‘No turning back’ : Determining the Best Backbone architecture for Contrastive Learning

Rishabh Sheoran rishabh.sheoran@u.nus.edu
Saisha Ajay Chhabria saisha.chhabria@u.nus.edu
Sourav Dey sourav.dey@u.nus.edu
Umaiyal Ramanathan e0838374@u.nus.edu
Qin Ziheng e0823059@u.nus.edu

Abstract

Of late, transformers have been portrayed to outperform CNN in the area of self-supervised clustering. We find this recommendation on the winning network architecture was drawn from unfair experimental settings. In this paper, we redo the test in fairer conditions: The objective is to compare the performance on a common unsupervised downstream classification task. For this, we establish the preferred backbone architecture by training all backbones under test with an identical pretext task. With (a) the appropriate choice of optimization techniques, (b) learning rates determined automatically from Superconvergence range tests and (c) coupling with aggressive scheduling techniques, an improvement of atleast 6% in CIFAR test accuracy was obtained with the CNN backbone(Resnet) vs that reported in literature. Moreover, the transformer architecture (ViT) suffered more than 4% loss in accuracy in this new test. ‘ConvNext’ is a recent ‘transformer’ inspired CNN architecture that is quickly gaining traction and was the third backbone put through this test. On CIFAR10 , Resnet50 now emerged the clear winner @ 85.1X% accuracy in this revised self- supervised contrastive learning contest; followed by ConvNext @81.7% and finally ViT.@ 79.3%. More than just a hypothetical experiment, a genuine requirement for such self- supervised classification based on minimizing contrastive loss alone is identified in the clustering of remote sensing images. Again, Resnet50 emerged to have the highest test accuracy when tested with a dataset from this domain (AID dataset). We hope this work can help the community better understand the extreme sensitivity of deep learning performance metrics to variations in optimization and scheduling. It is also reassuring to know that CNN will be here to stay, for longer than anticipated. Also, this work is a gentle reminder that domain expertise will need to go hand in and with deep learning knowledge so as to develop solutions tailored for each specific domain.

1 Introduction

Owing to the insufficient number of labelled imagery in many domains, a wide variety of real-world image classification tasks are still not automated yet, despite the availability of many deep learning network architectures. To address this shortcoming, in this paper, we have chosen to focus on the task of self-supervised learning on unlabelled images.

Recent results in [1] suggest that the Transformer architecture has superseded Convolutional Neural Network (CNN) in many areas of image recognition, including self-supervised learning(SSL) [2] and clustering. Meanwhile, to revive CNN, several new architectures [10], have been developed, including ConvNeXt[3]. ConvNeXt is a CNN architecture based on closely resembling the merits

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Table 1:

Method	Backbone	Unsupervised Classification on CIFAR10
SimCLR	ResNet56	78.75%±0.24
SiT	Vision Transformer(ViT)	83.50%±0.11

of the robust hierarchical Swin Transformer[12]. It retains the translation invariance property from CNN whilst incorporating the multi-headed self-attention mechanisms in Transformers. Though the ConvNeXt has been put to test against other backbone architectures for supervised classification[3], it has yet to be tested as a backbone in self-supervised classification.

In this paper, we do a comparison of the self-supervised classification performance after a systematic optimization procedure applied to the contrastive learning, across three different classes of backbone architectures : Resnet50, Vision Transformer and ConvNeXt and recommend the backbone of choice for contrastive learning on a specific image domain. We will then discuss and demonstrate self-supervised classification performances on a sample dataset from this domain.

