A Study of the Possible Advantages of Vision Transformers in the Automated Diagnosis of Diabetic Retinopathy from Fundus Images

Candidate Number: KRQV5

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Master of Research

of

University College London.

Institute of Health Informatics

University College London

September 1, 2021

I, Candidate Number: KRQV5, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Transformer models have represented the state-of-the-art (SOTA) approach in natural language processing since 2017. Whereas in computer vision, convolutional neural networks (CNNs) have dominated the domain since the emergence of deep learning. Recently developed, vision transformers (ViTs) have challenged this split by showing competitive results on the benchmark ImageNet dataset.

Through the use of transfer learning, CNNs have shown impressive performance in automated medical imaging tasks such as classifying cases of diabetic retinopathy from fundus images. However, challenges in explainability, as well as data and computational efficiency, of these models remain. The aim of this dissertation is to evaluate if ViTs offer a solution to these problems whilst maintaining a high classification performance. This is the first work to evaluate the performance of a pure transformer architecture when finetuned to a medical imaging classification task.

Models, pretrained by supervised training on ImageNet21k or using the unsupervised student-teacher DINO methodology on ImageNet-1k, have been finetuned to the public eyePACs dataset. When compared to a ResNet50 CNN, the top-performing small ViT model has been found to have a lower precision-recall AUC on the external Messidor-1 validation set (0.917 vs 0.957), inference speed (89.9 vs 205.2 images/s) and data efficiency. However, visualisation of the last attention layer provides a more precise saliency map than the SOTA ResNet50 GradCAM. In addition, the use of ViTs in this domain may allow for new multi-modal, large-scale pre-trained or high-resolution models in the future.

Keywords: Vision Transformers, Automated Medical Image Classification, Diabetic Retinopathy

Acknowledgements

I would like to my supervisors, (names retracted for anonymous marking), for the time and effort they have dedicated to steering me through this project. The experience of proposing, executing and writing up a dissertation whilst working remotely has undoubtedly been a challenge. Without their guidance though, the work presented here would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transformer models have represented the state-of-the-art approach in natural language processing since 2017 [15]. Whereas in computer vision, convolutional neural networks (CNNs) have dominated the domain since the emergence of deep learning [16]. Recently developed, vision transformers (ViTs) [6] have challenged this split by showing competitive results on the benchmark ImageNet dataset [17].

A key feature of CNNs is their ability to perform transfer learning [18]. This is the process by which a model's weights are fine-tuned to a specific task with a small dataset after large scale pre-training on a general dataset such as ImageNet [17]. Hence, this allows models to be trained for tasks where large scale data collection is prohibitively expensive or infeasible, for example many medical imaging tasks. Therefore, transfer learning has been widely used to successfully fine-tune CNN's for use cases such as classifying cases of diabetic retinopathy (DR) from fundus images [19].

It has been shown that, given enough training data, the task of DR classification can be performed with near-perfect accuracy by a CNN [10]. However, challenges in the data and computational efficiency of these models remains. Furthermore, saliency maps, produced using techniques such as GradCam [3], are being proposed as means to provide clinicians with an explanation of the classification decision made by a CNN. However, there is growing evidence that these maps are unreliable [20]. Therefore the aim of this dissertation is to evaluate if ViTs offer a solution to these problems whilst maintaining a high classification performance. This is the first work to evaluate the performance of a pure transformer architecture when finetuned to a medical imaging classification task.

This dissertation has finetuned several ViTs and CNNs, which show SOTA performance on the ImageNet dataset, to the largest publicly available DR dataset called eyePACs [21]. The performance of these models has then been evaluated in three key areas; classification performance, efficiency and

explainability.

This report begins by reviewing the relevant background and related work to the dissertation. It then moves on to describe the methodology used to train and evaluate both the ViT and CNN models. The results of this evaluation process are then presented. Finally, a discussion of the results and the future potential of ViT models concludes the report.

Chapter 2

Background

This chapter outlines the background work on which this dissertation builds upon. Specifically, it reviews the success and progression of deep learning computer vision models over the last decade. It then moves on to look at the impact this has had on the development of automated medical imaging tools, using DR classification as a case study.

2.1 Deep Learning Models

The emergence of deep learning over the last decade has led to the development of several SOTA domain-specific models [22, 23, 5]. This has been enabled by the increased availability of parallel computing and large datasets on which the effective training of such models depends [24]. Computer vision (CV) is an example of one such domain.

2.1.1 Convolutional Neural Networks

CV models are commonly benchmarked by their ability to accurately classify images from the ImageNet dataset [17]. AlexNet [16] was the first deep learning model to demonstrate SOTA performance on this task using a Convolutional Neural Network architecture. Such a model is made up of repeated convolutional, activation, and pooling layers [24], as shown in Fig 2.1.

The convolution operation takes as input a volume (such as an RGB image), I, and a kernel, K, of a small height and width, l. It is of note that the kernel has a depth, D, equal to the depth of the input image. The output of the operation, S, is found by sliding the kernel across the height and width of the input volume and computing the dot product between the kernel and input volume at each position to

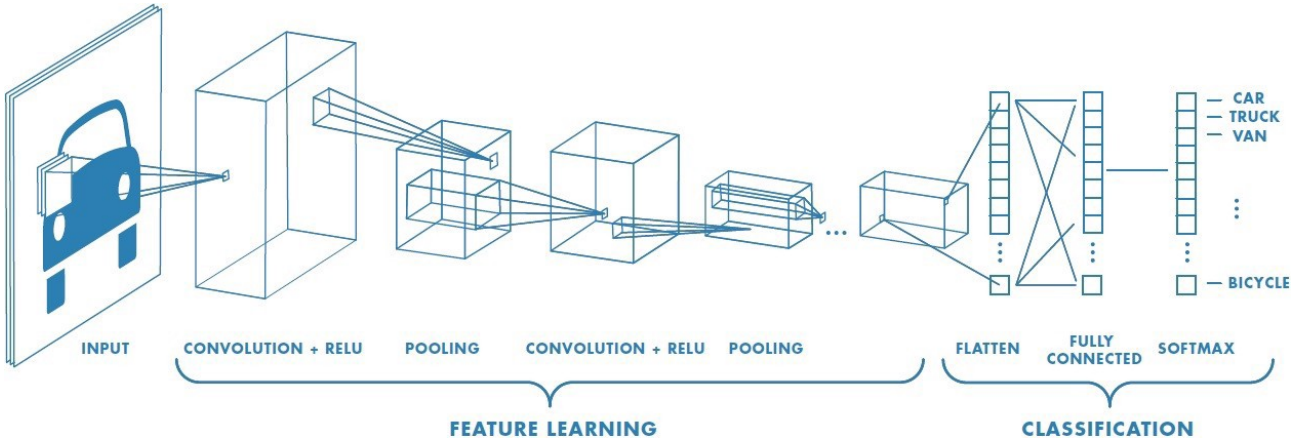


Figure 2.1: Visualisation of a simple CNN architecture made up of repeated convolution, ReLU activation and pooling layers which develop a low-level representation of the image. This is then flattened and a logit outputted by a fully connected layer. This logit can then be converted to a probability distribution using the softmax function from which predictions can be made [1].

give a 2D activation map. Formally, an element in the output map, $S(i,j)$, is calculated as

$$S(i,j) = \sum_{m=-\frac{l}{2}}^{\frac{l}{2}} \sum_{n=-\frac{l}{2}}^{\frac{l}{2}} \sum_{d=0}^{D} I(i+m, j+n, d) K(m,n,d). \tag{2.1}$$

The use of multiple kernels results in the layer outputting a volume such as that seen in Fig.2.1.

Convolutions give the CNN architecture the following properties. Firstly, they give an inductive bias that neighbouring pixels are of greater relative importance than those further away. Secondly, they make the model equivariant to the translation of the input. Finally, the use of convolutional kernels is equivalent to weight sharing between neurons in a fully connected network. Therefore the number of parameters in the model are greatly reduced, this is important due to the large size of images.

The activation layers apply a non-linear function to the output of the convolutional layers of the model, such as Rectified Linear Unit (ReLU),

$$ReLU(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}. \tag{2.2}$$

This allows for the stacking of repeated layers which is commonly referred to as increasing the depth of the model.

The dimensionality of the input is gradually reduced after each convolution using pooling layers. This operation applies a given function to a receptive field. The function commonly used is the maximum as it is found to significantly remove outliers. The use of pooling also gives the model the property of scale invariance.

Repeated stacking of these layers results in a deep network. After a final convolution and pooling layer, the model's reduced representation of the input is flattened and a fully connected neural network then outputs the model's prediction. In the case of a classification model, this is a vector or logit, l, of length, K, which is equal to the number of possible classes. This logit is then converted into a probability distribution, P, using the softmax function,

$$P_i = \frac{exp(l_i)}{\sum_{k=1}^{K} exp(l_k)}. \tag{2.3}$$

The parameters of the model can then be learnt by gradient descent. The individual parameter gradients are calculated by back-propagating from a pre-defined loss function and ground truths, $\hat{y}$. In the case of classification this function is commonly the cross entropy loss,

$$Loss = -\sum_{i=1}^{K} P_i log(\hat{y}_i), \tag{2.4}$$

Building on this base architecture, a great number of CNN variations have been proposed [25, 26]. A notable iteration is the use of residual connections between layers to give a family of models called ResNets[27]. These skip connections mitigate against the vanishing gradients effect during training and hence allow for benefits of more expressive deeper networks to be realised. Such improvements have led to the size and complexity of CNNs increasing [23] and resulted in a rapid improvement in the CV SOTA, as shown in Fig.2.2.

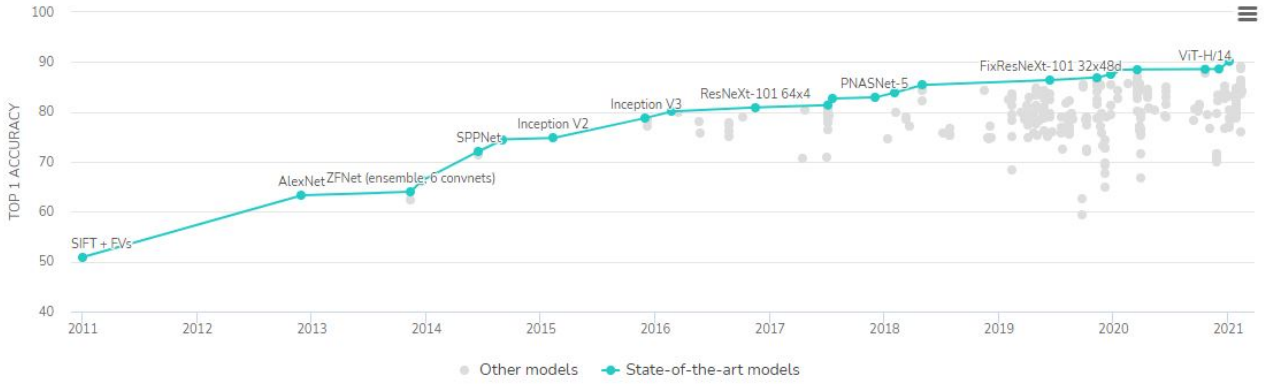


Figure 2.2: Plot of the increase in SOTA accuracy on the ImageNet classification challenge since. Note the rapid improvement since the re-emergence of deep learning in 2013 [2].

2.1.1.1 Grad-CAM

A common criticism of deep learning models is their so-called 'black box nature' due to their large number of parameters [28]. Hence, work has been conducted to develop explainability methods that quantify the effect that each part of the input has on the model's output [29, 30]. In the case of CV models, these methods commonly output saliency maps, such as those shown in Fig 2.3.

Figure 2.3: Example of a saliency map produced from a ResNet model using Grad-Cam on a dog cat classification example [3]

The development of explainability methods is critical for the application of deep learning models in the medical domain. Unlike many applications, the decisions made by these models could be life critical and will be used with a human clinician in the loop. Therefore, clinicians need to build trust and understanding in the model's outputs for them to be used in clinical decision making. This is unlikely to be achieved only through the presentation of impressive classification results. A commonly proposed method is to present clinicians with a saliency map in addition to the classification decision

and hence explain the decision made by the model [31].

A popular and widely used CNN explainability method is Grad-CAM [3]. This method produces a saliency map, $L^c_{Grad\_CAM}$, for a given class, c, from the weighted average of the k layers of the final feature map, A,

$$L^c_{Grad_CAM} = ReLU(\sum_k \alpha^c_k A^K). \tag{2.5}$$

The ReLU function is used so only positive values are emphasised. The weights, α, are found as the globally average pooled value of each features maps effect on the prediction of the class of interest, y^c,

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\delta y^c}{\delta A^k_{ij}}, \tag{2.6}$$

where Z is the number of values in A. $L^c_{Grad\_CAM}$ is then upsampled to the size of the input image. Examples of the output of Grad-CAM applied to a ResNet architecture trained on ImageNet are shown for two different target classes in Fig 2.3.

The systematic and quantitative evaluation of such saliency map methods is difficult. Some studies have treated the output saliency maps as weakly supervised segmentation maps. These works compare the saliency maps to ground-truth annotations through metrics such as DICE and AUC. Such studies have found the reliability of methods such as Grad-CAM, questionable especially on downstream medical imaging tasks.

Arun et al. [32] proposed a framework to assess the robustness of eight different saliency map techniques for a CNN trained on two large publicly available radiology datasets. The evaluation assessed the localisation utility, sensitivity to model weight randomisation, repeatability and reproducibility of each approach. These properties were quantified by calculating the area under precision-recall curves and structural similarity index under different perturbations to the model or inputs. They showed that all eight techniques failed at least one of the criteria. Therefore the conclusion of the paper was that due to the risks of medical imaging classification if localisation is a desired output then a separate model should be trained to do this. However, the labelling cost of a high- quality segmentation dataset is significantly higher than that of a labelled dataset. Moreover, using a separate model would not explain the classification decision of the original model.

Further work by Van Craenendonck et al. [4] developed a similar framework for the quantitative assessment of saliency maps in the context of DR in fundus images. Due to the small size of retinal

lesions that characterise DR, see Section 2.2.1, a new Explainability Consistency Score (ECS) was proposed. This is required as a naive comparison between the saliency and ground-truth maps would be biased towards the larger lesions. Therefore, each map is first discretised into a grid, see Fig 2.4. Each patch in the ground truth grid is then weighted by the inverse frequency of each lesion type. Evaluation of these weighted discretised maps showed there was a large variation in performance depending on the model and saliency map technique used. Furthermore, they found considerable disagreement between the saliency and ground-truth maps.

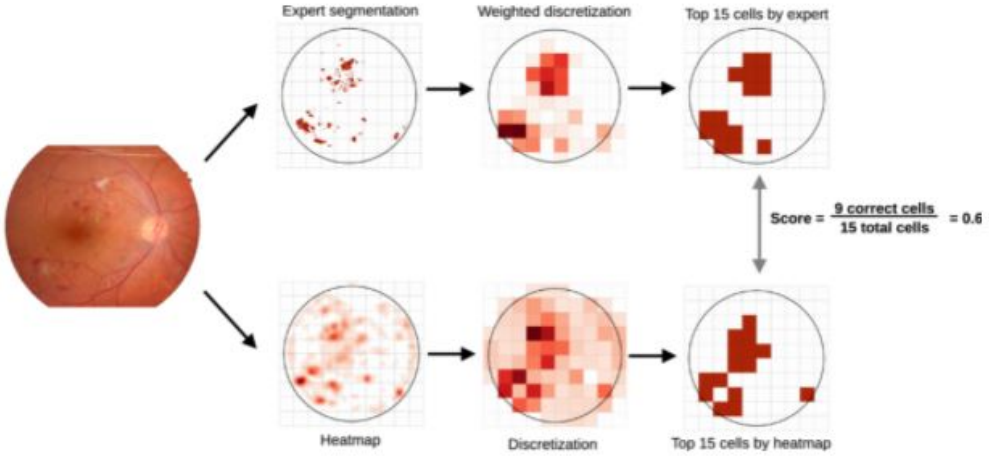


Figure 2.4: Illustration of the ECS metric developed by Van Craenendonck et al. [4]. Each map is discretised and the expert map is weighted by the inverse of each lesion type's frequency. Comparison is then made between the top 15 highest scoring patches of each map.

