Channel Attention-Based Temporal Convolutional Network for Satellite Image Time Series Classification

Pengfei Tang, Peijun Du, *Senior Member, IEEE*, Junshi Xia, *Senior Member, IEEE*, Peng Zhang, and Wei Zhang

***Abstract*— Satellite image time series classification has become a research focus with the launch of new remote sensing sensors capable of capturing images with high spatial, spectral, and temporal resolutions. In particular, in the field of crop classification, time dimension information is particularly important. Although some advanced machine learning algorithms, such as random forests (RFs), can achieve good results, they often ignore the time series information. To make full use of temporal and spectral information in multitemporal remote sensing images, a channel attention-based temporal convolutional network (CA-TCN) is proposed in this letter. Specifically, the proposed method is composed of two main modules: temporal convolutional network and attention block. The temporal convolutional network can capture long-range dependence by using a hierarchy of temporal convolutional filters. To capture relevant information inside the sequence and enhance the important information, the attention block is used to enhance the important features in the channel dimension since not all bands contain equal information in crop type classification. The proposed CA-TCN can excavate deeper phenological characteristics. Compared to the temporal attention-based temporal convolutional network and other deep learning-based models, the proposed CA-TCN has achieved state-of-the-art performance in the Breizhcrops dataset with fewer parameters.**

***Index Terms*— Crop type mapping, self-attention, temporal convolutional networks (TCNs), time series classification.**

I. Introduction

IN RECENT years, with the launch of several high-spatial-resolution remote sensing satellites, such as Sentinel-2A/B and Landsat-8, free optical data opens up an unprecedented opportunity for large-scale, long-term environmental and agricultural monitoring [1]. Crop classification, as one of the most important components in agricultural environmental monitoring, has attracted extensive attention [2]. Previous studies have shown that adding time dimension information can effectively improve classification accuracy. High temporal resolution data are capable of capturing specific crop growth stages, which is very useful to distinguish different crop types [2]. However, the data abundance also presents significant challenges to machine learning models. In the era of Big Data, new requirements are put forward for the selection of classification algorithms.

Machine learning algorithms [e.g., support vector machines (SVMs) [3] and random forests (RFs)] perform well in crop type classification [4]. However, these algorithms only treat the time dimension information as features and ignore the sequential structure in the time dimension. In this case, the temporal sequence of image presentation cannot be used to improve the results. Without exploring the order and interdependence of the time dimensions, it may cause a loss of temporal information for classes. There are some solutions to alleviate the problem by generating temporal sequence features using some preprocessing [5] but it is often time-consuming and ineffective [6].

Deep learning-based algorithms, which have been proved superior to traditional algorithms, have shown unprecedented advantages in time series modeling. On satellite image time series classification tasks, recurrent neural networks (RNNs) are the dominant approaches [7]. However, the RNNs have some inherent drawbacks, including the “gradient disappearance and explosion” and the inability to process in parallel. Recently, some new architectures have been proposed to replace the RNNs, which mainly includes attention mechanism [8] and convolutional neural networks (CNNs) [9].

The attention mechanism is a solution that selectively focuses on some related things while ignoring others. Simply, it is to quickly screen out high-value information from a large amount of information. It has been widely used in building extraction [10] and image inpainting [11]. The transformer [8], a self-attention-based network, initially designed for text translation, has been successfully used to crop classification tasks [1]. Then, some modifications on transformer have been proposed to adapt the characteristic of remote sensing data,

Manuscript received June 4, 2021; accepted July 5, 2021. Date of publication July 19, 2021; date of current version January 6, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41631176 and in part by the KAKENHI under Grant 19K20309. *(Corresponding author: Peijun Du.)*

Pengfei Tang, Peijun Du, Peng Zhang, and Wei Zhang are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science and the Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing University, Nanjing, Jiangsu 210023, China (e-mail: sgos_tpf@smail.nju.edu.cn; dupjrs@gmail.com; pzhangrs@smail.nju.edu.cn; zhangwrs@163.com).

Junshi Xia is with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo 103-0027, Japan (e-mail: junshi.xia@riken.jp).