The main contributions and findings of this study are summarised as follows:

- We perform a comparison across three backbone architecture performances for contrastive learning (for use in self-supervised classification). They are : Resnet50, Vision Transformer and ConvNeXt. These are chosen to represent one each of the broad classes of Neural networks: ie Full CNN, Full Transformer and CNN mock-up of Transformer. We are not aware of any other publications on backbone architecture comparison for contrastive learning yet.
- This paper is also the first demonstration of ConvNeXt for use in self-supervised classification.
- We propose a systematic optimization and scheduling procedure that attempts to reduce the sensitivity of Neural Network performance to the choice of arbitrary parameters applied. Thereafter, this optimization will allow for robust comparison of performances.
- We reason why some of the existing pretext tasks are not relevant and argue that contrastive learning alone as a pretext task should be adequate for application to self supervised classification of Remote Sensing images.
- From experiments conducted on two sets of data (CIFAR10 and AID), we conclude that CNN(Resnet50) over Transformer(ViT) is consistently the backbone of choice for contrastive learning. This is in contrast to the conclusions reached in [3] and [7]. Reasons for this discrepancy are postulated in this paper.
- The code and models are publicly available at https://github.com/rumaiyal/CS5260Project

2 Related Work and Observations

Challenging the relevance of CNNs after the onset of Transformers is not new and has been ongoing in the area of supervised imagery classification[8]. As for self-supervised classification, experiments in [3] again concluded on the superiority of the transformer backbone to CNN. Comparison was performed across the Self-supervised Vision Transformer (SiT) vs Resnet50 used in SimCLR. Both were trained on various camera collected imagery datasets, eg CIFAR10 and Imagenet, without using the ground truth label information. Result comparison presented in [3] for CIFAR10 are reproduced below in Table 1. In both cases, the downstream task was to perform unsupervised classification of CIFAR10 test images.

We find that such a conclusion has insufficient basis as (i) Training was not optimized in both cases (ii) The number of pretext tasks trained in SimCLR vs SiT are different. There was only one pretext task trained in SimCLR[11] which was to minimise contrastive loss whereas in SiT[3], three pretext tasks were trained simultaneously, first to perform image inpainting, second to predict image rotation and third to minimize contrastive loss. When trained with more pretext tasks, this will naturally translate to better robustness in the downstream task. Hence, this is not a fair test;

3 Methodology

We will be using the SimCLR framework for all backbone experiments in this project. SimCLR leverages on contrastive learning to achieve SOTA results in self-supervised and semi-supervised learning benchmarks by learning representations from unlabeled images based on heavy data augmentation by utilizing contrastive learning to maximize agreement between 2 augmented versions of the same image. For details, reader can refer to [11]

The SimCLR architecture block diagram is shown in Figure 1.

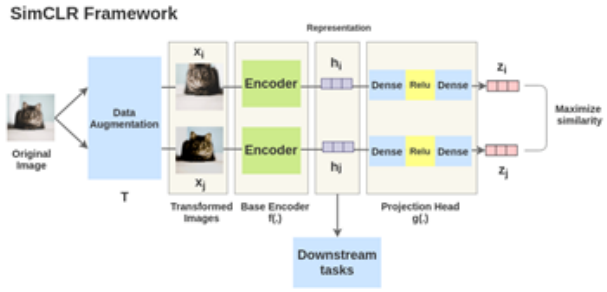


Figure 1: High level schematic of SimCLR[11]

The SimCLR framework allows for various choices of network architecture in the place of the base encoder, which is commonly referred to as the "backbone". In this paper, the experiments present models trained using ResNet, ConvNeXt-tiny and ViT-B as the backbone coupled with the projection head to maximize agreement using contrastive loss. The output of the base encoder is directly streamed to the downstream tasks.

Each backbone was further tested with different optimizer functions including RAdam, Adam and AdamW and LAMB. Methods for auto hyperparameter selection eg learning rate, were drafted for application on relevant optimizers to further improve the model performance.