Using a saliency map as a form of explanation is questionable for some use cases. For example, if a model is using the absence of a feature to make a classification decision this cannot be explained clearly through the use of a saliency map. However, saliency map methods have proved useful in detecting model failure cases in applications such as autonomous driving [33].

2.1.2 Transformers

In the domain of NLP, a similar deep learning revolution, as in CV, has occurred but using transformer models [15] instead of CNNs. The performance of transformers has also been greatly increased by training larger and more complex models [5, 22]. A recent development has been the proposal of the vision transformer [6], which takes image patches instead of words as an input, bridging the gap between CV and NLP models.

To describe and motivate the origins of the transformer the next section outlines the development

of the architecture in the domain of NLP. The adaptions required to alter the model to CV are then explained in the subsequent section.

2.1.2.1 NLP Transformer

NLP tasks vary from simple sentiment analysis classifications to complex word generation tasks such as question and answering. However, all tasks require a model to create an internal representation that incorporates the sequential contextual information of free-form text. For example, the conjugation of a verb depends on the plurality of the noun in the same sentence. These contexts can also be long-range such as the name of a character at the start and end of a chapter.

These long-range interactions proved challenging for the previous SOTA recurrent neural nets (RNN) [34] which had evolved mechanisms such as long short-term memory (LSTM) [35] to model this contextual information. These models were fundamentally limited by the issue of vanishing gradients. This is where the size of the gradient when back-propagated through the model becomes vanishingly small due to repeated fractional multiplications. The consequence of this is that the update to the model weights due to the representation of long-range interactions is small and so the model fails to incorporate this context [36].

The transformer model was introduced [15] to provide a solution to this vanishing gradient problem through the use of a mechanism called self-attention. This allows the model to choose the amount of attention each input should pay to all other inputs at each step. Crucially the model has no inductive bias that closer words have a higher importance. Instead, this is learnt. Hence, the vanishing gradient problem is overcome

The transformer model was initially proposed as an encoder-decoder model for language translation. The encoder created a representation of the input sentence which was then used as input to the decoder which outputted the sentence in the chosen language. The use of attention was found to be highly beneficial in this task as the word order is not necessarily maintained across languages. Hence being able to pay attention to inputs non-sequentially leads to a boost in performance.

Further iterations built upon this idea to create classifier models requiring only the encoder section of the model [5]. This model is shown in Fig 2.5 for a simple sentiment analysis task in which an input sentence is labelled as positive or negative.

The input is firstly converted from a string to an integer by tokenisation. This is done by looking up the input words in a pre-determined dictionary and returning an ID. This then converted to a vector

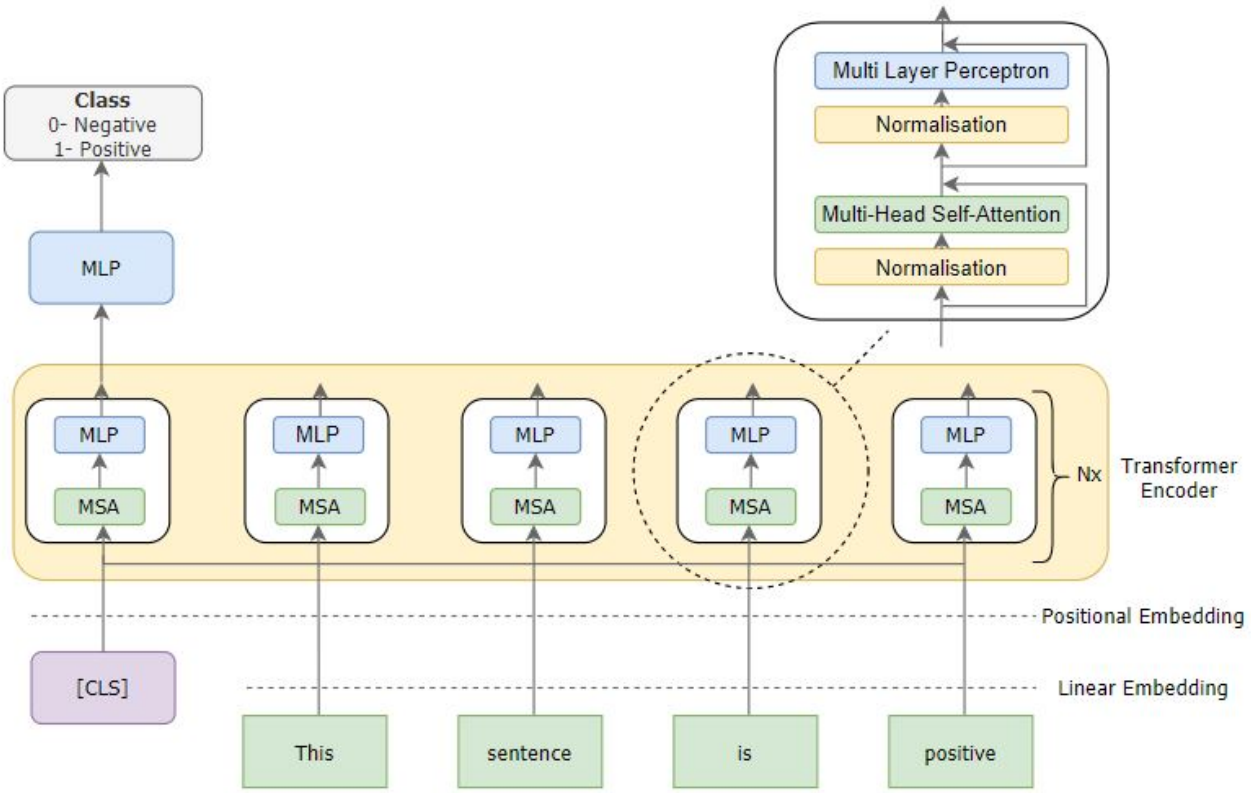


Figure 2.5: Illustration of the NLP BERT classification model for a trivial sentiment analysis input [5]. Each word is linearly embedded and a learnable class token ([CLS]) is prepended. A positional embedding is then added to all input tokens. The input tokens are then passed through N layers of MSA and MLP units. The final classification is performed using the outputted class token as the input to a shallow layer MLP

or token of a standard length by a learnt linear embedding. Similar to the ideas of Word2Vec [37], this moves the IDs into a latent space in which tokens with similar meanings are closer together. This is commonly quantified by their cosine similarity. A learnable class token of the same dimensions as the input vector is then prepended. Next, a positional embedding is added to each token. This encodes the position of the token in the sentence. This is the only positional information the model receives. These embeddings are either pre-defined by cosine functions or learnt during training.

These inputs tokens including the class token then enter the transformer encoder. This is a multi-layer architecture, as shown in Fig 2.6, made up of a single repeating unit. This unit contains a multi-layer perceptron (MLP) and a multi-head self-attention (MSA) layer as well as layer normalisation and residual connections. The same unit weights are shared across each layer. The MLP is a shallow 2 layer architecture and uses Gaussian Error Linear Units (GeLU) activation.

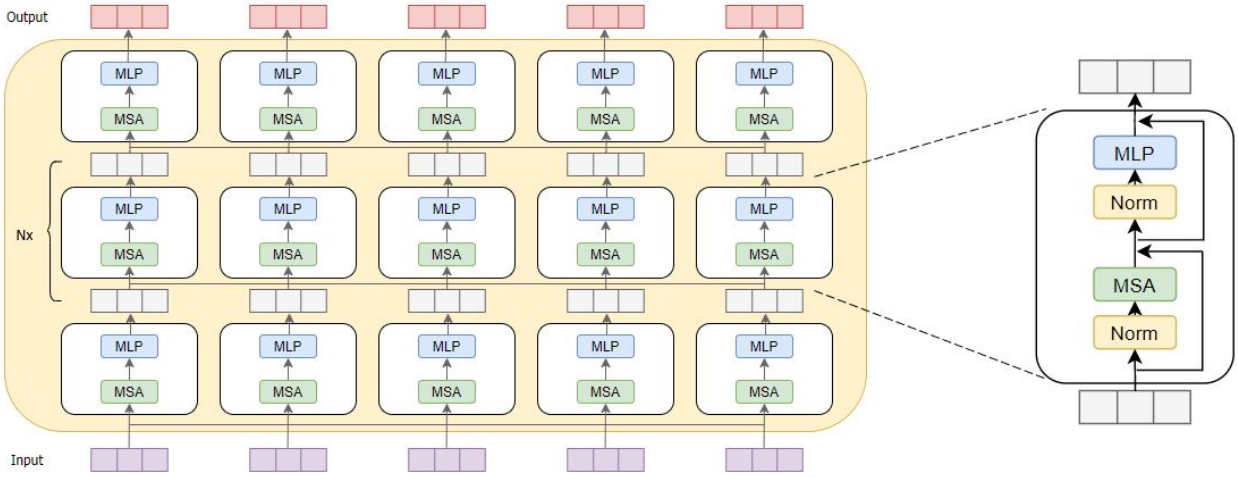


Figure 2.6: Illustration of the transformer encoder showing the repeated MLP and MSA units that make up the architecture. The main diagram simplifies the process by not showing the normalisation and residual connections. N is the number of repeated layers in the encoder.

The MSA contains the self-attention mechanism which separates the transformer architecture from a standard deep learning model. An intuition of this mechanism can be built up by considering the output of each attended token, z_k, as the weighted sum of each of the N input tokens, x,

$$z_k = \sum_{i=1}^{N} attention \times x_i. \tag{2.7}$$

It stands to reason that similar tokens may hold contextual clues and so should pay attention to each other. The similarity between two vectors can be quantified by their dot product,

$$z_k = \sum_{i=1}^{N} (x_k . x_i) \times x_i. \tag{2.8}$$

However, the output of each layer must be normalised to ensure the magnitude of the vectors does not explode or vanish. This is achieved using the softmax function across all N tokens,

$$z_k = \sum_{i=1}^{N} softmax(x_k . x_i) \times x_i. \tag{2.9}$$

This notation can be simplified to a matrix formulation where the rows of X are the input tokens and

the rows of Z the output tokens,

$$Z = \sum_{i=1}^{N} softmax(XX^T)X. \tag{2.10}$$

However, it is not only the similarities in words that are important to building a contextual representation of the input but also the differences. For example, representing negation is imperative to successful understanding an input sentence. This is modelled through a number of learnable weight matrices, W_Q, W_K, W_V,

$$Z = \sum_{i=1}^{N} softmax(W_Q X(W_K X)^T)W_V X, \tag{2.11}$$

these matrices are commonly referred to as the query, key and value matrices respectively. To condense the notation further the product of the weight matrices are combined as Q, K and V. Also a scaling factor given by the length of the rows of K is added for numerical stability

$$Z = \sum_{i=1}^{N} softmax(\frac{QK^T}{\sqrt{d_k}})V. \tag{2.12}$$

Multi-head self-attention extends this idea by capturing additional contextual detail by performing multiple sets of the self-attention calculation with the same input tokens but different learnable weight matrices. This results in the attention mechanism giving multiple different outputs, each of which captures different contextual details at the same layer of abstraction. To maintain the dimensionality of the tokens used throughout the model, each output is concatenated and linearly projected by an additional matrix W_0 to give the final MSA output. This process is shown in Fig 2.7.

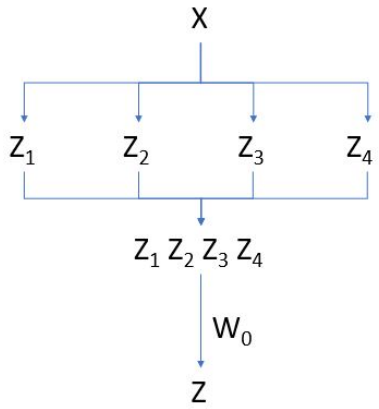


Figure 2.7: Visualisation of multi-head self-attention for an input X. In this example, four heads are used to give four different outputs $Z_1 - Z_4$. These are then concatenated and projected to the same size as X by the learnable matrix W_0 to give the output Z.

For a classification transformer, it is only the final state of the class token that is used from the encoder. This is fed into a shallow layer MLP which outputs a final classification probability distribution using the softmax function. Then given a set of labelled input images this allows the parameters of the model to be learnt using gradient descent by back-propagating from the cross-entropy loss of the model's predictions.

2.1.2.2 Vision Transformer

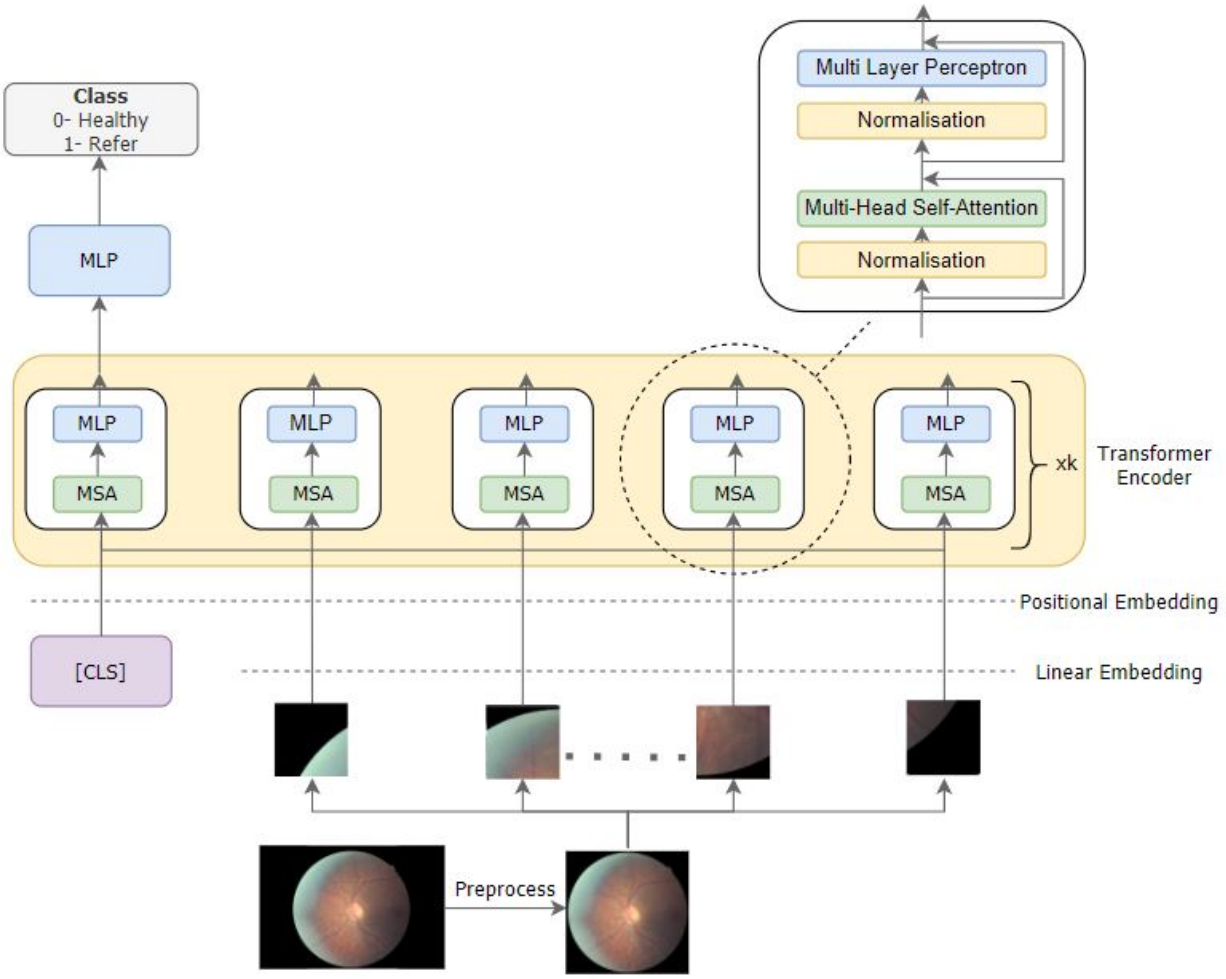


Figure 2.8: Illustration of the VIT model architecture [6]. Note the model is nearly identical to the NLP model shown in Fig 2.5. The only changes required are the patching of the input and linear embedding mechanism.

