult due to the tremendous number of unknown out-of-distribution data. Data augmentation is another line of research work. However, we've observed that the power of data augmentation algorithms diminishes when dealing with limited medical domain data. Many of these methods rely on transfer learning (TL) algorithms, which need to access to target datasets. However, accessing the target domain data in real-world scenarios is challenging, and additional fine-tunig of the models may be inefficient.

Recent work study on self-attention mechanisms [7,13,30,31] provides an alternative solution to boost the model's generalizability under zero-shot setting (no access to the target domain data). In particular, self-attention mechanisms in transformers achieve stronger robustness over conventional convolutional networks, indicating their essential role in improving visual recognition by capturing global dependencies among input patches. Concurrently, MLP-like models, such as MLP-Mixer andViP [26,11], utilize the simple fully-connected (FC) layers to encode spatial information of the input, which is efficient to capture global dependencies [23,7]. However, MLP-like models are only compatible with a fixed input resolution. When processing larger-resolution inputs, the model size must be increased correspondingly, leading to an unaffordable increase in computational memory ($\mathcal{O}(n^2)$).

To address this issue, we introduce a novel adaptive fully-connected (AdaFC) layer. We rethink their conventional design and utilization strategies to overcome these limitations. The key insight is that FC layer weights can be reused in a proper way to generate new weights of adaptive shapes for different inputs, rather than being fixed as in conventional designs. Specifically, the proposed

AdaFC takes the pre-defined weights as basis and learns to dynamically generate the weights of new shapes on-the-fly to adapt to the inputs of various shapes. Therefore, it is easy for AdaFC to process inputs with arbitrary resolutions while saving computational costs. We investigate multiple strategies for weight generation, including the basis size and selection, and present a simple and effective method for dynamic weight generation. Based on the proposed AdaFC, we design a new MLP-based network architecture, named MedMLP. MedMLP adopts the classic pyramid structure used in CNNs. The proposed AdaFC and the fully-connected layer are respectively used in each building block for spatial information encoding and channel information mixing. As shown in Fig. 1(a), MedMLP offers superior efficiency and accuracy compared to the state-of-the-art mobile CNNs, ViTs and recent MLP-like models. We conduct extensive experiments to evaluate the proposed MedMLP. Specifically, MedMLP achieves 20.1% higher accuracy on SINDI dataset than the ResNet-50 model with comparable model size and computations.

In short, we make the following contributions: We exploit the potential of MLP-like models to improve efficiency by developing a new family of MLP-like models, providing superior efficiency and zero-shot robustness. In particular, we propose a novel adaptive fully-connected layer that solves the fundamental limitation of traditional fully-connected layers in adapting to inputs of different sizes while maintaining GPU efficiency. Further, we propose a new MLP-like model, MedMLP, that significantly outperforms the previously widely used CNN-based MobileNetV2 and all the existing MLP-like models. We are the first to reveal the usability of MLP-like models under the mobile settings. In addition to competitive efficiency and accuracy, the proposed MedMLP offers stronger robustness than CNN-based models [20], making it more suitable for mobile applications.

2 Method

In this section, we describe in detail how the proposed adaptive fully-connected (AdaFC) layer works and show how it can be used to solve the above issue of the conventional fully-connected layer. Based on AdaFC, we also designed a new MLP-like model for mobile devices, named MedMLP.

2.1 Preliminary on All-MLP Network

Given an input image I, the MLP-like model (*e.g.*, MLP-mixer [26]) first uniformly partitions it into S patches (*a.k.a.* tokens). Each patch is linearly encoded into a C-dimensional feature vector and these encoded patches form a feature matrix X of size $C \times S$. The MLP-like model typically uses two all-MLP components, with weights $W \in \mathbb{R}^{S \times N}$ and $V \in \mathbb{R}^{C' \times C}$, to conduct the spatial token mixing and feature channel mixing to generate new features X':

