Multi-resolution Fine-Tuning of Vision Transformers

Kerr Fitzgerald[1(✉)], Meng Law[2], Jarrel Seah[2], Jennifer Tang[3], and Bogdan Matuszewski[1]

[1] University of Central Lancashire, Preston, UK
kffitzgerald@uclan.ac.uk
[2] Department of Radiology, Alfred Health, Melbourne, Australia
[3] St. Vincent's Hospital, Melbourne, Australia

Abstract. For computer vision systems based on artificial neural networks, increasing the resolution of images typically improves the performance of the network. However, ImageNet pre-trained Vision Transformer (ViT) models are typically only openly available for 224^2 and 384^2 image resolutions. To determine the impact of using higher resolution images with ViT systems the performance differences between ViT-B/16 models (designed for 384^2 and 544^2 image resolutions) were evaluated. The multi-label classification RANZCR CLiP challenge dataset, which contains over 30,000 high resolution labelled chest X-ray images, was used throughout this investigation. The performance of the ViT 384^2 and ViT 544^2 models with no ImageNet pre-training (i.e. models were only trained using RANZCR data) was firstly compared to see if using higher resolution images increases performance. After this, a multi-resolution fine-tuning approach was investigated for transfer learning. This approach was achieved by transferring learned parameters from ImageNet pre-trained ViT 384^2 models, which had undergone further training on the 384^2 RANZCR data, to ViT 544^2 models which were then trained on the 544^2 RANZCR data. Learned parameters were transferred via a tensor slice copying technique. The results obtained provide evidence that using larger image resolutions positively impacts ViT network performance and that multi-resolution fine-tuning can lead to performance gains. The multi-resolution fine-tuning approach used in this investigation could potentially improve the performance of other computer vision systems which use ViT based networks. The results of this investigation may also warrant the development of new ViT variants optimized to work with high resolution image datasets.

Keywords: Computer vision · Vision transformer · ViT · Fine-tuning · Transfer learning · Medical data · RANZCR CLiP

1 Introduction

When developing artificial neural networks for computer vision tasks, increasing the resolution of images used for training and inference often improves the performance of the network. Intuitively, this is because higher resolution images contain more information that can be used by the network. However, once a certain image size is reached the

G. Yang et al. (Eds.): MIUA 2022, LNCS 13413, pp. 535–546, 2022.
https://doi.org/10.1007/978-3-031-12053-4_40

performance gained from increasing image resolution will plateau. For EfficientNet [1] and EfficientNetV2 [2], models pre-trained on ImageNet are available that can use 224^2 image resolutions (B0 model variants) to 600^2 image resolutions (B7 model variants). These Convolutional Neural Networks (CNNs) provide good examples of increased classification accuracy on the ImageNet benchmark [3] when using higher image resolutions and also how accuracy begins to plateau once a given image resolution (600^2) is reached. It should be noted that the image resolution at which performance begins to plateaus will likely be different depending on the dataset.

Image resolution has also been shown to have an important effect on CNNs when evaluating their performance on test datasets and for transfer learning applications. Touvron et al. [4] used a light-weight parameter adaptation of a CNN to allow larger image resolutions to be used while testing the network (the main aim was to fix the resolution discrepancy seen by CNNs between training and testing). Touvron et al. showed that test performance increased when using higher resolution images (up to a plateau value) than those the CNN was trained on. Kolesnikov et al. [5] investigated methods to improve the generalization of CNNs for transfer learning tasks by altering network architecture (e.g. replacing Batch Normalization with Group Normalization). They also showed that fine-tuning CNNs to the test dataset resolution can improve transfer learning performance.

In the original Vision Transformer (ViT) paper [6] the authors fine-tuned the ViT network at higher resolution (384^2) than that used in pre-training (224^2) and attained higher accuracies on popular image classification benchmarks (including ImageNet) when using 384^2 ViT models when compared to 224^2 ViT models. This was achieved by keeping the image patch size the same, which results in the ViT network having a larger sequence of patches. However, ViT networks for image resolutions larger than 384^2 were not created and trained/fine-tuned in the original paper.