3.1 Comparison of Backbones

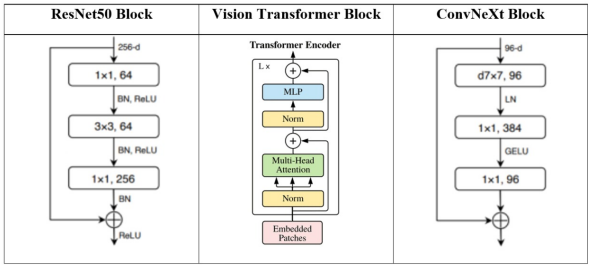


Figure 2: Network Architecture for the 3 Backbones under test with SimCLR

ResNet

Resnet50 is a variant of the ResNet model with 48 convolutional layers along with 1 MaxPool and 1 Average pool layer. The distinguishing feature of ResNet blocks is the residual learning framework with skip connections that perform simple identity mappings. This prevents deeper networks from producing greater error while not adding any additional parameters and keeping the training time in check. ResNet50 gives the advantage of depth in the network while still maintaining the computational expense.

Vision Transformers

ViT introduces a “patchify” layer, which splits images into a sequence of flattened patches. The ViT trains its parameters by measuring the relationship between input image patch pairs. Research has shown that with the help of larger models and dataset sizes, Transformers can outperform standard

ResNets by a significant margin. The ViT model used in this paper is based on the ViT-B/16 model first proposed by Alexey in [6]

ConvNeXt

ViTs were still observed to faces difficulties when applied to general computer vision tasks such as object detection and semantic segmentation, because of lack of the inductive bias of translation equivalence. The ConvNeXt model is fundamentally based on ResNet, but incorporates several key layer level design features from Vision Transformers. ConvNeXt competes favorably with Transformers in terms of accuracy and scalability, whilst preserving the simplicity and efficiency of standard ConvNets.

3.2 Systematic Optimisation per Backbone

A number of optimizers from the Adam 'family' were experimented with and we finally chose the RAdam optimizer as it demonstrates Super Convergence. Adam optimizers were chosen for their ability to perform non convex optimisation. One of the most important hyperparameters in training and tuning the models is the learning rate (lr). Having a poorly set learning rate can result in extremely slow training, poor convergence, and unstable networks. The super-convergence learning rate range test is adapted from [16]

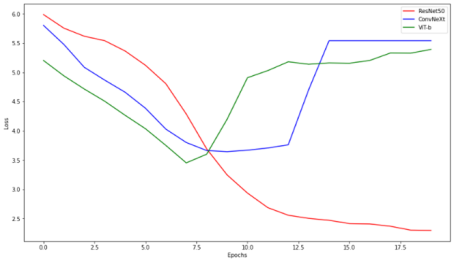
Learning rate range tests were first conducted to identify the optimal learning rate for training a particular backbone and optimizer combination. For each combination, the model was trained with an exponentially increasing lr, via the LambdaLR scheduler. The objective of this was to identify the optimal lr to maximize the training speed but also avoid the exploding gradients problem.

To obtain even more accurate learning rate(lr), the Super-Convergence algorithm was implemented twice – initially to obtain a broader lr range and subsequently to fine tune and obtain the lr with the lowest training loss. Upon obtaining the lr corresponding to the lowest loss from the second super-convergence stage, we multiplied the lr with an aggressive learning rate rule of thumb factor of 0.375[17] and set it as the maximum lr for input to the one cycle learning rate scheduler. A batch size of 128 was used.The test accuracy is averaged across 2 runs per backbone in an attempt to reduce statistical variations. Temperature constant, T= 0.1;Number of Epochs=500;ViT patch size=8

This optimization was conducted for all three backbones (ResNet, ConvNext and ViT) for the RAdam and also AdamW optimizers.

