TransNext: A vision enhanced Transformer model for accurate electricity load forecasting

Ayushman Mishra
Department of Electrical Engineering

Himanshu Jindal
Department of Electrical Engineering

Jayant Jethi
Department of Electrical Engineering

Bhavnesh Jaint
Department of Electrical Engineering

Garima
Department of Electrical Engineering

***Abstract*- Energy forecasting on loads plays the vital role for the energy suppliers and the customers as it allows them for building an effective plan to fulfil demands. In energy industry, decision-makers and specialists provide reliable estimation of future energy demand/load at the aggregate and particular site levels which are crucial. Through this paper, we provide a unique model TransNext and have conducted extensive experiments on UCI Power consumption dataset. Our model outperforms other SOTA models and achieves a RMSE of 0.1881, MAE of 0.07678 and MAPE of 0.15716.**

Keywords – transformers, CNN, positional encoding, multi head attention, load forecasting

1 Introduction

Predicting electricity usage is a critical responsibility in energy management, as it allows energy planners and operators to optimize energy production, distribution, and storage. Accurate and timely predictions of electricity consumption enable energy providers to avoid overloading the power grid, minimize energy waste, and reduce costs. However, predicting electricity consumption is a challenging problem due to its high variability and complexity.

Forecasting future trends is a method of applying statistical knowledge to analyse past data in relation to current events and the results collected over time are used to predict future events. Forecasting problems generally deal with time series data which is a sequence of historical data corresponding to groups or observations of data collected over time in consecutive periods and depending on the use case, the data may be presented in a daily, weekly, monthly, quarterly, or yearly manner. [1]

To address forecasting issues, numerous load forecasting techniques have been developed. These techniques fall into three categories: (i) techniques through which the thermal dynamics and energy behaviour of the buildings are estimated; (ii) statistical methods used for investigating the energy consumption and various components like climate data and occupancy; and (iii) machine learning techniques, which concentrate more on identifying distinct patterns in energy consumption from historical data [2].The majority of traditional techniques were built using regression and statistical models such as ARIMA (Auto-regressive Integrated Moving Average), SARIMA (Seasonal ARIMA), SVR and Random Forest. These provide interpretable and computationally efficient solutions but they are incapable to seize the sequence information present in the time series.

To capture the sequence information efficiently, sequence models examples involve LSTM, RNN and GRU, based architectures have been proposed. These models have shown superior performance in various time series forecasting problems, including electricity consumption prediction. Transformer networks are a recent proposal which aim to overcome the parallelization problem encountered in LSTM [3]. Transformer-based models may efficiently enhance the dynamics of complex time series data, which are difficult to extract for conventional systems like RNN [4]. This model has made great strides in terms of precision and improved performance with its parallelization capabilities for natural language processing (NLP). However, this architecture serves the purpose for linguistic data and time series data which can only

be applied when it is converted as an appropriate input to the transformer. Transformer usage with time series data has the following restrictions:(i) large memory utilization, (ii)quadratic computation in time, and (iii) slow processing speed. Through this study, we proposed TransNext, a robust and efficient Transformer+CNN based forecasting model that can preserve long range dependencies in the input time series.

2 Related Work

The challenge of load forecasting has increased dramatically due to technological change [1]. Both statistical methods and models based on deep learning and artificial intelligence can be used for load forecasting. Statistical models used for predicting energy demands include techniques such as ARMA [2] [3], ARIMA [4], grey models [5], Kalman filtering algorithms [6]. [7] reviewed the traditional techniques and AI techniques for predicting electricity usage. [8] reviewed the benefits, drawbacks and intended use of several AI-based strategies for predicting household energy consumption at the urban and rural levels. [9] [10] investigated neural networks for forecasting by considering LSTM cells. [9], [11], [10] investigated neural networks for forecasting by considering LSTM cells. By modifying Seq2Seq(sequence to sequence) model with the comparable attention mechanism, [12] proved that their technique surpassed models based on ARIMA and LSTM, that anticipate influenza prevalence. Deep learning has a strong ability for generalization, facilitates unsupervised learning, and has high learning ability for performing on vast scale dataset. It is capable of handling advanced applications like indoor object detection [13], fatigue detection [14] and forecasting problems [15].Gated recurrent network with temporal distribution was designed by Qu et al. [16] for forecasting solar power. [17] uses models such as artificial neural network (ANN) and support vector regression (SVR) which shows promising results particularly in short term forecasting. In response of this, many literature reviews on AI-Based models have been published. [18] designed a modified ANN prediction model using the Chaotic Particle Swarm Optimisation strategy (CPSO).The outcomes demonstrated that the method outperformed ANN in terms of prediction accuracy and nonlinear fitting strength. An assessment of traditional and artificial-intelligence-based approaches for calculating energy use found that 77% models were ANNs and remaining were SVR [19], Wavelet decomposition [22], Fourier transformation [23], and fast Fourier transformation [24], to name a few, have all demonstrated excellent competition in pre-processing time series jobs so far. Among these, Wavelet decomposition fixes Fourier analysis primary flaw [25]. Transformer networks have recently been suggested as a solution to the parallelization issue with LSTM [26].Transformer based models are capable of accurately capturing complex time-series data dynamics that cannot be computed conveniently using classical sequence models similar to recurrent neural network (RNN) [27].