More recently, the rules determining how ViT models scale have been investigated by scaling ViT models and characterizing the relationships between error rate, data requirements and computing power [7]. This resulted in the creation of the ViT-G/14 model variant [7] which was trained on extremely large proprietary datasets (e.g. JFT-300M) using 224^2 image resolutions before being fine-tuned using the same extremely large proprietary datasets with 518^2 image resolutions. The ViT-G/14 model, which contains approximately two billion parameters, attained previous state-of-the-art on ImageNet with 90.45% top-1 accuracy (top-1 accuracy relates to where the highest class prediction probability is the same as the target label). However, for many users the current hardware requirements needed to train the ViT-G/14 network or use it for transfer learning tasks would be prohibitively expensive. It would be even more challenging to use ViT-G/14 as an encoder for dense prediction (e.g. segmentation or monocular depth estimation) tasks due to even more parameters being needed within the models.

The work presented in this paper details the results of an investigation to take a multi-resolution fine-tuning approach (whereby networks are trained through transfer learning, initially on low resolution images before being fine-tuned on higher resolutions versions of these images) and apply this directly to transfer learning applications relating to medical image analysis using ViT systems. To the best of the authors knowledge, this is the first time that multi-resolution fine-tuning has been applied directly to medical imaging for ViT systems. The medical image dataset used in this investigation consists of over 30,000 chest X-rays (taken to evaluate the positioning of multiple catheters) and allows the multi-label classification performance of ViT models to be evaluated. Firstly, a performance comparison between ViT 384^2 and ViT 544^2 networks with no prior training (i.e. models were only trained using RANZCR CLiP data) was conducted. After the initial performance comparison showed that using larger image sizes is beneficial, a multi-resolution fine-tuning approach was applied directly to the RANZCR CLiP transfer learning task. This was achieved by transferring learned network parameters (via a tensor slice copying technique) from ImageNet pre-trained ViT 384^2 models, which had undergone further training on the 384^2 RANZCR CLiP data, to newly initialized ViT 544^2 models which were further trained on the 544^2 RANZCR CLiP data. Results provide strong evidence that this approach increases multi-label classification accuracy and that using higher image resolution can improve network performance.

2 Method

2.1 Image Dataset Selection

The Royal Australian and New Zealand College of Radiologists (RANZCR) Catheter and Line Position (CLiP) challenge dataset consists of over 30,000 high resolution (typically greater than 2000^2) labelled chest X-ray images [8]. The aim of the original dataset challenge [9] was to detect the presence and position of different catheters and lines within chest X-ray images. The positions of the inserted catheters/lines are important since if they are poorly placed, they can worsen the patient's condition. There are four types of catheters/lines: Endotracheal Tube (ETT), NasoGastric Tube (NGT), Central Venous Catheter (CVC) and Swan-Ganz Catheter (SGC). The ETT, NGT and CVC can be categorized as 'Normal', 'Borderline' or 'Abnormal' and the SGC is either 'Present' or 'Not Present' hence making this a multi-label classification problem with 11 classes. The metric used to evaluate the multi-label classification performance in the original challenge was the 'One vs Rest Area Under Curve Receiver Operator Characteristic' (AUC-ROC) and this metric is used to evaluate performance of models within this investigation. The RANZCR CLiP dataset was selected for this transfer learning investigation due to the high resolution of the images and because classifying the placement of catheters/lines likely requires analysis of fine detail within the images (Fig. 1).

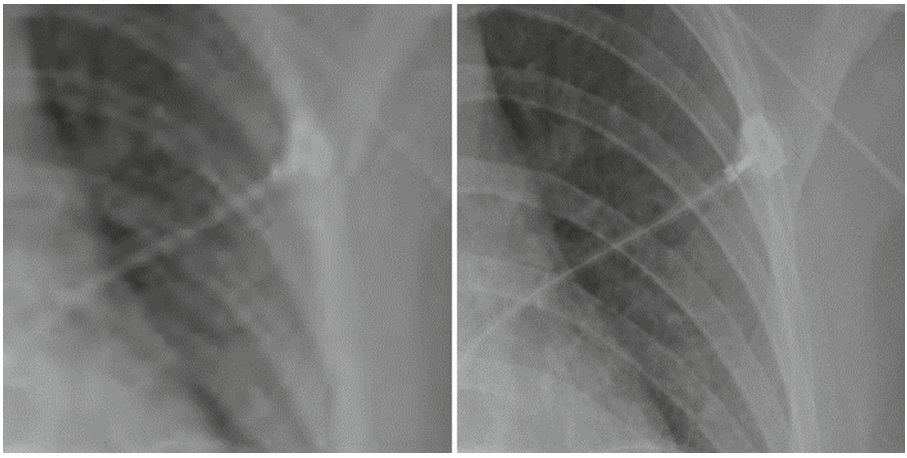

Fig. 1. Example of a cropped X-ray image region (from the RANZCR CLiP database) [8] generated using two different original image resolutions. This demonstrates potential information loss as the image size decreases.

