Earth Systems and Environment (2023) 7:525–540
https://doi.org/10.1007/s41748-023-00343-3

ORIGINAL ARTICLE

Two-Speed Deep-Learning Ensemble for Classification of Incremental Land-Cover Satellite Image Patches

Michael James Horry[1,2] · Subrata Chakraborty[1,3] · Biswajeet Pradhan[1,4] · Nagesh Shulka[5] · Mansour Almazroui[6,7]

Received: 23 September 2022 / Revised: 20 December 2022 / Accepted: 10 March 2023 / Published online: 24 March 2023

Abstract

High-velocity data streams present a challenge to deep learning-based computer vision models due to the resources needed to retrain for new incremental data. This study presents a novel staggered training approach using an ensemble model comprising the following: (i) a resource-intensive high-accuracy vision transformer; and (ii) a fast training, but less accurate, low parameter-count convolutional neural network. The vision transformer provides a scalable and accurate base model. A convolutional neural network (CNN) quickly incorporates new data into the ensemble model. Incremental data are simulated by dividing the very large So2Sat LCZ42 satellite image dataset into four intervals. The CNN is trained every interval and the vision transformer trained every half interval. We call this combination of a complementary ensemble with staggered training a "two-speed" network. The novelty of this approach is in the use of a staggered training schedule that allows the ensemble model to efficiently incorporate new data by retraining the high-speed CNN in advance of the resource-intensive vision transformer, thereby allowing for stable continuous improvement of the ensemble. Additionally, the ensemble models for each data increment out-perform each of the component models, with best accuracy of 65% against a holdout test partition of the RGB version of the So2Sat dataset.

Keywords Incremental data · Satellite image classification · So2Sat LCZ42 · Sentinel-1 · Sentinel-2 · Deep learning · Vision transformer

1 Introduction

Since the first Landsat mission launched on 23 July 1972, industry, intelligence, and policy-making bodies have used satellite imagery as a primary source of information relating to land-use and land-cover change (USGS 2021). According to the Committee on Earth Observation Satellites database, as of September 2022 there are over 197 current earth observation satellite missions with a further 138 planned missions over the next 17 years ("The CEOS Database," 2022). The volume of geospatial data that is collected from satellite missions is huge. For example, the Landsat archive contains over 10 million images as of December 2021 ("Landsat Archive Adds Its 10 Millionth Image," 2021). Newer missions from Europe, USA, China, Brazil, and India create as much data each year as does Landsat over a decade with most agencies allowing open access to this data (Câmara et al. 2016), resulting in a vast volume of earth observation data available for analysis. The question of how to store and analyze this huge volume of data has been a popular field of research over the past decade, with various highly scalable computer

✉ Subrata Chakraborty
Subrata.Chakraborty@une.edu.au

[1] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia

[2] IBM Australia Limited, Sydney, NSW 2000, Australia

[3] School of Science and Technology, Faculty of Science, Agriculture, Business and Law, University of New England, Armidale, NSW 2351, Australia

[4] Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

[5] Griffith Business School, Griffith University, Nathan, QLD 4111, Australia

[6] Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[7] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

architectures proposed in the literature (Zhao et al. 2022). These architectures solve the volume problem by distributing processing over clusters of high-performance compute nodes (Sedona et al. 2019) using parallel processing computing paradigms such as Hadoop/MapReduce (Boudriki Semlali and Freitag 2021; Rajak et al. 2015; Tho et al. 2020), Spark ("Apache Sedona," 2022; Ge et al. 2019), Data Cubes (Appel and Pebesma 2019; "Open Data Cube," 2022; Simoes et al. 2021), and scalable array databases (Câmara et al. 2016; Cudre-Mauroux 2018; Joshi et al. 2019). The most popular of these is the Google Earth Engine (Gorelick et al. 2017) which is based around a parallel processing Hadoop/MapReduce architecture. A recent and comprehensive survey of analytical tools used in addressing volume/scale considerations of “big earth” data is provided by (Yang et al. 2019) and the reader is referred to this survey for a comprehensive review of current approaches.

Data volume is not the only complexity encountered in satellite image data analysis. A key characteristic of satellite imaging is that the collected image data are non-stationary. New, incremental images are constantly being added as satellites orbit the earth and transmit new data back to ground stations. This incremental data is an example of the “velocity” dimension of the so-called five Vs of “big data” ("NIST Big Data Public Working Group," 2022), with the other dimensions being volume, veracity, variety, and value. Whilst the volume dimension is frequently considered in the literature with several viable approaches outlined above, velocity is typically overlooked, and satellite image datasets tend to be presented and studied as static snapshots. Incorporation of non-stationary data into analytical models is acknowledged as an open challenge for deep learning algorithms (Najafabadi et al. 2015) and there is no consensus as to how to learn from streaming data, with a range of different techniques used depending on the application (Gomes et al. 2019). The body of research into classification of high-velocity data is largely concerned with dynamically adapting supervised models to add or remove features in response to concept drift caused by incremental data, using techniques such as single-feature ensembles (Parker et al. 2012), embedded denoising auto-encoders (Vincent et al. 2008; Zhou et al. 2012), and deep belief networks (Calandra et al. 2012). Such models are adaptable, but they are complex and computationally demanding, thereby limiting uptake by framework developers (Gomes et al. 2019).

In this work we present the novel idea of a two-speed network ensemble that can quickly incorporate new incremental data via a fast-training component, whilst preserving model stability using a slower-training but higher accuracy base component. The fast-training component model is a high-speed convolutional neural network (HS-CNN). The slower-training base component model is a Vision Transformer (ViT) (Dosovitskiy et al. 2020). As an ensemble, these two complementary network architectures produce a scalable, accurate, and adaptable computer vision model for land-use/land-cover analysis of satellite image patches as represented by standard climate zone labels. The two-speed network allows incremental data to be incorporated into the ensemble model more quickly than retraining the base component in isolation would otherwise allow. Although the presented technique can be applied to any “big data” computer vision task, a very large satellite imaging dataset has been chosen for the study as a useful demonstration of the usefulness of the proposed method in solving the real-world problem of efficiently processing high velocity earth observation data.

This paper is organized as follows: Section 2 outlines related work and describes the principles of operation for the classifiers. Section 3 describes the satellite image dataset, component and ensemble classifier network architectures, and the staggered training schedule that is used to simulate incremental data over time. Experimental results are presented in Sect. 4, tabulated for both component and ensemble models. Results are discussed in Sect. 5, including a detailed analysis of the superior classification performance of the ensemble model over each of the component models. We conclude with Sect. 6 describing limitations of the study and detailing planned future work.

2 Related Work

Initially, automated methods for the analysis of this data were based on pixel analysis since the coarse-grained image pixels contained features of interest within the pixel boundary (Richards and Jia 2006), and comparative changes of discrete pixel values within the same image patch could be used to indicate land-use/land cover changes (Shakya et al. 2021). Modern satellite images have pixels that are much smaller than typical objects of interest. Meanwhile, computational methods for the analysis of satellite images have evolved from hand-crafted feature extraction techniques such as histogram analysis, GIST descriptors (Oliva and Torralba 2001), scale-invariant feature transform (SIFT) (Lowe 2004), and histogram of oriented gradients (HOG) (Dalal and Triggs 2005), through machine learning methods such as principal component analysis (PCA) (Hotelling 1933), Random Forest (RF) (Du et al. 2015), Support Vector Machines (SVM) (Niknejad et al. 2014), and K-Means clustering (Rekik et al. 2009), to supervised deep learning systems. Deep learning systems combine automatic feature extraction and classification using multi-layer neural networks, typically variants of the convolutional neural network (CNN) first proposed by (LeCun et al. 1989).

CNNs are a refinement of the basic feed-forward multilayer perceptron (MLP) class of artificial neural network. Whilst MLP is comprised of at least three fully connected layers of neurons (Rumelhart et al. 1986), the CNN uses convolution filters panned across the input layers to extract feature maps which are then pooled to achieve dimensional reduction (LeCun et al. 1989). Layers in a CNN are sparsely connected since only convolved patches of the input image are connected to lower layers, as shown in Fig. 1.

The combination of convolutions, sparse connections, and pooling allows CNNs to train very efficiently via backpropagation. The efficiency of CNNs allows for their network architecture to consist of many layers. For example, VGG-16 and VGG-19 are CNN networks commonly used for computer vision tasks, consisting of 16 and 19 layers respectively (Simonyan and Zisserman 2015). The CNN implements a hierarchy of filters that extracts course-grained features such as edges at the top level with progressively finer feature extraction (such as colors and textures) occurring at deeper levels of the network (Bau et al. 2020). The ability of CNNs to extract finer features more efficiently than MLP architectures has led to the dominance of CNNs in many computer vision tasks including image classification, image enhancement, video processing, semantic segmentation, and object detection (Bhatt et al. 2021). An in-depth review of the history and wider applications of CNNs is beyond the scope of this paper, and the reader is referred to the recent comprehensive work by (Alzubaidi et al. 2021) for further reading.

