HYBRID CNN-VIT MODELS FOR MEDICAL IMAGE CLASSIFICATION

Dimitrios Pantelaios, Paraskevi-Antonia Theofilou, Paraskevi Tzouveli and Stefanos Kollias

National Technical University, Athens, Greece

ABSTRACT

The success of Transformers in the field of Natural Language Processing and Computer Vision inspired their use in various image analysis tasks. Vision Transformers capture long-range global dependencies through attention layers, but lack inductive biases, making it challenging for them to generalize when trained on small datasets. As a result, larger datasets are required for their training, which poses a significant obstacle specifically in the medical image classification task. This study focuses on the classification of chest X-ray images corresponding to different diseases affecting the lungs, such as COVID-19. To address the aforementioned challenges, hybrid models were explored, aiming to incorporate some advantages of CNNs into Vision Transformers, enabling the training of models on smaller datasets. So, in this work, we compare the hybrid models pre-trained on ImageNet-1k with the traditional Vision Transformer pre-trained on ImageNet-21k using both a subset and the entire available COVID-QU-Ex dataset, while we also explore training the models from scratch. The results obtained demonstrate the superiority of the hybrid models in terms of accuracy, training time and the number of data required for training.

***Index Terms*—** ViT, CNN, Hybrid ViT-CNN models, Medical Image classification, COVID-19

1. INTRODUCTION

Vision Transformers (ViTs) [1] have garnered considerable interest in tasks related to the analysis of medical images [2]. Traditional approaches have relied on Convolutional Neural Networks (CNNs). However, ViTs offer an alternative approach by utilizing transformer-based architectures [3][4]. ViTs' self-attention mechanisms help them to capture relationships between different elements in an image, allowing them to understand the global context and long-range dependencies. This is in contrast to CNNs, which typically rely on local receptive fields to extract features. Also, they can handle variable-sized inputs processing images of different resolutions and aspect ratios without the need for resizing or cropping. Another crucial aspect of ViTs is their interpretability which is highly valued in the medical field where understanding the reasoning behind model predictions is very important. Despite that, ViTs need larger datasets than CNNs to be trained on. For this reason, the combination of the advantages of CNNs and ViTs is necessary in order to introduce more promising architectures, efficient and reliable systems.

Therefore, the aim of this work is to study different hybrid deep learning systems, CNN-ViT models, and apply them to a crucial problem of our days, COVID-19 detection via chest X-ray images. Until now, there are not specific studies related to this problem.

The main contributions of this paper are:

- We investigate a diverse range of hybrid CNN-ViT models in order to enhance the performance and capabilities of the image classification task.
- We apply these carefully selected models on the COVID-QU-Ex dataset, a valuable benchmark for evaluating image classification performance in medical imaging and COVID-19 detection.
- We compare CNN-ViTs' and simple ViTs' performances when fine-tuned and when trained from scratch. In both cases hybrid models achieve better results in terms of accuracy, training time and computational costs.
- Our study demonstrates enhanced COVID-19 detection capabilities, yielding robust and reliable results.

2. BACKGROUND

In this study, we aim to examine the advantages of ViTs and hybrid CNN-ViT models, compare them with each other and evaluate their impact on the task of medical image classification. Specifically, the models of interest are ViT and six hybrid models named CCT, DeiT, CeiT, LocalViT, Conformer and CvT respectively. A small description for each one of the aforementioned models is presented below.

ViTs, introduced by Dosovitskiy et al. [1], divide images into fixed-size patches, transforming them into one-dimensional embeddings processed by a transformer encoder. A class token is added, serving as the image's representation and it is passed through the final MLP for prediction. Encoder blocks include normalization, a Multi-Head Attention Layer (MHA) for global dependencies, another normalization layer, a Feed-Forward network (MLP), and residual connections for enhanced performance. Positional information is conveyed via one-dimensional positional embeddings based on the raster order of image patches.

To tackle data limitations in ViT training, Compact Trans-

formers, and specifically the Compact Convolutional Transformer (CCT), devised by Hassani et al.[5], use convolutional layers instead of patches to add inductive bias. This model diverges from ViT Lite, which constitutes a smaller version of ViT and Compact Vision Transformer (CVT), which replaces the final MLP with a Sequence Pooling structure.