2.2 Multi-resolution Fine-Tuning for Transfer Learning

When using the PyTorch deep learning framework for transfer learning, it is necessary to load weights from pre-trained models into your current model. This commonly requires that the tensors containing the parameters of the models match in name, shape, and size. Therefore, using larger image sizes as input into a ViT model which has been trained on smaller sized images would not be immediately possible. To overcome this limitation, it is possible to copy parameters (in the form of tensor slices) from pre-trained ViT models and insert these into the tensors (which are either the same size or are larger) of a new ViT model capable of processing higher resolution images. This tensor slice copying technique also allows other network design features of the new ViT model to be changed whilst still making use of the original pre-trained ViT model parameters. Examples of such network design features include: fully connected layer ratio, image embedding size, network depth and number of attention heads.

As an example, comparing a standard (i.e. ViT-B/16) ViT 544^2 model to a standard pre-trained ViT 384^2 model shows that only the size and shape of the Layer 2 tensor changes, while for all other layers the size and shape of tensors is identical. Therefore, the learned ViT 384^2 model parameters can be transferred via tensor slice copying to every layer of the 544^2 ViT model. Specifically, for Layer 2 it is possible to either: (1) ignore the ViT 384^2 Layer 2 tensor learned parameters and leave the Layer 2 tensor of the ViT 544^2 with its original initialization state; or (2) transfer the ViT 384^2 Layer 2 tensor learned parameters to the Layer 2 tensor of the ViT 544^2 which only partially fills the tensor.

In this investigation parameters from ImageNet pre-trained ViT-B/16 384^2 models [6, 10] were inserted via the tensor slice copying technique into a newly created ViT 544^2 ViT-B/16 models which had an increased fully connected layer ratio (4.25 compared to 4 in the original ViT 384^2 model). All possible parameters were transferred meaning

that some ViT 544^2 layer tensors would have been only partially filled (i.e. option (2) from the previous paragraph) (Table 1).

Table 1. Comparison of the first four layers of a ViT 544^2 network and ViT 384^2 ViT network. Only the size of Layer 2 changes between the models, all other layer sizes match.

Image input size	(384^2)	(544^2)
No. parameters	86,094,341	86,539,781
Layer 1 size	[1, 1, 768]	[1, 1, 768]
Layer 2 size	[1, 577, 768]	[1, 1157, 768]
Layer 3 size	[768, 3, 16, 16]	[768, 3, 16, 16]
Layer 4 size	[768]	[768]

2.3 Fold Selection and PyTorch Model Training

The RANZCR CLiP data was split into twenty folds (using a typical K-Fold random stratified sampling approach) with care also being taken to ensure that no data leakage occurred (e.g. data from a given patient was always contained in the same fold). Due to hardware limitations (all training and validation was run on a single Nvidia 3090 GPU) and the need for repeat runs using different random number seeds, it was not possible to use all twenty folds for cross validation in the transfer learning investigation. Instead six folds consisting of the three highest scoring and three lowest scoring AUC ROC validation scores were selected after the twenty-fold cross validation study was conducted using an ImageNet pre-trained ViT 384^2 network [6, 10] which underwent additional training on the RANZCR CLiP data. This found that the highest scoring validation folds were 14, 8 and 10, with the lowest scoring validation folds being 2, 20 and 12. After six epochs of training overfitting began to occur. The results of the twenty-fold cross validation study are displayed in Fig. 2. No data augmentation or image pre-processing was conducted (Fig. 2).

Before conducting the transfer learning investigation, a performance comparison between ViT 384^2 and ViT 544^2 models with no prior ImageNet training was conducted (i.e. only RANZCR CLiP data was used for training and validation) for each of the six folds selected.