The main obstacle to using images as inputs to a transformer model is the quadratic scaling of the complexity of self-attention with input size. Therefore, previous approaches to applying attention mechanisms to images have either only applied it to local neighbourhoods [38] or developed complex

attention architectures which theoretically scale well with input size but are difficult to implement efficiently [39].

The recent adaption of the NLP transformer architecture to CV proposed by Dosovitskiy et al. overcomes the issue of attention scaling by inputting the image as non-overlapping 16x16 pixel patches instead of pixelwise. This dramatically reduces the number of input tokens allowing the forward pass through the transformer to be computationally tractable.

The only difference between the ViT model, visualised in Fig 2.8, and the original NLP transformer encoder, see Fig 2.5, is the implementation of the image patching and linear embedding mechanism. A pre-defined dictionary is no longer suitable. Instead, a convolutional layer is used with a stride of 16 and a kernel depth equal to the dimensionality of the input tokens. This does mean that the ViT is not a technically a convolutional free architecture.

Investigations into the behaviour of ViTs have shown that the attention mechanism looks globally from the first layer of the encoder, see Fig 2.9. This is a feature not seen in CNNs in which initial layers look at only local features due to the small receptive field of the kernels used. A global representation of the image is then built up in the subsequent deeper layers.

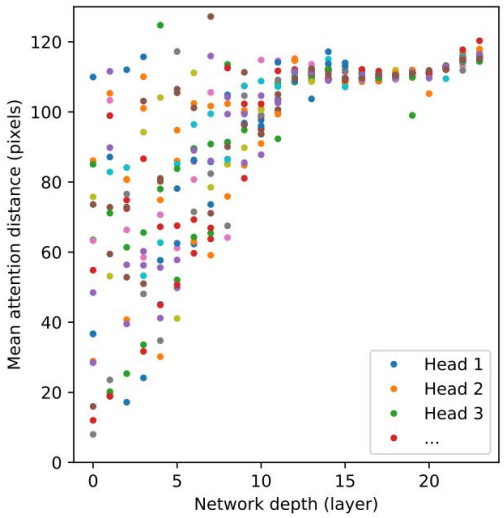


Figure 2.9: Plot of attention distance against network depth for a ViT for a 224x224 pixel image. Attention distance is calculated as the average distance between the query pixel and all other pixels weighted by attention. It can be seen even from the first layer the model is globally attending to patches [6]

2.1.2.3 Attention Rollout

It is possible to use the ViT's attention mechanism to visualise the patches of greatest importance to the model's classification decision. Hence, the use of ViT's has allowed for the development of a new

set of explainability methods.

One such method proposed to do this is attention rollout [40]. Given the attention matrix of each layer, $\mathbf{A}$, where $\mathbf{A}_{xy}$ is the attention token x gives to token y averaged across each head. The attention rollout, $\tilde{\mathbf{A}}$, from layers i to j is

$$\tilde{\mathbf{A}} = \begin{cases} (\mathbf{A}(l_i) + \mathbf{I})\tilde{\mathbf{A}}(l_{i-1}) & \text{if } i > j, \\ \mathbf{A}(l_i) & i = j. \end{cases} \tag{2.13}$$

The addition of the identity matrix accounts for the use of residual connections. However, there is qualitative evidence that the averaging of the attention at each layer is sub-optimal. Instead, taking the maximum attention for each token over each head and thresholding gives improved saliency maps [7], as shown in Fig 2.10.

Figure 2.10: Attention rollout of the ViT model, on a cat dog example from ImageNet. Two approximations of multi-head self-attention are shown. Using discard and max fusion gives visually better results [7]

2.2 Automated Medical Image Classification

A key ability of CNNs is their ability to achieve high performance on downstream tasks with small datasets after large scale pretraining on general datasets such as ImageNet [18]. This process of transfer learning has allowed the application of CNNs on a range of tasks including human pose estimation [41] and autonomous driving [42].

CNNs have also been successfully finetuned to perform medical image analysis tasks. Examples

include detecting malignant melanoma on skin photographs [43], Alzheimer's from MRI scans [44] and tuberculosis from chest X-rays [45]. Such models aim to increase the accuracy and efficiency of clinical practice. Several of these systems, based on CNNs, have been approved as medical devices by regulators including the FDA [46].

As ViTs are a new model in CV, their application to medical imaging problems is in its infancy. Since work on this dissertation began, several studies have used transformers in conjunction with CNNs to achieve SOTA performance on medical image segmentation tasks [47, 48, 49]. These studies report that the increase in performance is due to ViTs ability to model long-range interactions between pixels. However, pure transformer architectures are not used in these studies as the authors argue that the medical datasets available are too small to train a custom data-hungry ViT from scratch. However, classification tasks require less specialised model architectures and so can be finetuned from pre-trained models. Using this principle, this is the first work to evaluate the performance of a pure transformer architecture when finetuned to a medical imaging classification task.

2.2.1 Diabetic Retinopathy

One medical domain in which a great deal of automated medical image analysis work has been focused is ophthalmology [50]. As the busiest outpatient speciality in the UK with a predicted 30-40% increase in patient numbers over the next two decades, demand will soon outstrip the supply of physicians [51].

Therefore, the application of technologies such as CNNs which could increase the throughput of patients whilst maintaining a high quality of care is a potentially important development. This is especially true for large scale screening programs. These programs although effective in identifying cases and hence the opportunity for sight-saving interventions are an increasing drain on resources, both in terms of manpower and finance, for even the most developed healthcare systems [52].

DR is an example of one such rapidly increasing screening program. This disease is caused by the chronically high blood sugar levels of diabetic patients damaging the blood vessels in their retina. This can result in the blood vessels leaking or haemorrhaging onto the thin retinal layer which is responsible for sensing light and signalising this stimulus to the brain. Hence these retinal lesions, which characterise DR, can be potentially sight-threatening [53].

DR is most commonly diagnosed from retinal colour fundus images which pictures the retinal surface allowing for the detection of the aforementioned lesions, see Fig 2.11. This allows for

widespread screening programs to effectively diagnose cases of DR in the diabetic population and refer them for treatment including metabolic control and in extreme cases laser photocoagulation [54].

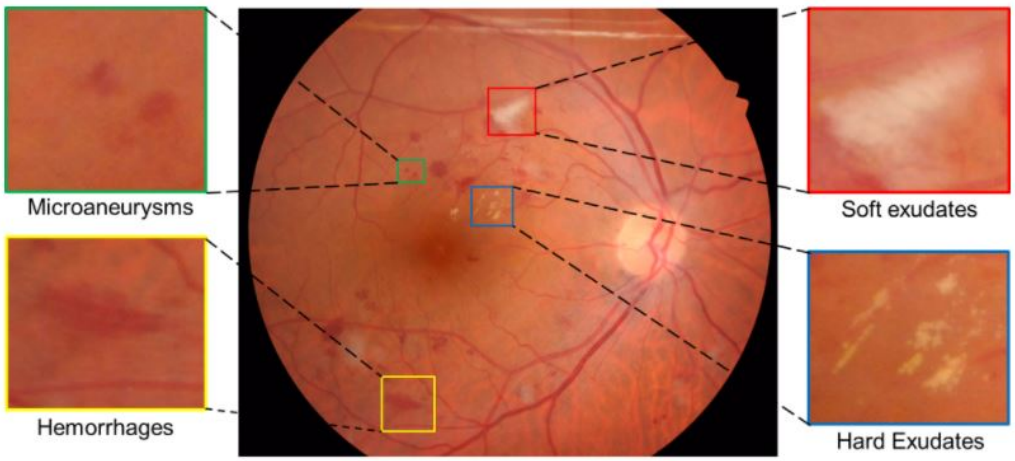


Figure 2.11: Example of a colour fundus image of a severe case of DR with retinal lesions highlighted [8]

It is predicted that by 2040, 600 million people globally will have diabetes and a third of these will have DR [55]. Therefore, DR will become an increasing pressure on ophthalmic services worldwide in the near future. Hence an effective automated DR screening tool could be of high utility in this domain.

2.2.1.1 Deep learning approaches

The development of automated diagnosis tools for DR has closely followed the evolution of CV models. Over the past decade, this has meant moving from using hand-crafted features as inputs to fine-tuning fully end-to-end pretrained CNNs, such as those outlined in Section 2.1.1.

An early handcrafted model was proposed by Acharya et al. [56] using a support vector machine (SVM) to classify images as normal, mild DR, moderate DR, severe DR, and prolific DR. The model was trained on a small dataset of 300 subjects at various stages of DR and had a reported accuracy of 82%, sensitivity of 82%, and specificity of 88%.

The same authors also proposed a similar SVM architecture [57] in which the number of haemorrhages, micro-aneurysms, exudates, and blood vessels were first extracted from the raw image and fed in as the model input. This model achieved an accuracy of 85.9%, a sensitivity of 82%, and a specificity of 86%.

As the field of CV found CNNs to outperform handcrafted SVM feature models [16] the development of DR CNN models followed. Nayak et al. [58] developed an early CNN model to classify DR fundus images into 3 classes. However, they still used hand crafted features, found by morphological processing and texture analysis techniques, to first detect features such as blood vessels and hard exudates. They reported an accuracy of 93%, a sensitivity of 90%, and a specificity of 100% when the model was trained on 140 subjects.

Pratt et al. [59] were one of the first groups to apply modern CNN techniques to the DR classification problem. Using a large training dataset of 80,000 images and data augmentation they showed a CNN could be successfully trained on the task at scale through the use of GPUs. The proposed CNN had a reported accuracy, sensitivity, and specificity of 75%, 30%, and 95%, respectively.

Dekhil et al. [60] developed these ideas by using a more sophisticated VGG-16 CNN, named after the Oxford Visual Geometry Group who proposed it, which had been pre-trained on ImageNet. When finetuned on the same dataset as Pratt et al. an accuracy of 77% and quadratic weighted kappa score of 78% was reported.

The development of models from this point has been on two fronts. Firstly, using different iterations of CNN architectures [61]. This is usually motivated by these models showing SOTA performance on the ImageNet classification task. However, it has been shown that for high-performance models the correlation between performance on ImageNet and a medical imaging task is weak [9], see Fig 2.12. Therefore, progress in this regard has been limited. This is further complicated by the wide range of training datasets, labels and validation metrics used to evaluate these models making the comparison between studies difficult.

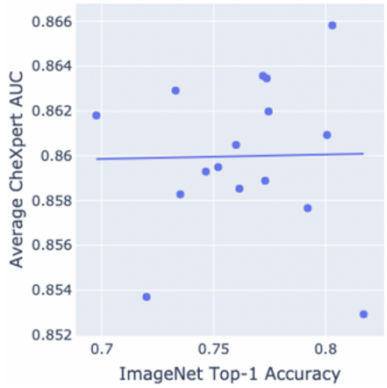


Figure 2.12: Plot showing the weak correlation between a CNN's performance on ImageNet and medical image classification, in this case chest X-rays [9]

The second improvement has been the size and quality of data used to train the CNN [62, 63, 64]. The best performing models on publicly available datasets have been trained on large scale private datasets. For example, Gulshan et al.'s model [10] was trained on a private dataset of 118,419 images each graded by multiple ophthalmologists. This model achieves impressive performance when evaluated on two public validation sets of 0.990-0.991 ROC AUC, 0.870-0.903 sensitivity and 0.981-0.985 specificity.

Gulshan et al.'s work also points to a primary challenge in developing a high-quality DR dataset. The agreement between clinicians when diagnosing edge cases of DR is low, as shown in Fig 2.13. This is exacerbated by the inability to use a later downstream anatomical test as a gold standard label as is the case in other domains such as biopsy of breast cancer tumours after X-ray screening [65]. If this is coupled with suboptimal image quality, the datasets produced can be very noisy. Therefore, when the training methodology of Gulshan et al. was repeated using only a smaller noisy publicly available dataset [21] in which each image is labelled by a single expert the ROC AUC of the model was found to be much lower at 0.951-0.853 [11].

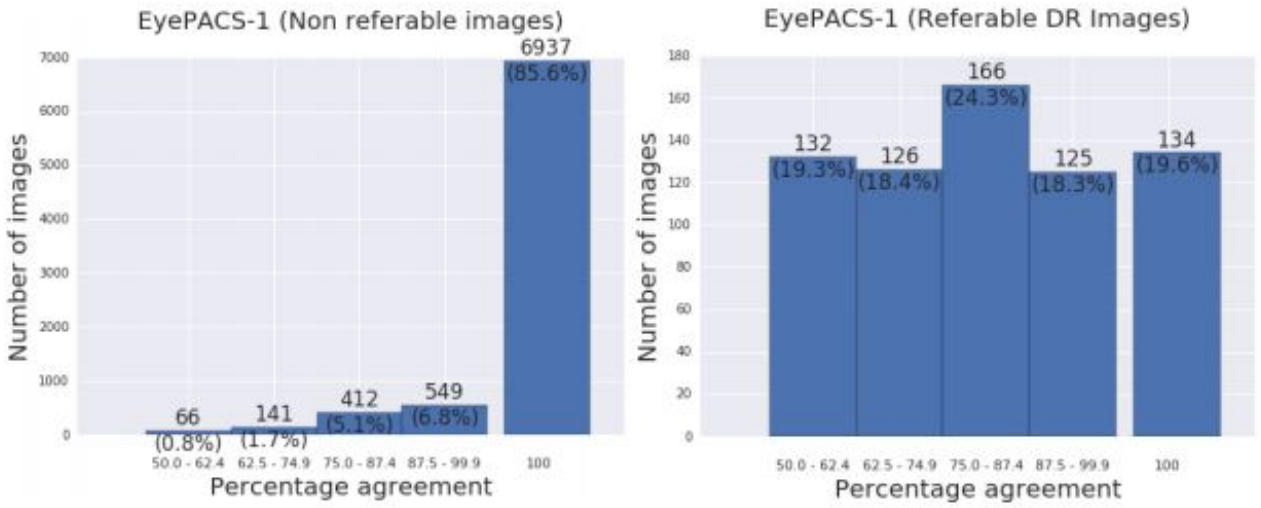


Figure 2.13: Plot showing the poor rate of agreement between 8 ophthalmologists when classifying the eyePACs dataset, see Section 3.1.1, as healthy or referable [10]

This being said DR fundus imaging has several features which make the use of deep learning methods attractive. Firstly, the input is a simple 2D RGB image for which models such as CNNs are designed to take as input and large general datasets exist for pretraining. Secondly, relatively large public DR datasets exist. Finally, retinal lesions are clear features within the image for the model to

identify. Therefore the use of saliency maps to explain the classification decision is valid. However, studies have questioned the reliability current SOTA methods [4] causing a barrier to the clinical deployment of CNN models. Therefore, this provides an excellent case study in which to explore the potential benefits of visualising the attention of ViTs in comparison to the current SOTA GradCam.

2.3 Summary

In summary, CNNs have represented the SOTA CV model since the introduction of AlexNet. The success of this architecture has led to finetuning of CNNs to automate medical imaging tasks such as DR classification. Such models have the potential to fill the predicted future shortfall between patient demand and the supply of medical professionals for services such as screening programs. However, challenges in the efficiency and explainability of these models remain.

Recently, ViTs have challenged this paradigm by showing competitive performance on the ImageNet dataset when rigorously pre-trained. However, the ability of a pure transformer architecture to be finetuned to downstream medical imaging tasks is yet to be explored. Therefore the aim of this dissertation is to evaluate the potential of ViTs compared to CNNs in this domain, using the DR classification task as a case study. This will be done in three key areas; classification ability, efficiency and explainability.

Chapter 3

Methodology

This chapter details the methodology used to train and then compare the ability of ViTs and CNNs to classify cases of DR from fundus images. Firstly, an outline of the datasets used in this dissertation is given before describing the exact model architectures and their pre-training protocols. The training methodology is then outlined. Finally a description of how each model is evaluated on the basis of classification performance, efficiency and explainability is given.

3.1 Datasets

Three independent publicly available datasets are used in this project; the Eye Picture Archive Communication Systems (eyePACs) dataset [21], Messidor-1 [66] and the Indian Diabetic Retinopathy Image Dataset (IDRiD) [8].