Digital Object Identifier 10.1109/LGRS.2021.3095505

which achieved higher classification performance [12], [13]. However, these self-attention-based methods are always limited to transformer architecture. Too much attention is paid to the temporal features, but the importance of band information is ignored.

CNNs always play a role as feature extractors mostly for spatial or spectral domains that are widely used in remote sensing studies [14]. Recently, some studies have been proposed to address the temporal dimension of remotely sensed time series with 1-D convolutional layers [2], [6]. Although the temporal convolutions yield better results than RNNs [2], the ability to account for long-term dependencies requires deeper architecture. In addition, due to the fixed architecture of CNNs, the same network is unsuitable for sequences of different lengths.

The temporal convolutional network (TCN) [9] is a new CNN-based sequence analysis model. By using the dilated-causal convolutions, the TCN's architecture can expand the receptive field and process sequences of arbitrary length [9] compared to other CNNs architecture. Although the TCNs have been widely used in time series tasks, such as speech separation [15], and other areas [9], there are few studies [16] on exploring the effectiveness of TCNs on satellite image time series classification tasks. In addition, there are many modifications on TCNs [17], [18]; most of them proved that, by adding a self-attention module on the time dimension, the TCNs can extract the internal dependence information of the input. However, there are few studies studying the influence of channel attention on TCNs.

Considering the problems above mentioned, this letter proposed a channel attention-based temporal convolutional network (CA-TCN) for temporal feature representation in crop classification. Compared to the temporal attention-based temporal convolutional network (TA-TCN) and other deep learning-based models, the CA-TCN has achieved state-of-the-art performance with fewer parameters in the Breizhcrops dataset.

II. Methodology

A. Overview

The overall architecture of CA-TCN is shown in Fig. 1. It consists of three main blocks: TCN block, channel attention block, and fully connected (FC) layers.

B. Temporal Convolutional Networks

The architecture of TCNs used in this letter is referred to the Conv-TasNet [18], for the great success in speech separation.

Unlike traditional CNNs, TCNs use dilated-causal convolution with residual connections. Within the residual block, TCNs have two layers of dilated-causal convolution and the rectified linear unit (ReLU). On the one hand, the causal convolution is a strict time-constrained model that the model has no access to future information. On the other hand, by using dilated convolution with residual connections, the model enables a large efficiently receptive field that can get long-time dependence. The illustration of dilated-causal convolutions is shown in Fig. 1(c).

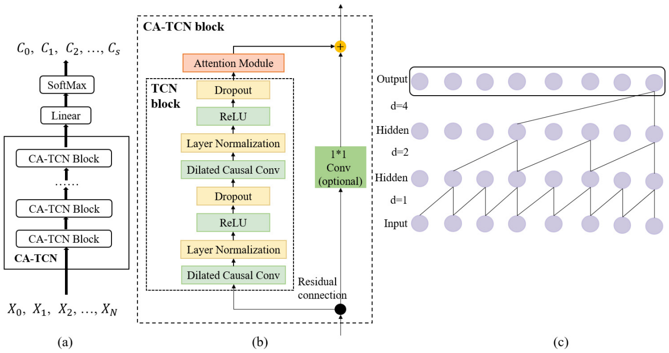

Fig. 1. (a) Overall architecture, where the X_i is the input time series data, C_i is the classes, N is the sequence length, and S is the class number. (b) Diagram of the CA-TCN module, and a TCN block is inside. (c) Overall dilated-causal convolutional network with dilation factors $d = 1, 2, 4$, and filter size $k = 2$.

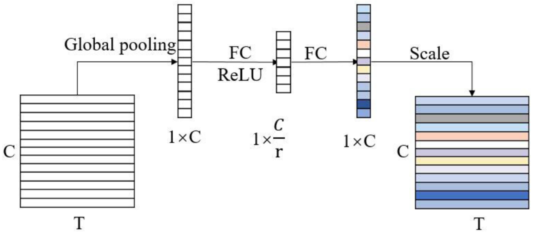

Fig. 2. Diagram of the SENet module. The different colors indicate different weights. C and T mean channel and time dimension, respectively. r means the reduction index.