Recently, competitive image classification results have been achieved by studies implementing a convolution-less architecture based on the transformer architecture commonly used in natural language processing (NLP) tasks (Vaswani et al. 2017). It was found by (Dosovitskiy et al. 2020) that transformer architectures could be applied to the task of image classification. Rather than feeding word embeddings into the transformer per NLP tasks, large images were broken into sequences of patches and fed into the transformer along with the corresponding patch position as training inputs to a sequence of attention modules denoted as ***L*** in Fig. 2. Output from the attention modules is used as input to a MLP header resulting in a classification output. This

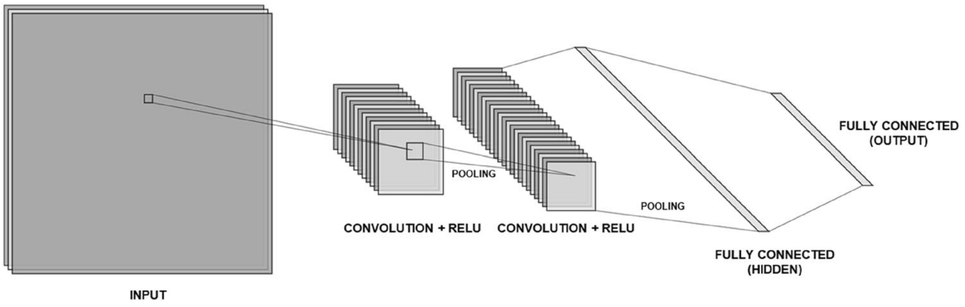
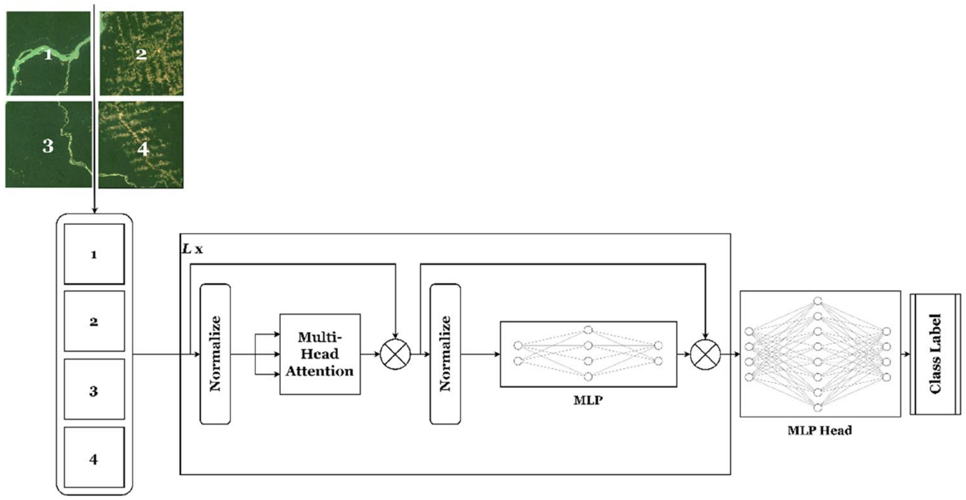


Fig. 1 Schematic for a simple two-layer CNN showing convolution, pooling, hidden and output layers. In this toy example, a 3 channel RGB image is first convolved to 12 channels followed by 24 channels. A pooling operation reduces dimensionality as input to a fully connected hidden layer. The hidden layer is fully connected to a smaller output layer, with each output neuron representing the probability of the input image containing a label

Fig. 2 Schematic for a simplified four patch Vision Transformer. Realistically, the image would be cropped into 16 or 32 patches. Image patches along with position embedding are input to a multi-MLP head self-attention modules denoted *L*. Final classification is by MLP to determine class labels and confidence score



architecture is known as a Vision Transformer (ViT). For small-scale training the ViT did not match classification metrics of modern CNN architectures, however with large-scale pre-training using hundreds of millions of images, the ViT was able to outperform CNNs on the ImageNet (Deng et al. 2009) classification task (Touvron et al. 2021) with state-of-the-art accuracy emerging from several studies (Chen et al. 2021; Dosovitskiy et al. 2020; Zhai et al. 2021). Some studies have also shown that the ViT architecture may be more robust than the CNN architecture, meaning that they are more stable when presented with adversarial images during training (Zhou et al. 2022).

An extensive study into the comparative performance of deep learning algorithms against hand-crafted feature extraction in the context of a large and diverse satellite image dataset was performed by (Cheng et al. 2017a). This study noted that handcrafted methods were typically evaluated against small datasets, resulting in unknown performance at scale. They evaluated hand-crafted, unsupervised learning, and deep learning algorithms against a large dataset (NWPU-RESISC45) consisting of 31,500 high-resolution image samples with an even distribution over 45 scene classes (Cheng et al. 2017b). It was shown that deep learning CNN models outperformed (on accuracy metrics) the tested handcrafted and unsupervised learning algorithms by a margin of at least 30%. A further performance boost of over 6% was achieved by fine-tuning off the shelf CNN models with VGG-16 (Simonyan and Zisserman 2015) achieving the highest accuracy for this task of over 90%. For comparison, none of the tested hand-crafted or unsupervised learning algorithms in this study achieved accuracy greater than 45%.

Training very deep neural networks such as CNNs is time-consuming and resource-intensive since multiple passes through large volumes of training data are needed to establish the optimum parameter values for millions of neurons in the network (Li et al. 2018). This has resulted in several researchers calling attention to the energy consumed in training and retraining these models (Dhar 2020; García-Martín et al. 2019). Incorporating new data into a deep learning model typically requires the model to be either fully retrained on a revised release of the entire training data corpus, which is time-consuming and resource-intensive, or fine-tuned with new training data as new samples become available. Unfortunately, fine-tuning using a limited set of new samples may lead to biasing the model to new sample data (Gavrilov et al. 2018; Li and Zhang 2021) unless great care is taken to appropriately weigh these samples during the training process, or alternatively impose weight constraints as a regularization measure (Sarle 1996).

A final complication encountered in fine-tuning pre-trained models is that new data may also bring new labels requiring a revised deep learning network architecture with the number of output neurons matching the revised number of class labels. In this case, a transfer learning approach (Chollet 2020) is not feasible, since this requires a match between the neural network architecture of the source and target models. Knowledge distillation (Hinton et al. 2015) techniques using the teacher/student paradigm provide a means of incorporating limited new data into student models (Nayak et al. 2019), but these methods are in their infancy and not proven at scale (Abbasi et al. 2020; Czyzewski 2021). Although adaptable to new data, one practical hindrance to adopting teacher/student models is the complexity involved in managing large numbers of resultant specialized student models and the question of how student model label scores are best combined into a domain level prediction.

3 Materials and Methods

New data acquisition is simulated by splitting the very large So2Sat dataset into four increments representing four points in time. The 25% data split represents a point where only the HS-CNN is fully trained. Therefore, the classification model for this smallest data increment is the HS-CNN alone. At the 50% data split the ViT and the HS-CNN are both trained with 50% of the total data. At the 75% split the ViT is still trained with only 50% of the full data and the HS-CNN is trained with 75% of the data, representing the real-world experience in which the 25% incremental data is rapidly included in the HS-CNN, but not in the ViT, which is slower to train. At the 100% data split, both the ViT and the HS-CNN are trained on all available data. The ensemble model is tested at each 25% increment against a holdout partition of the So2Sat data using a weighted average of the HS-CNN and ViT outputs. This experiment flow is depicted in Fig. 3.

It is envisaged that in a real-world implementation, this process of staggered training would continue indefinitely, using retraining of the HS-CNN to rapidly incorporate new data into the ensemble model whilst the ViT “catches-up” at a slower speed. For example, the HS-CNN could be trained hourly with the ViT trained on a daily or weekly basis as a real-world implementation of the scheme in Fig. 3. The most effective schedule would be determined empirically, using factors such as the rate of new data acquisition, compute resource availability, and cost.

3.1 Datasets

The scalability advantages of the ViT architecture over the CNN architecture are emergent only for large-scale datasets (Dosovitskiy et al. 2020; Steiner et al. 2021). Additionally, this investigation is primarily concerned with how deep learning models can adapt when incremental data are added to large datasets. Therefore, the proposed two-speed ensemble is trained against the very large So2Sat LCZ42 (So2Sat)