3 Proposed Model

This section presents the suggested model Trans Next which is explained in detail.

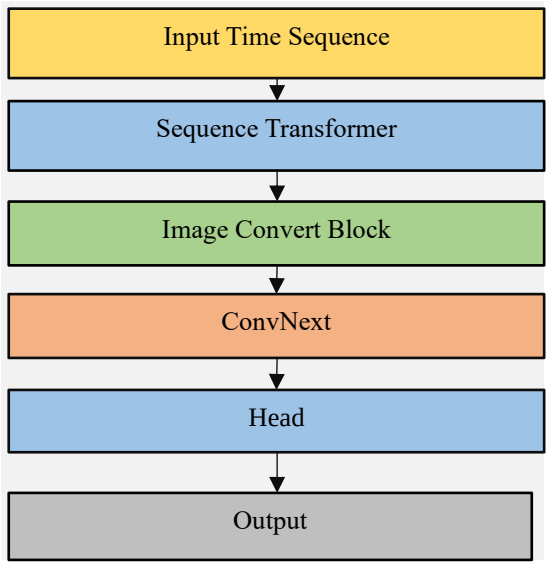

Fig. 1 The overall framework of TransNext

3.1 Sequence Transformer Block

Vanilla Transformers[20] outperform other sequence-based models like LSTM [21], encoder-decoder models [22] , RNN, etc. in time series tasks and natural language processing. The secret to their greater performance is a self-attention mechanism that enables a transformer to concentrate more on a sequence of input that is more crucial for prediction. Several identical blocks make up both the encoder and decoder. A position-wise feed-forward network and a multihead self-attention module make up each encoder block. Positional encoding is used to feed the transformer encoder with positional information about the input sequence.

Fig. *2* demonstrate the layout of the sequence transformer block.

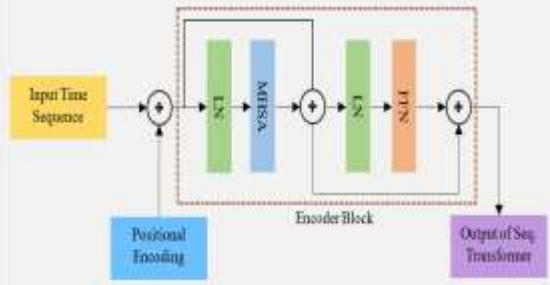

Fig. 2 The Sequence Transformer Block

3.1.1 Positional Encoding

Positional encoding is a tool to denote the location of an entity within a sequence so that each location gets a unique representation. A transformer has no recurrence. The transformer fixes this by including a positional encoding vector in each input embedding. The model learns a pattern from these vectors that allows it to estimate the position of each component or the separation between them in the input sequence.

3.1.2 Multi Head Attention

An attention mechanism uses many heads to process attention in concurrently. The individual attention outputs are then linearly combined to obtain the anticipated dimension. Multiple attention heads enable for diverse attention to be paid to different sequence elements, intuitively

$$\boldsymbol{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = [head_1, \ldots., head_h]\boldsymbol{W_0}$$

Where $head_i = \boldsymbol{Attention}(\boldsymbol{QW_i^Q}, \boldsymbol{KW_i^K}, \boldsymbol{VW_i^V})$

Here W_i are learnable parameters learnt during backpropagation.