In each residual block, two dilated-causal convolutional layers are stacked with ReLU, which is utilized as the activation function. In addition, spatial dropout layers are added to avoid overfitting. The 1 × 1 convolutional layer in the residual connections is used to ensure that the number of channels of the TCNs block input is the same as output. Depthwise separable convolutions have been shown to improve representation efficiency while being computationally and memory cheap [18] and are, therefore, used in this letter to replace the standard convolution of each convolution block. The normalization operation used in this letter is a cumulative layer normalization (cLN) [18], which is specially designed for causal configuration.

C. Channel Attention Module

As we all know, not all bands or features contain equal information. Irrelevant features' information often brings unnecessary noises. On the other hand, the crops have specific channel-temporal patterns. However, the generic TCNs think that different bands contribute equally. In view of this, a channel attention module has been added to TCNs for extracting the contribution degree of different bands.

For channel dimension, a lightweight module, squeeze-and-excitation network (SENet) [19] was used in this letter. The SENet can automatically obtain the importance of each feature channel and, then, according to this importance, enhance the useful features and suppress the features that are not of much use to the current task.

Fig. 2 illustrates the structure of SENet that is used in CA-TCN. For time series data, $z = (z_1, z_2, \ldots, z_T)$, whose dimension is $T * C$ (T means the time, and C means the channel). First, the features have been compressed on time dimension by average pooling and turn all time-dimensional feature channel into a $1 \times C$ vector. The vector has a global receptive field. It represents the global distribution of responses on the feature channel. In order to adjust the weight of each channel feature, excitation operation is applied in the networks. Here, a simple activation gate is added to control the calculation of the entire weight sequence

$$s = F_{\mathrm{ex}}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \mathrm{ReLU}(W_1 z)) \quad (1)$$

where $W_1 \in R^{(C/r) \times C}$, $W_2 \in R^{C \times (C/r)}$, σ refers to the sigmoid function, ReLU refers to the ReLU function, and r is the scaling parameters. Finally, the readjusted weight sequence s has returned to the original input z to get readjusted time series data $\tilde{z}$

$$\tilde{z} = F\mathrm{scale}(z_c, s_c) = s_c \cdot z_c. \quad (2)$$

The whole excitation operation can be understood as an encoding–decoding process, where the $1 \times C$ weight sequence is squeezed to $1 \times (C/r)$ and then excites it to the $1 \times C$ weight sequence. In this process, the weight coefficient of each channel was learned so that the model could distinguish the features of each channel better.

D. Advantages of the Proposed Network

1) Parallelism: In RNNs, predictions of future time steps must wait for their predecessors to complete, whereas convolutions and attention architecture can be completed in parallel. Therefore, in both training and evaluation, CA-TCN can process a long input sequence as a whole, rather than sequentially as in RNNs [9].

2) Attention: In generic TCNs, it is thought the different bands have the same contribution to classification, while that is contrary to our experience. By adding a channel attention module, the CA-TCN can focus on important bands, which can improve the classification performance.

3) Hybrid Configuration: The CA-TCN is a hybrid configuration that combines CNNs and attention. As [12] finds, in crop mapping tasks, the hybrid model always acquires the best performance.

III. Experimental Setup

A. Datasets

The CA-TCN has been tested in Breizhcrops [20], which has been introduced very recently as a common benchmark remote sensing dataset with an emphasis on the temporal dimension. The Breizhcrops dataset contains Sentinel-2 L1C acquired over the Brittany region in 2017. The dataset covers an area of approximately 27 200 km^2 and consists of 768 175 samples, each of which represents the average clustered spectral response of a field package. In order to standardize their time length, the sequence length of Breizhcrops samples was set to 45 in the original article [20]. In this dataset, it contains 328 crop labels that are grouped into 23 groups. For testing models, 9 and 13 crop categories have been selected in the BrezhCrops dataset. In addition, Table I shows the number of packages for each crop type. In Breizhcrops datasets, the class distribution is clearly unbalanced, as a challenge to the classification models.