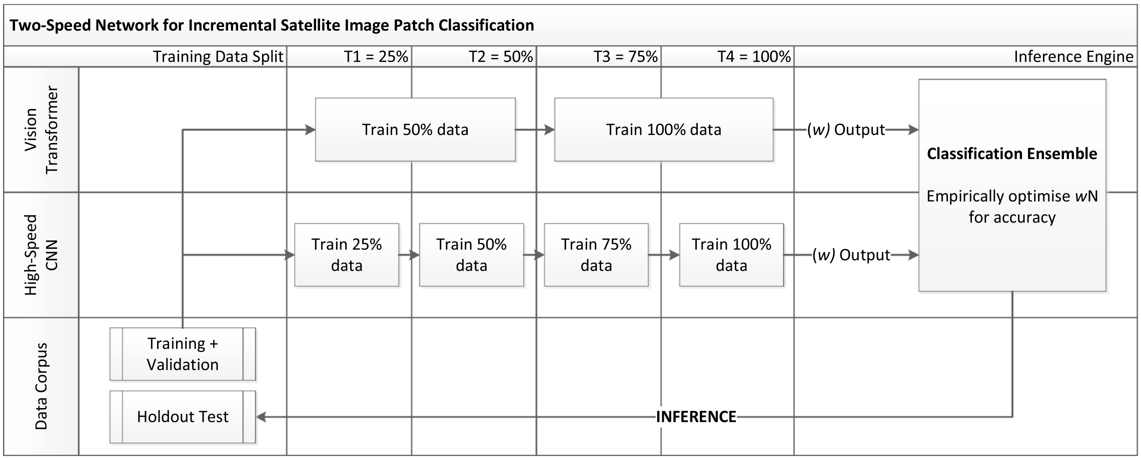


Fig. 3 Two-Speed network ensemble process flow for incremental satellite image patch classification using staggered training over four simulated time intervals

dataset consisting of 400,673 multispectral image patches from 42 cities at resolution of 10 m per pixel acquired from the Sentinel-1 and Sentinel-2 missions (Zhu et al. 2019a, b). The RGB subset of So2Sat used for this study contains 376,485 image patches since it is based on the first edition of So2Sat which does not contain an additional 24,188 multi-spectral test images from the second edition. So2Sat is an order of magnitude larger than other frequently cited sources of satellite/aerial data, such as UC-Merced (Yang and Newsam 2010), AID (Xia et al. 2017), Optimal31 (Wang et al. 2018), NWPU-45 (Cheng et al. 2017a), WHU-RS19 (Xia et al. 2010), RSSCN7 (Zou et al. 2015), and SIRI-WHU (Zhao et al. 2015). The second largest is NWPU-45 (Cheng et al. 2017a) with 31,500 images.

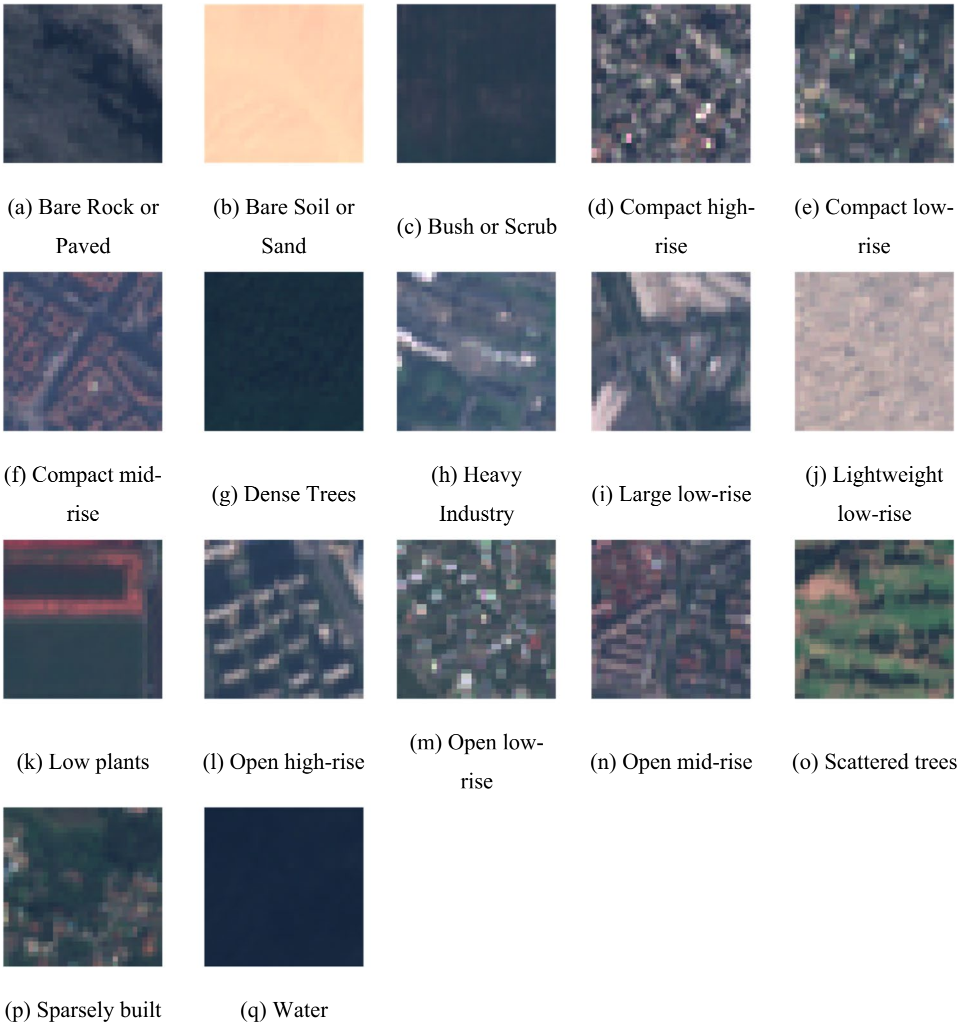
So2Sat image patches are classified using 17 local climate zone (LCZ) labels assigned by a team using a rigorous workflow including peer verification and quantitative evaluation resulting, in general, in human label confidence of 85% (Zhu et al. 2019b). Examples of each of the 17 LCZ labels in this dataset are shown in Fig. 4. Note that each So2Sat image patch measures only 32 × 32 pixels resulting in a small and pixelated appearance of the sample image patches, each being real-world dimensions of 320 m × 320 m.

The So2Sat authors established a baseline classification overall accuracy for several machine learning algorithms including RF, SVM, and attention augmented variation of ResNeXt (Xie et al. 2016). Overall accuracy metrics of 0.51 and 0.54 were achieved by RF and SVM classifiers respectively against the RGB version of So2Sat. The best overall accuracy in the source paper was 0.61 achieved using the ResNeXt based classifier, and this metric has been used as a baseline for this study. Similarly, this study makes use of the RGB subset of So2Sat to preserve a fair comparison to the So2Sat baseline metrics, since this study is focused on incorporating incremental data into deep learning models rather than state-of-the-art multi-spectral classification. It should be noted that classification accuracy was not the primary focus of (Zhu et al. 2019b), and that supervised machine learning classifiers including Maximum likelihood (ML), RF, and SVM have been employed with much higher accuracy in other studies of automated land-use classification. For example, ML and SVM classifiers have been used to classify land-use from multi-spectral Landsat 5 images with accuracies of 0.80 and 0.87 respectively over 13 land-use classes (Abbas et al. 2015). Similar results have been achieved for Sentinel-2 multi-spectral images using a RF classifier with various atmospheric correction techniques (Valdivieso-Ros et al. 2021) with best accuracy of 0.80 over 10 land-use classes. A recent benchmarking study on the multi-spectral version of So2Sat by (Qiu et al. 2020) achieved the best overall accuracy of 69% using a complex multi-level fusion CNN with 16 filters for the width of the first block and a network depth of 17 layers.

3.2 Network Architectures

We chose two distinct network architectures as a complementary pair for this study. Firstly, a ViT network architecture was selected as a highly accurate and scalable image classification network. Secondly, a HS-CNN architecture was handcrafted as an image classifier for high-speed training with few parameters, ideal for the incorporation of new data into a computer vision model.

CNNs tend to provide excellent performance on small to medium-sized datasets due to the relative ease with which CNNs identify inductive biases by automated feature extraction. For larger datasets, the scalability of the ViT architecture outweighs the inductive bias advantage of CNNs resulting in better classification performance at a large scale (Dosovitskiy et al. 2020). For this reason, ViT architectures



(a) Bare Rock or Paved

(b) Bare Soil or Sand

(c) Bush or Scrub

(d) Compact high-rise

(e) Compact low-rise

(f) Compact mid-rise

(g) Dense Trees

(h) Heavy Industry

(i) Large low-rise

(j) Lightweight low-rise

(k) Low plants

(l) Open high-rise

(m) Open low-rise

(n) Open mid-rise

(o) Scattered trees

(p) Sparsely built

(q) Water

Fig. 4 So2Sat samples showing 17 land-use/land-cover types (using standardized LCZ labels) with a mix of local and global features

have recently been proven highly effective in remote sensing applications using satellite imagery achieving state-of-the-art (Bazi et al. 2021) results across four datasets, UC-Merced (Yang and Newsam 2010), AID (Xia et al. 2017), Optimal31 (Wang et al. 2018), and NWPU-45 (Cheng et al. 2017a). For this study, we selected a 16 patch ViT architecture using 12 encoder layers, a hidden size of 768, MLP size of 3072, and 12 self-attention heads, resulting in a model with 85.7 million trainable parameters. This ViT architecture was selected as the smallest and least training resource-intensive option given the small size of the So2Sat images. The ViT used was pre-trained on ImageNet classes and shared with the community by (Morales 2021).

Although the ViT architecture is efficient and scalable, it requires a large number of samples before overtaking traditional CNNs in terms of classification metrics (Dosovitskiy et al. 2020). In the case of new data added to an already large data corpus, complete retraining of the ViT would be resource expensive and time-consuming. For this reason, we have handcrafted a high-speed CNN (HS-CNN) classifier with few parameters, derived from the VGG architecture (Simonyan and Zisserman 2015) but with only three layers, each comprising two convolution layers. This network was designed with the objective of minimizing training time while maintaining good accuracy for classifying new data.

