Multi-temporal Multi-spatial Scale Transformer for Fine Scaled Cotton Yield Prediction

YU QIU

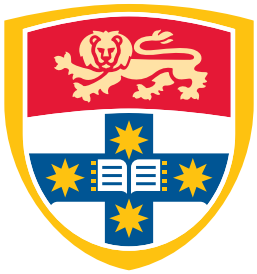
SID: 450199176

Supervisor: Zhiyong Wang

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Information Technology (Honours)

School of Computer Science
The University of Sydney
Australia

5 December 2022



THE UNIVERSITY OF
SYDNEY

Student Plagiarism: Compliance Statement

I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Yu Qiu

Signature: **Date**:

Abstract

Crop yield prediction plays an important role in critical decision making in precision agriculture. Fine-scale prediction, in particular, can be used in addressing within-field crop management decisions. Conventional crop yield predictions involve complicated in-field data collection which is usually expensive to perform. Remote sensing images provides an alternative solution for yield prediction, which are often used to monitor physical characteristics of farmlands. While remote sensing images contributes significantly in recent crop prediction studies due to the development of deep learning in computer vision tasks, yet very few of them has made approach in high resolution, pixel-wised prediction within the field.

Transformer has shown great potential in multiple computer vision tasks in recent years however it has not been applied in yield prediction. In this thesis, series of mainstream deep learning models were experimented on Llara dataset which is a farmland located in northern New South Wales, Australia.

Also, a novel inexpensive and accurate pixel-level cotton yield prediction algorithm, multi-temporal multi-spatial scale transformer(MTMSST), is proposed which purely based on publicly accessible satellite images. MTMSST is a transformer based deep learning architecture that is enhanced by multi-scaled patching over both temporal and spatial dimension for extra information to be extracted, followed by a two-phase cross attention fusion for global information exchange.

The experimental results are comprehensively compared with and outperformed multiple deep learning algorithms used in yield prediction. The result shows that transformer-based models such as Video Vision Transformer outperforms CNN models on prediction task with RMSE of 7.41 unit/ha. MTMSST further improves the result by 16 percent with RMSE of 6.18 unit/ha.

Acknowledgements

I would like to express my deepest appreciation to my supervisor Associate Professor Zhiyong Wang for all the guidance and supports through the semesters. I would also like to express my deepest gratitude to my mentor Kun Hu who always inspires me with knowledge and ideas through the journey of research. I wish to thank all my peer research students who shared their knowledge and ideas in every weeks meeting during the semester. I would also like to thank Ada Lee for her support during the research.

CONTENTS

List of Figures

CHAPTER 1

Introduction

1.1 Background

Agricultural management becomes increasingly important as the world's population grows.[8] The Food and Agriculture Organization (FAO) have made predictions where food demand will be increasing over 60 percent by 2050.[4] Considering the cost and environmental destruction for farmland expansion, achieving a stable or even higher yield of the existing farmland is one of the key aspects under such demand. To prevent fall of crop yield due to various causes such as change of the climate, critical decisions such as irrigation and use of fertiliser were made, however proper decisions could not be made without accurate, real-time monitoring of the crop and the idea of Precision Agriculture raised by ISPA (International Society for Precision Agriculture) under such challenge[3] where they officially define as "Precision Agriculture is a management strategy that gathers, processes and analyses temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production". Hence, crop yield prediction acts as an important role in the management process to monitor the current state of the crop in terms of the final yield which helps achieve a better yield in the harvest season.

1.2 Remote Sensing Images

Remote sensing here refers to obtaining a field phenomenon without direct contact with it. In this study particular, it refers to the use of satellite or other aircraft sensors to obtain on-field information. This differs from the traditional on-site collected data which normally requires higher cost in measurements[5]. Remote sensing suggests a possible solution to the given situation for its high cost-efficiency, semi-real time monitor of the field(depending on the cycling periods) and with well-preserved features. Studies

have proved that prediction models with the use of remote sensing data have shown promising improvement [10]

Sentinel-2 is an earth monitoring mission carried out by the Copernicus Programme to provide optical imagery at high spatial resolution from 10m to 60m.[2] Multiple studies in crop yield prediction tasks have used or partially used Sentinel-2 satellite imagery as the input of the prediction model for its high spatial resolution, easy to get access to and relatively short cycles(usually between 5 to 10 days) which is idea for ground monitoring.[12][13]

1.3 Crop Yield Prediction

Traditional yield predicting techniques such as linear regression and mechanistic approach has limitations due to their applicability and uncertainty[19]. Followed by the rapid development of computational technologies, machine learning algorithms have shown great strength in yield estimation tasks. Typical approach includes method such as Support Vector Machine and Random Forest etc.[20] However they are still limited by the complexity in data acquisition, since they often require complex input to the model includes weather components and soil conditions.

Recent advancement in deep learning algorithm which is a particulars branch of machine learning combing with the use of satellite images suggests a solution to the situation. Previously use of satellite images typically requires manual feature selection and extraction, for example the calculation over optical channels to obtain vegetation index. Deep learning algorithms such as CNN, on the other hand requires only the raw input of the image and features extracted through layers of artificial neural networks. Studies have demonstrated deep learning algorithm out performs traditional machine learning methods. [28] Carried on with this result, different deep learning algorithm was applied and contributed to the predicting task.Some of them are more advanced and specialised in using time-series image data such as 3D-CNN[13] and LSTM-CNN[25] that utilize the advantage of satellite imagery with multiple cycles during the growing seasons to have a better monitoring of crop growing.

Transformer is a special deep learning architecture that is originally invented for neutral language processing tasks with its unique multi-head self-attention mechanisms. [24] Following research has shown such architecture also have great potential towards computer vision tasks and outperforms traditional architectures such as CNN.[11] However, based on the current research, though there are research on use of transformer architecture with remote sensing images, such advanced method have not been used

in pixel-wised crop yield estimation. Hence, to investigate the performance of this architecture is one of the main researching goal in this thesis.

1.4 Contribution

Transformer architectures have out performed traditional CNN architectures in multiple computer vision tasks such as video classification and recognition as it was mentioned above. In this thesis we have developed a pixel-wise high resolution crop yield prediction framework, that clip the yield mask with multi temporal satellite imagery data and creates window for every measurement in spatial resolution of 10m which enables the interaction between a pixel and its neighbours. Also, we have applied transformer architectures such as vision transformer and video vision transformer that no studies have applied those on a yield prediction task with remote sensing images. What is more, we have proposed a novel transformer architecture multi-temporal multi-spatial scale transformer for fine scaled cotton yield prediction which outperform both CNN and transformer approach.

1.5 Thesis Outline

This thesis will start with related literature reviews to present a background on the research including the conventional crop yield prediction methods, deep learning and convolutional neural networks applied in this area and transformer architectures in computer vision and series input of data.

Followed by a detailed explanation on how are experiments are conducted over the provided dataset as well as a brief introduction on each deep learning models that was used in the experiments. Then we will introduce the proposed MSMTSST model and explain in detail as the main contribution.

In the next chapter, the result will be demonstrated and comprehensive comparisons will be made between the proposed method and others, also the ablation studies that was carried out. In the end, major discovery and future studies will be discussed.

