4G/5G

Introduction of Cell level multi-index Prediction project

Ironhead-5

China Mobile Jiangxi Ltd.

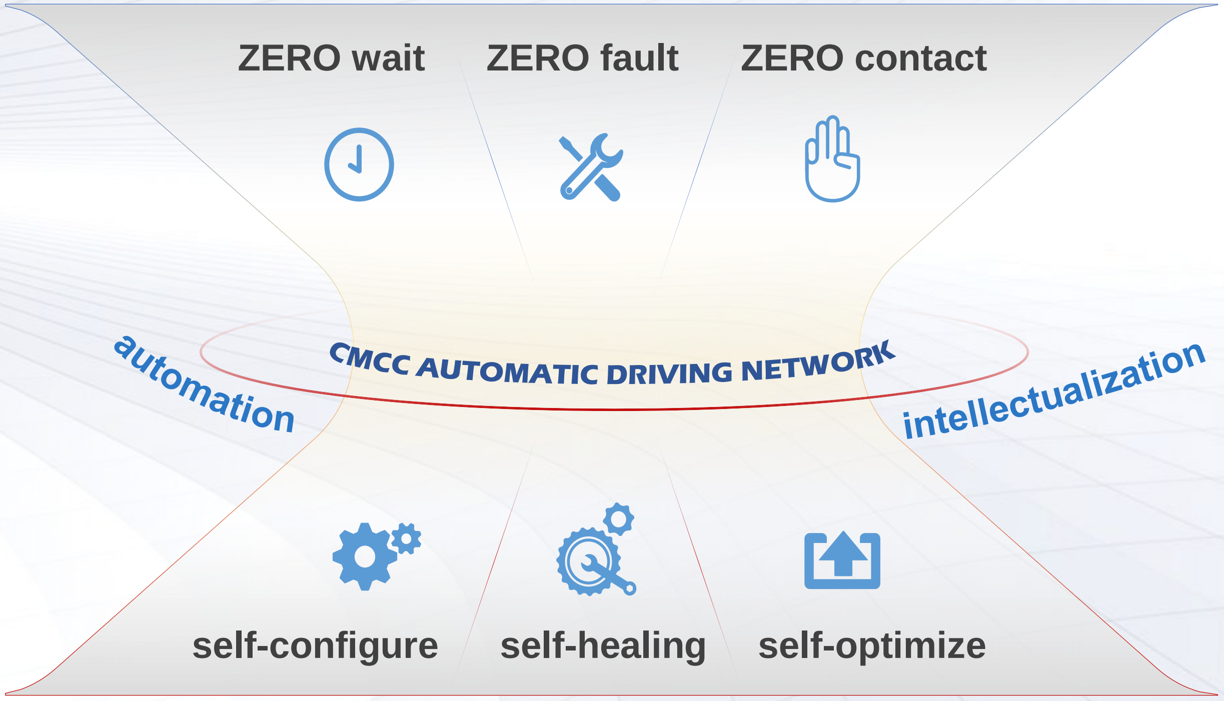
Sponsored by
과학기술정보통신부
Ministry of Science and ICT
XILINX
Organized by
ITU
AI for Good
Machine Learning in 5G Challenge

