Energy Consumption Modelling with Hybrid Feature Selection Aided Attention Mechanism

CAKE Team
Tingwei Chen, Yantao Wang, Hanzhi Chen, Zijian Zhao, Xinhao Li and Guangxu Zhu

***Abstract*—The fifth generation communication technology could provide marvellous services but its energy consumption increase correspondingly compared to 4G network. To subsequently optimize the energy efficiency of 5G network, modelling the affect of base station and servicing cell characters to the ultimate consumption measurement is vital. While the heterogeneous and coupled features pose barriers to feature selection and network architecture design. This paper proposed a hybrid feature selection method to process features from data science and physical meaning perspective. Following neural network utilizes attention mechanism to capture the impact of base station parameters and energy-saving method to its energy consumption. Our algorithm could achieve absolute low Mean Absolute Percentage Error (MAPE) across different base station products and configurations. And it is also lay the foundation for future optimization of energy efficiency of network.**

I. Introduction

The advanced features of 5G network [1] empower its service capacity to various applications including Virtual Reality, Vehicular Network and Internet of Things. However, its energy consumption is correspondingly higher than that of the 4G network. For future energy efficiency optimization frameworks of 5G networks, it is vital to model the effect of base station and servicing cell characteristics on the ultimate consumption measurement.

The heterogeneous and coupled features of 5G pose barriers to feature selection and network architecture design. To address these challenges, this paper proposes a hybrid feature selection method that processes features from both data science and physical meaning perspectives. This method aims to reduce the feature subset search space and eliminate redundant features while taking into account the relevance of similar features. Thereby, it can select more relevant features while reducing the feature space.

The attention mechanism [2] was first introduced in 2014 as a way to improve the performance of neural networks in natural language processing tasks. Since then, it has been widely adopted in various domains, including computer vision and speech recognition. Our solution leverages a similar mechanism in several modules to facilitate efficient feature extraction and decoupling.

To reduce energy consumption of base station, it is necessary to optimize its parameters and energy-saving methods which require a deep understanding of how parameters and methods impact the energy consumption. Therefore, accurate modelling of energy consumption is essential for achieving more energy-efficient network deployments

The authors are with the Shenzhen Research Institute of Big Data, Shenzhen, China. *(Corresponding author: Guangxu Zhu.)*

II. Related Work

The achievement of robust and scalable energy consumption management is a prerequisite for the potential for the high energy efficiency of the 5G network. Modelling the effect of base station settings and energy-saving modes on energy consumption is essential to reducing network energy consumption, and recent years have seen a lot of interest in this area of study. In [3], the authors offered a realistic power consumption model that takes into account a large number of MIMO components, the number of multiplexed users per physical resource block (PRB), and both downlink and uplink transmission. The work in [4] examined multi-carrier capabilities, including carrier aggregation and its various aggregation capacities, in conjunction with MIMO.

Some of these conventional methods are becoming insufficient due to the complex and varied service requirements of 5G networks, as well as their growing heterogeneity. Such large-scale network challenges with complex BS and user equipment distributions have been studied using certain machine-learning techniques to enhance the model's cross-equipment and cross-configuration generalization capacity. An artificial neural networks (ANN) power consumption model for 5G multi-carrier active antenna devices in [5].

However, there is little prior work on the generalisability of the learning-based energy consumption modelling, which has motivated this work.

III. System Model

Our system model is illustrated in following figure 1. After data pre-processing, hybrid feature selection module would produce two feature sets. Two networks are trained with corresponding feature sets then they are evaluated to generate energy consumption for sets w/o BS information.

$$WMAPE = \frac{\sum_{i=1}^{n}(w_i|y_i - \widehat{y}_l|)}{\sum_{i=1}^{n}(w_i|y_i|)} \tag{1}$$

The evaluation criteria is defined in equation 1. In order to evaluate the model's ability to generalize across different equipment and configurations, we use a weighted relative error evaluation method to estimate the accuracy of the test set. This method assigns a larger error weight, w_i, to samples corresponding to new devices and/or configurations in the test set.