3.1.1 eyePACs

The eyePACs dataset is the largest publicly available dataset of labelled DR fundus images with 88,702 scans [67]. The images have been expertly graded according to the International Clinical Diabetic Retinopathy severity scale (ICDR) [68]. This is a 5 stage grading system from 0, no DR, to 4, proliferative DR.

However, known issues exist with this dataset which significantly decreases the number of labelled images available. Firstly, a large percentage of the data is held out as a test set and so the labels are not publicly available. This lowers the available size of the dataset to 35,126. Secondly, the quality of the scans has previously been found to be highly variable. This is due to the data having been collected on multiple camera models by clinicians with varying experience levels. Previous studies have estimated the percentage of gradable images to be as low as 75% due to erroneous factors includ-

ing focus and exposure [69]. Further work has attempted to manually label these ungradable images, some examples of which are shown in Fig 3.1 [11]. Removing these images gives a final usable dataset size of 28,134 labelled images. However, a non-exhaustive search can easily find ungradable images still present within the dataset.

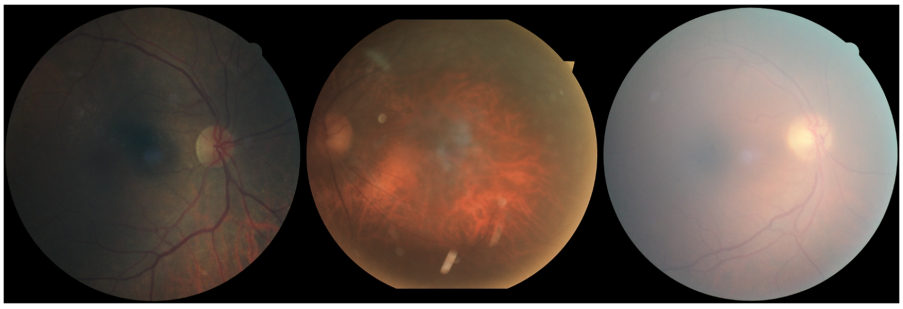


Figure 3.1: Examples of ungradable images from the eyePACs dataset [11] which are removed from the dataset. These examples are either out of focus, underexposed or overexposed.

Even with the unlabelled and ungradable images removed, the eyePACs dataset is an order of magnitude greater than any other public dataset and so it is used to finetune the models in this project. To facilitate the training methodology outlined in Section 3.3, the remaining eyePACs data is randomly split into training, validation and test sets in a 60:20:20 split.

3.1.2 Messidor-1

The Messidor-1 dataset [66] contains 1200 images acquired by three different ophthalmology departments using a colour video 3CCD camera mounted to a Topcon TRC NW6 non-mydriatic retinograph which have a 45-degree field of view. Two-thirds of the images were captured with pupil dilation. Once the previously reported duplicates and mislabels are accounted for the dataset contains 1187 high-quality gradable images. Although of higher quality than the eyePACs dataset the number of scans is an order of magnitude smaller. Therefore, the Messidor-1 dataset is used as an external validation set in this project.

However, a complication arises as the Messidor-1 dataset uses a 4 stage grading system rather than the 5 stage ICDR framework used by eyePACs. Hence to unify the labelling and allow external validation, the classification task is simplified to the still clinically applicable binary problem of healthy vs referral. This is achieved by defining all stages greater than 1 as requiring referral for both frameworks. This is a step previous studies have also conducted [10, 11].

3.1.3 IDRiD

The final dataset used is a subset from the IDRiD dataset which is the only publicly available dataset containing pixel-level annotations of retinal lesions. Eighty-one annotated images are available and highlight microaneurysms, haemorrhages, soft and hard exudates. An example is shown in Fig 3.2. These images were captured by a retinal specialist at an eye clinic in Nanded India on a Kowa VX-10 alpha digital fundus camera with a 50-degree field of view. This dataset provides a means of evaluating and comparing the validity of the saliency maps produced by the models trained in this project.

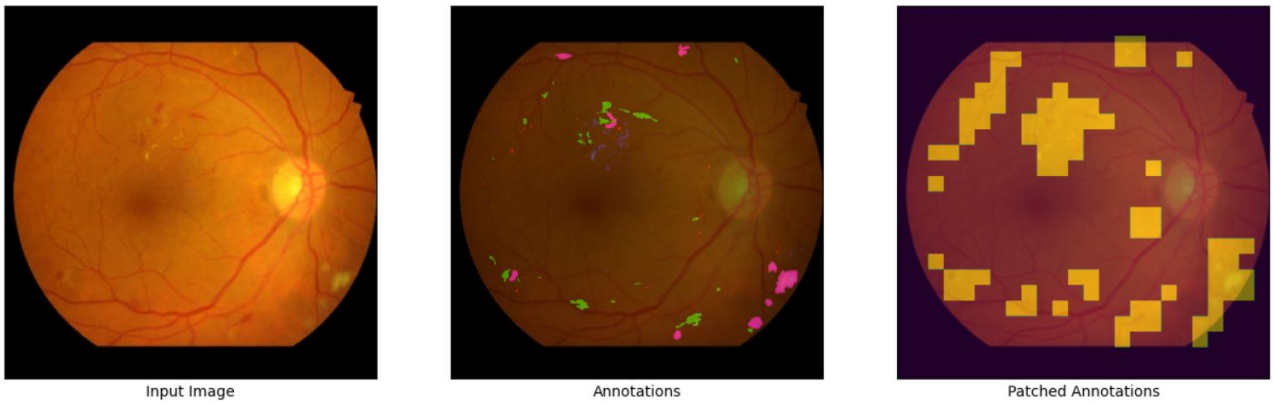


Figure 3.2: A preprocessed image from the IDRiD dataset [8]. Each annotation type is located in a different colour channel: red microaneurysms, green haemorrhages, blue hard exudates and pink soft exudates. The patched annotations pane is the ground truth map produced by the methodology outlined in Section 3.4.3

3.1.4 Pre-processing

The preprocessing steps used in this dissertation firstly square off the raw images and then downsample them to 448x448 pixels. The first step is important so that the images can be easily used to finetune either ViT or CNN models. The second step is required as many of the scans are of the order 1000x1000 pixels. Therefore loading such large images into memory is a bottleneck at training time unless considerable computational resources are deployed. However, the models trained in this dissertation accept 224x224 or 384x384 images as input to maintain a reasonable training time, see Section 3.3. Therefore, downsampling images offline decreases the computational requirement of loading the images at training time without affecting the model's performance.

The implementation of the preprocessing steps builds upon the work of Graham [70]. Firstly the radius and bounding box of the eye is found. As the background of each image is black a segmentation

mask of the retinal disc can be found by thresholding the image. Empirically an effective boundary is found to be $\frac{1}{5}\hat{x}$, where $\hat{x}$ is the average pixel value of the image. To increase the computational efficiency of this process the original image is strided every 100 pixels in both the x and y direction. The extreme x and y coordinates are then be found to give a square bounding box of a standard radius around the eye.

The image is then cropped according to the bounding box with additional padding of 100 pixels on each side to account for the error arising from using the striding technique. Finally, the image is rescaled to 448x448 pixels. This process is visualised in Fig 3.3.

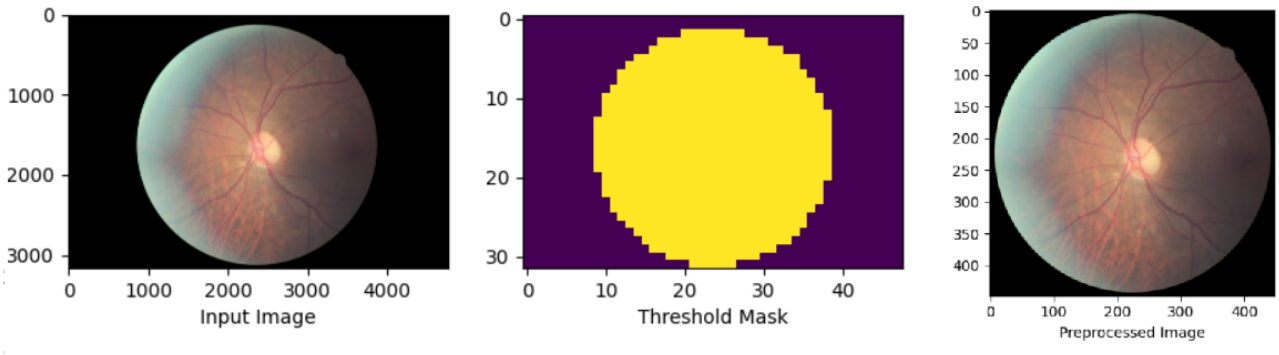


Figure 3.3: Example of a raw image, threshold segmentation map and output of the preprocessing methodology. Axis labels indicate the dimensions of the images

For the segmentation maps contained in the IDRiD dataset, the same process is followed but the bounding box and eye radius defined by the raw image is also used to preprocess each segmentation map. The result of this preprocessing is that all input images to the models are centred, squared and computationally cheap to load.

3.2 Models

The base experimental design used throughout this dissertation is to compare the performance of a pretrained ViT and CNN when finetuned on the DR classification task. For this comparison to be valid, the models chosen must be of a similar size and pretrained using the same methodology. This section outlines the selection of such models.

3.2.1 Architectures

Firstly regarding the base models chosen, the small ViT (ViT-S) [71] is selected as the transformer architecture. This is a transformer encoder, as described in Section 2.1.2, with 12 layers, 6 attention

heads, a patch size of 16x16 and an internal token length of 384 apart from in the hidden MLP layers where it is 1536. This is compared to a ResNet50 architecture [23], a ResNet model, as described in Section 2.1.1, made up of 50 layers.

As can be seen in Table 3.1, both these models have a similar number of model parameters and the ViT-S accuracy on the ImageNet dataset is highly competitive with that of the CNN. It is also of note that models of this size can fit on a single GPU whilst maintaining a reasonable batch size.

Model	Number Parameters	ImageNet21k	DINO Linear Probe
ViT-S	21.7M	81.4%[72]	77.0% [12]
Resnet50	23.5M	80.3%[72]	75.2% [12]

Table 3.1: Comparison of ViT-S model size and performance on the ImageNet classification task after two different pre-training methodologies

Both architectures can be finetuned at a range of resolutions. In the case of ResNet50, this is due to the global pooling used in the final layer [72]. For ViT, the only change required is the bicubic interpolation of the positional encodings [6], visualised in Fig 3.4. Therefore both models are trained at the standard ImageNet resolution of 224x224 as well as 384x384.

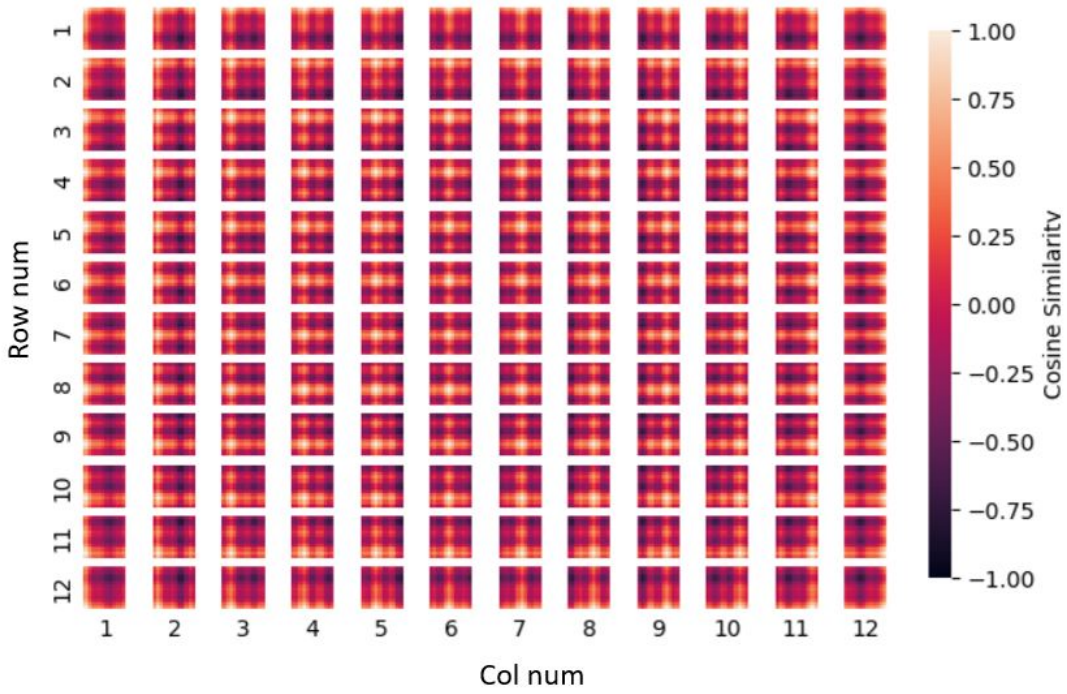


Figure 3.4: Visualisation of the cosine similarity between the positional embeddings of ViT-S-21k-384. Rows and columns 8-12 are interpolated.

3.2.2 Pretraining

Initial studies into ViTs have shown that as this set of models lack the strong inductive basis of CNNs, a more rigorous pretraining than the standard 1.28 million ImageNet images [17] is required to learn an effective model [6]. Two separate approaches to this challenge are considered in this project.

3.2.2.1 Supervised ImageNet 21k

Firstly, it has been shown empirically that strong performance can be gained by simply training on more data. For example, training on the 14.2 million images of the ImageNet-21k dataset [73] results in a highly performant model as shown in Table 3.1. Similar work has been conducted with large scale pretraining of ResNets [23] which too results in strong performance. It is of note though that a small architecture change is made by the authors to the model. The batch normalisation [74] operation is replaced by group normalisation [75] and weight standardisation [76] as the authors found this to aid performance in downstream tasks.

3.2.2.2 Unsupervised DINO Training

The other set of pretrained models considered in this dissertation have been pretrained using a self-supervised method called knowledge distillation with no labels (DINO) [12]. This training procedure is illustrated in Fig.3.5. In this framework, an input image is cropped into local views, V^l, ($< 50\%$ of the image) and global views, V^g, ($> 50\%$ of the image) to gives a set of V different views.

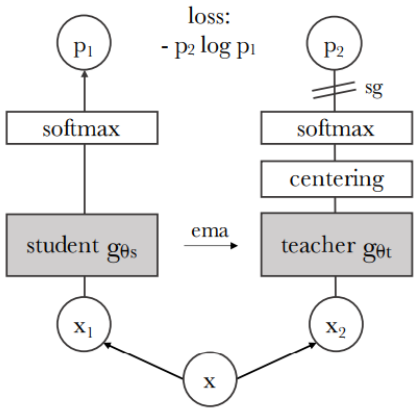


Figure 3.5: Illustration of the DINO training method. The student and teacher receive two different views of an image with the objective of predicting the same sharpened probability distribution. The success of this aim is quantified by the cross-entropy loss which is used to update the student's weights by gradient descent. The teacher's weights are updated by the exponential moving average (ema) of the student's. Hence gradients are stopped (sg) to this model. Also, the teacher uses centering to avoid training collapse.[12]

Two models with the same ViT architecture but different weights are then initiated. The first is referred to as the student network, g_{θ_s}, and receives all the image views as inputs. The second is the teacher network which receives only the global views. Each network is parametrised by θ_s and θ_t respectively. Given an image, x, each model outputs a logit of a fixed length, K, which is converted into a probability distribution, P, using the sharpened softmax function,

$$P(x)^i = \frac{exp(g_\theta(x)^i/\tau)}{\sum_{k=1}^{K} exp(g_\theta(x)^k/\tau)}, \tag{3.1}$$

where $\tau > 0$ and controls the sharpness of the distribution. The objective of the task is for the student network to output the same probability distribution as the teacher. Due to the difference in local and global views available to the two networks, this requires the student has to learn to interpolate global context from local features. Formally, this problem is can be encapsulated by a cross-entropy loss function, L,

$$L = \sum_{x \in V^g} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x')), \tag{3.2}$$

$$H(a,b) = -alog(b), \tag{3.3}$$

which is used to train the student network via gradient descent and backpropagation. The teacher network's weights are updated according to the exponential moving average [77] of the student weights,

$$\theta_t \leftarrow \lambda \theta_t + (1-\lambda)\theta_s. \tag{3.4}$$

The λ parameter controls the relative weighting of θ_t and θ_s and follows a cosine schedule from 0.996 to 1 during training. To avoid training collapsing into trivially predicting a uniform distribution or single dimension for all inputs two opposing mechanisms are employed. Firstly, the sharpening of the softmax function, see Eqn 3.1, prevents the prediction of a uniform distribution by extenuating small differences in the logit values. However, this encourages the domination of a single dimension. To offset this an additional bias or centering term, c, is added to the logits of the teacher network. This term is found as the exponential moving average of the teacher network's outputs at the end of

an epoch of B batches and is controlled by the hyperparameter m,

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g\theta_t(x_i). \tag{3.5}$$

Using the DINO methodology, it has been shown that a vision transformer can be successfully trained using only the ImageNet dataset [17]. The methodology is also independent of the backbone model and so a CNN such as ResNet50 can be trivially swapped into the framework. These pre-trained models can also be finetuned on a supervised task by simply changing the loss function.