CONTENTS

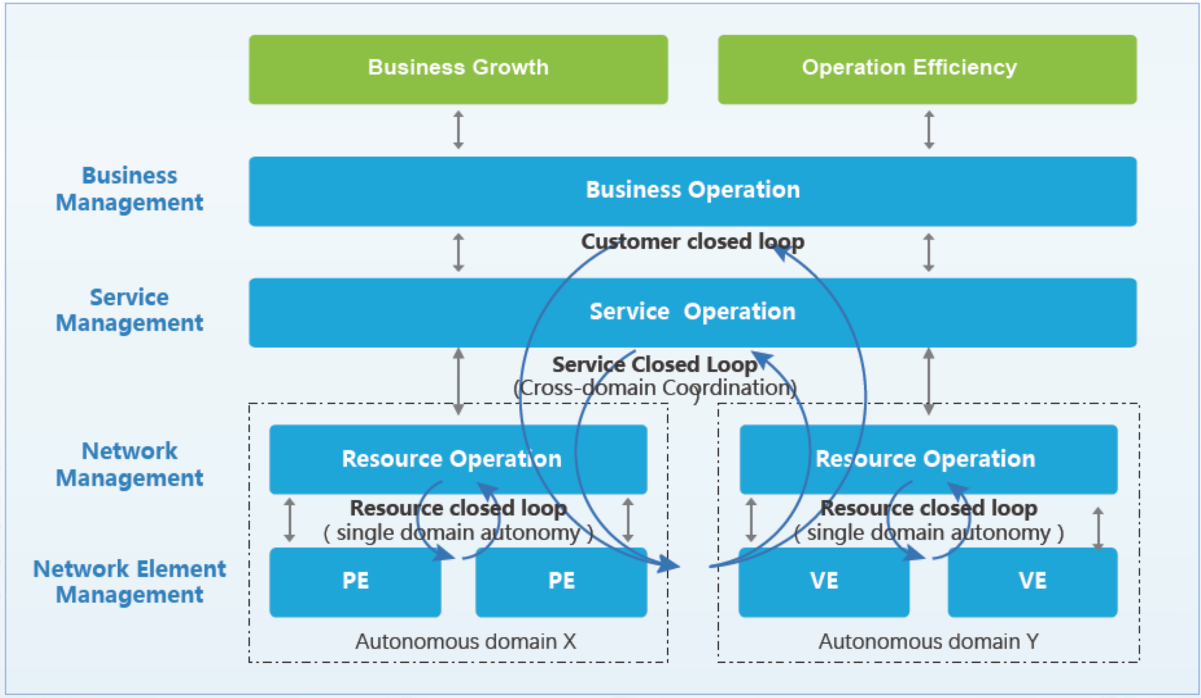
BACKGROUND Vision Goal of China Mobile's Autonomous Networks

- China Mobile issued **2021 Automatic Driving Network White Paper** for digitalization and intelligentization of network operation and maintenance.
- **Autonomous Networks(AN),** to provide consumers services of “zero wait, zero fault, and zero contact”, to build network autonomy with “self-configure, self-healing, and self-optimize” for intelligent operation and maintenance.

VISUAL GOAL



China Mobile’s 3-Zeros & 3-Selfs features Auto. Network



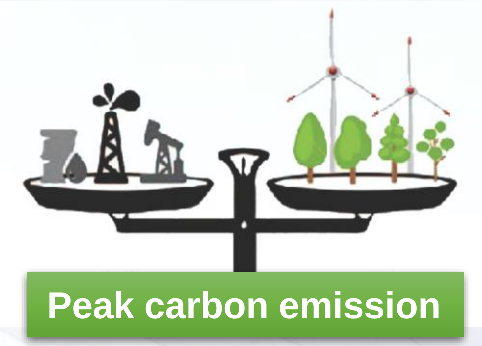
China Mobile’s Autonomous Network Framework

- Overall goal for China Mobile’s autonomous network to reach Level 4 by 2025, based on the TM Forum network autonomy framework.

BACKGROUND Importance of Multi-index Prediction in Radio Network Analysis

- This project mainly use six RF network indicators, including the uplink utilization rate of PUSCH, downlink utilization PDSCH, downlink utilization PDCCH, RRC connections, uplink and downlink traffic volume.
- By reasonable/proper tuning models in timing prediction, project has been applied in **intelligent power saving, capacity carrier automatic scheduling, 4/5G network cooperative working** in the network of China Mobile.

Intelligent Power Saving



- 5G wireless base station brings high electricity bills.Engineers used to tuning the parameters by manual.
- Through intelligent analysis based on timing forecasts, choosing to intelligently shut down cells can greatly reduce power costs and respond to peak carbon emission target.

Capacity Carrier Auto. scheduling

- Limited 4G wireless carrier resource and resource balance is required in newly area and residence.
- Through capacity auto scheduling system, automatic carrier scheduling is realized instead of manual, resource utilization rate is improved, and equipment expenses are saved.

4/5G network cooperative working

- There are tens of thousands of 45G base stations on the existing network, and mutual coordination and optimization must be combined with intelligent analysis methods to meet customer perception needs in a timely manner.

The main difficulty

Model

How to choose a large number of time prediction models, from traditional linear models, BOOST and other machine learning models, to deep learning CNN/RNN/GRU/LSTM and various variants.