TABLE I

Class Frequencies Per Region on the Breizhcrops Dataset. (a) 9-Class Partition. (b) 13-Class Partition

(a)

Crop ID	Crop Type	FRH01	FRH02	FRH03	FRH04
1	barely	13046	10733	7148	5978
2	wheat	30368	15005	27189	16993
3	rapeseed	5593	2347	3557	3236
4	corn	43990	36593	41992	31333
5	sunflower	1	6	10	2
6	orchards	944	350	1223	553
7	nuts	10	18	10	11
8	permanent meadows	32650	36513	32534	26117
9	temporary meadows	52011	39084	52728	38391
	Total	178613	140649	166391	122614

(b)

Crop ID	Crop Type	FRH01	FRH02	FRH03	FRH04
1	barely	13046	10733	7148	5978
2	wheat	30368	15005	27189	16993
3	corn	43990	36593	41992	31333
4	fodder	6514	4329	7639	4541
5	fallow	1521	3267	2814	4555
6	miscellaneous	17659	12126	21194	15571
7	orchards	944	350	1223	553
8	other cereals	6276	3660	4516	5784
9	permanent meadows	32650	36512	32534	26117
10	protein crops	1107	461	1079	655
11	rapeseed	5593	2346	3557	3236
12	temporary meadows	52011	39082	52728	38391
13	vegetables or flowers	8538	14266	3679	3851
	Total	220217	178730	207292	157558

B. Compared Methods

1) Temporal Convolutional Neural Network: The temporal convolutional neural network (TempCNN) [6] is specifically designed for crop classification tasks. It is similar to TCNs that apply convolutions on the temporal domain instead of the spatial domain.

2) Temporal Attention Encoder: The temporal attention encoder (TAE) is a part of the pixel-set encoder and temporal attention encoder (PSE-TAE) [12], which has achieved the best performance in a spatial–temporal crop dataset. Since the Breizhcrops dataset only uses the temporal dimension for classification tasks, we only used the TAE for comparison.

3) Temporal Attention-Based Temporal Convolutional Networks: Many studies [17] have proved that the generic TCN does not learn the dependence of the internal distance position of the sequence, nor does it extract the internal relevant information of the input time series. Thus, by adding temporal attention to TCNs, the new architecture, temporal attention-based temporal convolutional network (TA-TCN) can achieve state-of-the-art performance. Therefore, TA-TCN has been compared with our model. The temporal attention used in TA-TCN is multihead attention [8].

4) Other Models: The RF classifier is used as a traditional algorithm for comparing deep learning methods. For deep learning, two popular sequence modeling architectures, RNNs-based multilayer bidirectional LSTM (Bi-LSTM) [21] and attention-based transformer [8], have been used for comparing to our method.

C. Implementation Details

In this experiment, some experimental setups have been followed in [20] for the sake of fairness and comparability. Therefore, samples of regions FRH01, FRH02, and FRH03 are used for training, and samples of region FRH04 are used for testing. For the model selection, as [20] sets, samples of FRH01 and FRH02 regions were used for training, and samples of region FRH03 were used for validating.

The best configuration is given as follows. CA-TCN uses a kernel size of 3 and stacking four layers with 64 hidden units. The CA-TCN has channel attention, in which the reduction is 4. The dropout rate is set to 28%, the learning rate $= 5.85 \cdot 10^{-4}$, and weight decay $= 1.26 \cdot 10^{-5}$. In order to avoid the refluence of network depth and width on the results, the TA-TCN and TCN have the same kernel size, hidden units, and layers as CA-TCN. The TA-TCN has a temporal attention model, which has one head with a key size of 256. The dropout is 20% with a learning rate of $8.35 \cdot 10^{-4}$, and the weight decay is $6.38 \cdot 10^{-6}$. The TCN has a learning rate of $9.74 \cdot 10^{-4}$ with a weight decay of $4.88 \cdot 10^{-5}$.