4 Experiments on CIFAR10

4.1 Learning Rate Optimization



	Max Learning Rate chosen for One Cycle LR
Resnet50	0.1
ConvNeXt	0.01562
ViT	0.00024

Figure 3: Super Convergence Training Loss as a function Learning Rate

In the case of the ConvNeXt and ViT-B, we observe they demonstrate super-convergence with a particular learning rate generating the lowest training loss. By setting the max learning rate of the one cycle lr scheduler as a function of the lr corresponding to min loss, we obtain a much better starting point for the training process without having to run multiple lengthy training cycles to derive an optimal learning rate. On the other hand for ConvNext, the Training Loss vs Epochs plot did not exhibit a minimum. In this instance, we chose the one cycle lr max scheduling rate for the model training to be 0.1.Transformations employed :RandomResize,RandomHorizontalFlip, RandomApply, RandomGaryscale

4.2 Test Accuracy on CIFAR10 as a function of Backbone Architecture

Results of the backbone comparison on CIFAR10 unsupervised classification clearly show that Contrastive loss minimization using Resnet50 backbone has the best performance followed by ConvNeXt and lastly ViT. These results only go to show that in a fair test where contrastive learning is the only pretext task, CNN outperforms transformers. This alludes to the fact that the reason why the Transformer outperformed CNN in [3] is because of the unequal number of pretext task allocation between ViT and Resnet50, not because of the the transformer backbone per se.

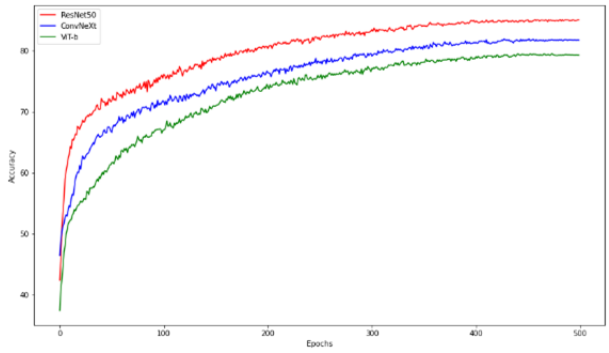


Figure 4: Test Accuracy on CIFAR 10 as a function of Backbone Architecture

4.3 Test Accuracy Results across different optimizers

	LAMB	RAdam @ 3/8	AdamW @3/8	AdamW @ 1/10
		With super convergence and one cycle LR		
ResNet	86.87	84.885	88	82.9
ViT	80.98	79.43	78.92	79
ConvNeXt	82.55	81.7	85.46	85.2

Figure 5: Test accuracy across various optimizers

A variety of optimizers were attempted before choosing RAdam. Their test accuracy results on CIFAR10 have been included in the table above to note that the ranking of the backbone performance did not change whether we took it from RAdam or took the best performance across all optimizers. All optimizer variant results on the Resnet backbone in this new test supersede that in [3].

4.4 TSNE Plots

These plots serve as a sanity check that the self-supervised clustering is able to find intra-class similarity even without labelled training. Classes 8,9 and 1 are well separated regardless of the backbone and the inter class proximity of many classes eg classes 6,4 and 7 appear to be maintained across the backbone. It is also interesting to note that the TSNE plots of the ConvNeXt and ViT, which fundamentally evolved from the Transformer Family of architectures exhibit more similarity to one another vs that from Resnet50 CNN Architecture

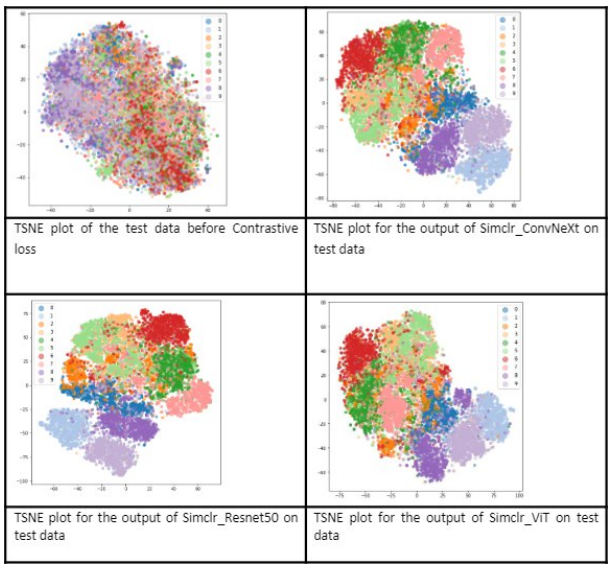


Figure 6: TSNE plots

5 Application on Remote Sensing Imagery

An exponential increase (>10000 PB) is forecasted in aerial imagery [4] in the next decade owing to the increasing number of high-resolution constellation satellites.