Hardware performance

Training servers have limited GPU usage time and memory, so need choose a reasonable performance and optimization plan.

Training data selection

The original data has 6 months of all the indicators among all the cells, how long is the selection of time period and how many of the training data of the cells are reasonable.

Outliers

(irrational or sudden indicators)

The indicator may have negative value and also exceeds 100%; And also encounter people sudden gathering/ group activities caused sudden rises by system fault or congestions.

Missing data

There are different numbers of time granularity indicators missing in most of the cells due to OMC Integrity not reach 100%.

MODEL PROCESS (reasons for WAVENET model selection)

- Compared to some other trial models, the single-feature WAVENET model has faster training speed and higher accurancy in calculation.

```
model = keras.models.Sequential([
    keras.layers.LSTM(10, return_sequences=True, input_shape=[Non
    keras.layers.LSTM(10, return_sequences=True),
    keras.layers.LSTM(10, return_sequences=True),
    keras.layers.LSTM(10, return_sequences=True),
    keras.layers.TimeDistributed(keras.layers.Dense(7))
])
early_stopping_cb = keras.callbacks.EarlyStopping(patience=3,
                                                  restore_best_we
def last_time_step_mse(Y_true, Y_pred):
    return keras.metrics.mean_squared_error(Y_true[:, -1], Y_pred
model.compile(loss=creat_mape_handel(ins_med,1), optimizer="adam"
history = model.fit(X_train_dnn, Y_train_dnn, epochs=50,
                    validation_data=(X_valid_dnn, Y_valid_dnn),
                    callbacks = [early_stopping_cb])
```

4-layer LSTM over 250us

```
Train on 82740 samples, validate on 35460 samples
Epoch 1/50
82740/82740 [==============================] - 30s 363us/sample -
723 - val_last_time_step_mse: 8.3372e-05
Epoch 2/50
82740/82740 [==============================] - 24s 286us/sample -
```

```
model = keras.models.Sequential()
model.add(keras.layers.InputLayer(input_shape=[None, 1]))
for rate in (1, 2, 4, 8) * 2:
    model.add(keras.layers.Conv1D(filters=80, kernel_size=2, pad
                                  activation="relu", dilation_ra
model.add(keras.layers.Conv1D(filters=7, kernel_size=1))
early_stopping_cb = keras.callbacks.EarlyStopping(patience=3,res
model.compile(loss="mae", optimizer="adam", metrics=[last_time_s
history = model.fit(X_train_dnn, Y_train_dnn, epochs=50,
                    validation_data=(X_valid_dnn, Y_valid_dnn),
                    callbacks = [early_stopping_cb])
```

