Precipitation Forecasting Using Transformer: A Comparative Study with Unet

Sining Ang

School of Intelligent Systems Science and Engineering, Jinan University, Guangdong, China

angsining@stu2020.jnu.edu.cn

Abstract. Accurate precipitation prediction has a huge socio-economic impact. With the development of algorithms, architectures, and hardware technologies in the field of machine learning, the disadvantages of traditional numerical weather prediction methods are becoming more and more obvious. Noting that related studies have attempted to use the latest machine learning architectures for precipitation prediction, this work uses Unet and Transformer for precipitation prediction based on cloud layer information and compares them. The results indicate that Transformer can achieve 94.13% accuracy, while Unet has 96.89%. Finally, in conducting the data comparison, a conjecture related to the lagging of Transformer data is proposed based on the results of this experiment.

Keywords: Transformer; Unet; Precipitation Forecast.

1. Introduction

Accurate weather forecasting will undoubtedly bring great social and economic value. Currently, the most widely used weather forecasting methods are still numerical weather prediction (NWP) [1]. For decades, there is growing evidence that the use of methods such as data mining or neural networks in traditional weather forecasting systems will help improve forecast accuracy [2], and many of these models have been well applied under specific regions [3].

Now, as Transformer is widely used in more and more artificial intelligence fields [4][5], it is not difficult to expect its concrete effect in weather forecasting. In addition to Transformer, Unet has also shown excellent results in medical image, and semantic segmentation [6][7]. Currently, many researchers have also used Unet for weather prediction and achieved some success [8]. As of the writing of this paper, there is already some ongoing work that applies the combination of Transformer and Unet with some success [9][10].

In this study, two neural network models, Transformer and Unet, are used to model and predict weather information based on cloud maps. And the differences in accuracy and computation time between them are compared. The data used are from the open-source dataset cds. climate (classify precipitation into 20 grades (3-hour precipitation ranging from 0 to 30 mm), and make classification predictions). Although it is well known that clouds do not completely affect precipitation, this does not deny the important role of clouds in precipitation. In fact, by using cloud maps, it can at least make rough predictions for some gridded areas.

Of course, it is undeniable that, regardless of the model used, the complexity of the weather makes it difficult to achieve a high level of prediction from cloud information alone, but it is sufficient to complete a preliminary evaluation of both models and to provide ideas for subsequent studies.

2. Methods

2.1 Data Processing

For this work, the data used are all from the open-source data site "https://cds.climate" which is named *ERA5 hourly data.* It is worth noting that ERA5 is often used to analyze global climate and weather over the past four to seven decades and is the fifth generation of ECMWF reanalysis. Here, choose Ensemble for the Product type option instead of the recommended Reanalysis, which contains too much-unwanted data. It is worth noting that all data used in this experiment have been

preliminarily processed, and they were divided into 0.5-degree uncertainty estimates and 0.25-degree regular latitude grids for reanalysis. In addition, we only use cloud information and precipitation from two years of data from 2018 to 2020 for model training and testing here.

It is not wise to perform exact numerical estimation in this simple small model. To simplify the model appropriately, we do not use regression but classification, we classify precipitation into 20 grades (3-hour precipitation ranging from 0 to 30mm), and make classification predictions. On the one hand, it is used to reduce the transport volume. On the other hand, it will likely be difficult to end up with better results in both Transformer and Unet (in other words, they are both equally bad), which is not in line with this study's original.

The raw cloud data and the precipitation (labeled results) are shown below.

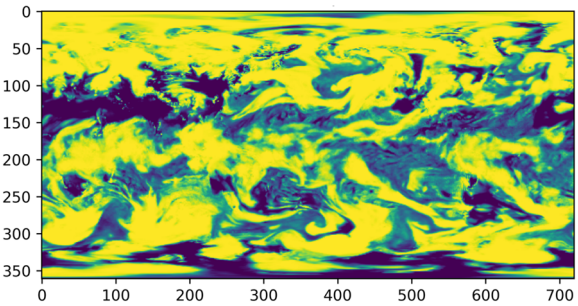

Fig 1. Raw cloud data (Coordinates indicate position)

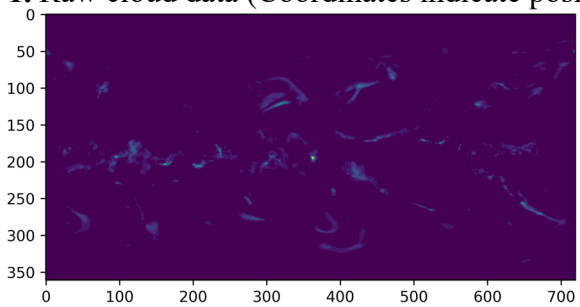

Fig 2. Precipitation (Coordinates indicate position)

2.2 Transformer

The Transformer was originally used in natural language processing (NLP). Subsequent work has shown that Transformer-based pre-trained models (PTMs) can achieve very good results on a very large number of cross-domain tasks.