The number of trainable parameters for the HS-CNN is 2.8 M. For comparison, other commonly used CNN architectures such as VGG16 (Simonyan and Zisserman 2015), ResNet18 (He et al. 2016), and ResNeXt (Xie et al. 2016) have 138.4 M, 11.5 M, and 25 M trainable parameters, respectively. To minimize overfitting to this very sparse network architecture, each pooling layer is followed by a dropout layer, and fully connected layers were regularized using an L2 regularization penalty (Ng 2004). The HS-CNN architecture used in this paper is shown in Fig. 5. The HS-CNN is initialized with random weights and biases and trained from scratch using the So2Sat image patch dataset.

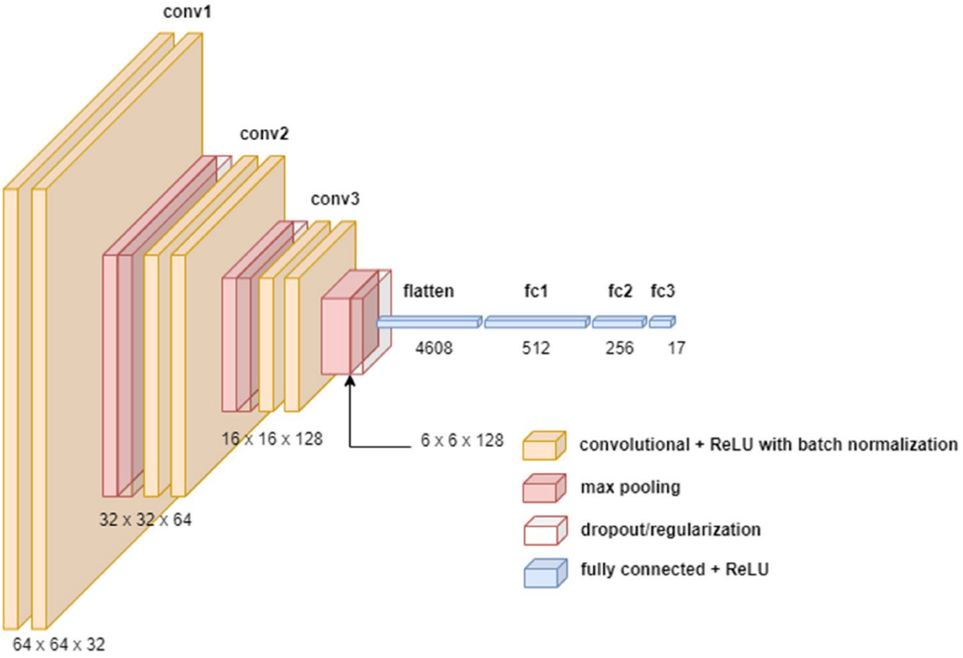


Fig. 5 HS-CNN architecture overview based on a VGG-like structure with three layers and additional regularization designed to minimize training time whilst avoiding overfitting

3.3 Ensemble Architecture

Since the ViT breaks an image into patches (16 for this study) and then encodes each patch with positional embedding as an input to the transformer encoder, the ViT learns global features of an image simultaneously with pixel values (Raghu et al. 2021). In contrast, since a CNN is trained by learning the co-relationships of overlapping small arrays of pixels, the CNN learns pixel-based local features first, with long-range global features becoming emergent as training proceeds. We expect that the contrasting learning strategies of ViT and HS-CNN make these models good candidate components for an ensemble model (confirmed in results Sect. 3.3), whereby outputs from each component model are combined via a weighted averaging algorithm to arrive at a final prediction according to Eq. 1.

$$\overline{p} = \frac{\sum_{i=1}^{n} pi \times wi}{\sum_{i=1}^{n} wi} \tag{1}$$

Here, p is the predicted score for samples from the i classifier and w is the weight assigned to predictions from that classifier. As we are combining outputs of two classifiers, $n=2$. In this study, each classifier has been assigned a weight ranging from 0.1 to 0.9 with steps of 0.1 with each classifier's weights adding to unity on each test. An industrial-strength implementation of the proposed staggered learning scheme would include the classifier weights as a learnable parameter to automatically optimize the predictive value of the ensemble.

3.4 Staggered Training Schedule

Since this study is concerned with additional data at four points in time, four classification models are used in testing as detailed in Table 1. After an initial time interval, CNN-25 is a HS-CNN, trained and validated on 25% of the data and representing a point in time (T1) where there has been sufficient time to train the HS-CNN but not the ViT. ENS-50 represents the point in time (T2) where the ViT has completed training on 50% of the data along with the CNN also having been trained on 50% of the data. ENS-75 represents a point in time (T3) partway through the next ViT training cycle where the ViT model is still only available as trained on 50% of the data, but the high-speed CNN has been trained on 75% of the data. Finally, ENS-100 represents a point in time (T4) where both classifiers are fully trained on 100% of the data.

Table 1 A staggered training schedule was used to mimic the availability of new data at four points in time denoted as T1 to T4

Model name	Time increment	Training data (%)	CNN training (%)	ViT training (%)
CNN-25	T1	25	25	None
ENS-50	T2	50	50	50
ENS-75	T3	75	75	50
ENS-100	T4	10	100	100

The high-speed CNN is trained every incement. The ViT is trained every two increments. Ensembles are created at each increment using the most completely trained component model

3.5 Experiment Setup with Incremental Data Simulation

The So2Sat corpus is available as a TensorFlow dataset providing both multi-channel and JPEG encoded red, green, and blue (RGB) images. For this study, we selected the RGB subset to allow for a fair comparison with the deep learning classifier results from the So2Sat source paper (Zhu et al. 2019b) and also to generalize the approach to other three-channel, visible spectrum computer vision tasks. The So2Sat dataset includes a standard split for model training and testing purposes. This split provides a total of 352,366 images for training/validation and 24,119 for holdout testing. Each model was trained and validated on increasing 25% increments of the training data but tested against the entire holdout testing corpus to provide a fair comparison of predictive capability at each simulated time increment. Training data was augmented with random left/right/up/down flipping along with random brightness, contrast, and saturation operations. Testing data was not augmented in any way. All images were shuffled before being used to train/test classifiers to eliminate sampling biases that may have been caused by data collection order, for example local geographical confounders such as a regional standard for roofing materials, building, and industrial layouts.

Table 2 Summary of hardware and software configuration for this study

Hardware	Software
CPU: Intel Xeon E-2288G @ 3.70 GHz	OS: Red Hat Enterprise Linux v7.9
Memory: 64GiB	Python: 3.8.12
GPU: Nvidia Quadro RTX 6000 with CUDA v 11.2	TensorFlow: 2.7.0

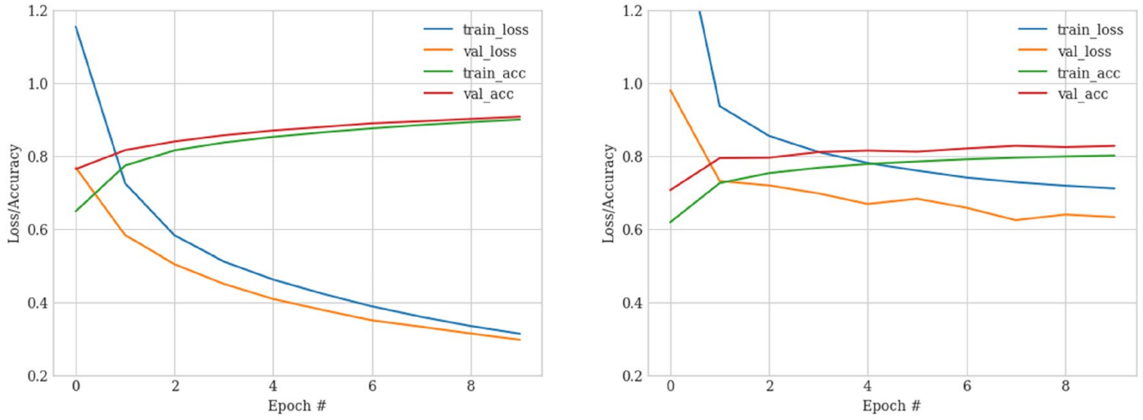
3.6 Compute Configuration

All experiments were executed on the University of Technology Sydney Interactive High Performance Compute environment, using hardware and software as described in Table 2.

4 Results

4.1 Model Training and Validation

Training curves for the ViT and HS-CNN classifiers, when trained against the complete So2Sat training data set, are presented in Fig. 6a and b. The ViT classifier training chart shows good convergence without overfitting with excellent validation accuracy of 0.92. The HS-CNN training curve also shows good convergence, especially for a scratch-trained network, but with an overall lower validation accuracy of 0.82. Training convergence was essentially identical regardless of the data split used, with the only noticeable difference being a slower convergence for the ViT classifier when trained with the 25% split.