$$X'_S = XW, \quad X' = VX'_S. \tag{1}$$

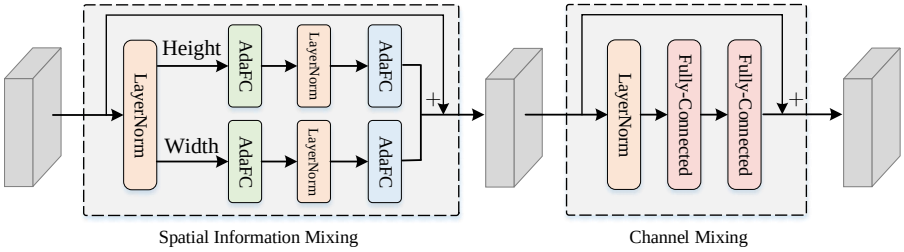

Fig. 2. Building block structure of the proposed MedMLP. Functionally, it consists of two MLP components, which aim at encoding spatial information and channel information, respectively. For spatial token mixing, it applies AdaFC to encode the information along the height and width dimension separately. For channel mixing, it adopts the standard FC layers.

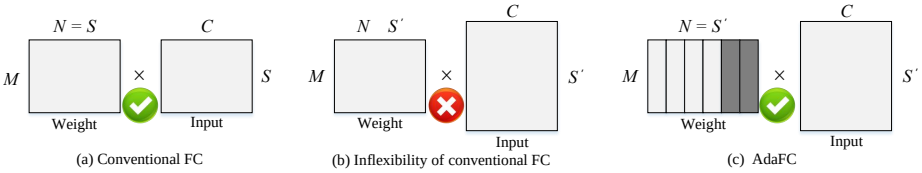

Fig. 3. Comparison to the standard fully connected layer. Given the weight matrix with shape $M \times N$, where N and M are the input dimension and the output dimension respectively. For an input tensor of shape $S \times C$, when we use the fully connected layer to encode the spatial dimension, S should be identical to N as shown in (a). However, when the input image resolution changes, *i.e.*, $S \rightarrow S'$ and $S' \neq N$ as shown in (b), the fully connected layer fails. Our AdaFC splits the weight matrix into multiple basis and uses them to generate new weights (deep gray boxes) according to the spatial dimension of the input to make the new weight matrix adaptive to the input, *i.e.*, $N = S'$ as shown in (c).

The all-MLP component provides global receptive field and less inductive bias than convolution kernels, which are benefiting the model's learning capacity. However, it also brings a severe limitation—the spatial resolution of the input must be fixed as S. Otherwise, the spatial mixing component has incompatible size with the input features and cannot be applied as illustrated in Fig. 3. To conquer this issue, we reconsider the way to conduct matrix multiplication in the fully-connected layer. Our design principles are based on the following two aspects: (1) It should be as simple as the standard fully-connected layer and no extra computation or memory overhead should be added; (2) It should be able to fit input tensors with arbitrary spatial sizes and the computations should be proportional to the spatial size of the input tensor. To meet both of the above principles, we propose our adaptive fully-connected layer in the following section.

Table 1. Architecture definition of AdaMLP-B0 model. We use 'h' to denote the number of heads and 'e' the expansion ratio in channel mixing MLP.

Stage i	Operator A_i	Resolution $H_i \times W_i$	#Filters C_i	#Layers L_i
1	PatchEmbed 4x4	224×224	32	1
2	AdaFC, e4, h1	56×56	42	2
3	AdaFC, e2, h2	28×28	56	4
4	AdaFC, e3, h4	14×14	96	4
5	AdaFC, e6, h8	14×14	112	4
6	AdaFC, e6, h32	7×7	224	4
7	Head & LayerNorm	1×1	1000	1

2.2 Adaptive Fully-Connected Layers

Let $M \times N$ be the shape of the weight matrix A, where N and M are the input and output dimensions of the fully connected layer, respectively. Let X be the input tensor of size $C \times S$, where S is the spatial dimension (*i.e.*, number of tokens) and C is the channel dimension. When we use the fully connected layer to encode the spatial dimension, if S is not identical to N, the weight matrix A does not match the input X and cannot be applied directly. Thus, the weight matrix A should be adjusted according to the spatial dimension S of the input X to make N equal S. To this end, we propose to take A as the basis for generating a new weight matrix A' via an adaptive weight generation process.