The Transformer follows an encoder-decoder structure architecture. In brief, the model uses an encoder to process the input data to generate a sequence of information and a decoder to decode the mapped sequence and output the desired result. In both the encoder and decoder, using stacked self-attention and point-wise, fully connected layers, which will be shown in the left and right halves of Figure 3.

In addition to the above encoder-decoder structure architecture. One of the most important functions in the Transformer lies in the attention function. Intuitively, the attention function is used to extract features and place "attention" on the most likely output choices, this is also one of the reasons for using the final SoftMax function. In practice, it is described as mapping a set of query vectors and key-value pair vectors to output vectors. Similar to the common machine learning processing, it is common to pack the above query, key, and value vectors into three matrices named K, Q, and V for parallel operation. The final expression of the attention function used is as follows:

$$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Related work points out that using the same linear projection for queries, keys, and values of d_k, d_k and d_v dimensions is less effective than using different projection parameters for h times linear projection. In each projection, the keys and values (also matrices Q and K) will perform the attention function in parallel and they will total to produce the d_v dimensional output. These values are concatenated and projected again to obtain the final output value.

If only the existing model is used, it is obvious that it cannot focus on the information from different locations in different subspaces, in other words, the information in the subspaces is weakened. For this reason, multiple attention mechanisms will be employed for better preservation of subspace information.

There are at least three places in the Transformer that use the multi-head attention: in "encoder-decoder attention" layers; in the self-attentive layer in the encoder; in the decoder to prevent leftward information flow.

Going a little deeper, a fully connected feedforward network in the Transformer is used in each layer of the encoder and decoder. This network will be the same at every position. The activation function here is chosen to be ReLU.

Last, transcode and output the results using SoftMax and positional encoding. The full network structure is shown in Figure 3.

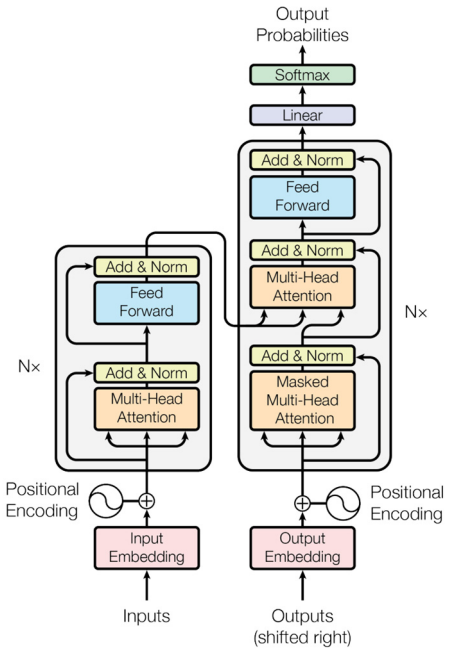

Fig 3. The Transformer - model architecture

2.3 Unet

Similar to Transformer, Unet was also not originally used in conventional computer vision and was originally applied to medical image segmentation. However, note the use of Unet in other areas in recent years. And it has been noticed that related studies are trying to use Unet for weather prediction. Naturally, it will be interesting to compare with Transformers that also generalize to other domains.

The network architecture of Unet is illustrated in Figure 4. A contraction path (left side) and an expansion path (right side) are included in the architecture. Typical architectures of convolutional networks are used in the contracting path. It consists of repeatedly applying two 3x3 convolutions (without increasing the convolution). The rectified linear unit (ReLU) will be used for the 3x3 convolution, after which it will be downsampled and a 2x2 maximum pooling operation with a span of 2 will be performed. In the downsampling step, the number of feature channels is doubled. Logically, each step in the extended path includes upsampling of the feature maps, and to halve the number of feature channels, connected to the corresponding feature maps in the compressed path, a

2x2 convolution ("up-convolution") will be performed, and thereafter two 3x3 convolutions using ReLU will be performed. Considering the loss of boundary pixels in the convolution, the model had to use cropping. In the last layer, the feature vectors of each 64-component are mapped to the required number of classes using a 1x1 convolution to obtain the final output.