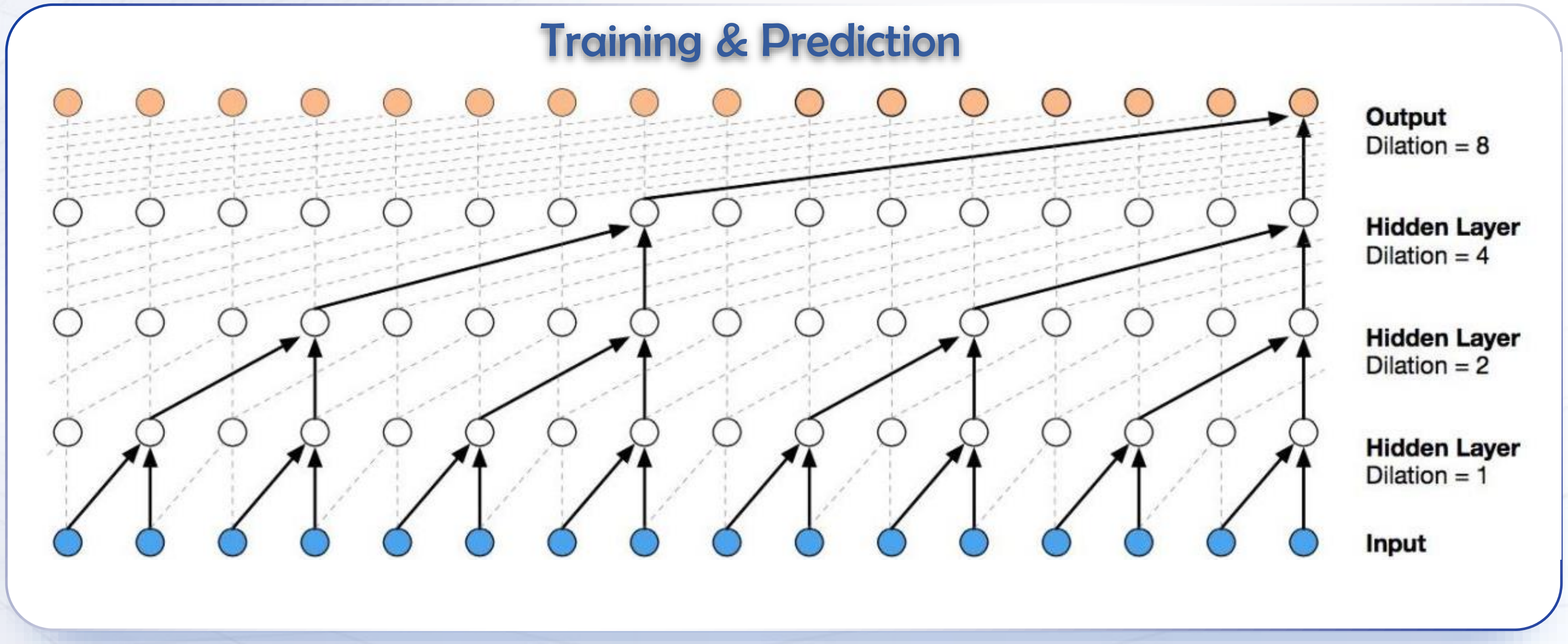
wavenet is close to 130us

```
Train on 82740 samples, validate on 35460 samples
Epoch 1/50
82740/82740 [==============================] - 14s 165us/sample
843e-04 - val_loss: 0.0047 - val_last_time_step_mse: 7.4082e-05
Epoch 2/50
82740/82740 [==============================] - 11s 130us/sample
215e-05 - val_loss: 0.0046 - val_last_time_step_mse: 7.5378e-05
Epoch 3/50
82740/82740 [==============================] - 11s 131us/sample
493e-05 - val_loss: 0.0050 - val_last_time_step_mse: 8.1760e-05
Epoch 4/50
82740/82740 [==============================] - 11s 129us/sample
894e-05 - val_loss: 0.0047 - val_last_time_step_mse: 7.3894e-05
```

✓ In tensorflow, LSTM/GRU has been optimized to be faster than RNN, but it is still much slower than conv1d. When using 1 V100, LSTM/GRU exceeds 3 layers and each sample is about 200-300us, while wavenet is about 100-150us. Use the 4-layer LSTM model training under the same conditions, which took more than 12 hours, and the result is close to WAVENET (0.2569)

✓ Tried to introduce multi-feature prediction, such as CNN+ WAVENET, CNN+LSTM, LSTM+COV2D model, but did not improve the predicted results.

Model process (data selection and cleaning)



Remarks

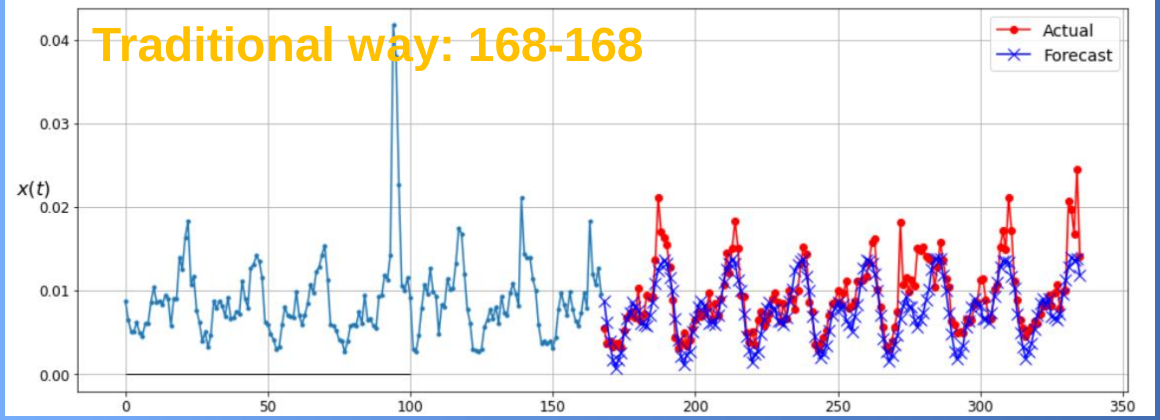
- In the entire model process, the actual cleaning time is longer than training and prediction. The data of 4 cities and 2 formats for 2 months requires 4 cores of 32G + 1 V100 for a total of 6 hours. When the actual project is implemented, daily data can be stored at any time. Cleaning is pre-processed to change the model training prediction lead time to 0.
- Training data input model WAVENET, using two layers of 1/2/4/8 diffusion rate model, input and output is sequence to sequence mode.
- Training data characteristics from May 11th to May 31st,total 21 days data, training data label as July 1st - June 7of 7 days data, predict data characteristics from June 10th - June 30th, 21 days . Finally, predict data on July 1 - July 7.