CHAPTER 2

Literature Review

Common Framework of machine learning involves sections such as data processing, feature selection and modelling process, however as different approaches were made by different research groups, the actual content could vary significantly. Following reviews analyzes different approaches towards the prediction task. Since my research is based on the remote sensing data, the review will particularly focus on the use of the remote sensing data, even though multi-sourced data was used in some of the research. The literature review part will focus on three major fields, the conventional approach where the experiment framework, use and process of datasets can be referenced in the experiment. Second, the use of deep learning algorithm on crop yield prediction where their architecture can be used to compare with. Third, different types of advanced vision transformer models in recognition and classification tasks that provide ideas on how they have different method comparing with the conventional CNN methods. Note that comparison between the models could be difficult, especially around pixel-wised prediction since their is no public accessible pixel-wised yield dataset available, many experiment are carried out in different field data over different year, location and type of crop.

2.1 Conventional Machine Learning Approach

Fillip et al. (2019)[15] has proposed the use of satellite image combined with the on-field data as the input to predict the yield of wheat, barley and canola. This is definitely not the earliest machine learning approach in crop yield estimation, but series of researches were conducted by this group to give a well demonstrated framework on how the prediction task is conducted.2.1

The research uses multiple sources of data including the on-field collected soil and nutrition, meteorology information as well as the Moderate Resolution Imaging Spectroradiometer(MODIS), a multi-spectral satellite image data. Remote sensing data here is not directly used as the input of the prediction model. Instead, the remote sensing data is converted to a layer of enhanced vegetation index which was

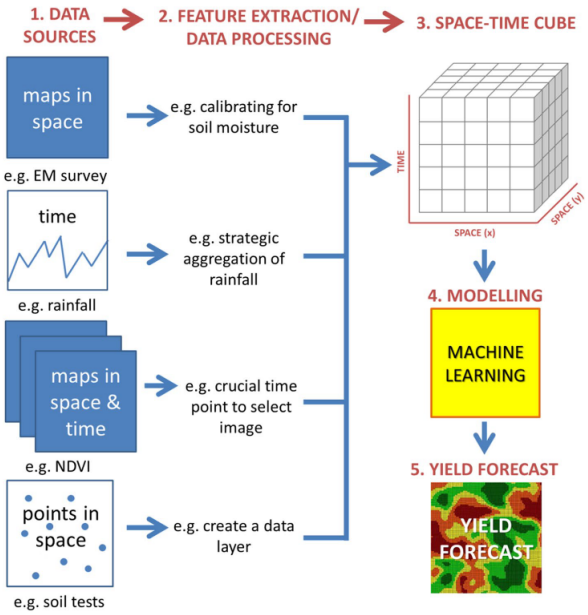


FIGURE 2.1: Data cube and work flow Fillip et al. (2019)

derived from some of the spectral at spatial resolution of 250m. The major relation of the paper with the research is that it demonstrated the data-cube structure that is well-fused on field data and remote sensing data. A data-cube is mapping the features with respect to geological information. Multiple features including the MODIS image at a critical time period are stacked to form a layer of feature, and different time stacks together formed the final data-cube structure. This data structure can preserve both spatial and temporal feathers, fusing them together as the input of the prediction model which can be used in my own research. The datacube is then proceeded with a random forest model with 10-fold cross validation. Another major contribution of the paper is that it demonstrates the importance of each co-variate in the data cube, where MODIS EVI layer data has great impact on the prediction, showing great potential of prediction with remote sensing. Such analysis demonstrated the importance of satellite imagery input and its potential in further feature extraction.

Fillip et al. carried out several studies that provided valuable information on the yield prediction tasks on different crops with aspects of both spatial and temporal. However, there still exists a gap in their research. Though they have stated the importance of remote sensing data several times, the research carried out with the mixed using multiple data sources. Whether satellite images can be the only or the major input of the investigation still remains un-investigated, hence the previously mentioned high cost-efficiency prediction model is not fully achieved. Also, current existing remote sensing data usually

contains a large number of spectral but researchers only selected one or few spectral derived data such as EVI and NDVI, ignoring a lot of other information or features that satellite images provided. Such a gap could be the intended outcome of my own research. Overall, Fillip et al(2019), Fillip et al. (2020) and Al-Shammari et al. (2021) has provided a research framework that my research could possibly follow, and different input and use of deep learning models instead of random forest could be the starting point of my own research.

Han,Bishop, Fillip. (2022) also carried on Fillip et al. (2019) research but this time focusing on the different prediction time period to predict the yield of sugarcane. One major difference from the previous studies is that this research is purely based on the open accessed data rather than a mixture with on-field data. The paper demonstrated different prediction accuracy of early-season and late-season. As a result, though late-season have better overall performance, early season prediction is very meaningful considering the actual purpose of the prediction which is to assist crop management during the growing season. Hence, this research is very practical compared to the previous studies which made a big step towards a realistic, cost-efficient model.

2.2 Deep Learning Approach

You et al.(2017)[28] has made an approach to predict country-level yield of corn with the use of remote sensing images with deep learning algorithms which can be considered as the state of the art at the time of publishing. The research used 9 bands MODIS satellite image, 7 bands for the surface reflection and two for the surface temperature. For each country level yield data point, 32 satellite images with 8 days between each are taken, hence the raw input is a list of satellite images. The raw input is then proceeded with a dimensional reduction method which is converting images to a 3D histogram. The Y-axis indicates the mapping of occurrence of 255 pixel values and reduces to 32, X-axis indicates time, the 32 images readings and Z-axis represents different bands.2.2 This process is performed under the assumption of permutation invariant which means that spatial relation between each pixel in the image does not contain any features. Instead of using handcrafted features like EVI and NVDI like previously mentioned in Fillip et al.(2019), This research uses deep learning algorithm convolutional networks to automatically extract features from the manipulated images. Following image is a demonstration of the structure of a simple network used in paper. The result shows a satisfying accuracy of RMSE of 5.7 in country-level performance.

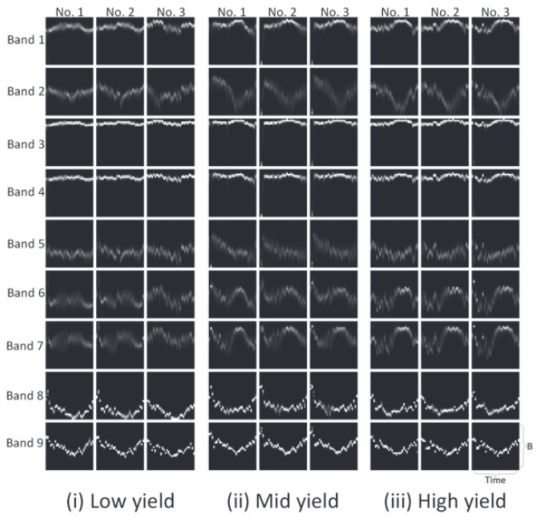


FIGURE 2.2: Dimension Reduction by You et al(2017)

The use of dimension reduction offered a choice of fusing spatial and temporal data other than the data cube in the previous research. However, the assumption of permutation in-variance may not hold in real-world on field situations due to the fact that the pixel around a data point does have physical meaning. In this case, use of dimension reduction may lead to spatial features and could have a negative impact on the prediction result. What is more, You et al.(2017) made yield predictions on a county-level, which may be not that meaningful for precision crop management where each location should be considered separately. Still their research has shown the strength of deep neural networks, where it marked the beginning of using of convolutional networks in yeild estimation.