Furthermore, by visualising the self-attention paid by the class token in the final layer accurate segmentation maps can be produced, as shown in Fig 3.6. This is a feature not seen by the supervised training of ViTs. This approach is simpler and faster than the attention rollout mechanism, outlined in Section 2.1.2. Therefore it is of particular interest to the explainability analysis of the vision transformer carried out in this project.

Figure 3.6: Visualisation of the class token's self-attention in the final layer of vision transformers trained by supervision and the DINO methodology. Visually the DINO maps are superior to their supervised counterparts. [12]

The large scale and self-supervised pretraining outlined here are currently extremely costly in terms of both time and compute. Therefore the use of these techniques is restricted to a few large industrial research and development (R&D) laboratories with access to such resources. However,

all the models discussed in this section have been open-sourced [72, 78]. This means that the cost to a practitioner to fine-tune a pretrained model, compared to one from scratch, is negligible whilst providing a large boost to performance [71].

To simplify referral to the chosen pretrained architectures and resolutions, models names will be appended with their pretraining and resolution. For example, ViT-S-21k-384 refers to the ViT-S architecture pretrained on ImageNet 21k and finetuned at a resolution of 384. Whilst ResNet50-DINO-224 refers to a ResNet50 model pretrained using DINO and finetuned at a resolution of 224.

3.3 Training

The training protocol used in this dissertation closely follows that of Steiner et al. [71] and is identical for all models.

Firstly, a linear layer is attached to the output of each pretrained model's backbone and random initialised by sampling from the uniform distribution $U(-\sqrt{K}, \sqrt{K})$, where k is the number of input dimensions to the layer. Logits of length 2 are outputted by this layer and represent the model's binary classification decision. These are converted into a probability by the softmax function.

The model weights are then finetuned on the eyePACs dataset, see Section 3.1.1, by backpropagating from the cross-entropy loss of the model's outputs on batches of the training data [24]. Followinf Steiner et al.'s training protocol [71], this is achieved by using a stochastic gradient descent optimiser with a momentum of 0.9, no weight decay and gradient clipping at global norm 1. After each epoch, the model's loss on the validation set is assessed and the model weights are saved if the loss has decreased.

To mimic the large batch size used by Steiner et al. [71], whilst still working within the single GPU training regime, gradient accumulation is used. This is where gradients from multiple small batches are summated and then backpropagated. This allows for an effective batch size of 512 whilst using a batch size of 64 for 224x224 inputs and 16 for 384x384 inputs.

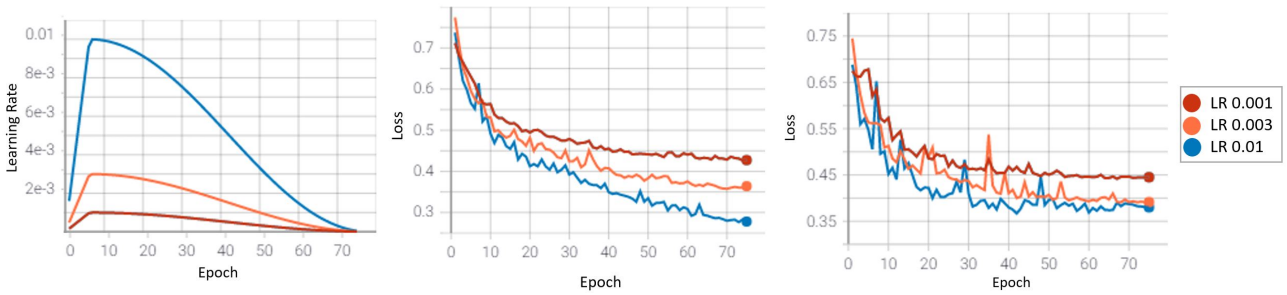
As referable cases make up only 18.8% of the eyePACs dataset the loss function is weighted for each class i as

$$W_i = \frac{N_s}{N_c N_i}, \tag{3.6}$$

where W_i is the weight of class i, N_S the total number of samples, N_C the number of classes and N_i the number of samples of class i [79]. This prevents training from collapsing to trivially predicting the

majority class for all inputs.

The learning rate evolves according to a cosine schedule over 2500 steps after a linear warm-up of 200 steps to a maximal learning rate. This rate is found by a hyperparameter search over the values $\{0.01, 0.003, 0.001\}$. An example of this learning rate search is shown in Fig 3.7 for ViT-S-21k-384.



(a) Learning Rate **(b)** Training set loss **(c)** Validation set loss

Figure 3.7: Plots of the evolution of the learning rate and losses during training of ViT-S-21k on the eyePACs dataset

Data augmentation is used heavily during training to prevent the model from remembering the relatively small number of referable cases instead of learning to identify clinically important abnormalities in the image. This takes the form of permutations randomly drawn from a list of augmentations based on those previously used by Mustafa et al. [80] to finetune large-scale pretrained ResNets to medical datasets. Namely these are; $\pm 180°$ rotation, $\pm 5\%$ translation, $\pm 5\%$ scaling, $\pm 0.1\%$ brightness, $\pm 0.2\%$ contrast, $\pm 0.2\%$ saturation, $\pm 0.02\%$ hue, Gaussian blur with a kernel size 5 and $0.1 < \sigma < 2$, horizontal and vertical flips. This training protocol is validated by the ablation study presented in Chapter 4.

3.4 Evaluation

Each model trained in this dissertation is evaluated on three criteria; its ability to classify cases of DR, its efficiency in terms of data and compute and the validity of its explainability methods. This section outlines how these are assessed.

3.4.1 Classification Performance

Evaluation of each model's ability to classify cases of DR begins with finding the threshold probability at which the model will classify an image for referral. This is done through the analysis of the model's precision and recall at different thresholds on the validation set [81] defined as,

$$precision = \frac{TP}{TP+FP}, \tag{3.7}$$

$$recall = \frac{TP}{TP+FN}, \tag{3.8}$$

where TP is the number of true positives, FP is the number of false negatives and FN is the number of false negatives. An optimal threshold is then found as the choice that minimises the Euclidean distance, d, between the maximum precision and recall scores of 1,

$$d = \sqrt{(1-precision)^2 + (1-recall)^2}. \tag{3.9}$$

A graphical representation of this process is shown in Fig 3.8. The precision-recall trade-off is considered rather than the true and false positive rate, used in the popular receiver operating characteristic (ROC) curve, as these metrics do not consider the number of true positives (TP). This is an advantage when evaluating a problem with class imbalance, such as DR classification, as the number of TP of the majority class would skew the evaluation.

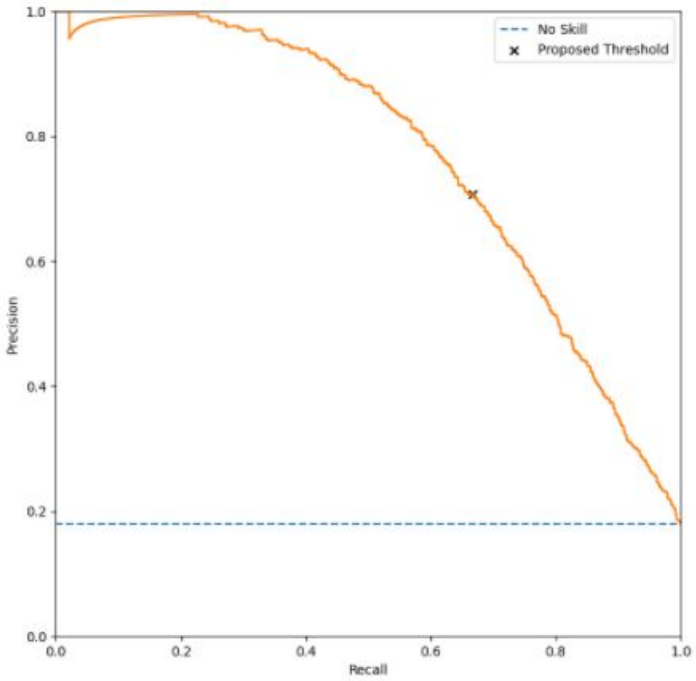


Figure 3.8: Graphical representation of the search for optimal probability threshold for ViT-S-DINO-384 using a Pre/Rec curve. This is defined as the point on the curve with the minimum Euclidean distance to the top right corner of the graph (1,1). The optimal threshold for this is example was found to be 0.70.

Using this validation threshold, predictions on the held out eyePACs test set and Messidor-1 dataset are then made. This allows for the calculation of a range of metrics that assess the performance of the classifier including; precision, recall and f1 score. Where the f1 score is the harmonic mean of precision and recall,

$$f1 = \frac{2(precision \times recall)}{precision + recall}. \tag{3.10}$$

The probability outputs of the models are also used to find the area under the curve of the precision-recall curve (Pre/Rec AUC). This provides a metric of the model's performance which is independent of its operating point.

3.4.1.1 Artificial Lesion Detection

During the development of this methodology, there was concern that as many of the retinal lesions are smaller than the patch size of the ViT-S these features would not be able to be identified. Therefore an experiment using artificiall lesions' was devised.

In this setup, ViT-S and ResNet50 models are trained only on the healthy eyePACs images. However, 50% of the images have a random placed 4x4 pixel white mask mimicking a retinal lesion and are labelled as positive, see Fig 3.9. The rest of the dataset is unmodified and assigned a negative label. The same 60:20:20 training, validation, test set split is used as in the case of the full eyePACs dataset.

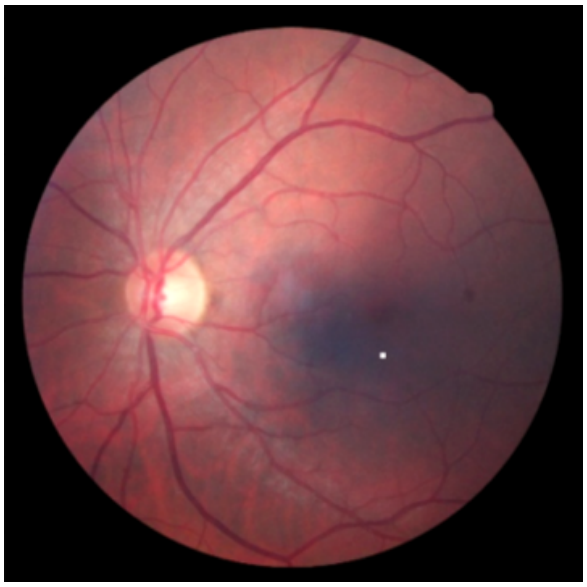


Figure 3.9: An example of an artificial DR lesion image generated by taking a healthy eyePACs image and adding a randomly placed 4x4 pixel white mask.

Both models were then trained according to methodology set out in Section 3.3 at a resolution of 384x384 pixels. Only the training of the ImageNet21k pretrained models was required to obtain a strong result from this setup. The model's classification ability has been evaluated using their F1 score on the predictions on the held-out dataset. The finding of optimal probability thresholds is not required to obtain strong results on this task.

Although a crude simulation of a DR fundus images, if ViT models do struggle to identify sub patch size features comparison of the two model's classification performance on this task would show this.

3.4.2 Efficiency

The efficiency of each model's compute is assessed by finding the average image throughput per second for a forward pass at inference time. This is done by averaging the results for 1000 batches of random inputs with a batch size of 16 for 384 inputs and batch size 64 for 224 inputs.

Data efficiency is evaluated by training each model on a fractional subset of the eyePACs dataset. The fractions used are $\{0.25, 0.5, 0.75, 1\}$. To reduce the computational and time expense of this experiment the maximum learning rate is fixed at 0.1 and only inputs of 224x224 are considered. All other parts of the training protocol outlined in Section 3.3 are followed. The size of the test set is held constant to allow comparison of the resultant models using the Pre/Rec AUC metric.

3.4.3 Explainability

ViT and CNNs have two different methods of producing saliency maps; attention visualisation [82, 12] and Grad-CAM [3]. These allow visualisation of which pixels in the input image are of greatest importance to the classification decision made by the model. The validity of these maps is evaluated by comparison to the huma- annotated retinal lesion maps of the IDRiD dataset.

In their raw form, the annotations of the IDRiD dataset are at the pixel level. However, the size of the retinal lesions varies significantly based on their type. For example, microaneurysms tend to be smaller than exudates. This means that a pixel-level comparison would be biased towards the detection of larger lesions. Therefore, building on the approach of Van Craenendonck et al. [83] each annotation map is max pooled into a grid of 16x16 patches. Explicitly, if any pixel within a given patch is labelled as a lesion the whole patch gains a positive label. An example of this process can be seen in Fig 3.2.

The patch size of 16x16 is selected as it is larger than the majority of lesions and also matches the input patch size of the ViT. To allow for comparison, the ResNet GradCam maps are also max pooled to give 16x16 patches. Furthermore, all saliency maps are normalised to sum to 1.

Several metrics are used to evaluate the validity of each model's saliency map. The first of these is hit rate which is defined as the fraction of inputs for which the pixel with the greatest intensity in the saliency map is also annotated as a lesion in the ground truth IDRiD annotations [20].

The second metric used is Pre/Rec AUC, as described in Section 3.4.1. Here multiple thresholds are used to convert the saliency maps into binary maps. This then allows for the calculation of precision and recall.

The third metric used is weighted sensitivity, W. This is found by summing over the output of the elementwise multiplication between the binary patched ground truth annotation map, G and a model's normalised saliency map, S,

$$W = \sum(G \odot S). \tag{3.11}$$

The intuition behind this metric is to show the sensitivity of the model whilst incorporating the amount of attention the model is giving to an individual patch. This information is lost in methods such as Pre/Rec AUC due to the use of thresholding.

Each model's saliency map technique is assessed by each of these metrics for an input resolution of 384. For the ViT-S models, both the attention rollout [82] and the last layer [12] methods are evaluated. Whilst for ResNet50, GradCam [3] is used.

3.5 Implementation

The methodology outlined in this section has been implemented in Python3.9 and extensively uses the Pytorch deep learning framework. All models have been trained on the UCL Computer Science High-Performance Computing Cluster. Following the principles of open and reproducible science, the results and figures presented in this report can be reproduced by running the iPython notebook found at this link.

3.6 Ethics

The main ethical consideration for this dissertation is accessing sensitive medical data. However, in the case of the eyePACs, Messidor-1 and IDRiD datasets the scans are anonymous and ethical

approval has been granted for their public release for research purposes [21, 8, 66].

Initially, it was thought that to access datasets of a large enough size and quality to conduct the work outlined in this Chapter access to Moorfields Eye Hospital private datasets would be required. Therefore permission was granted by the appointment of the researcher as an honorary research fellow at the Trust. This enabled the researcher to utilise a range of retrospective anonymised datasets for research including DR fundus images. Ethical approval for machine learning research on these datasets has recently been renewed for five years (IRAS 281957, HRA/20/2158), beginning May 2020 and so covered the entirety of the project. However, to allow the results of this dissertation to be easily reproducible it was decided to only use public data, especially as large labelled DR datasets are easily available.

3.7 Summary

To summarise, the eyePACs dataset is preprocessed and then used to finetune ViT-S and ResNet50 models at resolutions of 224x224 and 384x384. These models have then be pre-trained either on the large ImageNet-21k dataset or using the self-supervised DINO methodology.