(a)

(b)

Fig. 6 Training curves for component models trained for 10 Epochs for the full So2Sat training dataset. **a** ViT training converged, resulting in high validation accuracy of 0.92. **b** HS-CNN training converged, reaching a validation accuracy of 0.82

4.2 ViT and HS-CNN Training Metrics and Holdout Testing

Training and validation results for the ViT and HS-CNN are presented in Table 3. The HS-CNN achieved a best holdout test overall accuracy of 0.61 when using 25% of the training data and 0.60 when using the full training data set. This is lower than the benchmark of 0.61 set by (Zhu et al. 2019b) using complex attention augmented ResNeXt architecture but still a reasonable result given that the HS-CNN has an order of magnitude less training parameters (2.8 M vs ~ 25 M) than the ResNeXt architecture used in that study (Xie et al. 2016). The HS-CNN meets its design objective of good accuracy at high training speed, with the full training dataset of 317,129 images processed in 28 min. As expected, the ViT showed improved performance over the HS-CNN with peak overall accuracy of 0.63 using 50% of the training data, dropping to 0.62 when 100% of the training data was used. It is likely that this minor drop in accuracy at the 100% dataset increment is an indicator that the network has started to overfit, given that the validation overall accuracy (OA) metric showed a minor increase in accuracy for the same data increment. The ViT performance represents a marginal improvement on the benchmark overall accuracy of 0.61. The ViT took over 4 h to train with the full training set, which is over 8 times the training time of the HS-CNN.

4.3 Ensemble Model Holdout Testing Results

Three ensemble models were created using variously trained HS-CNN and ViT models as follows:

1. ENS-50 consisting of the HS-CNN and the ViT each trained on 50% of the training data,
2. ENS-75 consisting of the HS-CNN trained on 75% of the training data and the ViT trained on 50% of the training data, and
3. ENS-100 consisting of the HS-CNN and the ViT each trained on 100% of the training data.

Results of holdout testing for the models at each time increment are presented in Table 4. At time increment T1 using 25% of the training dataset partition, the only trained model is the HS-CNN. Therefore, results are identical to those obtained using HS-CNN at a 25% training split. For time increment T2, the ViT and HS-CNN are both trained using 50% training data. At time T3, the ViT and HS-CNN are trained on 50% and 75% of training data, respectively. At T4, both the ViT and the HS-CNN are trained on 100% of training data. This allows the HS-CNN and ViT to be combined with predictions used as inputs to the weighted averaging function described in Eq. 1. A scripted experiment varied the HS-CNN:ViT weighting by 10% from 10:90 to 90:10. Best, and identical, results were achieved using ENS-75 with weighting ratios of 40:60, 50:50, and 60:40 as shown in Fig. 7.

Table 3 Base model training metrics summary for each data increment

Data split (%)	Training image count	Validation image count	Training time (h:mm:ss)		Validation OA		Holdout test OA	
			HS-CNN	ViT	HS-CNN	ViT	HS-CNN	ViT
25	81,044	7,048	0:07:23	1:03:45	0.78	0.82	0.61	0.63
50	158,565	17,618	0:14:36	2:04:30	0.81	0.89	0.60	0.63
75	239,609	24,665	0:21:35	3:07:29	0.81	0.91	0.58	0.62
100	317,129	35,237	0:28:34	4:09:39	0.82	0.92	0.60	0.62

Holdout test results were obtained against the full So2Sat test set of 24,119 images regardless of training data increment

Table 4 Results of inference for each staggered training time interval

Model	Time increment	Total training time	Precision	Recall	F1	OA	Kappa
CNN-25	T1	0:07:23	0.59	0.61	0.58	0.61	0.57
ENS-50	T2	2:19:06	0.63	0.64	0.62	0.64	0.60
ENS-75	T3	2:29:05	0.64	0.65	0.63	0.65	0.61
ENS-100	T4	4:38:03	0.64	0.64	0.63	0.64	0.61

Holdout test results were obtained against the full So2Sat test set of 24,119 images. Ensembles were weighted 50:50 for each classifier

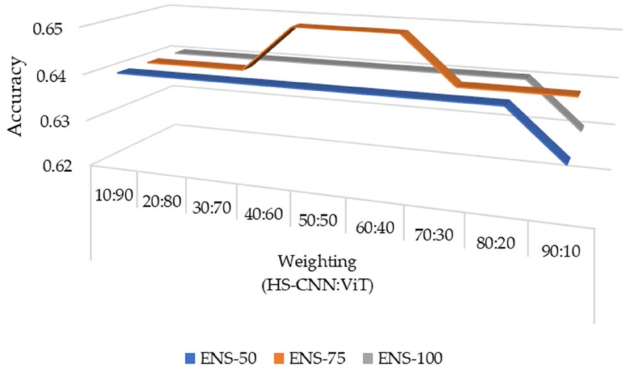


Fig. 7 Ensemble weighting results. Effect of different weight ratios for ensemble component models. Best accuracy is achieved using a balanced HS-CNN:ViT weighting ranging from 40:60 to 60:40

The result of holdout testing including the combined models is provided in Table 4. In general, the overall accuracy results at times T2, T3, and T4 given by the combined models are better than those of either component model at the same data partition. ENS-50, comprising HS-CNN trained on 50% data and ViT also trained on 50% data, achieved an overall accuracy of 64% on the holdout test. The highest overall accuracy was achieved by ENS-75, comprising HS-CNN trained with 75% data and ViT trained on 50% data. This ensemble achieved an overall accuracy of 65%, which is an improvement over the baseline overall accuracy of 61% (Zhu et al. 2019b). This result also represents an improvement in overall accuracy of the component HS-CNN and ViT classifiers (at the equivalent data split T3) being 58% and 62%, respectively. Such an improvement may be considered empirical to this study/dataset, and further investigation is needed to prove a more generalized link between the ensemble architecture employed and this small improvement in overall accuracy. ENS-100, consisting of an ensemble of fully trained HS-CNN and ViT achieved an overall accuracy of 64%. This represents a 1% reduction in overall accuracy at T4 compared to T3, and can be interpreted as a likely result of minor overfitting of the HS-CNN as indicated by the uptick in validation loss visible in Fig. 6b from epoch 7.

Where classes are highly imbalanced in object classification tasks, an algorithm may return artificially high accuracy metrics simply by classifying all samples as a majority class. For this reason, the measures of precision and recall are frequently used to report the quality and sensitivity of an algorithm, respectively. Precision is the proportion of true positive labels that are assigned by an algorithm against the sum of true positive labels and false positive labels. Recall is a measure of the correctness of the labels assigned for each class calculated as true positive labels divided by the sum of true positive labels and false negative labels. F1 score is the harmonic mean of the precision and recall metrics (Pedregosa et al. 2011). Finally, since precision, recall (and thereby F1) metrics do not take account of true negative the Cohen's Kappa coefficient of agreement (Artstein and Poesio 2008) is frequently employed in remote sensing studies to eliminate the role of pure chance from reported metrics, thereby providing a better real-world measure of the algorithms utility.

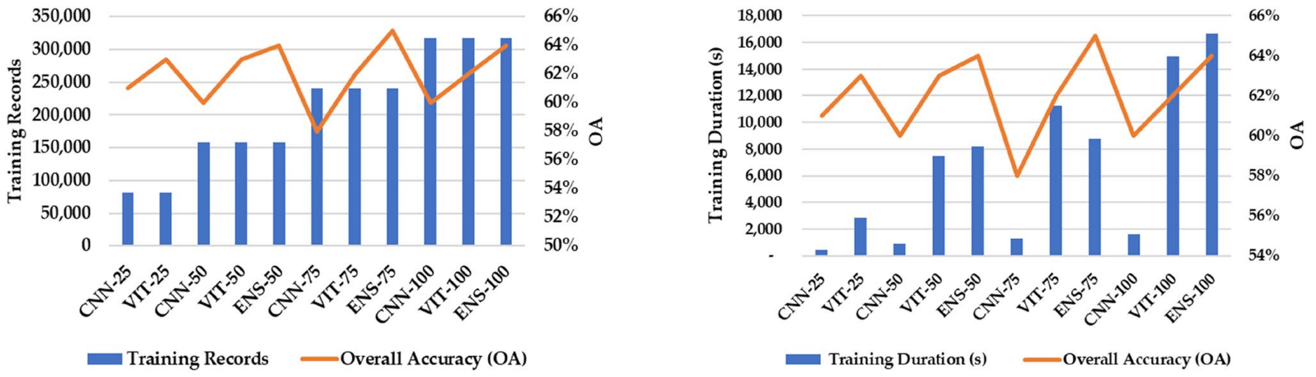
Precision and recall metrics were well balanced for all tests indicating that the accuracy was not achieved through simple over-classification of majority classes. The Cohen's Kappa score for the ensemble classifiers was in the range 0.60–0.61, indicating a moderate level of agreement between the predicted and true labels.

To illustrate the effectiveness of the ensemble approach in relation to both accuracy and efficiency, Fig. 8 presents comparative plots of all tested models. Figure 8a shows that the combined models improved accuracy over component models for all models at all data partitions. ENS-75 provided the highest accuracy of all tests with training time approximately half that of the fully trained ViT model as shown in Fig. 8b.

4.4 Ensemble Model Classification Analysis using Confusion Matrices

Confusion matrices associated with each classifier and the ensemble at training interval T3 were generated to analyze the relative strengths of each approach. The confusion matrices for HS-CNN, ViT, and ENS-75 are shown in Fig. 9a, b, and c, respectively.