Sagan et al.(2021)[23] has proposed another yield prediction framework purely based on the satellite image input. Image data comes from combination use of both WorldView-3 and Planet Scope. Both satellite images were preprocessed with calibration and pansharpening, then co-registorated together for spatial alignment. Four WorldView-3 images and 25 PlanetScope images were stacked to generate one input and match with the yield data with spatial information. The input is then fed to both ResNet2D and ResNet3D2.3 to generate results. This study is in the similar direction to my research and in the end 3D CNN architecture shows the best result that outperforms others even with plot level hand crafted feature extraction(those were unable to automatically learn image features like CNN). More importantly, this study is a demonstration of localized yield prediction at with-in field level. It split farmlands into several rectangles and treat each rectangle as a sample to predict the yield of a specific area. This is critical relating to the pixel-wised predcition since it considers a window as input rather than a single point as it

was in conventional approach.2.4 However, Worldview-3 and Planet Scope are high resolution satellite images which are not free and open to the public which is against the cost-efficient requirement.

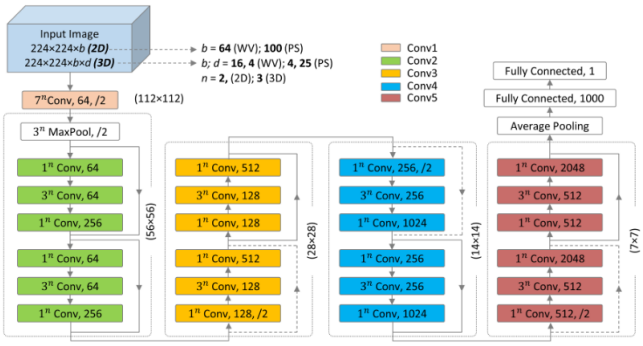


FIGURE 2.3: 3d and 2d CNN architecture By Sagan et al(2021)

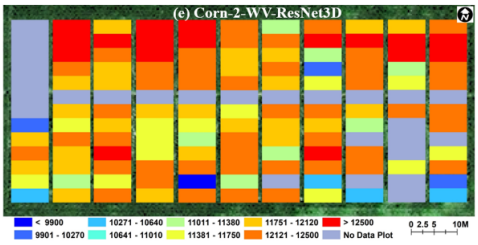


FIGURE 2.4: 3d and 2d CNN architecture By Sagan et al(2021)

Fajardo et al (2021)[12] has also made approach using convolutional neural network. Different from the previous ones, this research has achieved high resolution wheat yield forecasting through deep learning with the use of Sentinel-2 images which is similar to the desired structure.2.5 The research also shows with purely public accessible data can achieve satisfying forecasting result, and with further add-on data even better result could be achieved. The downside on this study is the CNN architecture that was applied was Mario net which is a very shallow and simple network which may limit its ability of feature extraction especially temporal relations of multiple satellite images input.

2.3 Transformer Architecture in Computer Vision

Transformer architecture is a special designed deep learning architecture towards neutral language processing tasks that was introduced in 2017 by Google research team which is quite recent comparing with other deep learning methods. [24] The major property that makes it different from the traditional

FIGURE 2.5: Sample Data Points location and window selection by Fajardo et al (2021)

deep neural networks is that it adopts a "self-attention" module. By patching words into vectors of specific dimension(the process is called word embedding), then adding the positional information through positional embedding, the self-attention mechanism is able to connect different parts of a single input sequence and generate new representation of the sequence. Applying attention layers through encoder for feature extraction and decoder for sentence reconstruction, transformer is able to accomplish translation tasks in its own way. By having a better global vision comparing with the tradition LSTM models, transformer showed outstanding performance which outperforms traditional models in NLP fields.

Soon it was found by Alexey,et al(2020)[11] surprisingly that the attention module is also compatible with computer vision tasks. Different from word embedding, input images is patched into 16*16 patches as representation of a input sequence. Vision transformer only uses the encoder part of the original transformer for feature extraction and output through a Multi layer perceptrons to output the classification result.

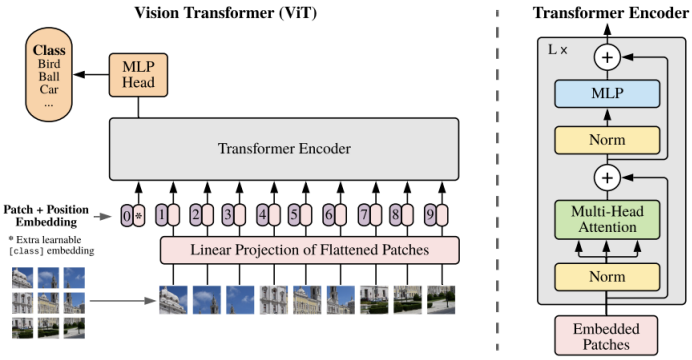


FIGURE 2.6: Vision Transformer Visualisation

In comparison with the traditional CNN methods, vision transformer is generally lack of inductive bias, which could be an advantage with sufficient training and data augmentation provided and as a result, vision transformer outperforms multiple tradition CNN models in classification in many popular data sets.

From those studies, transformer has shown its great potential in dealing with both time-series input and image input which makes it a popular backbone in many other models. Relate back to the content of this thesis, transformer's advantages is desired in crop yield forecasting using temporal remote sensing images which should have great potential but there is lack of studies related to this topic. Therefore this part of literature review focus on similar type of tasks such as action recognition and ideas could be shared and related to yield prediction.

Followed the great succeed of vision transformer, more studies was conducted to extend its performance. Chen et al[9] has came up with the idea of multi-scaled vision transformer. The critical contribution of his research can be concluded into two points, first, to patch the input image in two different ways, large ones and small ones where both interactions between small pieces and large chunks is considered. Second, the introduced a efficient cross attention module that allows information exchange between the large patch stream and small patch stream.

Figures2.8 shows the efficient information exchange between large scale tokens and small scale tokens. The class token of the large stream interaction with the patch tokens of the small stream. They also have performed projections to match the dimension of tokens in two streams.