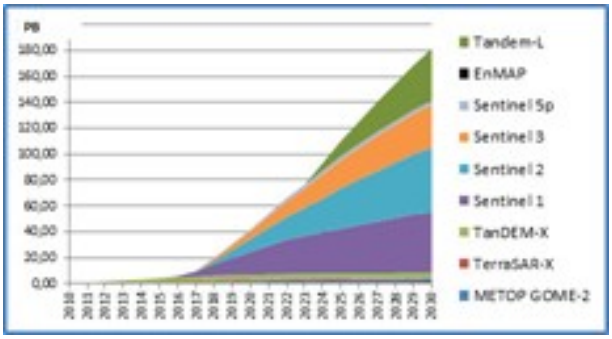


Figure 7: Trends and forecast of the data volume to be stored at D-SDA from earth observation missions (2010-2030), reproduced from [4]

Satellite sensors range across several spectral bands (Panchromatic, Multispectral, VHF, C band, X band Synthetic Aperture Radar-SAR) and their imaging geometries and resolutions are also vastly different. Their volume, variety and velocity make labeling and thus supervised classification an impossible task. Label free SSL with continuous retraining is the way to go to automate remote sensing imagery exploitation.

Satellite sensors image a 3 dimensional scene and project them to a 2 dimensional grid from various imaging geometry; Owing to the presence of tall structures in the scene eg trees, buildings etc, shadows cast are of different lengths and in different directions; Remote sensing images collected thus cannot be modelled as mere planar rotations of one another; Furthermore, signature of each object in the scene is a complicated characteristic of its shape, material as well as the sensor frequency, aspect and incidence angle[13]. Thus, it is not feasible to use contextual information to predict missing information in remote sensing images; For these reasons, training of pretext tasks like image inpainting and rotation estimation in [3] are not useful for remote sensing applications though data augmentation via random patch alterations and rotation are still useful to help reduce overfitting of the contrastive learning algorithm; Only the single task of contrastive learning as shown in [6] remains to

be useful in the context of remote sensing as opposed to the availability of several more useful pretext tasks for robust self supervised classification of images collected by a hand held camera. Remote sensing is thus one imaging domain identified to benefit from the techniques discussed in this paper.

5.1 Experiments with Remote Sensing Dataset: AID[14]

Figure 8: Image Samples from AID, reproduced from [14]

AID dataset consists of 5000 train and 5000 test sample images from Google Earth imagery belonging to 30 classes. Each sample is larger at 600x600 pixels to capture the contextual information necessary for remote sensing classification. The spatial resolutions of the images vary from 0.5 m/pixel to 0.8 m/pixel.