As shown in Fig. 3, we first divide A along the input dimension into G basis, yielding $[A_1, \cdots, A_G]$. The input dimension of each base weight thus becomes N/G. To make the input dimension of A match the input tensor X of spatial size S, the adaptive weight generation process generates SG/N different weight matrices A'_i of size $M \times N/G$ from the basis $[A_1, \cdots, A_G]$ as follows:

$$A'_i = \alpha_i A_{i,:,0:} + (1 - \alpha_i) A_{i,:,1:}. \tag{2}$$

Here α_i is the combination parameter for the i-th weight basis and is learned end-to-end. The generated weights W'_i with $i = 1, \ldots, SG/N$ are concatenated to form a new weight matrix $A' = [A'_1, \ldots, A'_{SG/N}]$ of size $M \times S$, which is thus compatible with the input of spatial dimension S. The generated new weights A', with input dimension S, is applied to the input for spatial encoding. A more detailed analysis of the exploration of the basis sampling strategy can be found in the supplementary material.

2.3 The MedMLP Model

The architecture of building blocks of the proposed MedMLP can be found in Fig. 2. It takes an image of arbitrary size $n \times n$ as input and uniformly splits it into a sequence of image patches (4×4). All the patches are then mapped into linear embeddings (or called tokens) using a shared linear layer as [26] followed by a layer normalization [1]. We next feed all the tokens into a sequence of Adaptive MedMLP block to encode both spatial and channel information.

The MedMLP Block. In classification, the input image resolution is often set to 224×224. Suppose the patch size is 16×16. The number of token, *i.e.*, the spatial size, should be $14 \times 14 = 196$. Such a large value will result in a large

Table 2. Zero-shot accuracy with scaling image resolutions. We compare the classification accuracy of MedMLP and MobileNetV2 when applying them directly to images of varying resolutions, without fine-tuning. Benefiting from the AdaFC, our proposed MedMLP can process images with varying resolutions and achieve higher accuracy than MobileNetV2 consistently.

Model	224×224	192×192	160×160	128×128
MobileNetV2	72.0	70.1	66.2	59.4
MedMLP-B0*	75.3 (**+3.3**)	73.5 (**+3.4**)	69.8 (**+3.6**)	63.3 (**+3.9**)

number of weight basis, consuming huge amount of memory. To mitigate this issue, we adopt a similar strategy to ViP [11] and encode the spatial information along the height and width dimension separately with the permutation strategy. Different from ViP, we further decompose the layer into two consecutive layers with a bottleneck structure. The channel mixing component is a normal MLP which consists of two fully connected layers with a non-linear activation. For spatial information mixing, we use two branches to encode the information along the height and width dimension, respectively, each of which has two AdaFCs. Suppose the input tensor has the shape of $C \times S'$. Without the bottleneck structure, to guarantee the spatial size of the output is still S', our basis sampling strategy should be applied to both the input and output dimensions of the weight matrix. This means the weight matrix would have a shape of $S' \times S'$ and the computation cost will be proportional to $C \cdot S' \cdot S'$, which is quadratic in S'.

3 Experiments

3.1 Implementation details

We use Pytorch for all model training. We use AdamW [14] optimizer with initial learning rate $1e^{-3}$ and weight decay of 0.05. We train the model for 300 epochs without cutmix and auto-augmentation, which are adapted by previous All MLP networks [26,27] reproduced in the timm [28] library. The reported results of MobileNetV2 are reproduced with the same training settings. When comparing with other SOTA models, we report the results with advanced training recipes with CutMix [29] and RandAug [5] added using same settings as previous methods [11,26,27].

3.2 Model Analysis and Ablation Studies

We first evaluate MedMLP's performance on natural image datasets since the pre-training is also an essential step for a good performance on medical datasets. We evaluate the performance of MedMLP on the ImageNet benchmark [6] and its variant ImageNet-Real [3], ImageNet-C [10], ImageNet-A and ImaegNet-R for the model's generalization performance under zero-shot settings.

Table 3. Top-1 accuracy comparison of our MedMLP with the recent MLP-like models on ImageNet [6] and ImageNet Real [3] ('Real'). All the models are trained without external data.