The finetuned models are then evaluated on the basis of their classification performance, efficiency and explainability. A range of metrics has been defined to do this. The Messidor-1 and IDRiD datasets are used as well as the held out eyePACs dataset for various parts of this evaluation.

This methodology has been implemented in Python and the code is publicly available to aid the reproducibility of the project. No major ethical issues were faced in the dissertation due to the use of public data.

Chapter 4

Results

This chapter presents the results of the experiments outlined in Section 3.4. This includes studies of each model's ability to classify cases of DR, efficiency and explainability methods as well as an ablation study of the training methodology used.

4.1 Methodology Ablation Study

The training methodology presented in Section 3.3 has been evaluated by the ablation study presented in Table 4.1. The removal of each aspect of the methodology results in a decrease in the performance of both the ViT-S and ResNet50 models. This validates the choices made in Section 3.3. The effect on both model types of each ablation is roughly equal.

	ViT-S-21k-384	**ResNet50-21k-384**
Baseline	0.864	0.821
No data augmentation	0.804	0.7554
No gradient accumulation	0.805	0.771
Ungradables included	0.889	0.845

Table 4.1: Pre/Rec AUC of models trained for ablation study of training methodology on eyePACs held out test set.

4.2 Classification Performance

The ability of each model to classify referable cases of DR has been evaluated through the calculation of the precision, recall, F1 and Pre/Rec AUC metrics. This has been done for both the held out eyePACs test set and the external Messidor-1 dataset. These results are presented in Table 4.2.

Model	Precision	Recall	F1	Pre/Rec AUC
ViT-S-21k-384	0.744/0.895	0.724/0.763	0.734/0.824	0.820/0.915
Resnet50-21k-384	0.786/0.882	0.778/0.897	0.782/0.889	0.864/0.954
ViT-S-DINO-384	0.763/0.871	0.690/0.768	0.724/0.816	0.811/0.914
Resnet50-DINO-384	0.740/0.895	0.661/0.549	0.698/0.680	0.787/0.846
ViT-S-21k-224	0.742/0.852	0.616/0.656	0.673/0.741	0.760/0.859
ResNet50-21k-224	0.775/0.898	0.671/0.763	0.720/0.825	0.807/0.913
ViT-S-DINO-224	0.678/0.834	0.669/0.694	0.673/0.758	0.767/0.876
ResNet50-DINO-224	0.608/0.835	0.628/0.487	0.618/0.615	0.702/0.782

Table 4.2: Classification metrics for models finetuned on the binary eyePACs classification task. Metrics calculated for model performance on eyePACs/Messidor-1 dataset

Table 4.2 shows that for nearly all ImageNet-21k pretrained models the ResNet50 model outperforms the ViT-S model but the converse is true for the DINO pretrained models. However, the ResNet50-21k models outperform their DINO counterparts with ResNet50-21k-384 showing significantly better performance than any other model. A the performance gap between the ViT-S 21k and DINO models is much smaller or even reversed in comparison the ResNet50 models.

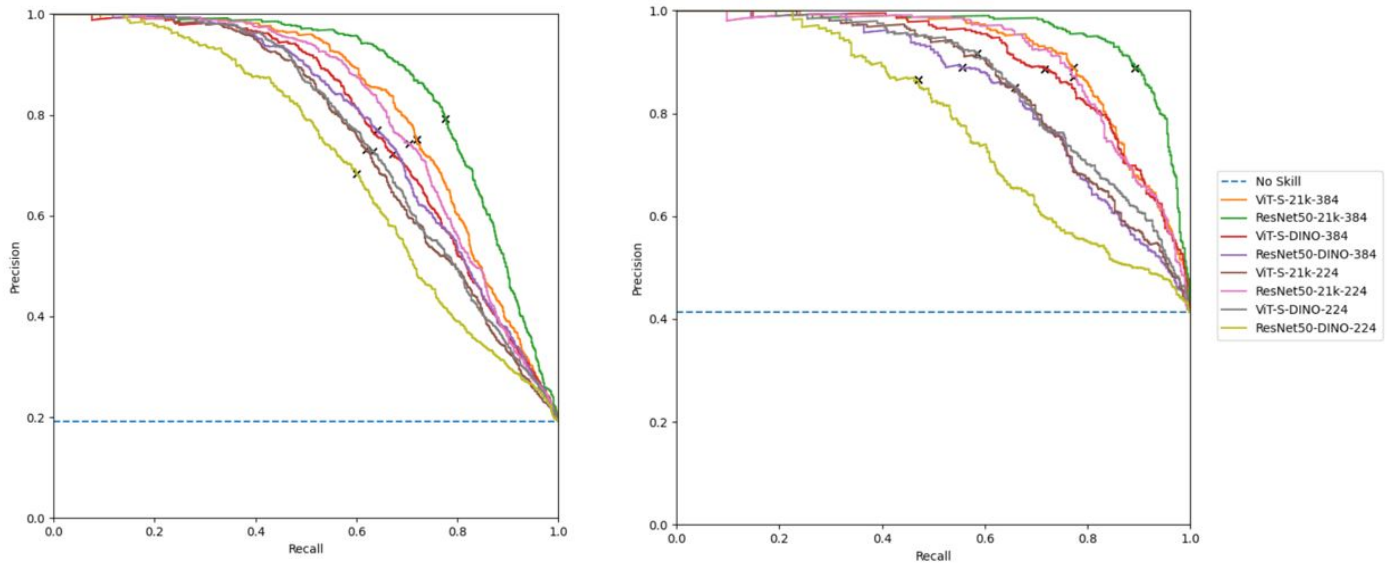
Furthermore, it can be seen that for all models increasing the image resolution results in a boost in performance. This is expected as the lesions present in the images can be relatively small especially in mild cases. Hence increasing the resolution increases the number of abnormal pixels inputted into the model making classification a simpler task. The cost of this performance gain is the decrease in image throughput as discussed in Section 4.3.

Another feature of the classification results is that for all models the Pre/Rec AUC metric increases from the eyePACs to the Messidor-1 dataset. This shows the models trained in this dissertation generalise well to unseen data. However, it may also point to the previous reported [11, 69] lower quality and ungradable images of the eyePACs dataset.

However, for the lowest-performing models, ViT-S-21k-224, ViT-S-DINO-224 and ResNet50-DINO-224, this increase in AUC may be misleading as this gain is not reflected in these model's recall and so F1 score. The reason for this is a large shift in the optimal probability threshold. This can be seen in Fig 4.1b in which the optimal thresholds found on the eyePACs validation set are clearly at a sub-optimal point on their respective curves.

This is a more subtle but still important loss of generalisation and has an impact on their potential

use as even though these models have the ability to classify cases of DR they require finetuning of their threshold on new datasets. For the high performing and more robust models, this is not required. Notably, all these models nearly require 384 input resolution.



(a) eyePACs test set

(b) Messidor-1

Figure 4.1: Pre/Rec curves for all finetuned models. Black cross shows operating point found on the eyePACs validation set

The results from the artificial lesion detection experiment, see Section 3.4.1.1, show that the difference in performance between the ViT and CNN models is not due to the failure of ViTs to recognise sub-token sized features. When trained both the ViT-S-21k-384 and ResNet50-21k-384 models could perform this task with near- perfect accuracy achieving F1 scores of 0.990 on the held-out test set.

4.3 Efficiency

Table 4.3 shows that for the two input sizes evaluated, ViT-S has a lower image throughput than ResNet50. The pretraining methodology used does not affect the downstream inference speed of the model so the results for only the ImageNet-21k models are shown.

Furthermore, as mentioned in the previous section the cost of increasing the classification performance by increasing the image resolution is image throughput. This is also shown by the results in Table 4.3. The decrease in ViT throughput is due to the quadratic increase of the number of input

Model	Image throughput (images/s)
ViT-S-21k-384	125.1
ResNet50-21k-384	195.8
ViT-S-21k-224	464.2
ResNet50-21k-224	578.3

Table 4.3: Inference throughput of the ViT-S and ResNet50 model at 384 and 224 image resolution.

tokens, N_t, with image resolution

$$N_t = \left(\frac{Image\_resolution}{Patch\_size} \right)^2. \tag{4.1}$$

For the ViT-S models trained in this dissertation which has a patch size of 16, a 224x224 image requires 196 tokens whilst a 384x384 image requires 576. Furthermore, a larger input image size increases the memory footprint of a forward pass on a GPU. Hence the results presented here use different batch sizes as outlined in Section 3.4.2.

For deployment in the clinic, even the minimum inference speed of 125.1 images/s is much greater than the throughput of a large screening clinic. Therefore the differences in the inference speeds of the models have little effect on their clinical utility. It is also of note that these image throughputs are dramatically higher than a human clinician performing the same task. However, inference speed does dramatically impact training time in which 10,000s of images are passed through the network multiple times. This meant that further increases to the input image resolution were unfeasible whilst staying within the single GPU training regime.

The second aspect of efficiency that has been explored in this dissertation is that of data. The results of the experiments outlined in Section 3.4.2 are visualised in Fig 4.2. It is clear that all the models' performances increase with increased training data. However, this is very modest for the ResNet50-DINO model.

The ResNet50-21k model has superior performance at all data fractions compared to any other pretrained model studied here. Furthermore, unlike the other models, its performance does not seem to be saturated by data set size. A caveat to these results is that no hyperparameter search has been performed and so the performance may be artificially low for some models.

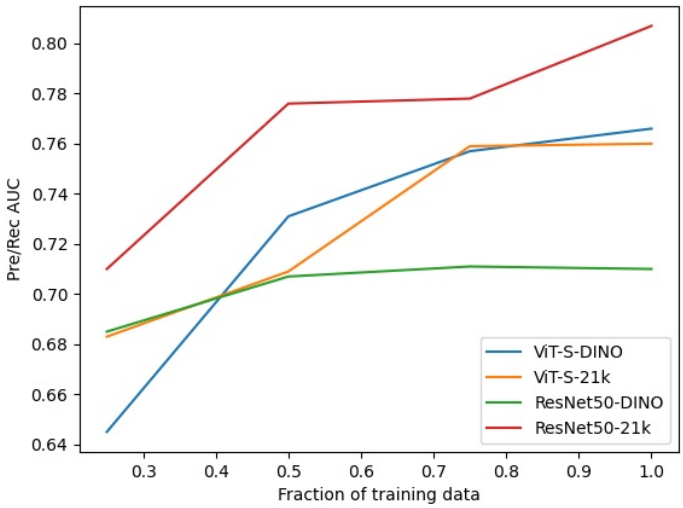


Figure 4.2: A plot of the Pre/Rec AUC of models finetuned on varying fractions of the eyePACs dataset at 224x224 image resolution. Evaluation conducted on the eyePACs held out test set.

4.4 Explainability

Fig 4.3 visualises of saliency maps produced by each model when applied to the same subset of samples from the IDRiD dataset. The raw and 16x16 patched ground truths are also shown for reference. Visual inspection of even this small sample reveals some distinguishing features of each method. Most evident is that the GradCam maps give a number of patches of high variance over the feature of interest. Whereas the attention maps highlight smaller discrete patches with low variance. It is also clear that no method or model highlights all the lesions present.

Both the attention methods of the Vit-S-21k model highlight regions other than retinal lesions, for example the fully black background. This is concerning as this may indicate the model is making its classification decision based on input features that are not clinically relevant.

It is also apparent that the last layer and attention rollout methods give different explanations of the same model's decision. In general, areas of high attention in the last layer map are also highlighted by attention rollout. However, attention rollout maps can also contain additional features as well as a higher level of background noise. Notably, one of these extra features is consistently the retina which is not an area considered by clinicians in the diagnosis of DR.

Finally, the GradCAM maps of 21k and DINO pretrained ResNet50 are generally similar. However, the DINO pretrained model maps have lower variance and can highlight additional features compared to the 21k model.

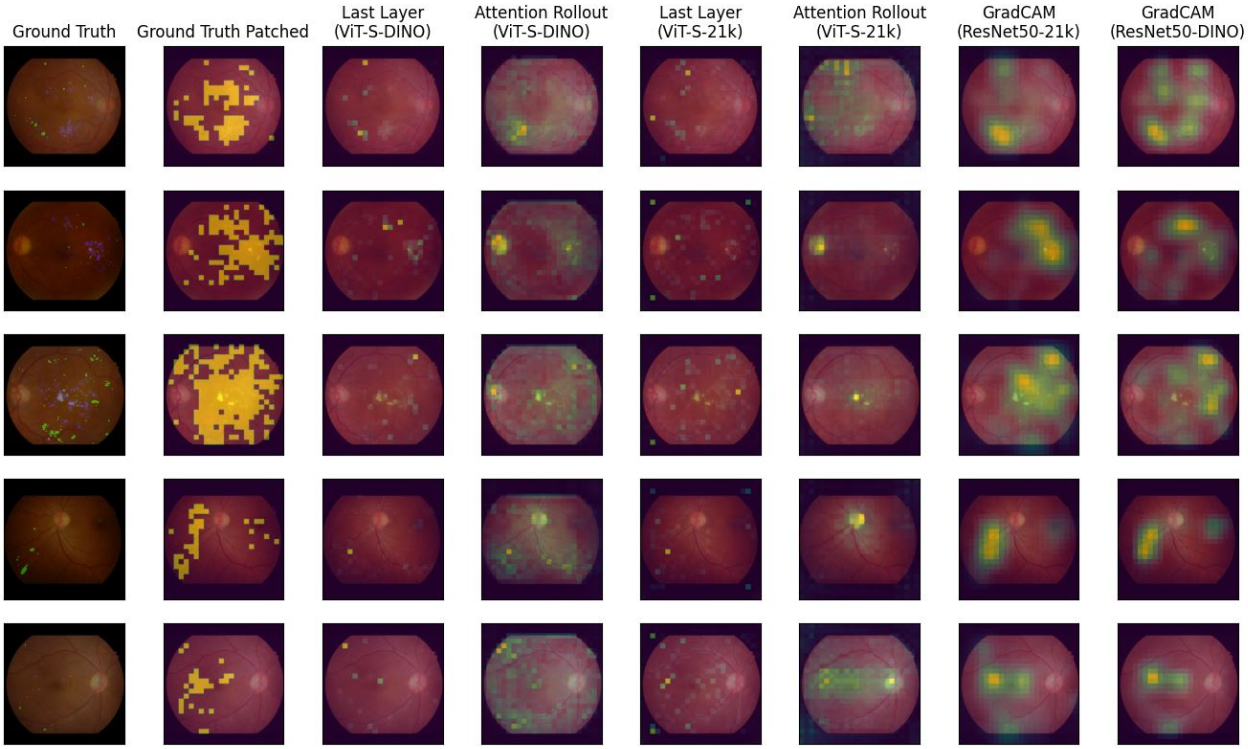


Figure 4.3: Visualisation of each model's explainability methods on a random subset of the IDRiD dataset. All input images are at 384x384 image resolution

To quantify the difference in the maps, weighted sensitivity and top pixel hit rate, as defined in Section 3.4.3, have been calculated for each model's explainability method, see Fig 4.4. The results show that the ViT-S-DINO model has the highest hit rate and weighted sensitivity followed by the GradCAM maps. The results for ViT-S-21k is the lowest for both methodologies but still above random. Furthermore, the last layer methodology is superior for the ViT-S-DINO using these metrics but it is unclear for ViT-S-21k.

Precision recall curves for each saliency map have been calculated by converting the maps to binary using a range of threshold. As can be seen in Fig 4.5, at low thresholds the ViT-S-DINO model has very high precision but low recall using the last layer method. However, as the threshold falls noisy low attention areas become positively labelled and hence the precision falls rapidly. This reaffirms the previous qualitative description that this method focuses with high sensitivity on a limited number of valid patches.