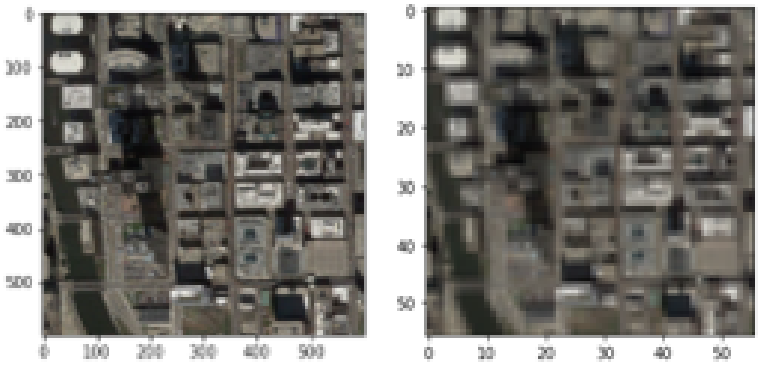


Figure 9: Resolution degradation of AID to meet GPU memory constraints

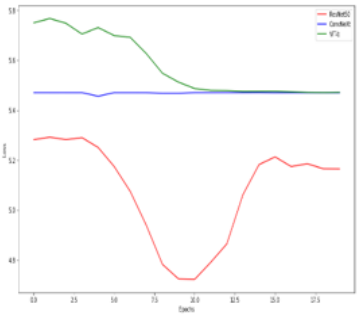
In order to fit a batchsize of 128 AID images into the 24GB RAM Single Nvidia RTX 3090 GPU that was being utilized for training, the original full resolution image had to be downsampled to 56x56 pixels; The effect of this massive resolution degradation can be witnessed through a sample in Figure 3.

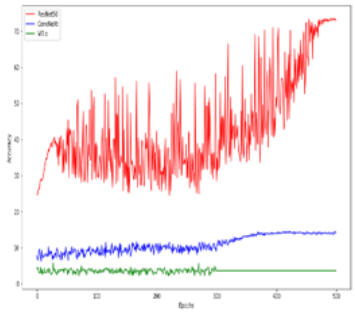
5.2 Results after Systematic Optimization

Besides the image size, all other algorithm parameters remain the same as in section 6. Results of the super convergence range test on AID in Figure 4a clearly show that ViT and ConvNeXt do not show any convergence as a function of learning rate and are also indicative of the classification results to be expected from these backbones. On the other hand, Resnet50 backbone shows a small amount of super convergence and there seems to be negligible overfitting. We proceed to optimize with the one cycle scheduler as described in section 5.2 over 500 epochs; Maximum learning rate for use in the one cycle scheduler is determined automatically for Resnet50 and arbitrarily for ViT and ConvNeXt. Only 1 run was achieved on AID for each of the backbones and the test accuracy performance as a

function of backbone is shown in Figure 4b . It is to be noted that we have not proceeded to do any further ablation study of the parameters that will work for ViT and ConvNeXt on AID.

Data is collected across different geographical locations, times, seasons and imaging geometries giving rise to tremendous intra-class variations in the AID dataset as opposed to CIFAR10. This can be witnessed in Figure 2. Attempting to be robust to these changes without the use of any labelled data is a huge challenge for most architectures. Coupled with the poor resolution of the degraded AID images and the smaller size of training data in AID ie 5000 vs 50000 in CIFAR10. it is clear that data-hungry ViT and ConvNeXt backbone architectures fail to offer useful classification. ViT performance comes close to that of a random throw of n=30 faced dice, yielding 3.3% test accuracy On the other hand, Resnet50 backbone after a slow start is able to reach more than 70% classification performance with 500 epochs. Supervised Classification performance using CNN for this dataset is reported to reach more than 85% in [15] Again Resnet50 has emerged as the backbone of choice for contrastive learning on AID.





	Max Learning Rate chosen for One Cycle LR
Resnet50	0.00179
ConvNeXt	0.000009
ViT	0.01

Figure 10: Results of SimCLR applied on AID dataset. (a) Super-convergence range test. (b) Test accuracy as a function of backbone. (c) Max LR table.

6 Conclusion

In this paper on the study of efficient backbone architectures for contrastive learning as applied to self-supervised classification, Resnet50 has proven to outperform ViT and ConvNext architectures consistently for both CIFAR10 and AID remote sensing dataset. These results convince us that as far as contrastive learning for self-supervised classification is concerned, CNN is still the desired backbone architecture vs Transformers.