For RF, 500 trees are at a maximum depth of 25. For deep learning method, the [20] has provided the optimal hyperparameters for Bi-LSTM, transformer, and TempCNN via a random search. For Bi-LSTM, the [20] used four stacked bidirectional long short-term memory layers with 128 hidden dimensions. The dropout rate equals to 57%, the learning rate to $9.88 \cdot 10^{-3}$, and the weight decay to $5.26 \cdot 10^{-7}$. For transformer, it uses N stacked modules of H multiple self-attention heads and d_{model} hidden states within the self-attention vectors. Rußwurm *et al.* [20] set a configuration with $N = 3$ layers, $H = 1$ head and $d_{\text{model}} = 64$, 40% dropout, learning rate $= 1.31 \cdot 10^{-3}$, and weight decay $= 5.52 \cdot 10^{-8}$. For TempCNN, Rußwurm *et al.* [20] used 128 hidden units with a filter size of 7. The dropout rate is set to 18%, the learning rate to $2.38 \cdot 10^{-4}$, and the weight decay to $5.10 \cdot 10^{-5}$. For TAE, the hyperparameters have been adopted the same as [12].

The Adam optimizer [22] has been used in all the above models. Furthermore, all models have been implemented with the PyTorch framework and trained on a single 3090 RTX with 24 GB of memory. All deep learning methods have been stopped in 100 epochs. Overall accuracy (OA), Cohen's kappa score (K), and mean f1 score have been used to measure the classification results.

TABLE II

COMPARISON OF MODELS ON THE BREIZHCROPS DATASET ON THE 9- AND 13-CLASS LAND COVER CATEGORIZATIONS. (A) ACCURACY ASSESSMENT ON 9-CLASS. (B) ACCURACY ASSESSMENT ON 13-CLASS

(a)

	Traditional methods	CNNs-based Models		Attention-based Models		RNNs-based Models	Hybrid Models	
ID	RF	TempCNN	TCN	Transformer	TAE	Bi-LSTM	TA-TCN	CA-TCN
1	75.24	83.28	72.33	89.37	91.76	93.66	93.32	**94.01**
2	90.77	96.67	93.64	97.31	95.34	95.73	96.20	**97.59**
3	86.71	95.67	86.37	97.06	94.68	96.72	95.30	**97.58**
4	97.03	96.47	97.21	97.72	97.23	96.20	97.38	**97.78**
5	0	0	0	0	0	0	0	0
6	4.15	5.24	3.43	1.44	14.82	7.41	9.76	**16.27**
7	0	0	0	0	0	0	0	0
8	34.97	54.30	37.19	61.79	**65.56**	52.06	50.28	57.92
9	**83.67**	73.09	82.40	69.61	68.65	77.39	77.26	73.80
OA(%)	76.98	79.10	77.41	80.35	80.58	80.32	80.20	**81.20**
Kappa(%)	69.61	72.64	70.39	74.46	74.70	74.34	74.13	**75.52**
Mean F1(%)	53.23	57.43	52.91	57.43	58.76	58.10	57.76	**59.24**

(b)

	Traditional methods	CNNs-based Models		Attention-based Models		RNNs-based Models	Hybrid Models	
ID	RF	TempCNN	TCN	Transformer	TAE	Bi-LSTM	TA-TCN	CA-TCN
1	82.15	87.85	77.43	87.61	89.92	92.09	90.56	**92.44**
2	79.67	93.60	87.45	**95.16**	92.60	92.78	94.01	94.25
3	94.70	95.71	95.85	96.49	95.90	96.24	96.12	**96.51**
4	22.35	21.64	22.36	19.49	**27.22**	23.92	24.40	26.08
5	0	0.93	0.11	0.13	0.44	**3.19**	0.96	0.10
6	51.34	41.67	51.02	51.44	**58.86**	54.62	53.22	55.00
7	1.62	2.89	2.80	3.59	**8.69**	7.50	7.83	6.98
8	21.66	36.53	35.19	44.61	50.97	56.28	48.56	**57.44**
9	40.17	46.19	41.22	**61.72**	51.34	53.36	44.67	55.29
10	29.61	51.60	39.93	50.24	57.02	55.17	56.42	**60.53**
11	82.07	94.49	85.84	95.57	95.45	95.24	95.67	**96.09**
12	72.54	**71.60**	71.53	62.26	67.35	67.05	71.03	67.06
13	69.15	65.82	66.40	76.32	75.22	75.19	73.12	**77.08**
OA(%)	64.89	67.41	66.23	69.51	69.94	70.11	69.19	**70.80**
Kappa(%)	57.83	60.87	59.50	63.64	64.19	64.36	63.16	**65.19**
Mean F1(%)	50.76	55.29	53.17	58.19	59.79	60.39	58.54	**60.98**