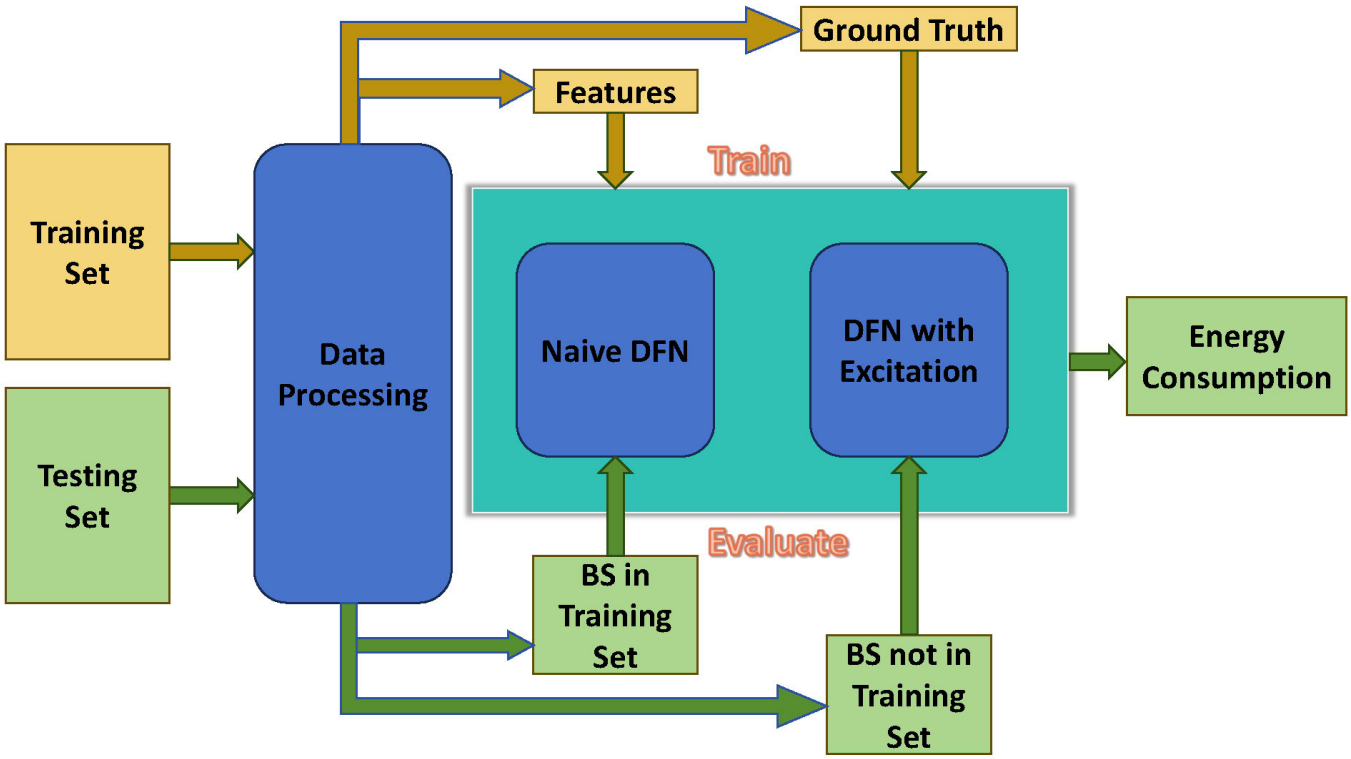


Fig. 1: System Model

IV. Feature Selection

A. Datasets

The raw dataset is segmented into three main components,

Base Station Information (BSinfo): Contains configuration and hardware attributes of each base station.

Cell-level Data (CLdata): Features hour-level metrics, including service compliance like load and energy-saving counters such as the energy-saving mode's duration.

Energy Consumption (ECdata): Provides hour-level details on the total energy consumption of base stations.

B. Data pre-processing

By analyzing the ECdata dataset, it became evident that the goal is to predict Energy Consumption for a given Time and BS. The BSinfo provides essential configuration information about the base station, including encapsulating attributes like the radio unit type (RUType), transmission mode (Mode), Frequency, Bandwidth, Antennas, and Maximum transmit power of the cell. Notably, for all given base stations, the RUType and Mode remain consistent, corroborating our findings from the dataset. The CLdata further offers insights into the load and energy-saving mode (ESMode) associated with each BS at diverse time. Consequently showcased in Table I, by merging BSinfo with CLdata, we obtain a unified dataset.