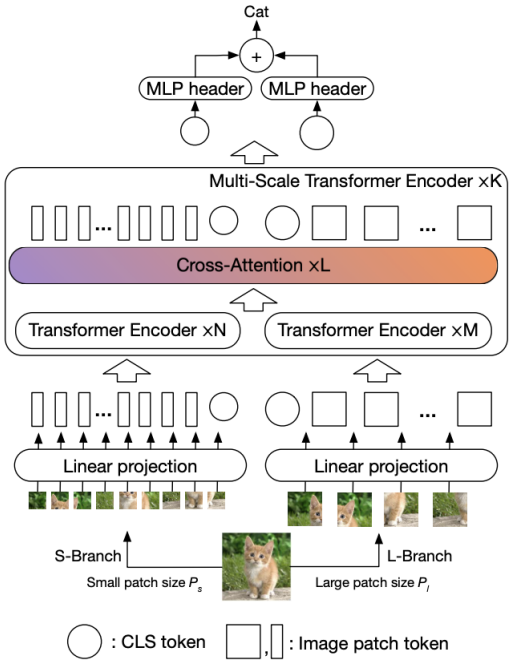


FIGURE 2.7: CrossViT architecture

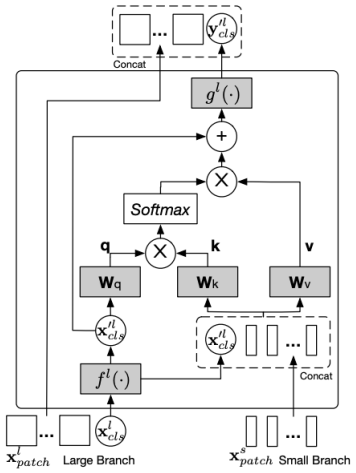


FIGURE 2.8: CrossViT Cross Attention module Example For Large Branch

Video feature extraction can be quite different from the single image extraction since both temporal and spatial dimensions needed to be considered. Anurag et al(2021)[6] has further extend Vision transformer to adopt video feature extraction tasks which is video vision transformer. The critical contribution of

this research is that video input is first patched with desired frame length, then first go through a spatial transformer encoder, and only the class token is then proceeded to a temporal transformer encoder as the figure shows2.9. Such process significantly reduced the computational resource by seprating the attention modules into two stages, since increasing number of tokens in a self-attention module is exponential. Video vision transformer was experimented on multiple data sets and achieved state-of-the-art performance in multiple video classification benchmarks.

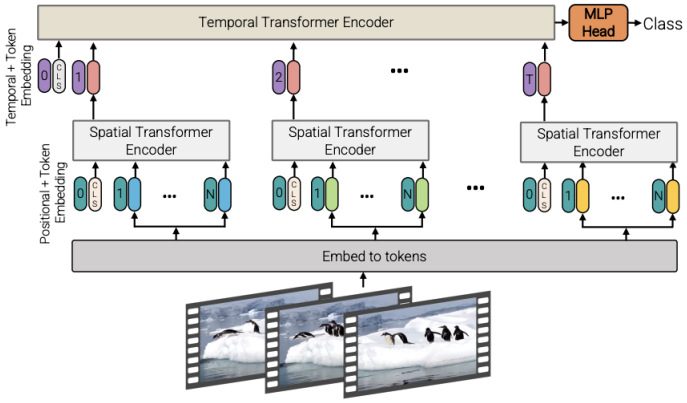


FIGURE 2.9: Video Vision Transformer

Video vision transformer is further extended by Shen et al(2022)[27]. Their research team has came up with multi-view transformer that is specialised in video recognition. The idea shares some levels of similarity with crossViT and since it is focusing on video recognition, multiple frame length of a patch is chosen as the different temporal scale. Multi-view transformer also performs cross attention between streams, but only update longer patches from the smaller ones, which is unlike the information exchange between large patches and small patches in crossViT. Figure gives a visualised MVT architecture2.10. This study still maintain the best performance in action recognition task which demonstrated its superior ability in catching the temporal relation shapes between images.

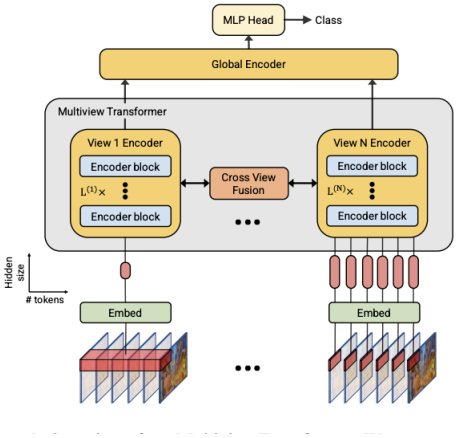


FIGURE 2.10: Multiview Transformers for Video Recognition

CHAPTER 3

Methods

3.1 Overview of Pixel-wise prediction Method

Pixel-wise prediction is performed by first spatially connect yield data, align with satellite image input. Then a overlapping patch method is performed to create a window for each data point for prediction with a specific type of deep learning algorithm. Figure3.1 visualise the overall framework to perform pixel-wised prediction. The motivation behind this architecture can be concluded to two major points. First, window centred at data point enable interaction of the data point with surrounding environment. Second, since we were limited by the size of dataset, this framework can maximise the use of every data point, create enough training set that enable data-hunger transformer architecture to learn.

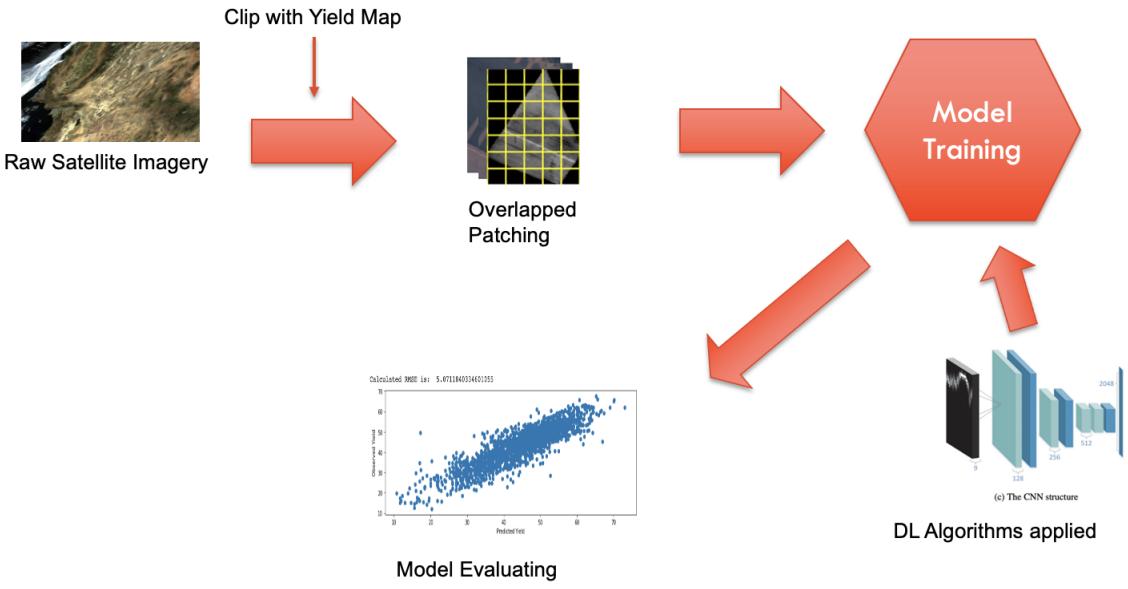


FIGURE 3.1: Overall Frame work of pixel-wised Yield Prediction

sectionMulti-temporal Multi-spatial Scale Transformer Inspired by multi view transformer[27] and crossVit[9], we have proposed a novel video vision transformer architecture called Multi-temporal Multi-spatial Scale Transformer. Both article have shown higher performance could be achieved with different input scale. As we deal with sequence image as the input of the model, advantage on extra information through multiple scale could be both acquired in time and spatial dimension.Also, it worth noting that regular image or video classification task are conducted over image size of 224*224[22], multi scale patching over multiple divisions could exhaust computational resources due increasing number in embedded tokens. On the other hand window patch lead to a much smaller input size and significant reduced.