As the ViT achieved higher overall accuracy than the HS-CNN we first consider the class labels contributing most to this accuracy delta. The top three such classes are "Open high-rise", "Bare rock or paved", and "Bare soil or sand". The HS-CNN failed to classify any "Open high-rise" correctly, and instead classified the majority (*n* = 442) of true "Open high-rise" as "Open mid-rise". Recalling Fig. 4l and n as examples of these two classes, the "Open high-rise" examples show regular building alignments that are not present in the open mid-rise. The pattern of these regular alignments is apparent as long-range diagonal features, explaining the ViT superior performance in classifying these classes. Similarly, the ViT outperformed the HS-CNN in separating the classes with sparse local features such as the "Bare" classes in Fig. 4a and b, where the visible features are long-range features across the image patch, such as topographical features in the case of bare rock or paved, or sand dune formations in bare soil or sand. The two classifier types provide similar performance for image classes that lack long-range features such as water, dense trees, and bush or scrub, as evident from the confusion matrices in Fig. 9.



(a) (b)

Fig. 8 Comparison of holdout test accuracy results for all models including HS-CNN, ViT and ensembles. **a** is Classification Accuracy by Training Data Size. Ensemble models composed of ViT and Lightweight CNN show higher accuracy with less training data than ViT or CNN trained on larger datasets. **b** is Classification Accuracy by Training Duration. Ensemble model consisting of ViT trained on 50% data and Lightweight CNN trained on 75% of data provides the best accuracy of 65% with training time approx. 40% lower than a fully trained ViT

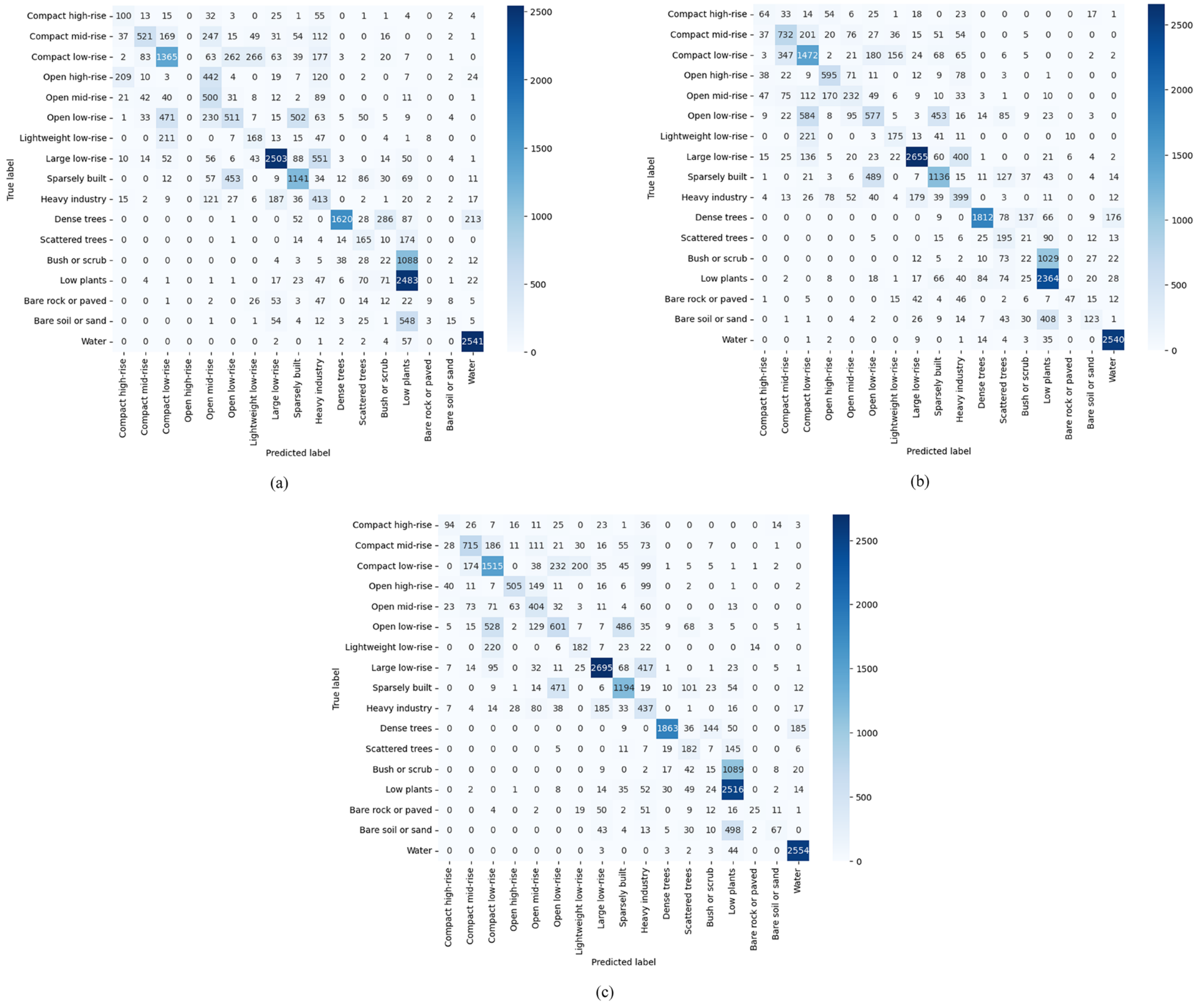
To further investigate the differences between the ViT and the HS-CNN, class activation maps were generated for the divergent classes are shown in Fig. 10a–i. Class activation mapping is a technique that provides a visual representation of the parts of an image that gain the attention of a deep learning algorithm (Zhou et al. 2016) via an image overlay of pixel intensity at the last convolution layer of the network.

Figure 10b illustrates the ViT attention to the long-range feature of building alignments for the Open high-rise class whereas the HS-CNN in Fig. 10c attends to less focused regions of pixels that are a mix of buildings and open space, resulting in the HS-CNN proving to be unable to distinguish between Open high-rise and Compact high-rise, Open mid-rise, and Heavy Industry. In a similar manner, Fig. 10e illustrates the ViT attending to the bare rock feature in the upper right corner of the image patch, which is an area of low attention to the HS-CNN 10(f). Finally, the ViT appears to have identified sand dune areas in Fig. 10h with the HS-CNN failing to attend to any feature at all in Fig. 10i. The "Bare soil or sand" image patch is featureless to the HS-CNN since it is poor at identifying the long-range sand dune edges when compared to the ViT.

5 Discussion

Incorporating new data into deep learning computer vision systems will remain a challenging problem, since complete re-training of such systems is resource-intensive, and alternate methods such as teacher-student modelling, and fine-tuning with new data are also problematic. Increases in computing power over time, particularly GPU processing, tend to be quickly consumed by the desire to train deep learning systems on more significant numbers of high-definition images, thereby instantly consuming compute improvements. The proposed two-speed ensemble network comprising a low-parameter HS-CNN combined with a slower but more accurate ViT provides a practical means of incorporating incremental data to a large dataset by leveraging a staggered training schedule, with our experiments confirming lower overall training time needed to reach maximum accuracy. Additionally, the complementary natures of these different deep learning architectures lead to improved classification metrics for the So2Sat dataset with accuracy of 65% achieved in holdout testing, using a fully trained HS-CNN and a ViT trained on 50% of the complete data corpus. This result improves on the overall accuracy baseline of 61% and is, to the best of our knowledge, the current state-of-the-art for the RGB version of the So2Sat dataset.

Although the objective of this study was to improve efficiency of incremental image patch classification for very large datasets, the contributing factors to our results were interesting. Image classes that were better separated by the ViT over the HS-CNN were identified, with network attention maps indicating that the ViT is superior to the HS-CNN in detection of long-range features, even in the small So2Sat image patches where such features are limited to around 10 m. The suitability of the ViT for identifying long-range features stems from the ViT inclusion of patch position relationships in training input, whereas the HS-CNN training input is limited to highly localized pixel arrays without position context. Therefore, the ViT can better train on features that span the image patch, such as building alignments and topographical features, making this architecture highly suitable for land-use classification.