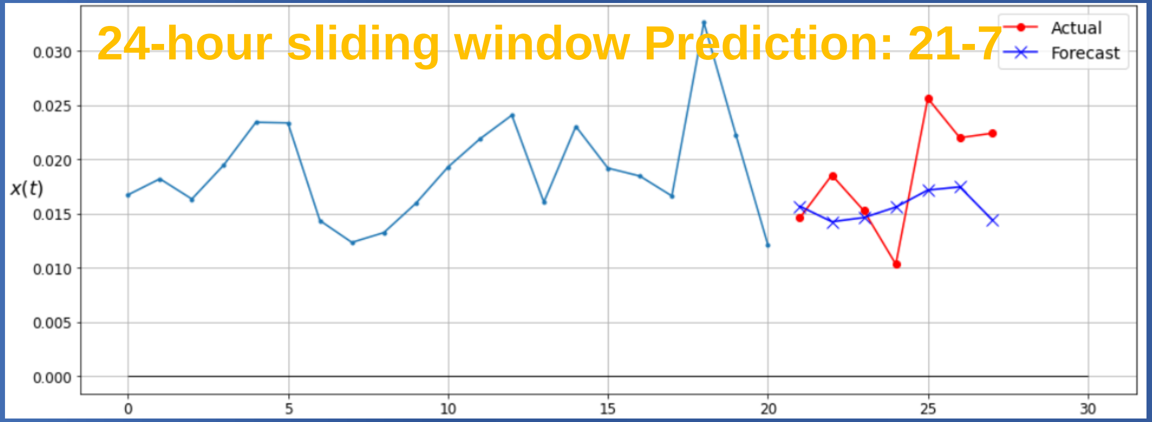
TECHNOLOGICAL INNOVATION (1.n*24-hour sliding window Prediction)

Our Method Traditional models using 24-hour * N days to predicted 24 hours * 7 days. Our models using 1 hour * n days in each fixed time point to predicts 1 hour * 7 days to get better result.

	Model Feature Section	Model Label Section
Traditional Model	24 Hours * n-days	24 Hours * 7-days
N*24 Hour Sliding Windows Prediction Model	1 Hour * n-days（0:00）	1 Hour * 7-days（0:00）
	1 Hour * n-days（1:00）	1 Hour * 7-days（1:00）
	1 Hour * n-days（2:00）	1 Hour * 7-days（2:00）
	1 Hour * n-days（3:00）	1 Hour * 7-days（3:00）
	...	...
	1 Hour * n-days（23:00）	1 Hour * 7-days（23:00）

Technological innovation (1.n*24-hour sliding window Prediction) cont.





Actual effect by using n = 21 days prediction mode

- Compared to the traditional model, in the case of Batch_Size, the memory occupancy per batch is greatly reduced, especially in sequence to sequence mode, improved, reducing the hardware requirements of the model arrangement.

- The 24-hour cycle distribution characteristics of the wireless capacity timing indicator are obvious, and the training error is smoother by directly comparing the feature attributes of the cycle to the model, and the prediction is more accurate. The comparison of WAVENET models were used for 168 hours -168 hours, and the MAPE score was only 0.2665 (about 4% difference).