Figure3.2 shows the architecture of proposed MTMSST model. MTMSST uses four different patch methods to tokenize the original image input into small short, large short, small long and large long patches to be embedded as token. Each set of token is input into a full video vision transformer layer for feature extraction and export series of tokens representing the features of current patches. Then cross attention modules is used for the first stage information exchange between patches of the same temporal resolution. The exchanged tokens are then aggregated as a full representation of features with short and long time resolution. The aggregated tokens are input into a second stage cross attention module, to exchange information on different temporal resolution. In the end, all the tokens together as a complete feature representation of image inputs and insert into a final global transformer encoder, output the class token and MLP layer to generate the final prediction result.

3.2 Weight Enhanced MSE Loss

Mean squared error(MSE) loss is the most commonly used loss function for regression tasks in deep learning area however it can be limited in some cases such as unbalanced data.

Focal loss is a special designed loss function that solves the problem related to unbalanced data in binary classification.[18] The idea of focal loss is to lower the weight of samples that are easy to learn and increase those that hard to learn.[7]

As the figure shows3.3, yield values generally follows a bell shaped normal distribution. To achieve better RMSE which is one of the major evaluation metric of this study, to better learn the data points that locate at two sides of the distribution can be critical, especially when there are outlier data in test set. This is the motivation of weight enhanced MSE loss.

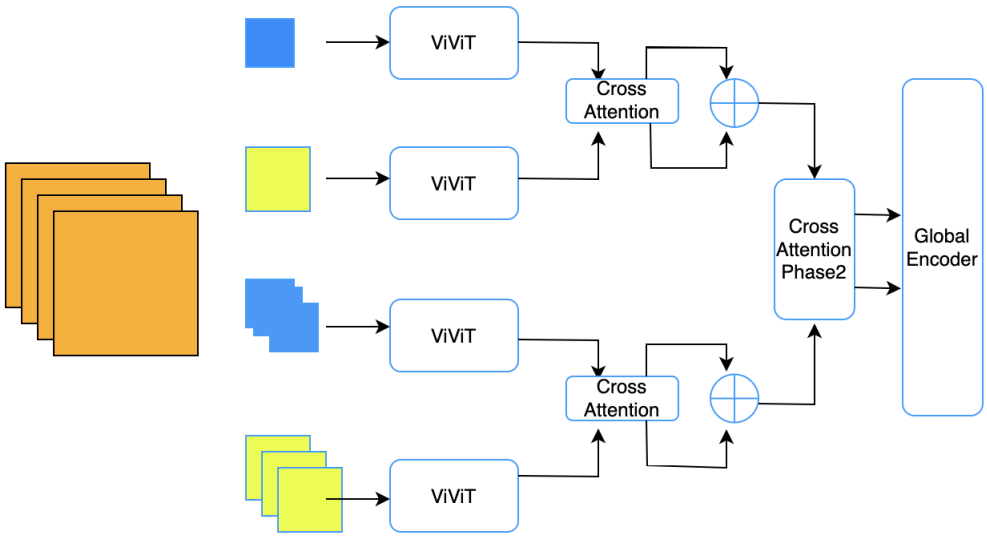


FIGURE 3.2: Multi-temporal Multi-Spatial scale Transformer architecture

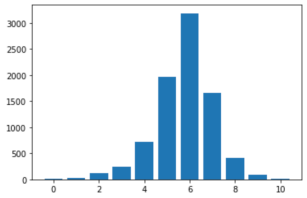


FIGURE 3.3: Yield Value Distribution

The traditional MSE loss is as follows.

$$\sum_{i=1}^{D}(x_i - y_i)^2$$

The proposed Weight Enhanced MSE loss is as the equation here.

$$\sum_{i=1}^{D}(1 - \alpha * p_y^i)(x_i - y_i)^2$$

The major difference is that the squared error is modified by the probability in the bell shaped distribution where more frequent showing values are considered as easy to prediction hence their weight is reduced. The reducing affect is also regulated by a factor alpha, where it supposed to lie between zero to one. Larger the alpha value is, the stronger regulation affect their would be.

CHAPTER 4

Experiments and Results

4.1 Task Definition

The yield prediction task in this thesis is to prediction the yearly cotton yield over an area of 10m using sentinel-2 satellite image data. A window of 21m*21m is cropped, the target 10m area located at the centre of the window and 10 images with the lowest cloud coverage over a 20 days cycle across the growing season of cotton was used as the input to predict the yield.

4.2 Dataset and Data Processing

Yield dataset is provided by our collaborator from Sydney Institute of Agriculture. Dataset is the high resolution yield mask of a farmland name "Llara" located at norther New South Wales. The spatial resolution of the yield mask is about 5m and the yield is cotton harvested in 2018. Figure4.1 is an visualized example of the dataset. As for data security reason, all yield unit is normalized on a scale from zero to one as a representation of the actual amount of yield. Yield data is provided in geo-raster format which comtains the geological information as well as the spatial resolution.

Sentinel-2 data used for crop yield prediction of Llara dataset is collected using the google earth engine platform. Although sentinel-2 has a relatively short cycle of five to ten days between each image, there is possibilities that cloud coverage affect the image capturing. Hence, to ensure that there is a valid satellite image to be collected each cycle with a cloud coverage below 20 percent, the collection cycle is extended to 20 days. Australia cotton growing season starts at October from previous year and harvest at May, hence data collected from the end of October 2017 until the end of April which is prior to the harvest season. Figure4.2 is an example of visualized image collected on 19 April 2018. The farm land area is then manually cropped to match the shape of yield mask. There are 13 bands(channels) in Sentinel-2b dataset, how ever they do not have uniform spatial resolution, hence manual selections were

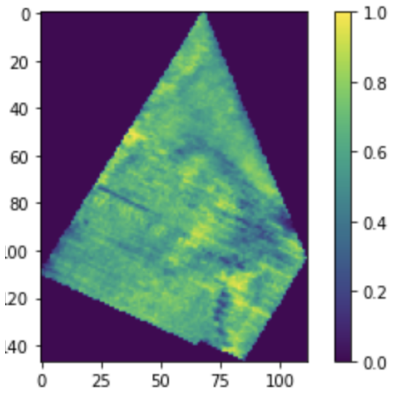


FIGURE 4.1: Yield Mask visualization of Llara dataset

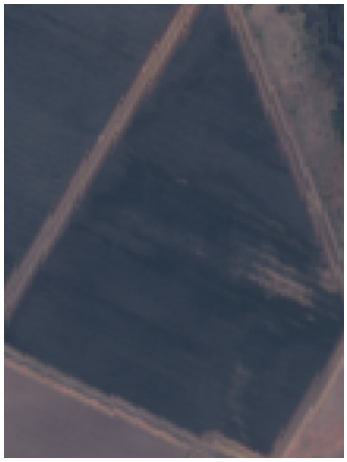


FIGURE 4.2: Llara satellite image Visualization

applied to remove band 1 Coastal aerosol band 9 water vapour and band 10 SWIR - Cirrus which all has spatial resolution of 60m and is not useful in the prediction task. The 10 remaining bands have 4 with spatial resolution of 10m 6 bands with 20m.