For the investigation into the multi-resolution fine-tuning approach which can be directly applied to transfer learning, the model states of ImageNet pre-trained ViT 384^2 models which underwent additional training on the RANZCR CLiP data were saved for each of the six folds investigated. The saved ViT 384^2 model states after epoch three of training were then transferred to the corresponding ViT 544^2 networks using the tensor slice copying technique. ViT 544^2 networks were then trained on the RANZCR CLiP data, hence allowing for a multi-resolution fine-tuning approach. For each fold, six ViT 544^2 model runs were then conducted using different random number generation seeds. An additional six ViT 384^2 model runs were conducted using different random

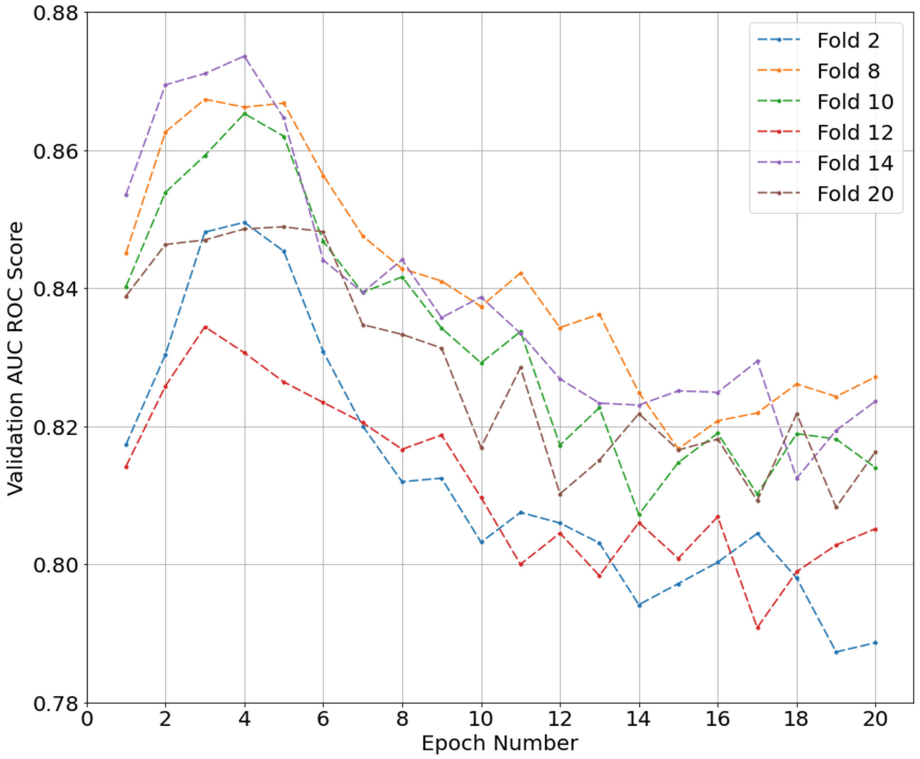

Fig. 2. Comparison of the three highest scoring and three lowest scoring AUC ROC validation scores from the twenty-fold cross validation scoping study conducted using a pre-trained ViT 384^2 network which underwent additional training on the RANZCR CLiP data. Overfitting begins to occur after approximately six epochs.

number generation seeds. These repeat runs were conducted to give confidence that improvements in performance are not down to the small random variability of network predictions. The different random number generation seeds impact the order of how image batches are loaded. In order to focus on the effects of image resolution, the PyTorch training settings and hyperparameters were kept the same between runs. However, it is likely that the training process followed in the original ViT paper [6] is heavily optimized compared to that used this investigation (Table 2).

Table 2. PyTorch training settings and parameters used in this investigation.

PyTorch training option	Selected setting
Optimizer	Adam
Loss function	BCEWithLogitsLoss
Learning rate	1.00e-05
Learning rate decay factor	0.95
Image batch size	4
MLP dropout rate	0.1
Residual path dropout rate	0.1
Attention dropout rate	0.1

3 Results

3.1 ViT 384^2 vs ViT 544^2 Model Comparison with no Prior Training

The results of the performance comparison between the ViT 384^2 and ViT 544^2 networks which had no pre-training for the six folds investigated (i.e. only trained using RANZCR CLiP data) are presented in Fig. 3. It can be seen that the average, maximum and minimum (shown with error bar range) AUC ROC validation scores of the six folds investigated are higher for the ViT 544^2 network (after eight training epochs) when compared to the 384^2 ViT network. Numerical values of the maximum achieved AUC ROC validation scores for each fold investigated are presented in Table 3.