The Data-efficient image Transformers (DeiT) model by Touvron et al. [6], employs a teacher-student paradigm with a CNN (usually RegNetY-16GF) as the teacher and a ViT as the student. The ViT model includes a distillation token, akin to the class token but without considering actual labels during training; instead, it leverages the teacher's predictions. The use of a CNN as the teacher introduces inductive bias through the distillation token, leading to improved performance on smaller datasets.

The Convolution-Enhanced Image Transformer (CeiT) by Yuan et al. [7] differs from ViT in three key aspects. It uses an Image-to-Tokens (I2T) unit for patch extraction, incorporating CNN components like convolutional layers. Locally-enhanced Feed-Forward (LeFF) layers replace MLPs in the ViT encoder, employing depth-wise convolutions. The model includes a last class token attention (LCA) layer at the encoder's end, combining outputs from previous layers.

LocalViT, by Yawei Li et al. [8], retains ViT's structure but introduces depth-wise convolutions in the Feed-Forward networks to integrate local image information, enhancing computational efficiency and reducing parameters. The modified Feed-Forward Network processes the two-dimensional image, which is later flattened and tokenized. The class token is treated separately from the FFN and doesn't contribute to the two-dimensional image. Some experiments explore the impact of non-linear activation functions after the depth-wise convolution.

Conformer, by Peng et al. [9], features two branches: a ResNet-style CNN branch and a ViT-style Transformer branch. Data undergoes initial processing in a stem, extracting local features for both branches. The Feature Coupling Unit (FCU) in each layer facilitates information exchange, enabling down-sampling for CNN blocks and up-sampling for ViT blocks. Outputs from each ViT block influence subsequent blocks in both branches, and the same applies to CNN blocks. Each branch ends with a classifier, and the final prediction results from combining both branches' outputs.

CvT, introduced by Wu et al. [10], enhances ViTs' performance on smaller datasets by incorporating the Convolutional Token and the Convolutional Projection. The Convolutional Token involves a two-dimensional convolutional process before inputting patches into ViT blocks, enabling representation of more complex patterns. The Convolutional Projection, applied at each block's start, models local spatial contents, enhancing efficiency by undersampling K and V matrices used in the attention layer. Positional embeddings are omitted as they don't contribute significantly to model performance.

Integrating CNN features into ViTs combines their strengths, overcoming ViTs' challenges like extensive training data requirements.

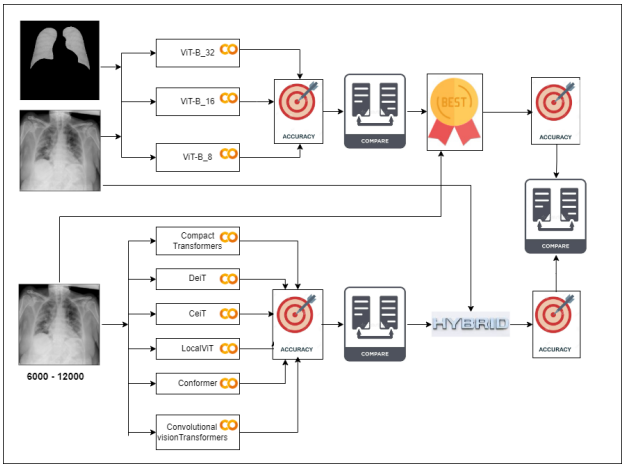

Fig. 1. Workflow

3. DATASET

We use the Dataset COVID-QU-Ex introduced by Tahir et al. in this study [11]. It consists of 33.920 chest X-ray images (CXR) that are divided into three categories. More specifically, it contains 11.956 COVID-19 images, 11.263 Non-COVID images that correspond to diseases such as Viral or Bacterial Pneumonia and finally 10.701 Normal images, i.e. images that belong to patients who have not been affected by any of the diseases mentioned above. For every image of this dataset a ground-truth lung segmentation mask is included, and a masked image is created by the application of the lung mask to the image, as shown in Figure 1 (top left corner).

4. METHODOLOGY

This section summarizes the experiment and model comparison process used, detailed in Figure 1. We start by testing traditional, pre-trained on ImageNet-21k, ViT variations on patch sizes, batch sizes, and fine-tuning steps. Some experiments use raw chest X-rays, while others use masked ones. To introduce inductive bias, we explore alternatives like ResNet models or added convolutional layers instead of patches.