3.1.3 Feed Forward Network and Residual Connection:

Each attention vector is subjected to this basic feed-forward neural network with the intention of transforming the attention vectors into a format that can be handled by the following encoder or decoder layer. Each sub-layer in an encoder (Self-Attention, Feed-Forward Network, etc.) has a residual connection all around it, and a layer-normalization step comes after it.

3.2 ConvNext

A ConvNext has the same straightforward construction as Convolutional Neural Networks and outperforms Vision Transformers in terms of accuracy, performance, and scalability. It is multi-staged, and each stage's design was influenced by Vision Transformers.

3.3 Image Formation Block

The information/feature map coming out of a transformer block was passed through a activation layer and then projected to form a 3-D feature map of dimension $\mathbb{R}^{B\ X\ 3\ X\ 16\ X\ 16}$ from a feature map of dimension $\mathbb{R}^{B\ X\ 768}$ containing sequence information.

4 Experimental Setup

4.1 Datasets

Measurements over a four-year period of the amount of electricity used in the same residence was taken. A sampling rate of one-minute was established. The dataset shows a total of 9 electrical readings and sub-metering measurements. This was sourced from Georges Hebrail, Senior Researcher, EDF R&D, Clamart, France (georges.hebrail@edf.fr). Master of Engineering Internship at R&D, Clamart, France, Alice Berard, TELECOM ParisTech.

This dataset is a time series dataset with multiple variables. The dataset was compiled within a 47-month span, from December 2006 to November 2010, and contains 2075259 unique cases that were collected in a home in Sceaoux, France (only 7 kilometres from Paris). Approximately 1.25% of the rows in the dataset have some measurement missing values. The dataset contains all calendar timestamps, however some of the timestamps measurement values are absent. An empty value is denoted by the space between two semicolons in an attribute separator. For instance, on April 28, 2007, the dataset displays missing values.

Table 1 Summarization of variables used in dataset

Index	Variables	Description
1.	Voltage	Average voltage (V)
2.	Global Reactive Power	Global reactive power utilized in a household (KW)
3.	Global active Power	Global active power utilized in a household (KW)
4.	Date Time	Date and time
5.	Global Intensity	Intensity of current measured intermediately (A)
6.	Sub_metering_1	Active energy measured for cookhouse (Wh)
7.	Sub_metering_2	Active energy measured for washing room (Wh)
8.	Sub_metering_3	Active energy measured for cooling systems (Wh)
9.	Sub_metering_4	Leftover energy consumption estimated for house(Wh)

Table 1 shows the description of household power consumption dataset variables.

4.2 Data Pre-processing

The null values in the dataset were imputed using mean and before training the model the dataset was standardised using MinMaxScaler to scale all the input features in the range [0,1].

4.3 Hardware

The models were trained on NVIDIA Tesla P100 GPU (16GB VRAM)

4.4 Hyperparameters

AdamW was used for training our model with an initial pace of learning at 1e-4 and CosineAnnealing learning rate scheduler was used. The model was developed to train for 20 iterations. The batch size was taken to be 48 and input window size was set to 16. Model training was accomplished using Pytorch library on NVIDIA P100 GPU.

4.5 Performance Metrics

The forecast obtained was evaluated on 3 parameters namely MAE, RMSE and MAPE.

RMSE: The standard deviation of the residuals (estimation error) is known as the root mean square error. The separation between the line of regression and the data points is then determined. Now the spread of the above mentioned residues is calibrated using the RMSE. Intrinsically it informs you of the strength of the data surrounding the line of best fit.

Mathematically RMSE is Calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{pred_i} - y_{actual_i}\right)^2}$$

MAE: Mean absolute error is the distinction between the measured value and the true value of a quantity which is to be calculated. Suppose a tape reads 19 cm but the true value is 18 cm, then the absolute imprecision of the tape is 19cm-18cm which equates to 1 cm.