These results provide further evidence of how increasing image resolution can increase the performance of deep learning image classification systems and that this relationship is valid for ViT systems. However, the maximum achieved AUC ROC validation scores for each fold are significantly lower for the ViT 544^2 network with no pre-training compared to those of the ImageNet pre-trained ViT 384^2 network shown in Fig. 2. This necessitates the need for multi-resolution fine-tuning which can be directly applied to transfer learning tasks (Fig. 3 and Table 3).

3.2 Multi-resolution Fine Tuning

The results of the multi-resolution fine-tuning approach directly applied to the transfer learning task of medical image multi-label classification are visualized for each fold using box plots (showing the minimum, maximum, quartiles and median validation AUC ROC scores) in Fig. 4. Apart from for fold 14, the maximum and median AUC ROC scores achieved using the ViT 544^2 network are higher than those obtained using the ViT 384^2 network. However, even though the maximum achieved accuracies are significantly higher when pre-training is used, the magnitude of the performance increase between ViT 544^2 networks and ViT 384^2 networks is smaller compared to when no pre-training was used.

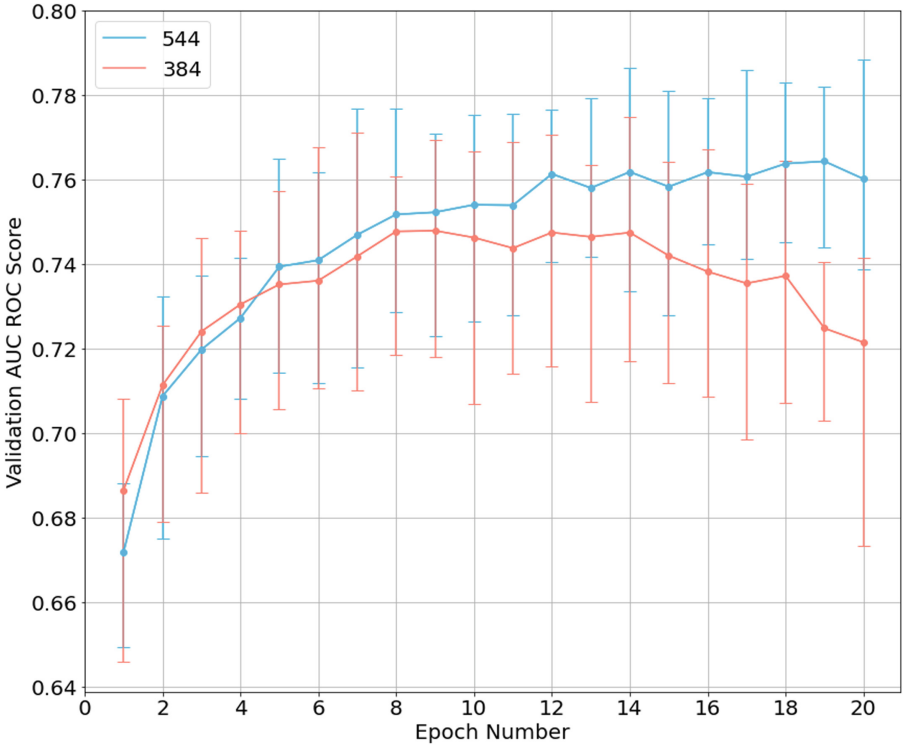

Fig. 3. Comparison of the average, maximum and minimum AUC ROC validation scores of the six folds investigated for the ViT 384^2 and ViT 544^2 networks.

Table 3. ViT 384^2 and ViT 544^2 maximum achieved AUC ROC validation scores for each fold when no pre-training is used.

Fold	384^2	544^2	Difference
14	0.7666	0.7812	0.0146
8	0.7749	0.7885	0.0136
10	0.7571	0.7651	0.0080
2	0.7506	0.7747	0.0241
20	0.7561	0.7602	0.0042
12	0.7185	0.7488	0.0303

These results provide further evidence that multi-resolution fine-tuning can improve network performance and that, importantly, this approach can be directly applied to transfer learning tasks using ViT systems (Fig. 4 and Table 4).

Table 4. Maximum achieved AUC ROC validation scores for each fold for ViT 384^2 pretrained networks and ViT 544^2 networks using the multi-resolution fine-tuning approach.