At this point, it's worth noting that a specific BS and Time can uniquely identify a data entry. However, multiple Cells can be associated with a single BS. To simplify this data, we aligned cell-related features with other corresponding features, leading to new attributes, as represented in Table II.

one-hot encoding: To enable features in the form of identifiers, as well as numerically insignificant real-numbered features, to serve as inputs to the neural network, it is essential to apply one-hot encoding to features with categorical attributes, including BS, Time, RUType, Mode, Frequency, Bandwidth, and Antennas. One-hot encoding transforms a multi-category feature into a binary vector where only one element is '1' and the rest are all '0'. This encoding not only retains the categorical information from the original data but also ensures that the model doesn't mistakenly infer any unintended ordinal relationships between the categories. To illustrate this concept formally, consider the following definition: Let X be a discrete feature with n categories, where each category is denoted as c_i for $i \in [1, n]$. Through one-hot encoding, we can transform c_i into a vector v_i of length n. For each element j of v_i, its value is defined as:

$$v_{i,j} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In essence, for each category c_i, its corresponding one-hot vector v_i has its i^{th} element as 1, and all other elements are 0.

C. Feature selection

In our feature selection phase, we undertook a systematic approach to refine the features used for training the model. A

TABLE I: Data Structure

BS	Time	CellName	...	Load
BS_0	1/1/2023 1:00	Cell_0	...	0.48793617
		Cell_1	...	0
		Cell_2	...	0
		Cell_3	...	0
...	...	...	...	...
BS_1019	1/1/2023 1:00	Cell_0	...	0.01993985
		Cell_1	...	0
		Cell_2	...	0
		Cell_3	...	0

key step was computing the standard deviation for each feature in the training set. Based on our evaluation:

Features with a standard deviation of 0 were identified as static, having identical values across the entire training set. These features inherently lacked variability and thus offered no discernible value during the neural network's training process. The inherent risk is the model over-emphasizing such static features while neglecting those with more valuable information, potentially undermining overall performance.

Features with a standard deviation close to, but not exactly 0 (e.g., 0.01), exhibited minimal variability within the training set. Such features could potentially compromise the model's generalization ability, leading to subpar performance on unseen test data.

To ensure optimal performance and mitigate risks of overfitting, columns with a standard deviation less than 0.01 were eliminated from our feature set. This refinement enabled our model to better focus on pertinent features, enhancing its adaptability to new data.

V. Methodology

In this section, we will introduce the architecture of our networks. First, we divide the problem into two different scenarios based on whether the base station to be predicted exists in the training set. As illustrated in Fig. 2, we then design two distinct networks to address each scenario respectively.

A. Naive DFN

In the scenario where the BS to be predicted exists in the training set, we employ a Deep Feedforward Network (DFN) as depicted in Fig. 2a. The DFN architecture consists of multiple layers, each comprising a feed-forward layer followed by a batch normalization layer to address overfitting and improve model performance. Additionally, we incorporate an activation layer, specifically the LeakyReLU function, to compute the output.

The DFN model can be represented mathematically as

$$X_n = \text{LeakyReLU}\left(\text{BatchNorm}\left(\text{Linear}\left(X_{n-1}\right)\right)\right), \tag{3}$$

where each component is defined as

$$\text{Linear}(X_{n-1}) = X_{n-1}^{m,d_{n-1}} W_{n-1}^{d_{n-1},d_n} + b_{n-1}^{d_n}, \tag{4}$$

$$\text{BatchNorm}(x) = \frac{x - \text{E}(x)}{\sqrt{\text{Var}(x) + \epsilon}} \odot \gamma + \beta, \tag{5}$$

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x \geq 0, \\ \alpha x, & \text{if } x < 0. \end{cases} \tag{6}$$

In particular, $X_{n-1}^{m \times d_{n-1}}$ represents the input of layer $n-1$, where m is the batch size and d_{n-1} is the feature dimension of layer $n-1$. $E(x)$ and $Var(x)$ denote the average and variation of x, respectively, with ϵ, γ, β and α denoting respectively the small number, the scale parameter, the shift parameter, and the leakage coefficient. In the proposed DFN model, we utilize a total of 5 layers, where the output of each layer serves as the input for the subsequent layer.