Mathematically MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred_i} - y_{actual_i}\right|$$

MAPE: The Mean Absolute Percent Error (MAPE) is used to gauge the accuracy of the forecast. It is commonly known as Mean Absolute Percent Deviation (MAPD).The accuracy is expressed as a percentage. It can be enumerated by multiplying the average percent inaccuracy each time by the absolute value minus the absolute value

Mathematically MAPE is Calculated as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{pred_i} - y_{actual_i}}{y_{pred_i}}\right|$$

5 Results and Discussion

5.1 Results on Benchmark Datasets

TransNext was able to achieve improved results on the UCI power consumption dataset and outperformed the existing models with a big margin. TransNext has achieved a RMSE of 0.1881, MAE of 0.07678 and MAPE of 0.15716.

5.1.1 Quantitative Analysis

The results attained for UCI power consumption dataset is tabulated in **Table 2**.

Table 2 Results of Proposed Model on UCI Power Consumption Dataset

Model	Metric	
TransNext	RMSE (kW)	0.1881
	MAE (kW)	0.07678
	MAPE (%)	0.15716

5.1.1.1 Comparison Against State-of-the-Arts

The suggested model has been evaluated against other SOTA models on UCI Power Consumption dataset and results are tabulated in **Table 3**.

Table 3 Analysis of performance of Trans-Next Model with alternative SOTA models in terms of RMSE (kW) on UCI Power Consumption Dataset

Model	Minutely
CRBMs [23]	0.9032
LSTM-Seq2Seq [24]	0.6670
FCRBM [23]	0.6663
CNN-LSTM [25]	0.6114
CNN-BLSTM [26]	0.5650
CNN-GRU [27]	0.4700
CNN-LSTM-DWT [28]	0.4645
Transformer-SWT [28]	0.3929
Proposed Model	**0.1881**

5.2 Qualitative Analysis

The ground truth time series and the time series as predicted by the proposed model have been plotted in ***Fig. 3***. It is evident to observe that the time series predicted by the TransNext is very close to the ground truth time series and thus proposed model has ability for learning the trends in the input time sequence.

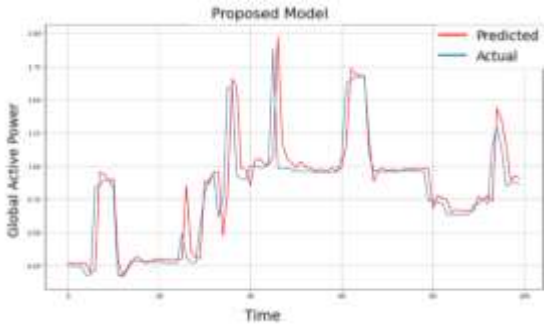

Fig. 3 Time Series Plots of Ground Truth Time Series and Time Series predicted by the proposed model.

6 Conclusion and Future Trends

We proposed a novel model TransNext and conduct extensive experiments on UCI Power consumption dataset. Our conceptualization outperforms other SOTA models and achieves a RMSE of 0.1881, MAE of 0.07678 and MAPE of 0.15716.

Image-based techniques may be merged with other modalities to formulate multi-scale and multi-domain structures for forecasting time series in the distant future.

7 References

[1] L. Li and K. Ota, "When Weather Matters:IoT-Based Electrical Load Forecasting For Smart Grid," *IEEE Communications Magazine,* p. 6, 2017.

[2] J.-F. Chen and W.-M. Wang, "Analysis of an adaptive time-series autoregressive moving-average (ARMA) model for short-term load forecasting," *Electric Power Systems Research,* vol. 34, no. 3, pp. 187-196, 1995.

[3] S.-J. Huang and K.-R. Shih, "Short-term load forecasting via ARMA model identification including non-Gaussian process considerations," *IEEE Transactions on Power Systems.*

[4] G. E.P and G. , "Some recent advances in forecasting and control," *Journal of the Royal Statistical Society,* vol. 17, no. 2, 1968.

[5] A. Foucquier, S. Robert and F. Suard, "State of the art in building modelling and energy performances prediction: A review," *Renewable and Sustainable Energy Reviews,* vol. 23, pp. 272-288, 2013.

[6] H. Al-Hamadi and S. A. Soliman, "Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model," *Electric Power Systems Research,* pp. 47-59, 2004.

[7] M. A. M. Daut and M. Y. Hassan, "Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review," *Renewable and Sustainable Energy Reviews,* vol. 70, pp. 1108-1118, 2017.

[8] A. Tanveer and C. Huanxin, "A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review," *Energy and Buildings,* vol. 165, pp. 301-320, 2018.