Networks	Param.	FLOPs	ImageNet (%)	Real (%)
ViP (scaled) [11]	6.7M	1.5B	70.3	78.4
MedMLP-B0 (Ours)	4.9M	0.6B	74.3	81.6
gMLP-tiny16	6.0M	2.7B	76.4	–
MedMLP-B1 (Ours)	8.4M	1.0B	76.2	83.0
Mixer-B/16 [26]	59.0M	11.6B	76.4	82.0
ResMLP-S12 [27]	16.0M	0.8B	76.2	83.5
MedMLP-B2 (Ours)	12.7M	2.1B	78.3	84.8

Scalability. The proposed MedMLP provides a modular architecture design by taking the AdaFC block as the basic block for spatial information encoding and the standard FC for channel information encoding. Such design makes it easier for MedMLP to scale up by removing the constraints of the strict match between the dimensions of the weights tensor and the feature tensor. To verify the effectiveness, we present four variants of MedMLP, termed -B0, -B1, -B2 and -B3, respectively, and evaluate their performance. The models are scaled based on -B0 model. Their performances on ImageNet are summarized in **Supplementary B**. By directly scaling up the model from -B0 to -B1, the accuracy of MedMLP can be improved from 74.3% to 76.2%, yielding a gain of 1.9%. Further scaling up the model to -B2 and -B3 results in 78.3% and 81.1% top-1 accuracy, respectively. These experiments indicate that our MedMLP indeed provides a series of MLP-like models that are not only applicable for mobile settings but also applicable for other settings with less computation resource constraint.

Zero-shot recognition on dynamic image resolutions. As mentioned above, it is impossible for existing MLP-like models, that rely on traditional fully-connected layers, to cope with images of various resolutions. Instead, the proposed MedMLP overcomes this limitation. Table 2 shows the zero-shot accuracy (without any model architecture change or fine-tuning) of our MedMLP and MobieNetV2 when classifying images of varying resolutions on ImageNet. As shown, MedMLP can deal with images of varying resolutions without fine-tuning the model and consistently perform better than MobileNetV2. The improvement over MobileNetV2 becomes gradually larger when the resolution goes lower. This implies the proposed dynamic weight generation approach in AdaFC effectively extracts distributed features from the inputs and enables the model to adapt well to scalable inputs.

3.3 Results on Zero-shot Retinal Image Classification

After verifying the performance on natural image datasets, we move to verify the zero-shot classification capability on medical datasets. We included anterior

Table 4. Domain generalization capability on medical image dataset

Model Name	Model Size	FLOPs	ImageNet-1k	SCES	SINDI	Pre-trained
ResNet-50	25M	4G	76.1	71.2	55	No
Swin-Tiny	29M	4.5G	81.3	81.5	72.9	Yes
MedMLP-B0	4.9M	0.6G	74.3	65.4	58.5	Yes
MedMLP-B1	8.4M	1.0G	76.2	70.5	63.1	Yes
MedMLP-B2	12.7M	2.1G	78.3	73.1	68.5	Yes
MedMLP-B3	25.7M	4.1G	81.1	83.6	75.1	Yes

segment photographs obtained from Singapore Epidemiology of Eye Diseases (SEED) Study, which includes the Singapore Chinese Eye Study (SCES) and Singapore Indian Eye Study (SINDI). We further included two clinical studies, namely the Iris Surface Features (ISF) and Irido-Choroidal Characteristics (ICC) study, on primary angle closure disease (PACD).

PACD is a spectrum of disease that is characterized in common by an obstruction to aqueous humor outflow. It may culminate in developing a more visually debilitating form of glaucomatous optic neuropathy. We randomly split sub-set of data from the Singapore Chinese Eye Study (SCES), the Singapore Indian Chinese Cohort (ICC), the Iris Surface Features (ISF), in total 4715 eyes into training, validation, and testing dataset following a ratio of 7:1:2. The other iris fundus photo dataset used for external validation is sub-set of the Singapore Indian Eye Study (SINDI) which contains 250 eyes.

Finally, we show that MedMLP achieves the best trade-off between computational cost and the zero-shot cross-domain generalization capability. The results are shown in Table 4. It is clearly observed that both Swin transformer and MedMLP achieve significantly better accuracy when tested on the out-of-distribution dataset SINDI.