(a)

(b)

(c)

Fig. 9 Confusion matrices relating to the best ensemble model ENS-75. **a** Lightweight CNN trained on 75% of data. **b** 16 patch ViT using 50% of training data. **c** Ensemble model ENS-75 taken at T3

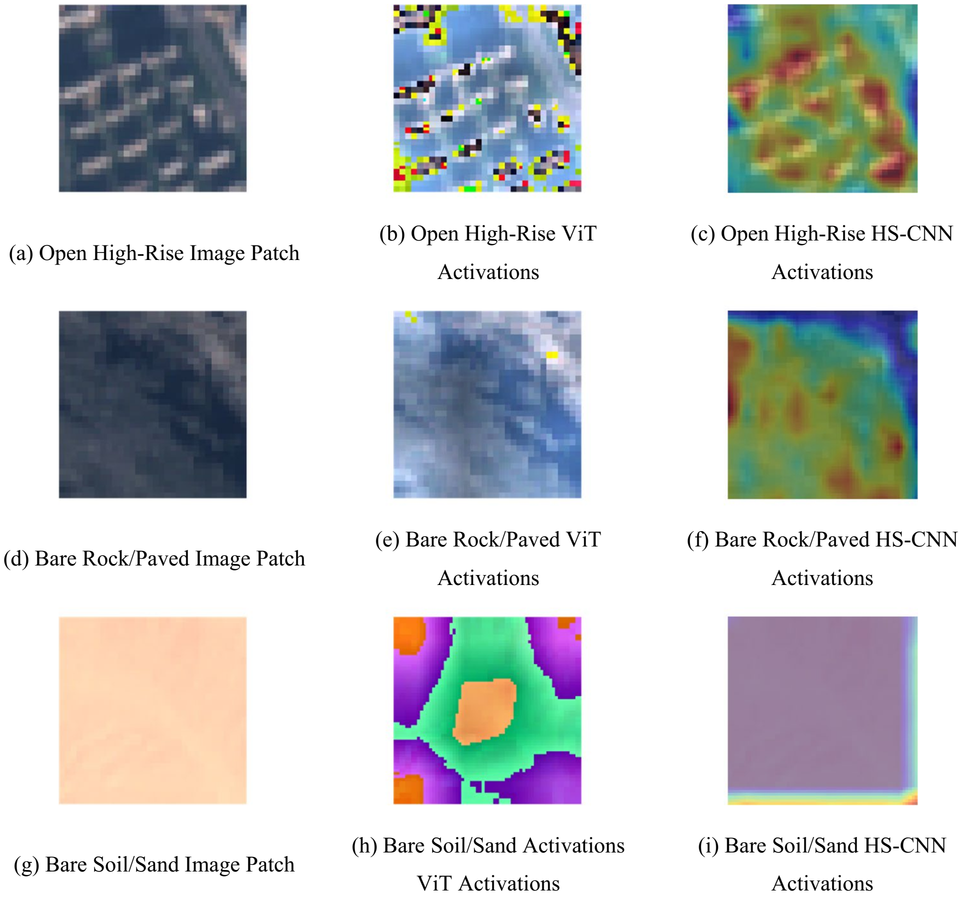
In summary, this study shows that the high resource cost/training time required by a ViT architecture can be mitigated by combining it with a low-parameter count HS-CNN that can quickly retrain with new incremental data, with better results than the ViT alone trained on the same dataset.

6 Conclusions

This study presents a first investigation into the use of a two-speed network as a means of incorporating incremental data into deep learning-based classification schemes. Our focus was on showing that the proposed method succeeds at this task, with potentially broad-ranging application to other domains where new image data is generated at high velocity. This limited the study in two ways. Firstly, although the So2Sat dataset provides multi-band data, we have restricted our experiments to the visible spectrum to facilitate reproducibility beyond the remote sensing use case. Secondly, we restricted the study to deep learning-based algorithms, rather than the combination of hand-crafted feature extraction with RF or SVM classifier commonly employed in remote sensing studies. Our next study will further investigate the remote sensing use case, using multi-spectral images to improve the goodness of fit along with side-by-side comparisons of ensembles composed of both deep learning and machine learning classifiers such as RF, SVM, and clustering.

In the future, we intend to progress the two-speed network to an industrial trial whereby empirical performance data will be used to tune the ViT and HS-CNN architectures,

Fig. 10 ViT activation maps for class labels more accurately separated by the ViT over the HS-CNN. **a** Open high-rise with **b** Open high-rise ViT activation map accurately tracking the building alignment, and **c** Open high-rise HS-CNN activation map also tracking the building alignment but at a much lower resolution. **d** Bare rock or paved with **e** ViT activations tracking long-range topographical features as a feature in the top right corner, and **f** HS-CNN also tracking some topographical features but with poor resolution resulting in widespread activation over the image and no attention to the bare rock in the top right corner, **g** Bare soil or sand with **h** ViT activations again tracking the long-range sand dune edges, and **i** HS-CNN with no relevant activations for this patch attributable to lack of long-range attention


(a) Open High-Rise Image Patch

(b) Open High-Rise ViT Activations

(c) Open High-Rise HS-CNN Activations

(d) Bare Rock/Paved Image Patch

(e) Bare Rock/Paved ViT Activations

(f) Bare Rock/Paved HS-CNN Activations

(g) Bare Soil/Sand Image Patch

(h) Bare Soil/Sand Activations ViT Activations

(i) Bare Soil/Sand HS-CNN Activations

hyperparameters, and classifier weightings, resulting in a domain-specific ensemble that is efficient to train and adaptable to new data. This will provide a valuable tool for strategic planning agencies to formulate actions in response to changes in the landscape. We also note that the proposed two-speed network approach allows the model production release to be undertaken using an agile software methodology/pipeline, whereby the resource costly ViT model is considered a major release, with the frequently updated HS-CNN model component considered to be a point release. Such as scheme would allow for stable continuous improvement of computer vision models in a manner that has not been previously reported in the literature. Finally, we are investigating how the two-speed network ensemble might be enhanced by the inclusion of a few-shot learning engine based on an edge-labelling graph neural network as suggested by (Kim et al. 2019) as a means of adding real-time classification capability for previously unseen image classes.

Acknowledgements The first author would like to acknowledge their employer, IBM Australia Limited for approving writing time to produce this manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research was supported by Defence Australia funding under the AI for Decision Making Initiative project titled “Effective updating of deep learning models with limited new data” (UTS Ref: 210018980).

Data availability The image data used in this study are publicly available from Zhu et al. (2019a, b)

Declarations

Conflict of interest The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

Abbas T, Fereydoon S, Amin M, Chamran Taghati Hossien P, Amir Hossein Esmaile S (2015) Land use classification using support vector machine and maximum likelihood algorithms by Landsat 5 TM images. Walailak J Sci Technol 12:681–687. https://doi.org/10.14456/WJST.2015.33

Abbasi S, Hajabdollahi M, Karimi N, Samavi S (2020) Modeling teacher-student techniques in deep neural networks for knowledge distillation. In: 2020 International conference on machine vision and image processing (MVIP). IEEE, pp 1–6

Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8:53. https://doi.org/10.1186/s40537-021-00444-8

Apache Sedona (2022) https://sedona.apache.org/. Accessed 6 Sept 2022

Appel M, Pebesma E (2019) On-demand processing of data cubes from satellite image collections with the gdalcubes library. Data 4:92

Artstein R, Poesio M (2008) Survey article: inter-coder agreement for computational linguistics. Comput Linguist 34:555–596. https://doi.org/10.1162/coli.07-034-R2

Bau D, Zhu J-Y, Strobelt H, Lapedriza A, Zhou B, Torralba A (2020) Understanding the role of individual units in a deep neural network. Proc Natl Acad Sci 117:30071–30078. https://doi.org/10.1073/pnas.1907375117

Bazi Y, Bashmal L, Rahhal MMA, Dayil RA, Ajlan NA (2021) Vision transformers for remote sensing image classification. Remote Sensing 13:516

Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K, Ghayvat H (2021) CNN variants for computer vision: history, architecture, application, challenges and future scope. Electronics 10:2470

Boudriki Semlali B-E, Freitag F (2021) SAT-hadoop-processor: a distributed remote sensing big data processing software for earth observation applications. Appl Sci 11:10610

Calandra R, Raiko T, Deisenroth MP, Pouzols FM (2012) Learning deep belief networks from non-stationary streams. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 379–386

Câmara G, Assis LF, Queiroz G, Ferreira K, Llapa E, Vinhas L, Maus V, Ipia A, Souza R (2016) Big earth observation data analytics: matching requirements to system architectures

Chen X, Hsieh C-J, Gong B (2021) When vision transformers outperform ResNets without pre-training or strong data augmentations. Preprint at arXiv:2106.01548

Cheng G, Han J, Lu X (2017a) Remote sensing image scene classification: benchmark and state of the art. Proc IEEE 105:1865–1883

Cheng G, Han J, Lu X (2017b) resisc45. https://www.tensorflow.org/datasets/catalog/resisc45. Accessed 2 Mar 2022

Chollet F (2020) Transfer learning & fine-tuning. Complete guide to transfer learning & fine-tuning in Keras. https://keras.io/guides/transfer_learning/. Accessed 22 Feb 2022

Cudre-Mauroux P (2018) SciDB. In: Sakr S, Zomaya A (eds) Encyclopedia of big data technologies. Springer International Publishing, Cham, pp 1–3

Czyzewski MA (2021) Transfer learning between different architectures via weights injection. Preprint at arXiv:2101.02757

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE, pp 886–893

Deng J, Dong W, Socher R, Li L, Kai L, Li F-F (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255

Dhar P (2020) The carbon impact of artificial intelligence. Nat Mach Intell 2:423–425. https://doi.org/10.1038/s42256-020-0219-9

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16x16 words: transformers for image recognition at scale. Preprint at arXiv:2010.11929

Du P, Samat A, Waske B, Liu S, Li Z (2015) Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. Int J Photogramm Remote Sens 105:38–53

García-Martín E, Rodrigues CF, Riley G, Grahn H (2019) Estimation of energy consumption in machine learning. J Parallel Distrib Comput 134:75–88. https://doi.org/10.1016/j.jpdc.2019.07.007

Gavrilov AD, Jordache A, Vasdani M, Deng J (2018) Preventing model overfitting and underfitting in convolutional neural networks. Int J Softw Sci Comput Intell 10:19–28. https://doi.org/10.4018/IJSSCI.2018100102