Analysis via these metrics shows that for both ViT-S-DINO and ViT-S-21k, the last layer has superior precision at high thresholds than attention rollout. However, the performance of ViT-S-

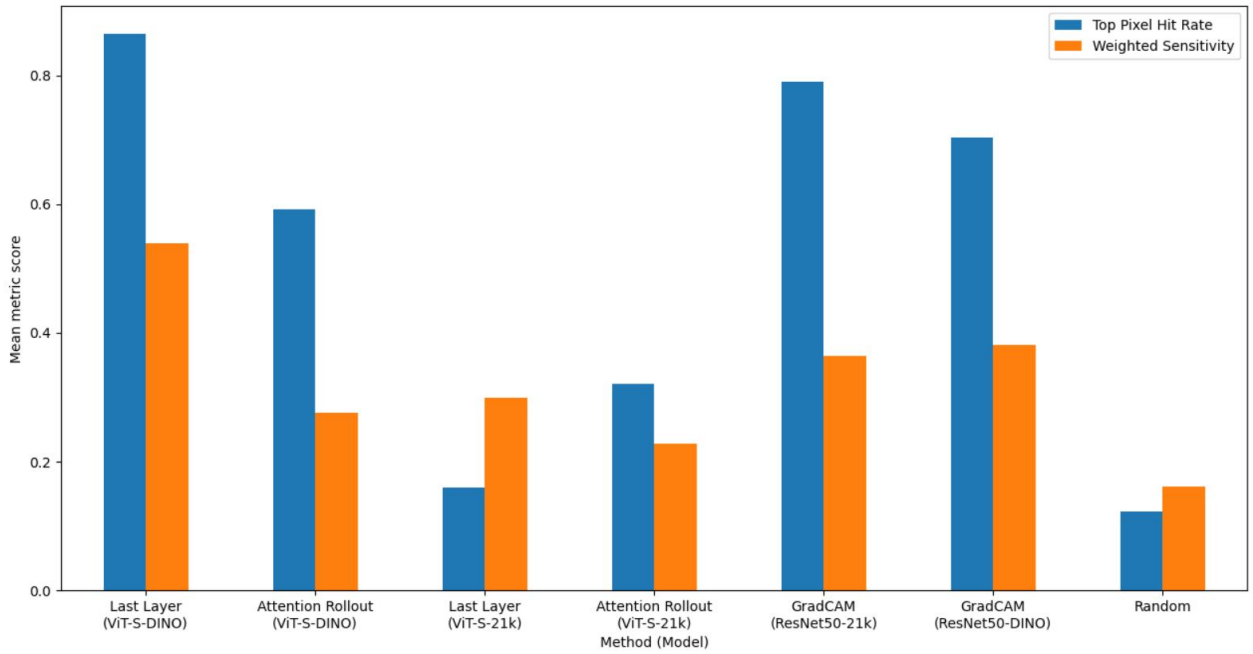


Figure 4.4: Graph of weighted sensitivity and hit rate analysis for each model's explainability methods. All evaluations have been done with 384x384 inputs.

21k is notably lower. For the Grad-CAM models as they have a more diffuse nature their precision although low remains stable with the drop in threshold.

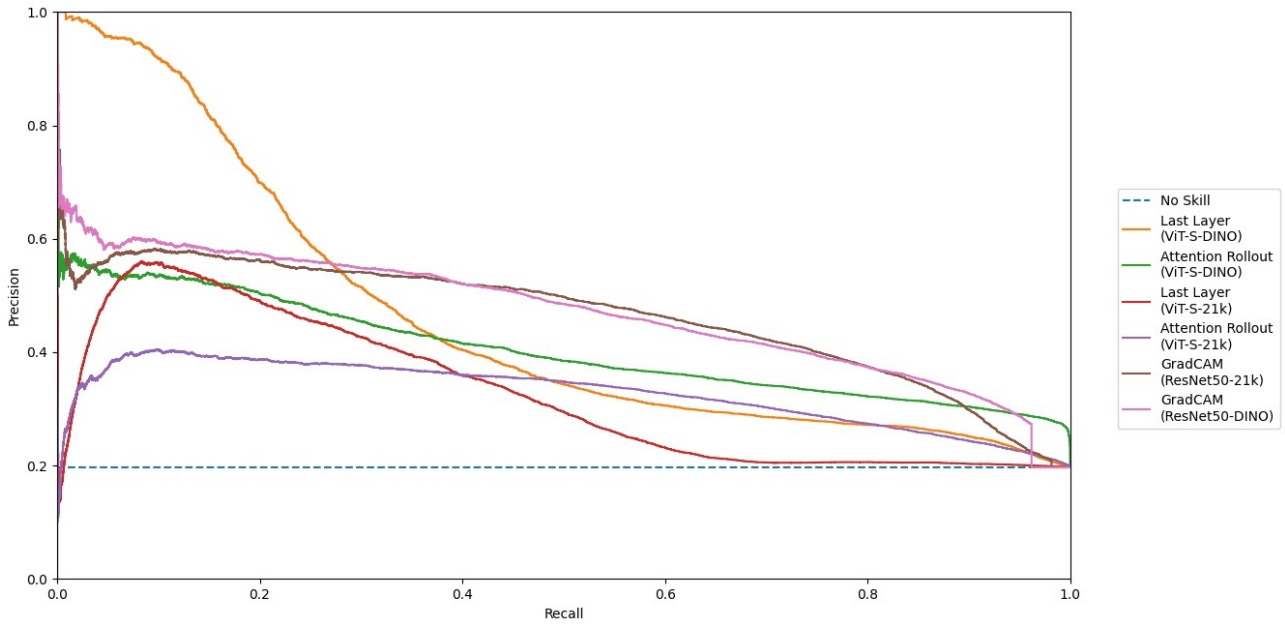


Figure 4.5: Precision recall curve of each model's explainability methods on the IDRiD dataset at a resolution of 384x384.

4.5 Summary

To summarise, the results presented here show that ViT-S has an inferior classification performance, image throughput and data efficiency when compared to ResNet50 on the eyePACs DR classification task. However, visualisation of the final attention layer gives a higher precision saliency map than is possible using GradCAM.

Chapter 5

Discussion

The aim of this dissertation has been to evaluate if ViTs provide a solution to the poor efficiency and explainability of CNNs while maintaining the high performance seen when such models are finetuned to medical imaging tasks. The results of Chapter 4, show the ViT-S model does not have a significantly improved classification performance or efficiency, when compared to the ResNet50 architecture, on the DR classification task. However, it has been shown that visualising the last attention layer of the ViT-S-DINO model gives higher precision saliency maps than those produced by GradCAM. An in-depth discussion of these results is presented in this chapter, as well as an exploration of the methodological limitations and potential future work resulting from this dissertation.

5.1 Classification Performance

The metrics presented in Table 4.2 clearly show that the ResNet50 models outperform the ViT-S model at the DR classification task for both 21k and DINO pre-training. This is despite the superior classification accuracy of the ViT-S model on the ImageNet dataset.

The reasons for this have not been conclusively found. However, it is evident that the domain shift between identifying objects such as cats and dogs in ImageNet to recognising retinal lesions on a fundus image is large. One of the biggest differences between these datasets is the size of the features of interest. In ImageNet the class object takes up a large percentage of the image whereas in a DR fundus image a lesion can be less than 5 pixels in diameter. Therefore, an initial hypothesis of the performance difference was that the ViT model struggled to identify the sub-patch size retinal lesions. However, experiments placing 'artificial lesions' of white 2x2 pixels on a random selection of healthy fundus images showed these could be identified with near-perfect accuracy by the ViT-S model.

Another explanation is that ViTs may be intrinsically more difficult to train than ResNets. Work by Chen et al. [13] to describe the loss landscape of the two model classes showed that ViTs have an extremely sharp local minima, see Fig 5.1. Hence convergence by gradient descent is more difficult. In their work, Chen et al. said this explained the need for large scale pretraining or could be overcome using a sharpness-aware (SAM) optimizer. They went on to show ViT-SAM had a superior performance on downstream tasks than a vanilla ViT. Work to characterise the loss landscape for the DR task and analyse the potential benefits of the SAM optimiser would be an interesting area for future work.

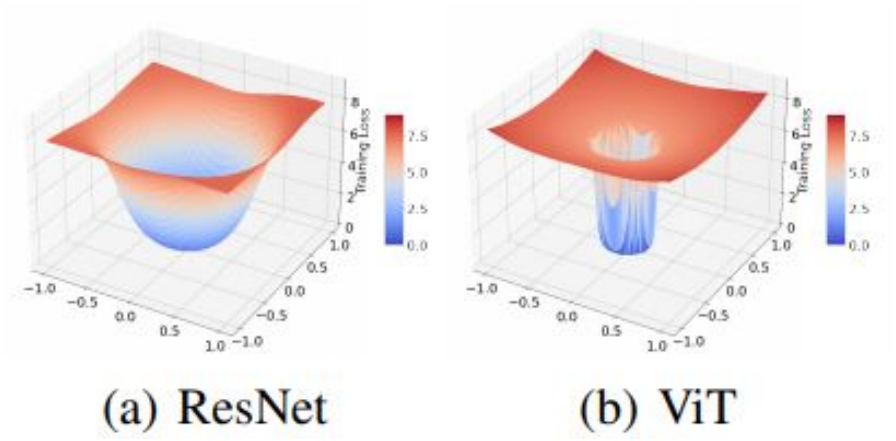


Figure 5.1: Loss landscape of ResNet and ViT when trained on ImageNet [13].

It is also worth considering the work of Ke et al. [9] which showed that the ability of a model to adapt to a new medical image domain is weakly correlated by the model's ImageNet performance, as shown in Fig 2.12. This is especially true when the performance of the models being compared are within a percentage of each other. For a given model, its ability to be finetuned to a new dataset with different feature sizes, label noise and pixel intensity distribution cannot be currently be predicted. Therefore, empirical research on individual datasets such as that carried out here is still necessary.

Although the ViT models trained in this dissertation have clear limitations it is clear that the model is still able to classify the majority of DR cases accurately. Furthermore, the ability of all the models to generalise to the unseen external validation Messidor-1 dataset is impressive. This shows the models' robustness to shifts in the data origin both in terms of population demographics and camera types. The noisy nature of the eyePACs dataset and heavy data augmentation although not without their issues are two of the main reasons for the generalisation seen across all models.

5.2 Efficiency

5.2.1 Computational

The results in Section 4.3, show that the inference speed of the ResNet50 model is superior to ViT-S at both 224 and 384 resolution. However, for both models increasing the image resolution dramatically decreased the inference speed. Furthermore, increasing the image resolution size increases the memory footprint of a batch when training. This means that increasing the image resolution size further requires a decreased batch size to less than 16 during training. Therefore, the effect of image size on training time is even greater than that shown in Table 4.3.

This constraint meant that investigations with greater resolution images have not been pursued even though the results suggest this could lead to greater classification performance. Increasing resolution tends to increase performance [23] but in the case of DR classification, it is further motivated when the small size of the retinal lesions, such as microaneurysms, are considered. It is highly likely that downsampling the input results in the loss of some of these small lesions as noise. This makes the mild cases of DR even more challenging and perhaps impossible in some cases as the model's input does not contain the lesion.

Developing computationally efficient deep learning models that use high-resolution inputs is an active area of research. For the ViT, the quadratic scaling of the attention matrix with the number of tokens is the bottleneck here. The recently proposed Perceiver model [14], see Fig 5.2, proposes a solution to this. The proposed model is a transformer that accepts a compressed form of the input. This compressed form is found by an attention layer, see Section 2.1.2, in which the key and value vectors are calculated from the large input but the query vector from a low dimensional latent array. Therefore, the output of this layer is the length of the smaller latent array. Hence, inputs that were previously too large to be processed by a transformer, such as high-resolution images, are compressed into a latent array which is computationally tractable. The Perceiver also generalises the transformer to accept multi-modal forms of data. In future work, such a model could be plausibly trained on full resolution fundus images.

5.2.2 Data

The evaluation of the data efficiency of both models showed that both the ResNet and ViTs model's performance is directly correlated with the size of the training dataset. Furthermore, considering the

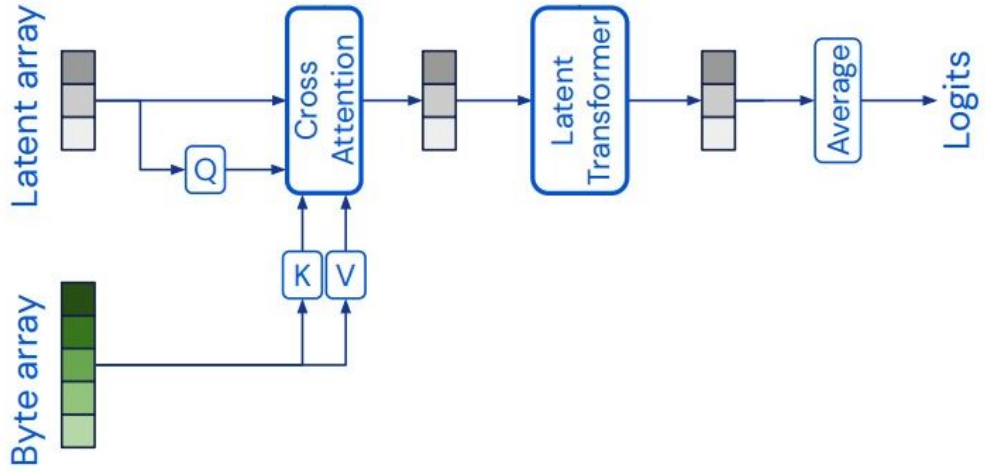


Figure 5.2: Single layer Perceiver architecture [14]. Q, K and V denote the query, key and value vectors respectively. The length of the Q vectors is much less than the length of K and V

superior results of studies that have trained models on datasets containing over 100,000 fundus images [10, 64] shows that this relationship is not saturated by the currently available public datasets.

On the one hand, it could be said that as there is the ability, funding and appetite to produce datasets of this size the data-hungry nature of the current SOTA models is offset. However, this is only the case for commonly seen diseases such as DR which are tested for regularly in the population i.e. a screening program exists. Furthermore, such an approach is highly laborious, requires large program coordination and raises issues of data privacy. Therefore, a model which could be finetuned on 100s of images would allow for a rapid expansion in the use cases and ease of deployment of medical CV models.

Although the ViT architecture does not solve the problem of data efficiency it may point to future solutions. The use of transformers in NLP has been successful not only due to the ability of the architecture to model long-range interactions but also the extensive use of self-supervised learning to train these models. For example, GPT-3 [22] is pre-trained using a language generation task on a text corpus of 499 billion tokens. This then allows the model to be finetuned to a downstream task using few-shot or even zero-shot learning. It is of note these models also tend to be extremely large.

In CV no such analogous model exists. The development of self-supervised training methods, such as DINO [12], means that steps in this direction are being taken though. However, as yet such techniques have only been applied to the ImageNet dataset. As empirical evidence shows supervised training tends to outperform self-supervised training if labels exist, the currently available self-

supervised pretrained models show inferior performance to their supervised peers.

It is also unclear if a ViT architecture would be advantageous for such a training methodology but self-supervised methodologies specific to the transformer architecture have been proposed [84]. Additionally, for DINO methodology the ViT showed superior performance to ResNet on ImageNet and similar performance on eyePACs. However, neither of these studies give conclusive evidence in favour of ViTs.

There are downsides to this large scale self-supervised training though. Firstly, this approach is extremely expensive. Hence only large industry R&D labs have the monetary and computing resources to develop such a model. The community then has a dependency for such a model to be open-sourced. Moreover, the environmental impact of developing such a model should not be underestimated.

Arguably such an approach mimics the learning process of humans more closely. For example, when a medical student is first learning to classify fundus images they have previously seen a vast number of 'images' if the sense of sight is considered as a continuous stream of input unlabelled input images. Then the student is presented with 10s of labelled textbook DR examples before reaching higher performance through exposure to 100s of nosier and more complex cases as a junior doctor. They are then licensed to perform autonomous classification as a consultant once a baseline level of competency has been met. Finetuning a CV model on 100s of labelled images after large scale pre-training would follow this learning process more closely than the current SOTA.

Another option to increase the data efficiency of CV models is to use richer data sources. Currently, the setup for a classification task is simply a dataset of integer labelled images. However, when compared to either the input or output of a human medical decision this is very sparse. For example, a doctor would not make a medical diagnosis based on a single fundus image. They would take into account an array of additional patient-specific information taken during a consultation including their family history and symptoms. Furthermore, once a diagnosis had been made it would not be a simple binary healthy vs refer as outputted by the models in this dissertation but a short report including information such as the location and type of lesions present.