The results reveal that the conventional ResNet-50 model achieves comparable accuracy on the SCES dataset with the Swin-Tiny model. However, it performs significantly inferior to Swin-Tiny when evaluated on out-of-distribution datasets (SINDI). Nevertheless, the performance of the Swin-Large model, which is scaled up from Swin-Tiny, does not exhibit an adequate level of improvement, likely because of underfitting, owing to the small amount of data available. To address this issue, we proceed to utilize the pre-trained Swin-Large model on the ImageNet-22K dataset. Remarkably, the classification accuracy improves by 13.3% using the same model architecture. Nevertheless, it is crucial to note that transformer-based models typically necessitate a substantial amount of computation, resulting in a high runtime latency. As a consequence, deploying such models to medical diagnosis devices might be not practical.

4 Conclusions

In this paper, we introduced the adaptive fully-connected (AdaFC) layers to address a fundamental limitation of existing MLP-like models that they cannot adapt to different input resolutions. Taking AdaFC as the basic building block for

spatial information encoding, we present the MedMLP model. We surprisingly find that MedMLP is a strong competitor to CNNs for mobile settings and performs much better than ViT models via extensive experiments. MedMLP reveals the great potential of MLP-like models and offers a promising alternative model for mobile applications. We believe our work inspires future works to explore the performance potential of MLP models further.

Acknowledgement This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-TC-2021-003). This work was supported by the Agency for Science, Technology and Research (A*STAR) through its AME Programmatic Funding Scheme Under Project A20H4b0141. Besides, this work is also partially supported by Career Development Fund (CDF) C233312010, and Taishan Scholars Program (Grant No. tsqn202312067).

Disclosure of Interests The authors have no competing interests in this paper.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 5
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021) 1
3. Beyer, L., Hénaff, O.J., Kolesnikov, A., Zhai, X., Oord, A.v.d.: Are we done with imagenet? arXiv preprint arXiv:2006.07159 (2020) 6, 7
4. Bi, W.L., Hosny, A., Schabath, M.B., et al.: Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: A Cancer Journal for Clinicians **69**(2), caac.21552 (feb 2019) 1
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 702–703 (2020) 6
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1, 6, 7
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 2
8. Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R., de Leeuw, F.E., Tempany, C.M., Van Ginneken, B., et al.: Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 516–524. Springer (2017) 2
9. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021) 2

10. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019) 6
11. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision permutator: A permutable mlp-like architecture for visual recognition. arXiv preprint arXiv:2106.12368 (2021) 2, 6, 7
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (may 2015) 1
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021) 2
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 6
15. Mehta, S., Rastegari, M.: Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021) 2
16. Morid, M.A., Borjali, A., Del Fiol, G.: A scoping review of transfer learning research on medical image analysis using imagenet. Computers in biology and medicine **128**, 104115 (2021) 2
17. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019) 2
18. Rajkomar, A., Dean, J., Kohane, I.: Machine Learning in Medicine. New England Journal of Medicine **380**(14), 1347–1358 (apr 2019) 1
19. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: AI in health and medicine. Nature Medicine **28**(1), 31–38 (jan 2022) 1
20. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018) 2, 3
21. Sapoval, N., Aghazadeh, A., Nute, M.G., et al.: Current progress and open challenges for applying deep learning across the biosciences. Nature Communications **13**(1), 1728 (apr 2022) 1
22. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022) 1
23. Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. arXiv preprint arXiv:2101.11605 (2021) 2
24. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. arxiv. arXiv preprint arXiv:1909.11575 **10** (2019) 2
25. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019) 2
26. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Keysers, D., Uszkoreit, J., Lucic, M., et al.: Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601 (2021) 2, 3, 5, 6, 7
27. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Joulin, A., Synnaeve, G., Verbeek, J., Jégou, H.: Resmlp: Feedforward networks for image classification with data-efficient training. arXiv preprint arXiv:2105.03404 (2021) 6, 7
28. Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019). `https://doi.org/10.5281/zenodo.4414861` 6

29. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6023–6032 (2019) 6
30. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Hou, Q., Feng, J.: Deepvit: Towards deeper vision transformer. arXiv preprint arXiv:2103.11886 (2021) 2
31. Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M.: Understanding the robustness in vision transformers. In: International Conference on Machine Learning. pp. 27378–27394. PMLR (2022) 2