IV. RESULTS AND ANALYSIS

A. Classification Results

Table II provides the classification results for all models. Table II shows that the RF performs as poorly as expected as it just takes the information of the time dimension as a multidimensional feature and ignores the sequence information. Both the CNN-based TempCNN and TCN are better than the RF model but also significantly lower than other deep learning models. This may be due to the difficulty of handling cloudy acquisitions in CNN-based models [20]. The difference in accuracy between the attention and RNN-based models is minimal. The modified transformer model-TAE is better than the transformer model. Compared to other models, the TAE model performs well in all categories. In the hybrid model, the TA-TCN model performs similar to the transformer model. Compared to the original TCN model, there is a significant improvement in accuracy, which shows that adding a time-dimensional attention module is very effective. The CA-TCN model significantly outperforms all models. In the crop classification task, attention on the channel dimension is more important than attention on the temporal dimension attention. In addition, as shown in Table II, the hybrid model has a clear advantage in crop classification.

B. Interpreting the Behavior of the CA-TCN

To further illustrate the role of channel attention in CA-TCN, a class activation mapping (CAM) is used to visualize the weights of the attention module in the final layer of the model. The CAM used in this letter is a guided grad-CAM [23]. All samples in the test dataset were CAM visualized by category, and the mean values of samples of the same category were taken as the result of cam visualization of samples in this category. Since the last layer of CA-TCN has 64 channels, the final results were compressed into 13 channels for the convenience of visualization.

Table III shows the final results. In the original TCN, the effect of different channels on the classification results was not taken into account, so their channel weights were all the same. However, in the CA-TCN model, all classes have different weights on different channels. Although most values are distributed between 0.4 and 0.6, classes with higher classification accuracy have very obvious differences in the weights of different channels. For example, in category 2, the highest and lowest channel weights were 0.75 and 0.47, respectively, while, in category 7, the highest and lowest channel weights were 0.57 and 0.46, respectively. These data

TABLE III

CLASS ACTIVATION WEIGHTS IN THE LAST CHANNEL ATTENTION LAYER OF CA-TCN

ID	OA	Class Activation Weights												
1	92.44	0.67	0.59	0.60	0.70	0.53	**0.44**	0.54	0.69	0.59	**0.73**	0.61	0.66	0.61
2	94.25	0.59	0.50	0.52	0.64	0.54	**0.47**	0.52	0.64	0.52	0.62	0.61	**0.75**	0.53
3	96.51	0.54	0.54	0.47	0.60	0.51	0.46	0.61	0.56	0.52	**0.64**	0.61	0.63	**0.38**
4	26.08	0.40	0.43	**0.36**	0.42	0.38	0.41	**0.56**	0.42	0.44	0.42	0.50	0.53	0.42
5	0.10	0.47	0.45	0.49	0.49	**0.41**	0.47	0.45	**0.53**	0.49	0.45	0.52	0.47	0.41
6	55.00	0.52	0.48	0.54	0.56	0.46	0.51	0.48	0.56	0.53	0.47	**0.57**	0.50	**0.43**
7	6.98	0.53	0.50	0.53	**0.57**	**0.46**	0.57	0.52	0.55	0.55	0.47	0.56	0.50	0.47
8	57.44	0.53	**0.61**	0.49	0.56	0.51	0.46	0.49	0.57	0.48	0.53	0.57	0.49	**0.42**
9	55.29	0.59	0.57	0.57	0.57	**0.47**	**0.61**	0.57	0.54	0.60	0.54	0.60	0.58	0.56
10	60.53	0.43	0.52	0.42	0.50	0.49	0.50	0.55	0.51	0.47	**0.40**	0.46	**0.55**	0.45
11	96.09	0.55	0.58	**0.69**	0.64	0.51	0.49	0.54	0.60	**0.47**	0.54	0.58	0.58	0.56
12	67.06	0.55	0.53	0.53	0.54	**0.43**	0.59	**0.61**	0.53	0.58	0.51	0.56	0.58	0.54
13	77.08	0.46	0.46	0.50	0.45	0.48	0.52	0.47	0.47	0.44	**0.42**	0.51	**0.57**	0.48