B. DFN with Excitation

As mentioned in the previous section, we use one-hot encoding to represent discrete input parameters such as the number of base stations. This allows our model to handle any number of base stations without limitations. For base stations that are not present in the training set, we set their one-hot vectors as zero vectors to ensure scalability.

To tackle the more challenging task of predicting base stations that are not in the training set, we propose a more efficient network structure. We draw inspiration from the attention mechanism [6], which has been widely used in time-series models, particularly in the field of Natural Language Processing (NLP). The attention mechanism has shown excellent capacity in capturing relationships between different elements in a sequence. Similarly, in the field of Computer Vision (CV), there are similar mechanisms such as the Squeeze-and-Excitation Network (SENet) [7], Spatial Transformer Network (STN) [8], and Convolution Block Attention Module (CBAM) [9], all of which have demonstrated good performance. In this work, we introduce an Excitation layer based on SENet and our data features to capture relationships between different attributes.

$$Y = E \odot X \tag{7}$$

$$E = \text{Sigmoid}(\text{Linear}(\text{ReLU}(\text{Linear}(X)))) \tag{8}$$

$$\text{ReLU}(x) = \max(0, x) \tag{9}$$

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

where E is the Excitation Vector which illustrates the relationship between different features.

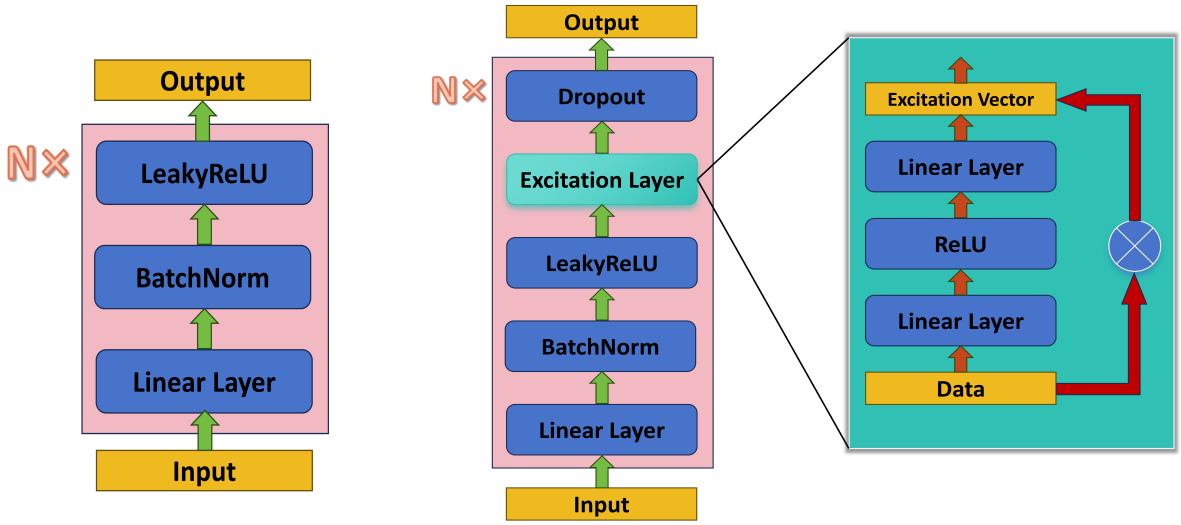
Specifically, we incorporate the Excitation layer after a dropout layer with a probability of 0.5, which enhances the generalization capability as shown in Fig. 2b. This Excitation layer is added after each layer in the Naive DFN. In total, our model employs 7 layers to further improve its performance.

VI. Experiment

In this section, we present a series of ablation studies conducted to demonstrate the efficacy of our model design and data processing techniques.

TABLE II: Merge by Cell

BS	Time	Mode1	Mode2	RUType1	RUType2	...	ESMode1_cell0	ESMode1_cell1	...	Load_cell0	...
BS_0	1/1/2023 1:00	0	1	1	0	...	0	0	...	0.48793617	...
BS_1	1/1/2023 1:00	0	1	0	1	...	0	0	...	0.3477	...
...	...	...	...	...	...	...	...	...	...	...	...
BS_1019	1/2/2023 1:00	1	0	0	0	...	0.95694444	0	...	0 01993985	...