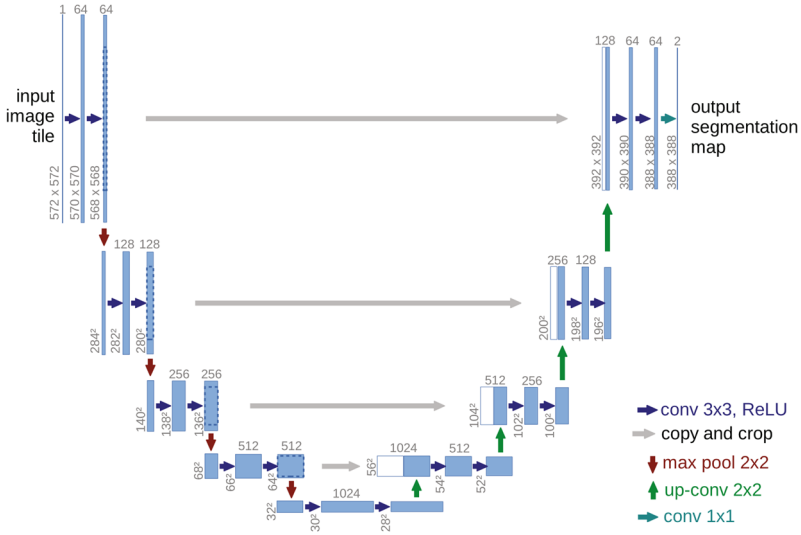

Fig 4. Unet architecture

3. Experiment

3.1 Details

As mentioned above, the data used in this work is the CDF4 type data provided by cds. climate. This experiment intercepts a portion of the ERA5 dataset for model training and testing. Specifically, the experiment selected all cloud cover and precipitation information from January 2016 to January 2018 with a three-hour interval (excluding marine regions) as data for training. Set January 1, 2018, set March 10, 2018 cloud cover information as the cross-validation set and information from January to June 2019 as a test set. To ensure that the comparison is as fair as possible, this work rewrites the Dataload function for handling datasets in CDF4 format and makes it available in both models (both in Unet and Transformer). For Unet, this work sets the learning rate to 0.00005 and trains 25 epochs. For the Transformer, this work is reluctant to modify its hyperparameters too much because the understanding of its internal mechanism is not yet perfect. Specifically, set the dimension of word embedding and location embedding to 512 (both dimensions must be the same), the number of hidden neurons in the FeedForward layer to 2048, the dimension of Query, Key, Value (Q, K, V) vectors to 64, the number of Encoder and Decoder to 6, and the heads in Multi-Head Attention is 8. (In fact, this work tries to modify some parameters, such as setting the heads to 16 and modifying other hyperparameters accordingly, and the test results and computing time consumption are extremely close). In addition, the traditional Transformer model for natural language processing (NLP) generally has a "mask" mechanism to prevent models from using subsequent information in advance. This step is not needed in this scenario, so this work does not use the "mask" mechanism. With a single 3090 GPU training, the Unet model took about 3 hours and the Transformer about 3.5 hours.

3.2 Results

The prediction results under all regions for 6 months indicated that Unet achieved 96.89% accuracy and Transformer had 94.13%. The figure below shows the forecast results for June 5, 2019, with the real data at the top, Unet forecast results at the bottom left, and Transformer forecast results at the bottom right.

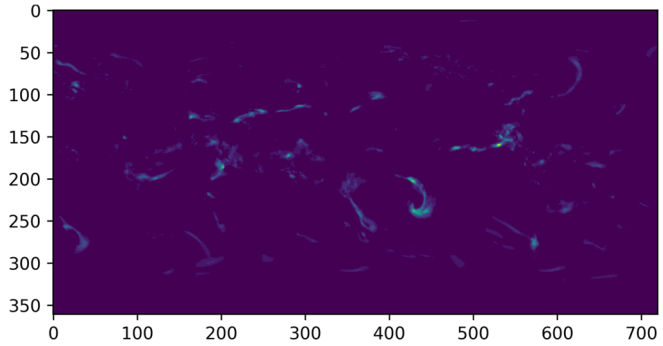

Fig 5. Precipitation (labeled results)