[9] Zhang and E. Patuwo, "Forecasting with artificial neural networks," *International Journal of Forecasting,* pp. 35-62, 1998.

[10] N. Kourentzes, "Intermittent demand forecasts with neural networks," *International Journal of Production Economics,* vol. 143, no. 1, pp. 198-206, 2013.

[11] Gers, S. and F. A., "Applying LSTM to time series predictable through time-window approaches," in *Springer,* 2001.

[12] K. Kondo and M. Kimura, "Sequence to sequence with attention for influenza prevalence prediction using google trends," in *Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics,* New York, 2019.

[13] M. Afif and R. Ayachi , "An evaluation of EfficientDet for object detection used for indoor robots assistance navigation," *Journal of Real-Time Image Processing,* pp. 651-661, 2022.

[14] R. Ayachi and M. Afif, "Drivers Fatigue Detection Using EfficientDet In Advanced Driver Assistance Systems," in *18th International Multi-Conference on Systems, Signals & Devices (SSD),* Monastir, 2021.

[15] N. Ayoobi, D. Sharifrazi and R. Alizadehsani, "Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods," *Results in Physics,* 2021.

[16] Y. Qu and J. Xu, "A temporal distributed hybrid deep learning model for day-ahead distributed PV power forecasting," *Applied Energy,* 2021.

[17] L. Yang, H. Yan and J. C. Lam, "Thermal comfort and building energy consumption implications- A review," *Applied Energy,* vol. 115, pp. 164-173, 2014.

[18] H.-d. He, "Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm," *Building and Environment,* vol. 78, pp. 111-117, 2014.

[19] N. Wei and C. Li, "Conventional models and artificial intelligence-based models for energy consumption forecasting: A review," *Journal of Petroleum Science and Engineering,* pp. 106-187, 2019.

[20] A. Vaswani and N. Shazeer, "Attention Is All You Need," in *Neural Information Processing Systems,* Long Beach, 2017.

[21] Y. Benjio, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks,* vol. 5, no. 2, pp. 157-166, 1994.

[22] Sutskever and O. Vinyals, "Sequence to Sequence Learning with Neural Networks," *arXiv,* 2014.

[23] E. Mocanu and P. H. Nguyen, "Deep learning for estimating building energy consumption," *Sustainable Energy, Grids and Networks,* pp. 91-99, 2016.

[24] F. Ullan and I. U. Haq, "Short-Term Prediction of Residential Power Energy Consumption via CNN and Multi-Layer Bi-Directional LSTM Networks," *IEEE Access,* 2020.

[25] T.-Y. Kim and B. S. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy,* pp. 72-81, 2019.

[26] D. L. Marino and K. Amarasinghe, "Building energy load forecasting using Deep Neural Networks," in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society,* Florence,Italy, 2016.

[27] M. Sajjad, "A Novel CNN-GRU-Based Hybrid Approach for Short-Term Residential Load Forecasting," *IEEE Access,* 2020.

[28] L. S. Saoud and H. AlMarzouqi, "Cascaded Deep Hybrid Models For Multistep Household Energy Consumption Forecasting," *arXiv,* p. 13, 13 10 2022.

[29] P. Chujai and N. Kerdprasop, "Time Series Analysis of Household Electrical Consumption with ARIMA and ARMA Models," in *International MultiConference of Engineers and Computer Scientists 2013 Vol I,* Hong Kong, 2013.

[30] N. Wu and B. Green, "Deep Transformer Models For Time Series Forecasting: The Influenza Prevalence Case," *arXiv,* p. 10, 2020.

[31] Y. Guangle and L. Tao, "A review of Convolutional-Neural-Network-based action recognition," *Elsevier,* vol. 118, pp. 14-22, 2019.

[32] A. Dhillon and G. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Springer ,* pp. 85-112, 2019.

[33] I. Sneddon, "Fourier transforms," *Courier Corporation,* 1995.

[34] W. T. Cochran, J. W. Cooley and D. L. Favin, "What is the fast fourier transform?," in *Proceedings of the IEEE,* 1967.

[35] D. B. Percival and A. T. Walden, "Wavelet methods for time series analysis," *Cambridge university press,* vol. 4, 2000.

[36] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (covid-19) cases: A data-driven analysis," *Chaos. Solitons & Fractals,* vol. 135, 2020.