The next step is to clip yield data and sentinel-2 image data and so that they can spatially match. Since the highest resolution of sentinel-2 data is 10m, all bands have a uniform scale in raster format. The first thing is to adjust the spatial resolution of the yield map. Fortunately, "rasterio" a geo-raster tool in python provides function of matching and stacking geo-rasters data.[1] Such process generates a data cube as it was introduced in literature, multiple satellite images is stacked over yield mask over the time axis in the order of date.

To create window for pixel-wised yield prediction, overlapping pathches are used. For each data point in yield mask, an area of +-10 along x and y axises is selected to create a window for the data point on

data cube. The window is cropped from each satellite image to generate the final input. The patching processes is show in Figure4.3.

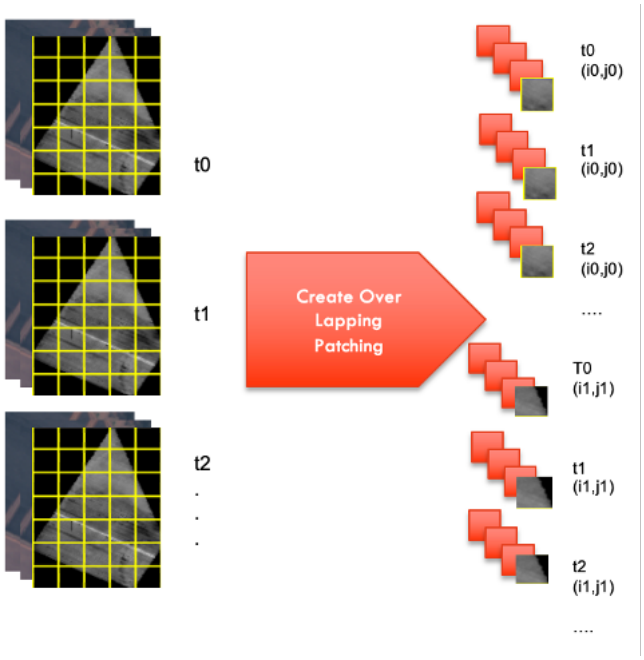


FIGURE 4.3: Data Patching

As patches are formed according to the central point, an area of 20m*20m is selected as the testing set which gives us 400 test set data. The rest are training data. Patches that overlaps with the test set data is also removed from the training set. In the end, 500 data were randomly spited as the validation set, left with around 4900 data points as the training set.

4.3 Evaluation Metrics

The research task of generating prediction of a data point given an area is very similar to a regression task. Typical evaluation metrics includes mean absolute percentage error(MAPE), root mean squared error(RMSE) and R^2 .

MAPE is calculated by the following equation.

$$MAPE = \frac{1}{n}\sum|\frac{(y_i - x_i)}{y_i}|$$

Advantage of using MAPE is that any unit in the prediction task is converted to percentage which enable certain level of comparison is made between different prediction systems.

RMSE is calculated by the following equation.

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$$

Root mean square error is a very commonly used evaluation metric in regression tasks, comapring with MAPE, though it does not convert to uniform percentage units, it punishes predictions that further to the true label harder than MAPE, where large errors are particularly undesirable is yield prediction task making RMSE a important metric in evaluating prediction results.

R^2 is calculated by the following equation.

$$R^2 = 1 - \frac{\sum_(y_i - x_i)^2}{\sum_(y_i - \hat{y})^2}$$

Statistically, R^2 tells us variation of a dependent variable is explained by the independent variable(s) in a regression model.[14] Usually used to describe how well does data fit the model.

4.4 Implementation

There are five different deep learning algorithms that was implemented in this thesis. Mario Net, the simple CNN architecture neural network that was used by fajardo(2021)[12] in yield prediction thorugh a window. 3D ResNet50, implemented reference to kenshohara's implementation. [16] Vision transformer used in the experiment is pre-trained with image net and downloaded through vit pytorch library. Implementation of video vision transformer is referenced to rishikksh20's implementation[21]. All experiments conducted using a batch size of 50, Adam as the optimizer with constant learning rate 0.001.

All codes implemented using pytorch modules. Implementation of weight enhanced loss function is done through extending an torch nn Module.

4.5 Quantitative Comparison of MTMSST

Quantitative Comparision			
Model	RMSE↓	MAPE↓	R Squared↑
Mario Net	0.0790	0.0971	0.1278
3D ResNet 50	0.0845	0.1080	0.0026
ViT	0.0871	0.0871	0.0871
ViViT	0.0741	0.0936	0.2344
MTMSST	**0.0618**	**0.0754**	**0.4671**

From the table4.5 presented, it is obvious that MTMSST model has out performed all other models in the experiment. It has the lowest RMSE and lower MAPE which showed its ability in dealing both minimising the large and small errors. On the other hand, even with highest R^2 value, the value of 0.46 is still considered as low, which means that data is relatively hard to fit in the model. However, the performance of the proposed model is still outstanding since it bested the ViViT model by reducing the RMSE value over 16 percent.

It was surprising to find out that even with the simplest structure, Mario Net is able to achieve a fair good result, only slightly worse than the complicated video vision transformer structure.

Performance of 3D ResNet is not good as expected since multiple studies have achieved good result using 3D CNN.

4.6 Prediction Result Visualisation

The numeric out could may not be intuitive and hence their prediction heatmap4.4 and in-field visualization4.5 could help in understanding how the predictions varies. The yellow color of heat map indicates the number of points locating at the same spot. x and y axis are the predicted yield and true yield respectively. Data points landing on the black line indicates a accurate prediction. The colour of prediction visualization shows whether it is over estimated(blue) or under estimated(red) or accurate predicted(white). X and y axis are the true geo location of the data point in the field.

From the heat map. we can see that Mario Net a) has many accurately predicted points alone the regression line, however its ablity is limited in handling with outlier data points which may be the main

reason its performance is not too well even the shape of distribution looks alright. From the in-field visualisation, it is more obvious that there exists large chaunk of blue which indicates over estimation.

3D CNN and Vit on the other hand had very poor shape alone the axis indicating a very weak correlation between data and model. This could also be reflected in the field visualisation picture where large chunk of blue locate in the middle and top left corner of 3D CNN.

Video vision transformer has better shape where data points lies around the black line but not as good as MTMSST. The most obvious observation is that data points lies very tight with the black line, and in the visualization image, it can be observed that MTMSST is able to accurately predict data points locating on the top left corner where all other models are struggle to give accurate prediction, showing a blue color.