In the course of comparison, a systematic optimization scheme has been demonstrated. Use of a one cycle scheduler ensures less sensitivity to any arbitrary choice of a fixed learning rate. A quick improvisation of the super convergence range test for contrastive loss is conducted to determine automatically the maximum learning rate input to the one cycle scheduler. This systematic workflow subjects the comparison to fewer adhoc parameter inputs. On top of this workflow, other optimizers like LAMB and AdamW have also been used to check the sanity of the performance figures obtained.

Results from the experiments in this paper show that the sensitivity of NN performance to the choice of optimizer, scheduler and learning rates cannot be under-stated. If not exhaustively tested, conclusions based on a difference of a few percentage points may not be robust. Tests to check the consistency of trends across more challenging and exemplary datasets is more than a good-to-have; it is a necessity.

Domain knowledge on the unique characteristics of the remote sensing dataset have helped in discerning the class of self supervised classification techniques appropriate for it.

7 Bibliography

[1] Dosovitskiy, A. et al(2021). “An Image is worth 16X16 Words: Transformers for Image Recognition at Scale, ICLR 2021.” from the arXiv database.

[2] Atito, S. Awais, M. and Josef Kittler, J. (Nov 2021). “Self-supervised vision Transformer.” from the arXiv database.

[3] Liu, Z. Mao, H. Wu, Feichtenhofer, C. Darrell, T. Xie, S. (Mar 2022). “A ConvNet for the 2020s.” from the arXiv database.

[4] “DLR-Earth Observation Center – 60 Petabytes for the German Satellite Data Archive D-SDA.” Startseite – DLR Portal, https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12632/22039$_{r}ead - 51751$

[5] Risojevi´c, V. Stojni´c, V. (Dec 2021) “The Role of Pre-Training in High-Resolution Remote Sensing Scene Classification.” from the arXiv database

[6] Li, H. Li, Y. Zhang, G. Liu, R. Huang, H. Zhu, Q. Tao, C. (Jan 2022) “Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images.” from the arXiv database.

[7] Mathilde, C. et al. “Emerging Properties in Self-Supervised Vision Transformers.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021,

[8]Bai,Y. et al. "Are Transformers More Robust Than CNNs?"Advances in Neural Information Processing Systems 34 (NeurIPS 2021)

[9]d'Ascoli et al(Mar 2021)."ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases" from the arXiv database

[10]Li, C. et al(Aug 2021)"BossNAS: Exploring Hybrid CNN- transformers with Block-wisely Self-supervised Neural Architecture Search" from the arXiv database

[11]Chen,T. et al(Feb 2020)"A Simple Framework for Contrastive Learning of Visual Representations", from the arXiv database

[12]Liu, Z. et al(Mar 2021)"Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", from the arXiv database

[13]https://www.geospatialworld.net/article/image-interpretation-of-remote-sensing-data/

[14]Xia,G. et al(2017)"AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification", IEEE Trans. on Geoscience and Remote Sensing, Vol. 55, No.7, 2017

[15]Li,Y. et al (Dec 2020)"Multi-Label Remote Sensing Image Scene Classification by Combining a Convolutional Neural Network and a Graph Neural Network". Remote Sensing

[16]Smith, L. Topin, N. (Aug 2017), "Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates". from the arXiv database

[17]https://github.com/sgugger/Deep-Learning/blob/master/Cyclical LR and momentums.ipynb

8 Acknowledgements

Sincere gratitude is owed to Assoc Prof Terence Sim Mong Cheng (NUS School of Computing) for approving the access to his research group's cluster of 4 NVIDIA RTX 3090 GPUs, on short notice without which the team will never have been able to complete the training. Also appreciation is conveyed to Sanjay Saha, System Administrator of this GPU cluster for assistance in the usage of these resources. Appreciation is also conveyed to NUS School of Computing CS5260 Module Lecturer Presidential Young Prof Yang You for initiating us to the breadth of topics covered in this module. We would also like to thank Assoc Prof Ng Teck Khim for inviting the guest speaker Zhuang Liu (from Facebook) to NUS to share his work on ConvNeXt.