Fold	384^2	544^2	Difference
14	0.8756	0.8725	−0.0031
8	0.8763	0.8830	0.0067
10	0.8653	0.8689	0.0036
2	0.8496	0.8559	0.0063
20	0.8496	0.8615	0.0119
12	0.8439	0.8519	0.0080

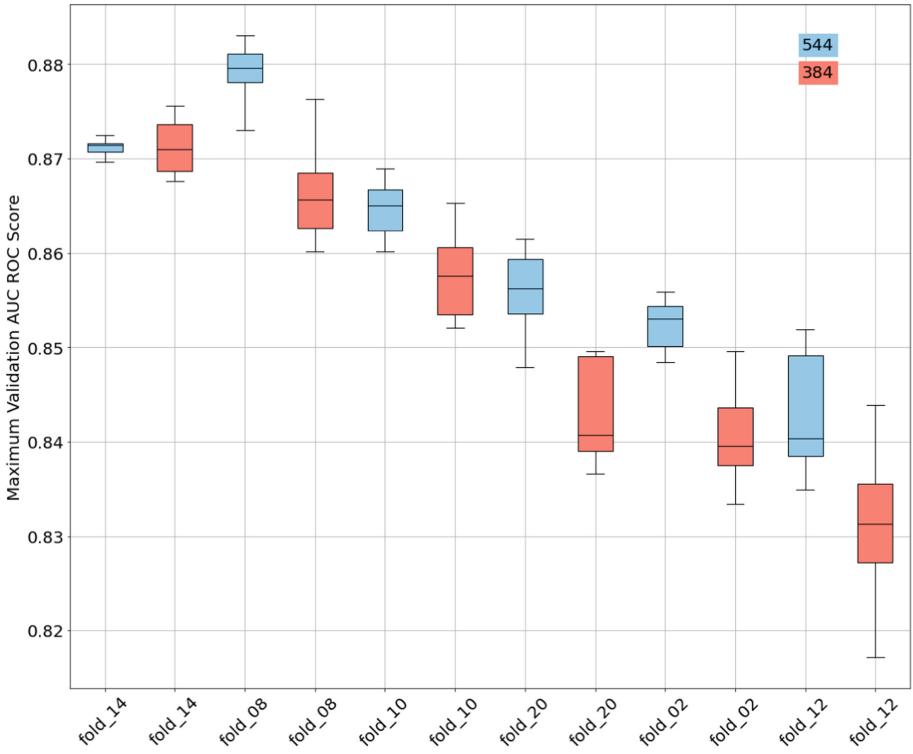

Fig. 4. Box plots showing the minimum, maximum, quartiles and median validation AUC ROC scores of the repeat runs of the six folds investigated for the ViT 384^2 (red) and ViT 544^2 (blue) networks. (Color figure online)

4 Discussion

The performance comparison between the ViT 384^2 and ViT 544^2 networks which had no pre-training strongly demonstrate how using larger image resolutions positively impacts

ViT network performance. This is further supported by the results of the multi-resolution fine-tuning approach which found that the ViT 544^2 network slightly outperformed the ViT 384^2 network for five out of the six folds tested.

The multi-resolution fine-tuning approach could potentially impact the performance of other computer vision systems designed for dense prediction tasks (e.g. monocular depth estimation) which use pre-trained ViT models as encoders. An example of such a system would be the Dense Prediction Transformer (DPT) [11] which previously held state-of-art performance on certain monocular depth challenges (such as NYU-Depth V2 [12]). The DPT used the original ViT 384^2 ImageNet pretrained network as the encoder starting point. In addition, the multi-resolution fine-tuning approach for direct transfer learning may also be applicable to new ViT systems (such as the Vision Longformer [13]) being developed.