Bold denotes the maximum and minimum values of channel weights

TABLE IV

MODEL COMPLEXITY AND COMPUTATION TIME

Models	Parameters	Time(s/epoch)
Bi-LSTM	1335331	34.27
TempCNN	3197449	9.31
Transformer	102025	27.85
TAE	151497	8.27
TCN	34833	12.39
TA-TCN	152455	34.20
CA-TCN	**40977**	**14.07**

make it clearer that channel attention does learn the weights of the different channels and improves classification accuracy.

C. Computational Complexity

Table IV shows the number of network parameters and computation time in all models. Since TempCNN is a fully CNN architecture, it takes less time to compute even though it has very large parameters. Similarly, the TCN has a relatively short calculation time. The Bi-LSTM is an RNN architecture so not only are the parameters huge but the computation time is also very long. In addition, the transformer model takes about the same amount of time as the Bi-LSTM model due to the fact that each Key and each Query are multiplied by two, and the time dimension of this dataset is relatively long. Therefore, the TA-TCN model with the addition of the self-attention module in the time dimension is also time-consuming. The TAE model modified by the transformer has a significant improvement in computation time. The CA-TCN model proposed in this letter not only has a smaller number of parameters but also requires less computation time.

V. CONCLUSION

In this letter, a CA-TCN network has been proposed for remote sensing time series classification. The proposed CA-TCN uses the attention mechanism to enhance the information of the original TCN network in the channel domain, which can excavate deeper phenological characteristics. This is the first attempt of using TCN together with attention mechanism in the context of remote sensing time series analysis. Experiments have been conducted in the Breizhcrops dataset. The proposed CA-TCN has achieved the best performance with fewer parameters. Experiments also prove that the hybrid configuration model is superior to a single model.

REFERENCES

[1] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 421–435, Nov. 2020.

[2] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, Feb. 2019.

[3] P. Du *et al.*, "Advances of four machine learning methods for spatial data handling: A review," *J. Geovisualization Spatial Anal.*, vol. 4, no. 1, pp. 1–25, Jun. 2020.

[4] Y. Chen *et al.*, "Mapping croplands, cropping patterns, and crop types using MODIS time-series data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 69, pp. 133–147, Jul. 2018.

[5] S. Valero *et al.*, "Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions," *Remote Sens.*, vol. 8, no. 1, p. 55, 2016.

[6] C. Pelletier, G. I. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens.*, vol. 11, no. 5, p. 523, 2019.

[7] A. Sharma, X. Liu, and X. Yang, "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks," *Neural Netw.*, vol. 105, pp. 346–355, Sep. 2018.

[8] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[9] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: http://arxiv.org/abs/1803.01271

[10] P. Zhang *et al.*, "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, p. 3764, Nov. 2020.

[11] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107448.

[12] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12325–12334.

[13] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2021.

[14] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the Art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[15] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3575–3584.

[16] M. Račič, K. Oštir, D. Peressutti, A. Zupanc, and L. Č. Zajc, "Application of temporal convolutional neural network for the classification of crops on sentinel-2 time series," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. XLIII-B2-2020, pp. 1337–1342, Aug. 2020.

[17] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1598–1607, 2020.

[18] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[20] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, "Breizhcrops: A time series dataset for crop type mapping," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. XLIII-B2-2020, pp. 1545–1551, Aug. 2020.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.