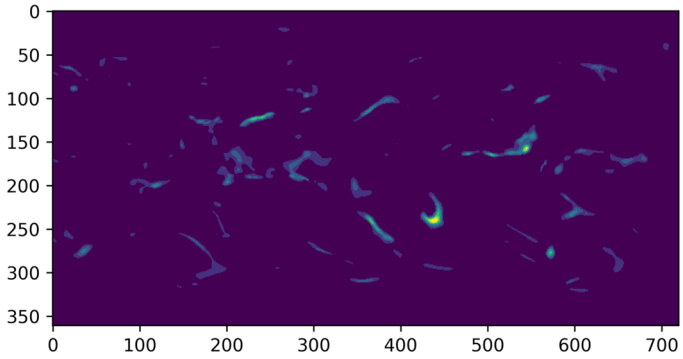

Fig 6. Precipitation with Unet

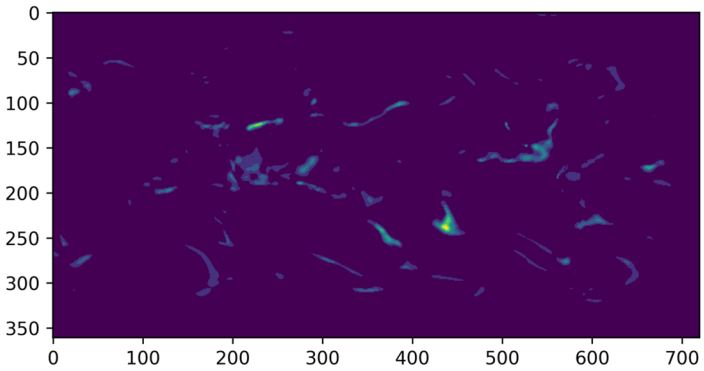

Fig 7. Precipitation with Transformer

Table 1. Hyperparameters settings and accuracy for Transformer and Unet

Transformer	Location embedding	Hidden neurons	Dimension of K, V	Encoder	Decoder	Multi-Head Attention	Training set	Accuracy
	512	2048	64	6	6	8	24 months	94.13%
	512	2048	128	6	6	4	24 months	93.77%
	512	2048	64	6	6	8	36 months	93.86%
Unet	Learning rate			Epoch			Training set	Accuracy
	0.00005			25			24 months	96.89%
	0.00001			25			24 months	93.04%
	0.00005			25			36 months	96.97%

It must be noted that weather prediction is a very complex task and the high accuracy shown here is somewhat confusing. We only predicted for precipitation and made a classification of 20 classes

between 0 and 30 mm, which allows the model to predict with very high accuracy the areas with no precipitation that occupy the majority of the area, which in turn leads to high overall accuracy.

The test results using different parameters are shown in the table 1.

3.3 Discussion

Honestly, the results of this experiment were somewhat surprising. I thought the Transformer-based prediction model should have had better results before the experiment started, but Unet performed better. It is known that the climate model is a time series (or has some time relationship), but Transformer with time series processing is not as good as Unet without any processing. Note the arc-shaped precipitation area down the right of the center in Figure 1. In Figure 6, Unet better predicts the trend of the arc, while in Figure 7, this is predicted to be a triangular area. In this scenario, although it has time series features, it does not introduce features such as "wind direction", "wind speed", etc., and the effect of each region of the time series alone is so heterogeneous that it is almost unpredictable. In this case, the processing of Positional Encoding in the Transformer pair may introduce bad data instead. Maybe this is one of the reasons for the poor performance of the Transformer.

4. Conclusion

This work regionalizes precipitation information based on EAR5 cloud information and focuses on comparing the accuracy of the models with Unet and Transformer architectures.

On the 6-month test results, Unet could achieve 96.89% correct prediction, and transform was 94.13%. Based on the results, it is conjectured that the Transformer is not suitable for processing time series of raw data in some scenarios, and if the features provided by the Transformer are too few, it will tend to perform worse than the traditional model without processing due to the introduction of time series.

References

[1] Lorenc A C. Analysis methods for numerical weather prediction[J]. Quarterly Journal of the Royal Meteorological Society, 1986, 112(474): 1177-1194.

[2] Schultz M G, Betancourt C, Gong B, et al. Can deep learning beat numerical weather prediction? [J]. Philosophical Transactions of the Royal Society A, 2021, 379(2194): 20200097.

[3] Sønderby C K, Espeholt L, Heek J, et al. Metnet: A neural weather model for precipitation forecasting[J]. arXiv preprint arXiv:2003.12140, 2020.

[4] Han K, Xiao A, Wu E, et al. Transformer in Transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[5] Jin W, Barzilay R, Jaakkola T. International conference on machine learning[J]. 2020.

[6] Huang H, Lin L, Tong R, et al. 3+: A Full-Scale Connected UNet for Medical Image Segmentation. arXiv 2020[J]. arXiv preprint arXiv:2004.08790.

[7] Li X, Chen H, Qi X, et al. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes[J]. IEEE transactions on medical imaging, 2018, 37(12): 2663-2674.

[8] Trebing K, Stańczyk T, Mehrkanoon S. SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture[J]. Pattern Recognition Letters, 2021, 145: 178-186.

[9] Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure Transformer for medical image segmentation[J]. arXiv preprint arXiv:2105.05537, 2021.

[10] Wang H, Xie S, Lin L, et al. Mixed Transformer u-net for medical image segmentation[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2022: 2390-2394.