It is likely that further improvements could be made to the multi-resolution fine-tuning approach used in this study. The training method used is likely to not be as optimized as that used in the original ViT paper [6] and any improvements made to the training process could further increase the performance of the ViT 544^2 networks. The tensor slice copying technique could also be improved as the approach used in this study directly copied Layer 2 learned parameters from the ViT 384^2 network to Layer 2 of the ViT 544^2 network. Layer 2 represents learned image embeddings with positional encodings and using a more complex approach to transfer these particular learned parameters to the ViT 544^2 network could help during fine-tuning. For example, in the original ViT paper [6] the authors performed 2D interpolation of the pre-trained position embeddings, according to their location in the original image, for resolution adjustment. This would also ensure that all parameters in Layer 2 are updated rather than some parameters keeping their randomly initialized value which could potentially be adversely impacting gradient calculations. However, preliminary investigations which left all Layer 2 parameters in their randomly initialized state had only marginally worse performance compared to partially filling the Layer 2 tensor, suggesting that adverse effects on the gradient calculations are minimal. Examining all twenty folds rather than the six selected folds would also reduce possible bias and conducting more runs for each fold would give even higher confidence in the results obtained. Applying the developed transfer learning approach to other imaging problems and investigating performance would allow external validation of the methods used.

Since the performance gains of the ViT 544^2 networks were essentially attained by changing the number of patches used, this might also justify the development of new ViT network variants designed specifically to work with larger image sizes but with network design parameters which are not as extreme as the ViT-G/14 variant (i.e. significantly reduced network depth and total parameter number). Pre-training and fine-tuning of these ViT networks with higher image resolutions (e.g. $>500^2$) than those used in the original ViT paper (224^2 and 384^2) using large image databases (e.g. ImageNet dataset variants) could lead to significant performance increases. Such models would likely have hardware requirements that would make them accessible to a large number of users/developers and also make them suitable for use in encoder-decoder style systems for dense prediction tasks.

5 Conclusion

The impact of using ViT networks with higher resolution images (compared to those typically used for training on ImageNet dataset variants) on a multi-label classification problem has been evaluated. The dataset used in this investigation was the RANZCR CLiP challenge dataset which consists of over 30,000 high resolution labelled chest X-ray images [8].

A performance comparison between two ViT-B/16 networks [6, 10], which had no pre-training, designed to work with 384^2 and 544^2 image resolutions has been conducted on the RANZCR CLiP medical image multi-label classification task. The ViT 544^2 network outperforms the ViT 384^2 network for all six of the data folds that were tested.

A multi-resolution fine-tuning approach was applied to ViT 544^2 networks directly for the RANZCR CLiP medical image multi-label classification task. To achieve this, ImageNet pre-trained ViT 384^2 model states, after three epochs of additional training on the RANZCR CLiP dataset, were saved. The ViT 384^2 model states were then transferred to ViT 544^2 models using a tensor slice copying technique and the 544^2 models were then trained on the RANZCR CLiP dataset. The results of this approach show that the ViT 544^2 network outperformed the ViT 384^2 network for five out of the six data folds that were tested.

The results obtained provide strong evidence that using larger image resolutions positively impacts ViT network performance. This may justify the development of new ViT network variants (with significantly less computational requirements than the current state-of-the-art ViT-G/14 variant) designed to work with higher image resolutions (i.e. greater than 384^2). Such variants would likely be more accessible to a larger number of users and may also be suitable for use in encoder-decoder systems for dense prediction tasks. The performance of existing encoder-decoder systems for dense prediction tasks, which use ViT based systems as encoders, may also receive transfer learning performance increases by using multi-resolution fine-tuning approaches similar to that used in this investigation.

References

1. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), Long Beach (2019)
2. Tan, M., Le, Q.: EfficientNetV2: smaller models and faster training. In: International Conference on Machine Learning (ICML), Virtual (2021)
3. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami (2009)
4. Touvron, H., Vedaldi, A., Douze, M., Jegou, H.: Fixing the train-test resolution discrepancy. In: Advances in Neural Information Processing Systems (2019)
5. Kolesnikov, A., et al.: Big transfer (BiT): general visual representation learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 491–507. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_29
6. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), Virtual (2021)

7. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. arXiv (2021)
8. Tang, J., et al.: CLiP, catheter and line position dataset. Sci. Data **8**, 1–7 (2021)
9. Law, M., et al.: RANZCR CLiP - catheter and line position challenge. Kaggle, 8 March 2021. https://www.kaggle.com/competitions/ranzcr-clip-catheter-line-classification/overview. Accessed June 2021
10. Wightman, R.: PyTorch image models. GitHub Repository (2019)
11. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: International Conference on Computer Vision (ICCV), Montreal (2021)
12. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
13. Zhang, P., et al.: Multi-scale vision longformer: a new vision transformer for high-resolution image encoding. In: International Conference on Computer Vision (ICCV), Virtual (2021)