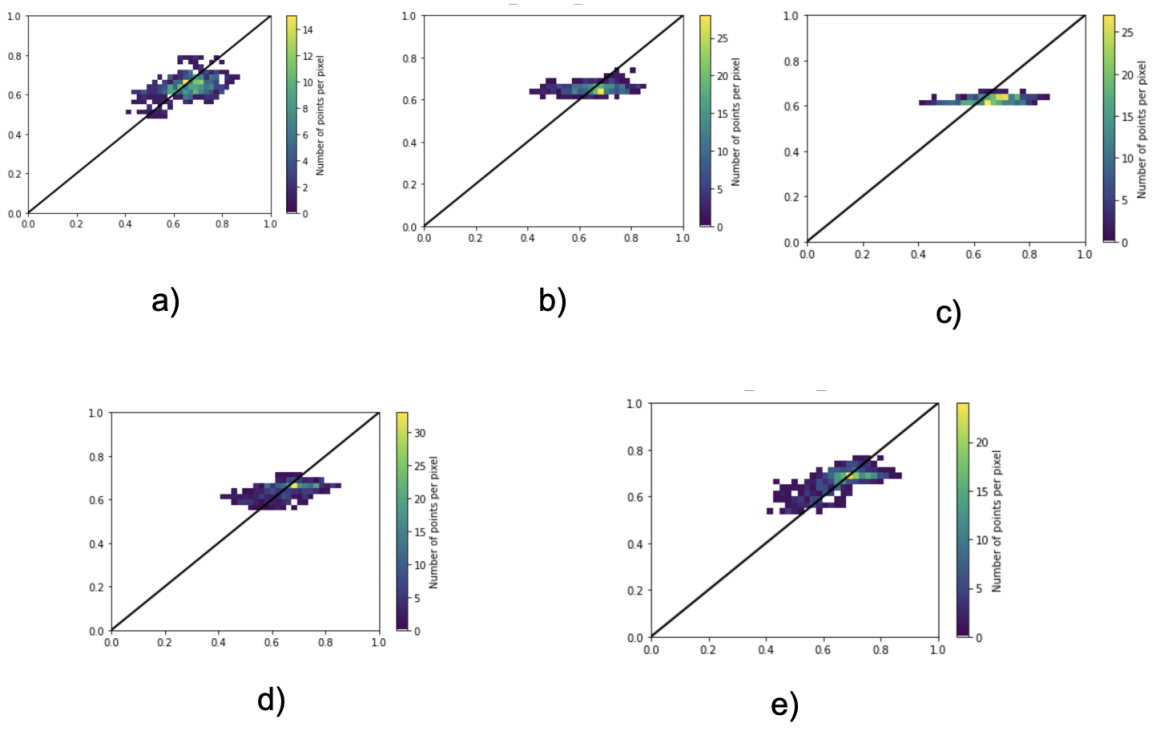


FIGURE 4.4: The prediction heatmap of a) Mario Net b)3D ResNet c)Vision Trasnformer d)Video Vision Transformer e)MTMSST

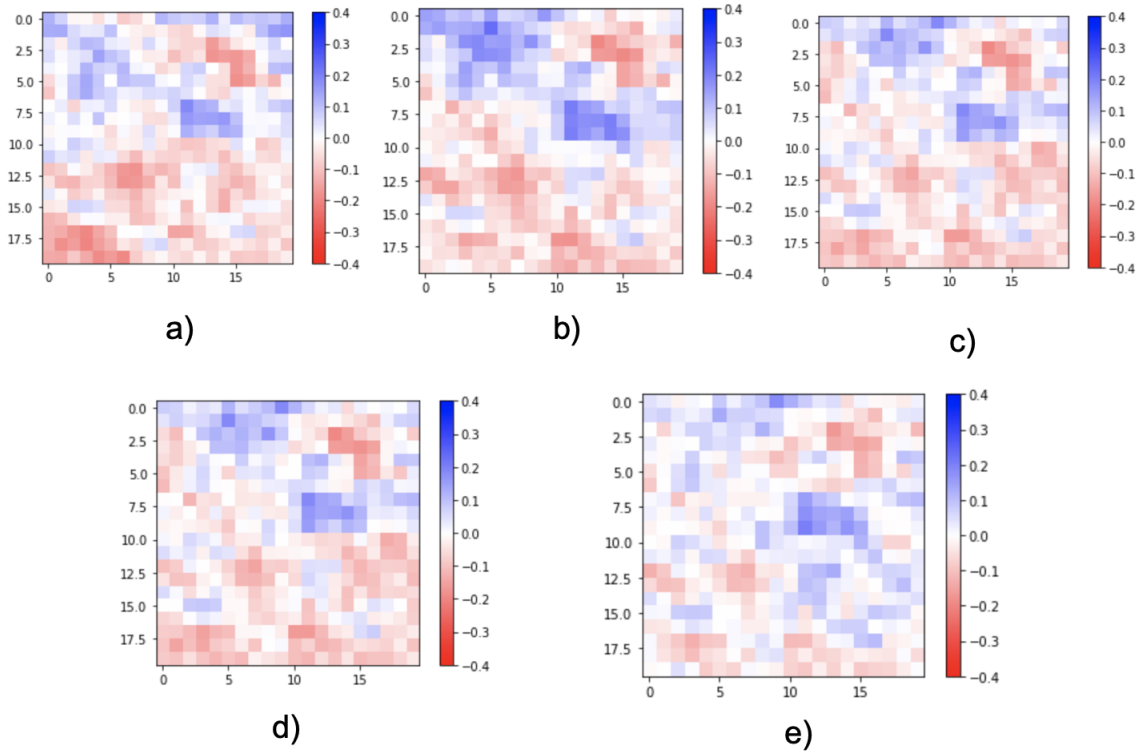


FIGURE 4.5: The in-field prediction visualization of a) Mario Net b)3D ResNet c)Vision Trasnformer d)Video Vision Transformer e)MTMSST

4.7 Experiments on Weight Enhanced MSE Loss

Weight Enhanced MSE Loss on MTMSST			
Model	RMSE↓	MAPE↓	R Squared↑
$\alpha = 0.2$	0.0798	0.0925	0.0399
$\alpha = 0.3$	0.06419	0.0813	0.4252
$\alpha = 0.4$	0.0810	0.0944	0.0385
Standard MSE	**0.0618**	**0.0754**	**0.4671**

Table above has shown the experiment result of using weight enhanced MSE Loss with MTMSST. At alpha equal to 0.3, the regulation have performed the best, but still slightly worse than the standard MSE loss function which indicate that Weight Enhanced MSE Loss could not help the model to achieve a better prediction.

4.8 Experiment on data prior to harvest season

In season Prediction with MTMSST			
No. Images Used	RMSE↓	MAPE↓	R Squared↑
6	0.0793	0.1325	-0.0126
8	0.0631	0.0790	0.4447
10(Full)	**0.0618**	**0.0754**	**0.4671**

The ultimate goal to achieve in yield estimation is that it could be accurately performed within the growing season. Therefore I have gradually reduced the number of satellite images used in estimation. Experiment has shown that the model could still provide accurate prediction discarding the last two images, however if the images reduced to 6, the perfomance has dropped significantly. This indicates that the last four sentinel-2 images are critical in MTMSST model. This also indicates, yet it still cannot make accurate in-season prediction.