Due to the scarcity of large medical datasets, we conduct experiments with new models designed for improved performance on smaller datasets, maintaining ViTs' advantages for larger datasets. Hybrid models, including Compact Transformers, DeiT, CeiT, LocalViT, Conformer, and CvT, are fine-tuned initially on 6.000 and 12.000 images, and later on the entire dataset. The best-performing ViT is compared with hybrid models on a small training set, evaluating accuracy for subsets and the entire dataset, with uniform image selection

from the three classes when fewer images are used. Finally, the models are trained from scratch on 6.000 images and their performances are compared.

5. EXPERIMENTS AND RESULTS

In this section, the results of the experiments conducted on the COVID-QU-Ex dataset, which was introduced in the Dataset section, are presented. It includes the process followed for selecting the models that would participate in the experiments, as well as the comparison of the models based on Top-1 accuracy (only metric used by all relative models). The experiments were performed in the Google Colab environment, utilizing the Tesla T4 GPU with 16GB VRAM.

The use of the Base version of ViT for all experiments (patch_size=8,16,32) and batch_size=32 for experiments with patch_size=8 was due to limited computational resources. From the experiments conducted on variations of traditional ViT on the whole dataset, we observe that as the batch size decreases and the number of steps used for fine-tuning increases, the model's performance improves. Additionally, we notice that as the patch size decreases, the accuracy and the training time increase. This is logical because smaller patch sizes result in more tokens being created, which increases the network's representation capacity but also increases the processing cost. Therefore, the model that exhibited the best behavior was ViT-B with a patch_size=8, finetuned on raw images and displaying accuracy **96.58%**. The same version displayed the best accuracy on the masked images, but the results were much worse, having an accuracy of **91.57%**. Increasing the number of steps significantly improved performance up to a saturation point at around 1.200 steps.

For most hybrid models, smaller versions were preferred due to computational resource limitations. Specifically, the models used for each type are as follows: cct_14_7x2_224, deit_tiny_patch16, ceit_tiny_patch16, cvt-13-224x224, Localvit_small_mlp4_act3_r384 and Conformer_small_patch16.

Some key hyperparameters that were used are the input size (224x224), epochs (30), optimizer (AdamW), initial learning rate ($55 \cdot 10^{-5}$ for cct, $25 \cdot 10^{-5}$ for cvt and $5 \cdot 10^{-4}$ for the rest), LR scheduler (Cosine), warmup epochs (5), weight decay ($6 \cdot 10^{-2}$ for cct, 0.1 for cvt and 0.05 for the rest), and the dataset used for pre-training (ImageNet-1k).

Next, in Table 1, the results of the experiments conducted on the Compact Transformers, DeiT, CeiT, LocalViT, Conformer and CvT models are presented using 6.000, 12.000 and 21.715 images respectively for their training. In the experiments using 6.000 and 12.000 images, the training set is chosen uniformly from the three classes, while in the experiments utilizing the whole dataset the model cct_14_7x2_224 was fine-tuned only on 15.000 train images due to computational restrictions.

Afterwards, we train the ViT that had shown the best results (patch_size=8, batch_size=32) in the initial experiments on smaller datasets in order to compare it with the hybrid models. The results are shown in Table 2.

From the results obtained, we observe that the hybrid models have fewer parameters, offering greater resistance to overfitting, and the majority of them require less training time compared to ViTs. Additionally, for 6.000 images, there are two hybrid models (CCT and CvT) that achieve higher accuracy than ViT-B_8. These specific models continue to outperform the Vision Transformer even when larger training datasets are used, including the entire dataset. The only exception is cct_14_7x2_224 for 12.000 images, which shows slightly lower accuracy compared to ViT. However, the other models exhibit similar results to ViT-B_8 and could be preferred due to their parameter advantages and shorter training time.