Recently, such information has become accessible due to the extensive use of electronic health records. Using such free language data in conjunction with imaging has previously been difficult due to the differing SOTA model architecture. However, with the convergence of vision and language

models in the form of transformers multi-modal inputs and outputs may now be possible [14]. Little work has been done in this area but it would seem logical that increasing the richness of the data source would decrease the amount required.

5.3 Explainability

When comparing the GradCam maps to the attention maps, in Fig 4.3, the two differing mechanics behind the classification decision are evident. The GradCam maps give a number of patches of high variance over the feature of interest. This is a result of the gradual dimensionality reduction of the architecture from a full image to a low dimensional representation through repeated max-pooling layers. In contrast, the ViT-S understanding of relative pixel location is completely learnt. Therefore, there is no inductive bias that patches adjacent to a patch with high attention also has high attention. This results in attention maps highlighting discrete patches with low variance. Furthermore, the patches in the last layer of the DINO model are shown to be selected with very high sensitivity as shown by the high hit rate.

The results of Section 4.4 shows that the models are highlighting the retinal lesions in the fundus images with varying levels of accuracy. Notably, no models highlight all retinal lesions present. Hence the AUC of the curves shown in Fig 4.5 are relatively low. This may be due to the nature of the task the models have been trained to perform. To classify an image for referral only the presence of a few lesions are required the rest are superfluous. However, if the models were trained for a multi-class task the number and severity of lesions would also be of importance to the decision. Hence a more complete saliency map may be produced.

However, when viewing the patched ground-truth maps in Fig 4.3 the image is extremely noisy. At a resolution of 384x384 resolution, a patch size of 16x16 is large especially compared to the average lesion size. Therefore, it must be questioned even if the ground-truth could be produced what clinical utility it would provide in this patched form. It is possible to simply reduce the size of the patches, for example an 8x8 DINO ViT-S exists [12]. However, this is at the cost of a four times increase in the number of patches and so slower inference time and this would still not provide the ideal pixelwise saliency map.

This challenge is highly task-specific though, as for most CV applications the feature of interest is of a size greater than 16x16, such as the objects in 3.6. So for medical applications with larger

features of interest, for example tumour detection, this would not be an issue. Another alternative would be to increase the input resolution, as discussed previously, decreasing the difference in lesion and patch size.

Although the clinical utility of the maps is questionable they do provide a sanity check that the models are focusing on valid biological abnormalities. For the ResNet50 models and ViT-S-DINO, this is true. However, the ViT-S-21k maps show evidence of paying the model is paying attention to the black corner areas. These areas do not have lesions and so this behaviour is concerning. Furthermore, the explainability metrics for all ViT-S-21k models are inferior to that of their DINO counterparts. This is despite showing equal or superior classification performance. The reasons for these superfluous connections are not clear as all models were trained using the same data and protocol. However, this does show the utility of such methods in checking the validity of a model's decision before potential deployment

It is also interesting to compare the maps created when accounting for all attention layers through attention rollout or just the last layer. Analysing the examples in Fig 4.3, it can be seen that the patches highlighted in the last layer also appear in the rollout map. This is to be expected as recursive rollout calculation includes the final layer. However, the additional information gained by looking at the initial layers seemingly adds a general layer of noise with increased attention paid to the retina. This suggests in the earlier layers the model is checking each patch for the presence of a lesion before highlighting the abnormal areas in the final layer before classification. The additional retinal attention could be explained due to the similarities in colour and shape to large exudates. However, pleasingly this is never highlighted in the last layer maps suggesting it is discarded as a healthy feature during the model's inference process. Therefore, surprisingly accounting for the earlier attention layers makes reduces the utility of the maps produced and can instead produce subtle artefacts. It is also of note the last layer method is simpler and therefore computationally faster.

5.4 Limitations

The ablation study presented in Table 4.1 validates many of the choices made in the training methodology. However, there are aspects of this protocol worth further exploration.

Firstly, heavy data augmentation is used when training each model on the eyePACs dataset, see Section 3.3. This prevents overfitting of the highly expressive models to the training data and results

in a boost to Pre/Rec AUC of more than 0.05 for both the ResNet50 and ViT-S models. However, a common criticism of such an approach is that the model is then trained on biologically implausible images. To an extent, this is true in this project. For example, as the fundus images commonly have a larger horizontal than vertical dimension the top and bottom percentile of the retina can be missed in the original scan. Randomly rotating these images after square cropping gives an input to the model that is impossible not anatomically possible. However in the defence of the methodology used, none of the augmentation or pre-processing methods that are applied to the input images remove or add retinal lesions. Therefore the ground-truth label, if correct, is always truthful to the input and so even if the image as a whole is implausible the target features to be learnt are still valid.

The second aspect worth consideration is the method used to find the probability threshold for each model. In this project, this is done by minimising the Euclidean distance of the model's precision and recall from their optimal values on the validation set, see Eqn 3.9. However, the models in this dissertation are tasked with classifying cases of a medical disorder. Therefore assuming that the cost of an FP and FN are equal may not hold. Explicitly the impact of a model incorrectly diagnosing a healthy case as DR compared to missing a case altogether is very different. Decreasing the threshold will result in fewer missed cases but at the cost of decreasing the discriminate power of the model and hence the utility of the screening process. Whilst increasing the threshold will have the opposite effect.

The evaluation of where the precision-recall balance lies for a real-world application is multi-factorial, for example if the model would be deployed as a second reader or fully autonomously. Therefore, previous studies have instead tuned this threshold manually. This discussion points to the utility of metrics such as Pre/Rec AUC which is independent of the threshold probability

In regard to the threshold probability, it is also of note that for weaker models this fails to generalise. This can be seen in Fig 4.1. For all models on the eyePACs test set, the optimal threshold from the validation set approximately holds. However, for the Messidor-1 dataset only the top-performing models, such as ResNet50-21k-384, show generalisation of their thresholds. For example, the ResNet50-DINO-224 validation threshold is 0.68 but when calculated for the Messidor-1 dataset this drops to 0.27. This failure of model generalisation is hidden by the AUC metric which increases for all models between the eyePACs and Messidor.

An explanation of these results could be that the proportion of referable cases is significantly

higher in the Messidor-1 dataset as shown by the higher 'no skill' boundary in Fig 4.1. The higher-performing models could adjust to this domain shift whilst the others could not. This result shows the importance of analysing models through multiple metrics.

Finally, the eyePACs dataset has been used to finetune all the models in this project. This decision was made on the basis of it being the largest publicly available dataset. However, as previously described the images are of questionable quality. Even though the ungradable images labelled by a previous study [11] are removed a non-exhaustive search can easily find images of questionable quality in the remaining dataset. Furthermore, the labelling protocol used to generate the dataset is unclear. However, many studies have employed multiple experts to relabel the data using systems such as majority vote [10]. Such studies in general show very good performance for models trained on this cleaner data source. However, such a method is costly due to the need for multiple expert annotators and so was not available or desired to be used in this project.

5.5 Future Work

During the preceding discussion, several areas of potential future work have arisen. Many of these relate to improved iterations of the ViT architecture or training methodology. These include using the SAM optimiser [13] to potentially improve the convexity of the model's loss landscape and so boost classification performance, reducing the patch size of the input tokens [12] to give higher resolution attention visualisations or using a Perceiver based architecture [14] to increase the input image resolution.

A further idea for future work is to use the convergence of NLP and CV architectures to train models using richer data sources. Instead of the current approach of using binary or integer labels to train a CV classification model the medical reports written by clinicians could be the target labels. This is a natural extension of the ViT architecture. The transformer used in this dissertation is the encoder half of the encoder-decoder model of the original transformer proposed for language translation [15]. Hence, treating the problem as an image to language translation task the image representation learnt by the ViT could be decoded into free text by a decoder. This richer nature of the label, which would include ideas such as the lesion type and position, could result in an increase in classification performance or data efficiency.

Further ideas on how to increase the data efficiency of deep learning models on downstream

tasks using ViTs tie into a larger movement within the domain to develop 'foundation models' [85]. Such models are pre-trained on immense amounts of general data which usually requires the use of an unsupervised task. The extensive pre-training then means that only extremely small datasets are required for finetuning. In the domain of NLP, these exist in the form of models such as BERT [5] and GPT3 [22] but no such CV models yet exist. If such models should take the form of a ViT or CNN is still an open question. Though the larger issue with such an approach is that the time and compute resources required to train such a model are prohibitive for all those apart from the largest industry R&D laboratories. However, the evaluation of how these models, if open-sourced, perform on critical downstream tasks, such as medical imaging, will be vital future work.

Chapter 6

Conclusion

In conclusion, this report has presented the performance of ViT-S [6] and ResNet50 [27] models when finetuned to a binary DR classification task using the eyePACs dataset [21]. These model's weights are initialised either from the result of pre-training on the large ImageNet 21k dataset [73] or unsupervised training according to the DINO methodology [12].

Comparison has shown that the ResNet50 model is superior in terms of classification accuracy, data and computational efficiency. Furthermore, the models pretrained using the supervised ImageNet-21k models are found to be superior to the unsupervised DINO models in terms of classification ability and data efficiency.

However, the visualisation of the ViT-S-DINO last attention layer highlights a few highly precise patches which robustly overlap with the ground-truth segmentation maps of the IDRiD dataset [8]. This contrasts with the high variance maps created by GradCAM when applied to the finetuned ResNet50 models. Although, the ViT-S-21k models showed a concerning tendency to highlight areas other than retinal lesions.

The results presented in this report do not fully answer the problems of explainability or data and compute efficiency of CNNs, there is the potential for future models to do so based on a transformer architecture. New methods of exploiting large unsupervised pretraining, multi-modal data sources and high-resolution inputs are now possible using a transformer architecture. Therefore, ViTs hold the potential to form the basis of the next generation of CV and automated medical imaging tools.

Bibliography

[1] Sumit Saha. A comprehensive guide to convolutional neural networks — the eli5 way. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd Accessed: 10-08-2021.

[2] Image classification on imagenet- paperwithcode. https://paperswithcode.com/sota/image-classification-on-imagenet. Accessed: 02-03-2021.

[3] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[4] Toon Van Craenendonck, Bart Elen, Nele Gerrits, and Patrick De Boever. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Translational Vision Science & Technology*, 9(2):64–64, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Jacob Gildenblat. Exploring explainability for vision transformers. `https://jacobgil.github.io/deeplearning/vision-transformer-explainability`. Accessed: 10-08-2021.

[8] Porwal Prasanna, Pachade Samiksha, Kamble Ravi, Kokare Manesh, Deshmukh Girish, Sahasrabuddhe Vivek, MacGillivray Tom, Sidibé Désiré, Giancardo Luca, Quellec Gwenolé, and Meriaudeau Fabrice. Indian diabetic retinopathy image dataset. `https://idrid.grand-challenge.org/`. Accessed: 03-08-2021.

[9] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 116–124, 2021.

[10] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[11] Mike Voets, Kajsa Møllersen, and Lars Ailo Bongo. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PloS one*, 14(6):e0217541, 2019.

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[13] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

[14] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[19] Andrzej Grzybowski, Piotr Brona, Gilbert Lim, Paisan Ruamviboonsuk, Gavin SW Tan, Michael Abramoff, and Daniel SW Ting. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*, 34(3):451–460, 2020.

[20] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021.

[21] Kaggle. Diabetic retinopathy detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection`. Accessed: 03-08-2021.

[22] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[23] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer*

Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pages 491–507. Springer, 2020.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[26] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[30] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[31] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021.

[32] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the (un) trust-

worthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv preprint arXiv:2008.02766*, 2020.

[33] Sina Mohseni, Akshay Jagadeesh, and Zhangyang Wang. Predicting model failure using saliency maps in autonomous driving systems. *arXiv preprint arXiv:1905.07679*, 2019.

[34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[35] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[36] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[38] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064. PMLR, 2018.

[39] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

[40] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[41] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[42] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving:

Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011.

[43] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[44] Joseph Giorgio, Susan M Landau, William J Jagust, Peter Tino, Zoe Kourtzi, Alzheimer's Disease Neuroimaging Initiative, et al. Modelling prognostic trajectories of cognitive decline due to alzheimer's disease. *NeuroImage: Clinical*, 26:102199, 2020.

[45] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

[46] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[47] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[48] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. *arXiv preprint arXiv:2102.10662*, 2021.

[49] Yundong Zhang, Huiye Liu, and Qiang Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. *arXiv preprint arXiv:2102.08005*, 2021.

[50] Daniel T Hogarty, David A Mackey, and Alex W Hewitt. Current state and future prospects of artificial intelligence in ophthalmology: a review. *Clinical & experimental ophthalmology*, 47(1):128–139, 2019.

[51] The Royal College of Ophthalmologists. Workforce census 2018. https://www.rcophth.ac.uk/wp-content/uploads/2019/02/RCOphth-Workforce-Census-2018.pdf. Accessed: 11-08-2021.

[52] Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.

[53] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[54] Massimo Porta and F Bandello. Diabetic retinopathy. *Diabetologia*, 45(12):1617–1634, 2002.

[55] Joanne WY Yau, Sophie L Rogers, Ryo Kawasaki, Ecosse L Lamoureux, Jonathan W Kowalski, Toke Bek, Shih-Jen Chen, Jacqueline M Dekker, Astrid Fletcher, Jakob Grauslund, et al. Global prevalence and major risk factors of diabetic retinopathy. *Diabetes care*, 35(3):556–564, 2012.

[56] Rajendra Acharya, Chua Kuang Chua, EYK Ng, Wenwei Yu, and Caroline Chee. Application of higher order spectra for the identification of diabetes retinopathy stages. *Journal of medical systems*, 32(6):481–488, 2008.

[57] Udyavara R Acharya, Choo M Lim, E Yin Kwee Ng, Caroline Chee, and Toshiyo Tamura. Computer-based detection of diabetes retinopathy stages using digital fundus images. *Proceedings of the institution of mechanical engineers, part H: journal of engineering in medicine*, 223(5):545–553, 2009.

[58] Jagadish Nayak, P Subbanna Bhat, Rajendra Acharya, Choo M Lim, and Manjunath Kagathi. Automated identification of diabetic retinopathy stages using digital fundus images. *Journal of medical systems*, 32(2):107–115, 2008.

[59] Harry Pratt, Frans Coenen, Deborah M Broadbent, Simon P Harding, and Yalin Zheng. Convolutional neural networks for diabetic retinopathy. *Procedia computer science*, 90:200–205, 2016.

[60] Omar Dekhil, Ahmed Naglah, Mohamed Shaban, Mohammed Ghazal, Fatma Taher, and Ayman Elbaz. Deep learning based method for computer aided diagnosis of diabetic retinopathy. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–4. IEEE, 2019.

[61] Zhentao Gao, Jie Li, Jixiang Guo, Yuanyuan Chen, Zhang Yi, and Jie Zhong. Diagnosis of diabetic retinopathy using deep neural networks. *IEEE Access*, 7:3360–3370, 2018.

[62] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017.

[63] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.

[64] LP. Cen, J. Ji, JW. Lin, and et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 12:4828, 2021.

[65] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[66] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.

[67] Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. A global review

of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 2020.

[68] CP Wilkinson, Frederick L Ferris III, Ronald E Klein, Paul P Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R Pararajasegaram, Juan T Verdaguer, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003.

[69] Alexander Rakhlin. Diabetic retinopathy detection through integration of deep learning classification framework. *bioRxiv*, page 225508, 2017.

[70] Benjamin Graham. Diabetic retinopathy detection competition report. `https://www.kaggle.com/c/diabetic-retinopathy-detection/discussion/15801`. Accessed: 03-08-2021.

[71] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[72] Ross Wightman. Timm pytorch image models. `https://github.com/rwightman/pytorch-image-models`. Accessed: 03-08-2021.

[73] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[74] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[75] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[76] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[77] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[78] Facebook Research. Self-supervised vision transformers with dino. `https://github.com/facebookresearch/dino`. Accessed: 03-08-2021.

[79] Gary King and Langche Zeng. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

[80] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

[81] David M W Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[82] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[83] Toon Van Craenendonck, Bart Elen, Nele Gerrits, and Patrick De Boever. Systematic comparison of heatmapping techniques in deep learning in the context of diabetic retinopathy lesion detection. *Translational vision science & technology*, 9(2):64–64, 2020.

[84] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[85] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang,

Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021.