Ge S, Isah H, Zulkernine F, Khan S (2019) A scalable framework for multilevel streaming data analytics using deep learning. In: Getov V, Gaudiot JL, Yamai N, Cimato S, Chang M, Teranishi Y, Yang JJ, Leong HV, Shahriar H, Takemoto M, Towey D, Takakura H, Elci A, Takeuchi S, Puri S (eds). 43rd IEEE annual computer software and applications conference, COMPSAC 2019. IEEE Computer Society, pp 189–194

Gomes HM, Read J, Bifet A, Barddal JP, Gama J (2019) Machine learning for streaming data: state of the art, challenges, and opportunities. SIGKDD Explor Newsl 21:6–22. https://doi.org/10.1145/3373464.3373470

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens Environ 202:18–27. https://doi.org/10.1016/j.rse.2017.06.031

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. Preprint at arXiv:1503.02531

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417

Joshi A, Pebesma E, Henriques R, Appel M (2019) Scidb based framework for storage and analysis of remote sensing big data. Int Arch Photogramm Remote Sens Spatial Inform Sci-ISPRS Arch 42:43–47. https://doi.org/10.5194/isprs-archives-XLII-5-W3-43-2019

Kim J, Kim T, Kim S, Yoo CD (2019) Edge-labeling graph neural network for few-shot learning. Preprint at arXiv:1905.01436

Landsat Archive Adds Its 10 Millionth Image (2021) https://www.usgs.gov/landsat-missions/news/landsat-archive-adds-its-10-millionth-image. Accessed 5 Sept 2022

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. Neural Comput 1:541–551. https://doi.org/10.1162/neco.1989.1.4.541

Li D, Zhang HR (2021) Improved regularization and robustness for fine-tuning in neural networks

Li Y, Zhang H, Xue X, Jiang Y, Shen Q (2018) Deep learning for remote sensing image classification: a survey. Wires Data Min Knowl Discov 8:e1264. https://doi.org/10.1002/widm.1264

Lowe G (2004) Sift-the scale invariant feature transform. Int J Comput Vision 60:91–110

Morales F (2021) vit-keras. https://github.com/faustomorales/vit-keras. Accessed Jan 10 2022

Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2:1. https://doi.org/10.1186/s40537-014-0007-7

Nayak GK, Mopuri KR, Shaj V, Radhakrishnan VB, Chakraborty A (2019) Zero-shot knowledge distillation in deep networks. In: International conference on machine learning. PMLR, pp 4743–4751

Ng AY (2004) Feature selection, L 1 vs. L 2 regularization, and rotational invariance. Proceedings of the twenty-first international conference on Machine learning, p 78

Niknejad M, Zadeh VM, Heydari M (2014) Comparing different classifications of satellite imagery in forest mapping (case study: Zagros forests in Iran). Int Res J Appl Basic Sci 8:1407–1415

NIST Big Data Public Working Group (2022) https://bigdatawg.nist.gov/home.php. Accessed 5 Sept 2022

Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vision 42:145–175. https://doi.org/10.1023/A:1011139631724

Open Data Cube (2022) https://www.opendatacube.org. Accessed 5 Sept 2022

Parker B, Mustafa AM, Khan L (2012) Novel class detection and feature via a tiered ensemble approach for stream mining. In: 2012 IEEE 24th international conference on tools with artificial intelligence, pp 1171–1178

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res 12:2825–2830

Qiu C, Tong X, Schmitt M, Bechtel B, Zhu XX (2020) Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: benchmark results on the So2Sat LCZ42 dataset. IEEE J Sel Top Appl Earth Obs Remote Sens 13:2793–2806

Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A (2021) Do vision transformers see like convolutional neural networks? Adv Neural Inf Process Syst 34:12116–12128

Rajak R, Raveendran D, Bh MC, Medasani SS (2015) High resolution satellite image processing using hadoop framework. In: 2015 IEEE international conference on cloud computing in emerging markets (CCEM), pp 16–21

Rekik A, Zribi M, Hamida AB, Benjelloun M (2009) An optimal unsupervised satellite image segmentation approach based on pearson system and k-means clustering algorithm initialization. Methods 8

Richards JA, Jia X (2006) Remote sensing digital image analysis: an introduction, 5th 2013 edn. Springer Berlin/Heidelberg, Berlin, Heidelberg

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Sarle WS (1996) Stopped training and other remedies for overfitting. Comput Sci Stat 352–360

Sedona R, Cavallaro G, Jitsev J, Strube A, Riedel M, Benediktsson JA (2019) Remote sensing big data classification with high performance distributed deep learning. Remote Sens 11:3056

Shakya AK, Ramola A, Vidyarthi A (2021) Exploration of pixel-based and object-based change detection techniques by analyzing ALOS PALSAR and LANDSAT data. Smart and Sustainable Intelligent Systems pp 229–244

Simoes R, Camara G, Queiroz G, Souza F, Andrade PR, Santos L, Carvalho A, Ferreira K (2021) Satellite image time series analysis for big earth observation data. Remote Sens 13:2428

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. Preprint at arXiv:1409.1556

Steiner A, Kolesnikov A, Zhai X, Wightman R, Uszkoreit J, Beyer L (2021) How to train your vit? data, augmentation, and regularization in vision transformers. Preprint at arXiv:2106.10270

The CEOS Database (2022) http://database.eohandbook.com/. Accessed 5 Sept 2022

Tho, Nam V, Nguyen D, Le HA (2020) A Big Data Framework for Satellite Images Processing using Apache Hadoop and RasterFrames: A Case Study of Surface Water Extraction in Phu Tho, Viet Nam

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jegou H (2021) Training data-efficient image transformers & distillation through attention. In: Marina M, Tong Z (eds). Proceedings of the 38th international conference on Machine Learning. PMLR, Proceedings of Machine Learning Research, pp 10347–10357

USGS (2021) What is the Landsat satellite program and why is it important? https://www.usgs.gov/faqs/what-landsat-satellite-program-and-why-it-important. Accessed 21 Feb 2022

Valdivieso-Ros C, Alonso-Sarria F, Gomariz-Castillo F (2021) Effect of different atmospheric correction algorithms on sentinel-2 imagery classification accuracy in a semiarid mediterranean area. Remote Sens 13:1770

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Vincent P, Larochelle H, Bengio Y, Manzagol P-A (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. Association for Computing Machinery, Helsinki, Finland, pp 1096–1103

Wang Q, Liu S, Chanussot J, Li X (2018) Scene classification with recurrent attention of VHR remote sensing images. IEEE Trans Geosci Remote Sens 57:1155–1167

Xia G-S, Yang W, Delon J, Gousseau Y, Sun H, Maître H (2010) Structural high-resolution satellite image indexing. ISPRS TC VII Symposium-100 Years ISPRS, pp 298–303

Xia G-S, Hu J, Hu F, Shi B, Bai X, Zhong Y, Zhang L, Lu X (2017) AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans Geosci Remote Sens 55:3965–3981

Xie S, Girshick R, Dollár P, Tu Z, He K (2016) Aggregated residual transformations for deep neural networks. Preprint at arXiv:1611.05431

Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, pp 270–279

Yang C, Yu M, Li Y, Hu F, Jiang Y, Liu Q, Sha D, Xu M, Gu J (2019) Big Earth data analytics: a survey. Big Earth Data 3:83–107. https://doi.org/10.1080/20964471.2019.1611175

Zhai X, Kolesnikov A, Houlsby N, Beyer L (2021) Scaling vision transformers. Preprint at arXiv:2106.04560

Zhao B, Zhong Y, Xia G-S, Zhang L (2015) Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. IEEE Trans Geosci Remote Sens 54:2108–2123

Zhao Q, Yu L, Du Z, Peng D, Hao P, Zhang Y, Gong P (2022) An overview of the applications of earth observation satellite data: impacts and future trends. Remote Sens (basel, Switzerland) 14:1863. https://doi.org/10.3390/rs14081863

Zhou G, Sohn K, Lee H (2012) Online Incremental feature learning with denoising autoencoders. In: Neil DL, Mark G (eds). Proceedings of the fifteenth international conference on artificial intelligence and statistics. PMLR, Proceedings of Machine Learning Research, pp 1453--1461

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929

Zhou D, Yu Z, Xie E, Xiao C, Anandkumar A, Feng J, Alvarez JM (2022) Understanding the robustness in vision transformers. In:

Kamalika C, Stefanie J, Le S, Csaba S, Gang N, Sivan S (eds), Proceedings of the 39th international conference on machine learning. PMLR, Proceedings of Machine Learning Research, pp 27378–27394

Zhu X, Hu J, Qiu C, Shi Y, Bagheri H, Kang J, Li H, Mou L, Zhang G, Häberle M, Han S, Hua Y, Huang R, Hughes L, Sun Y, Schmitt M, Wang Y (2019a) So2Sat LCZ42 30 August 2018 edn. TUM

Zhu XX, Hu J, Qiu C, Shi Y, Kang J, Mou L, Bagheri H, Häberle M, Hua Y, Huang R (2019b) So2Sat LCZ42: A benchmark dataset for global local climate zones classification. Preprint at arXiv:1912.12171

Zou Q, Ni L, Zhang T, Wang Q (2015) Deep learning based feature selection for remote sensing scene classification. IEEE Geosci Remote Sens Lett 12:2321–2325