4.9 Ablation Studies

Ablation Studies			
No. Images Used	RMSE↓	MAPE↓	R Squared↑
ViViT with small patch only	0.0741	0.0936	0.2344
ViVit with large patch only	0.0753	0.0953	0.2090
MTMSST w/o Cross attention	0.0626	0.0801	0.426
FULL MTMSST	**0.0618**	**0.0754**	**0.4671**

In the end, ablatiuon studies was conducted on MTMSST where the experiment indicates that simply with samll patch or large patch would only have sound prediction result. However, if patches are all used, even without the cross attention module, the performance increased quite significant. With the use of cross attention module, even better perforamnce could be achieved indicates the impact of information exchange during the feature extraction process.

CHAPTER 5

Disscussion

The lack of publicly accessible dataset on pixel-wised yield making it extremely difficult to have proper comparison with the existed prediction methods. The discussion will be focusing on the comparison of experiments conducted on the given dataset.

5.1 MTMSST

The result of quantitative research has shown that MTMSST have outperforms all other experimented methods significantly. It is worth noting that even with extra token as the input to the model, the training process was still quite efficient. An 100 epoch train only takes around 30 minutes which is no significantly increase comparing with other methods except the simplest Mario Net. This might be the parallelism advantages of transformer models. Vision transformer has the longest training time around 40 minutes per 100 epoch followed by 3D CNN model which both had very poor performance. This indicates that ether two models have limited performance in the given task, or some mistakes were made during the implementation. Further studies are required to find out.

All the experiments were conducted using google colab pro with RAM of 32G and GPU Memory of 16G on P100 model normally. The training process is generally efficient due to the small dataset and image input.

5.2 Weight Enhanced MSE Loss

Based on the experiment result, weight enhanced MSE loss did not have positive affect in the training result, there could be two main reasons that caused this. First, the assumption made was wrong, some more frequently showed values may not be easy to predict. Second, the selection of test set did not include many outlier data points, most of the are clumped with-in certain range in the middle where

enhance by weight does not help achieving a better prediction or even have opposite affect that may lead to easily over-fitting.

5.3 Within Season Prediction

The experiment has shown a very poor result when four images were removed from the dataset. However I believe that this is mainly due to the lack of data. If the model could be trained on the same crop wit past year data, an in-season prediction might be achievable with MTMSST.

5.4 Limitation of MTMSST

The studies carried out in this thesis could be extended in many perspectives. First, is to expand the dataset to multiple farm lands over several years. Then the model could be test against the geological transfer and time transfer or even transfer studies on different crops[25], where full potential of MTMSST could be tested. Secondly, more advanced components could be added to the current architecture such as a learn-able token selection mechanism[26] and relative positional encoding[17] to further improve the performance of MTMSST.

CHAPTER 6

Conclusion

Precision agriculture leads the way to future human society and fine scaled yield prediction will be more critical for correct decisions to be made in agriculture. Also deep learning predictions will be more powerful and reliable with the advancement of computational technologies.

In this thesis, we have suggested and made automation for a efficient and accurate prediction through remote sensing images. We have tested the compatibility of recent the recent development transformer models. We proposed Multi-temporal Multi-spatial Scale Transformer(MTMSST) that utilise the advantages of taking multiple view on spatial scale and on temporal scale.

Our experiment result on MTMSST has show great potential in fine scaled crop yield prediction task that is able to capture information that conventional methods could not. However, limitation still exists in the current study such as it is weak in in-season prediction. Further studies could be carried on to achieve between prediciton results.

Bibliography

[1] Access to geospatial raster data.

[2] Gearing up for third sentinel-2 satellite. https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-2/Gearing_up_for_third_Sentinel-2_satellite. Accessed: 2022-09-30.

[3] Ispa precision agriculture definition. https://ispag.org/. Accessed: 2022-09-30.

[4] Pathways to zero hunger: Un global compact.

[5] Putting a price on farm data, Jul 2019.

[6] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid. Vivit: A video vision transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[7] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[8] D. Cervantes-Godoy and J. Dewbre. Economic importance of agriculture for poverty reduction. (23), 2010.

[9] C.-F. R. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[10] P. Doraiswamy, J. Hatfield, B. A. T.J. Jackson, J. Prueger, and A. Stern. Crop condition and yield simulations using landsat and modis. In *Remote Sens. Environ.*, pages 548–559, 2004.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] M. Fajardo and B. M. Whelan. Within-farm wheat yield forecasting incorporating off-farm information. *Precision Agriculture*, 22(2):569–585, 2021.

[13] R. Fernandez-Beltran, T. Baidar, J. Kang, and F. Pla. Rice-yield prediction with multi-temporal sentinel-2 data and 3d cnn: A case study in nepal. *Remote Sensing*, 13(7):1391, 2021.

[14] J. Fernando. R-squared formula, regression, and interpretations, Nov 2022.

[15] P. Filippi, E. J. Jones, N. S. Wimalathunge, P. D. Somarathna, L. E. Pozza, S. U. Ugbaje, T. G. Jephcott, S. E. Paterson, B. M. Whelan, T. F. Bishop, and et al. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20(5):1015–1029, 2019.

[16] kenshohara. 3d-resnets-pytorch. https://github.com/kenshohara/3D-ResNets-PyTorch, 2018.

[17] J. Lin and S.-h. Zhong. Bi-directional self-attention with relative positional encoding for video summarization. *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2020.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[19] P. Muruganantham, S. Wibowo, S. Grandhi, N. H. Samrat, and N. Islam. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9):1990, 2022.

[20] B. Pejak, P. Lugonja, A. Antić, M. Panić, M. Pandžić, E. Alexakis, P. Mavrepis, N. Zhou, O. Marko, V. Crnojević, and et al. Soya yield prediction on a within-field scale using machine learning models trained on sentinel-2 and soil data. *Remote Sensing*, 14(9):2256, 2022.

[21] rishikksh20. Vivit-pytorch. `https://github.com/rishikksh20/ViViT-pytorch`, 2121.

[22] C. F. Sabottke and B. M. Spieler. The effect of image resolution on deep learning in radiography. *Radiology: Artificial Intelligence*, 2(1), 2020.

[23] V. Sagan, M. Maimaitijiang, S. Bhadra, M. Maimaitiyiming, D. R. Brown, P. Sidike, and F. B. Fritschi. Field-scale crop yield prediction using multi-temporal worldview-3 and planetscope satellite data and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:265–281, 2021.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] A. X. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon. Deep transfer learning for crop yield prediction with remote sensing data. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, 2018.

[26] J. Wang, X. Yang, H. Li, L. Liu, Z. Wu, and Y.-G. Jiang. Efficient video transformers withnbsp;spatial-temporal token selection. *Lecture Notes in Computer Science*, page 69–86, 2022.

[27] S. Yan, X. Xiong, A. Arnab, Z. Lu, M. Zhang, C. Sun, and C. Schmid. Multiview transformers for video recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[28] J. You, X. Li, M. Low, D. Lobell, and S. Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017.