At the end, we attempt to train from scratch all the aforementioned models on a subset of the train set (6.000 images) for a few epochs/steps (30/200), using the same batch_size as in the previous experiments, to observe their behavior. The results obtained are shown in Table 3. In the training of the models from scratch, even though it is for a few epochs, we observe that the results of five out of the six models for the 6,000 images are improved compared to ViT, showing better accuracy. Additionally, three out of six models show shorter training times than ViT-B_8. This may mean either that after several epochs they will continue to show better results or that they converge faster.

6. CONCLUSIONS AND FUTURE WORK

This paper compares ViTs to hybrid CNN-ViT models, trained on subsets and the entire COVID-QU-Ex dataset, both pre-trained and from scratch. Hybrid models, despite being pre-trained on smaller datasets like ImageNet-1k, outperform ViTs, highlighting their efficiency in addressing ViTs' need for large datasets. Specifically, for 6.000 and 12.000 training images, several hybrid models achieve better accuracy and reduced training time, and even in the entire dataset scenario, the top-performing models from the 6.000-image case (cct_14_7x2_224 and cvt-13-224x224) maintain their superiority. Notably, in the training-from-scratch scenario, five out of six hybrid models outperform ViTs in accuracy.

Future work includes testing models on diverse medical datasets, and experimenting with various hybrid model combinations for enhanced network metrics. Optimization could involve combining optimal characteristics from different models to create more effective networks. These hybrid models may also be integrated as backbones into other networks, emphasizing accuracy and interpretability.

7. COMPLIANCE WITH ETHICAL STANDARDS

Ethical approval was not required as confirmed by the license attached with the open access data.

Model	Batch size	Number of parameters	6.000 images		12.000 images		all images	
			Training time	**Accuracy**	**Training time**	**Accuracy**	**Training time**	**Accuracy**
cct_14_7x2_224	128	21.983.428	**22:05**	**0.948**	**41:10**	0.95	**46:30**	0.9663
deit_tiny_patch16_224	32	**5.524.995**	22:48	0.90	47:24	0.9342	1:18:30	0.9492
ceit_tiny_patch16_224	32	6.164.483	26:33	0.893	53:16	0.894	1:33:57	0.94
Localvit_small_mlp4_act3_r384	32	22.049.331	36:31	0.89	1:12:59	0.9356	2:02:03	0.9417
Conformer_small_patch16	32	36.267.654	58:54	0.9177	1:43:43	0.9434	2:59:07	0.9487
cvt-13-224x224	32	20.000.000	29:32	0.943	1:04:28	**0.9605**	1:36:57	**0.9694**

Table 1. Results of Hybrid CNN - ViT models fine-tuned on 6.000, 12.000 and 21.715 train images

Number of training data	Steps	Number of parameters	Training time	Accuracy
6.000	200	86.000.000	**29:33**	0.931
12.000	400	86.000.000	50:47	**0.9536**

Table 2. Results of ViT-B_8 using a subset of the train set

Model	Number of parameters	Training time	Accuracy
ViT-B_8	86.000.000	26:44	0.6067
cct_14_7x2_224	21.983.428	**22:02**	**0.821**
deit_tiny_patch16_224	5.524.995	22:28	0.53
ceit_tiny_patch16_224	6.164.483	27:23	0.69
Localvit_small_mlp4_act3_r384	22.049.331	34:11	0.66
Conformer_small_patch16	36.267.654	56:11	0.7436
cvt-13-224x224	20.000.000	26:16	0.6937

Table 3. Results of all models trained from scratch on 6.000 images

8. ACKNOWLEDGMENTS

No funding was received for conducting this study.

9. REFERENCES

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[2] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu, "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, pp. 102802, 2023.

[3] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, "Intriguing properties of vision transformers," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 23296–23308, Curran Associates, Inc.

[4] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 12116–12128, Curran Associates, Inc.

[5] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi, "Escaping the big data paradigm with compact transformers," 2022.

[6] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers distillation through attention," 2021.

[7] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu, "Incorporating convolution designs into visual transformers," 2021.

[8] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool, "Localvit: Bringing locality to vision transformers," 2021.

[9] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye, "Conformer: Local features coupling global representations for visual recognition," 2021.

[10] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang, "Cvt: Introducing convolutions to vision transformers," 2021.

[11] Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid, "Covid-19 infection localization and severity grading from chest x-ray images," *Computers in Biology and Medicine*, vol. 139, pp. 105002, 2021.