(a) Naive DFN

(b) DFN with Excitation

Fig. 2: Network Architecture

A. *Excitation Layer*

As mentioned in Section V-B, we propose an Excitation mechanism to address the scenario where the base station to be predicted does not exist in the training set. To demonstrate the effectiveness of this approach, we conducted an ablation study to compare the performance of two models: the Naive DFN (Fig. 2a) and the DFN with Excitation (Fig. 2b).

Due to time constraints and the unavailability of ground truth for the testing set, we evaluated our model's performance by submitting our results to the contest website. During testing, we ensured a consistent variable by keeping the results of other testing data the same, specifically for base stations present in the training data. For instance, when comparing the performance of the Naive DFN and DFN with Excitation on testing data where the base station is not in the training data, we maintained a constant result for the other base stations in the training data.

The experimental results indicate that the Naive DFN performs better when the base station to be predicted is present in the training set. Conversely, the DFN with Excitation achieves superior performance when the base station to be predicted is absent from the training set.

B. *One-Hot Encoding*

Certain parameters such as frequency, bandwidth, and antennas are continuous variables, but the dataset contains only a limited number of distinct values for each parameter. Therefore, we treat them as categorical variables and employ one-hot encoding, as described in Section IV-B, to represent these variables. In order to demonstrate the effectiveness of this approach, we conducted an ablation study by comparing the training results with and without one-hot encoding, as depicted in Fig. 3. The results clearly indicate that the absence of one-hot encoding, despite allowing the model to converge quickly, leads to inferior performance.

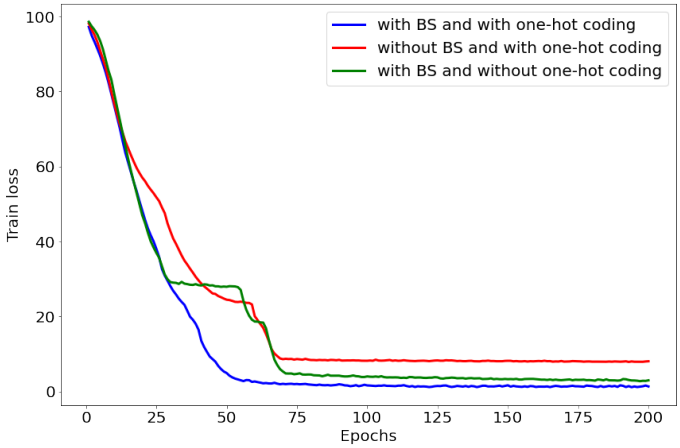


Fig. 3: Experiment Result

C. Base Station

As mentioned before, we noticed that some base stations in the testing set do not exist in the training set. As a result, we considered whether to use the information of the base station in training. As shown in Fig. 3, we found that using base station information can help to improve performance. We also evaluated it by submitting the results to the contest website and found that when using base station information, the model achieved a better score.

VII. Conclusion

In this paper, we proposed a hybrid feature selection method and an attention-based neural network to model the energy consumption of 5G base stations. Our method can capture the impact of different parameters and energy-saving techniques on the energy efficiency of the network.

We evaluated our method on different base station products and configurations. Our method achieved low mean absolute percentage error (MAPE) across all scenarios, demonstrating its accuracy and generalization ability.

Our method can provide a foundation for future optimization of the energy efficiency of 5G network. If we could get access to the ground truth of test set in the future, we plan to extend our method to reveal quantitative relation between base station parameters and energy efficiency and energy consumption.

References

[1] Mamta Agiwal, Abhishek Roy, and Navrati Saxena. Next generation 5g wireless networks: A comprehensive survey. *IEEE communications surveys & tutorials*, 18(3):1617–1655, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Emil Björnson, Luca Sanguinetti, Jakob Hoydis, and Mérouane Debbah. Optimal design of energy-efficient multi-user mimo systems: Is massive mimo the answer? *IEEE Transactions on Wireless Communications*, 14(6):3059–3075, 2015.

[4] D. López–Pérez, A. De Domenico, N. Piovesan, X. Geng, H. Bao, and M. Debbah. Energy efficiency of multi-carrier massive mimo networks: Massive mimo meets carrier aggregation. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 01–07, 2021.

[5] Nicola Piovesan, David Lopez-Perez, Antonio De Domenico, Xinli Geng, and Harvey Bao. Power consumption modeling of 5g multi-carrier base stations: A machine learning approach, 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[9] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.