TECHNOLOGICAL INNOVATION (2. Improvement of loss function)

Commonly used regression loss functions

MAPE, MAE, MSE, SMAPE

MAPE

The MAPE (average absolute percentage error) loss function for competition is too large to response to 0 and cannot be used directly. Actually, the predicted values of the actual verification training model are basically all 0;

MSE

MSE (mean squared error) squares the error, and penalizes too much for large errors, which is contrary to the concept of insufficient prediction of the bias of the use of MAPE in the competition, and the actual verification has poor accuracy of MAPE;

MAE

MAE (Mean Absolute Error) does not penalizes too much for outliers, but the gradient of MAE for neural network updates is still the same, even for a small loss value, the gradient is very large. It is also inconsistent with MAPE's notion of insufficient bias prediction. Actual verification is better than MSE, but there is still room for improvement.

SMAPE

SMAPE (Symmetric Mean Absolute Percentage Error) still has the problem of too large response to 0 value or extreme value, and the actual verification effect is very poor (off 0 value or extreme value).

Original MAPE formula

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

MAPE

Original MAPE Tensorflow representation

error = y_pred - y_true
mape =tf.abs(error)/tf.abs(y_true)

MAPE

Improved MAPE Tensorflow representation

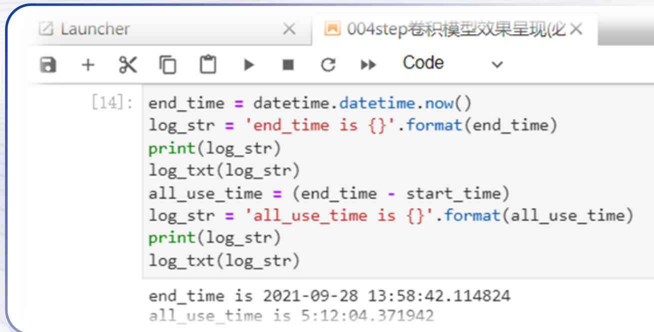
mape_handel =tf.abs(error)/tf.clip_by_value(y_true, low_discrete,1)

- Among them, low_discrete is set to np.median(X_train) by comparison, which can converge and achieve a good prediction effect.

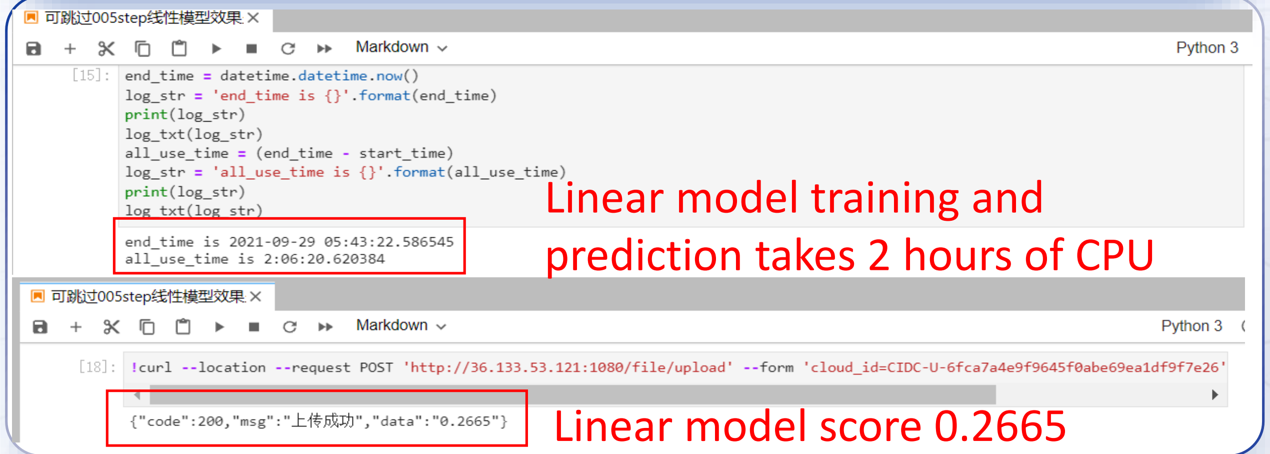
APPLICATION EFFECT

Through effective data cleaning and innovation techniques, even if the linear model can be obtained, it is also possible to obtain a better grade of MAPE 0.2665, and only the CPU training and predicting all the city, all the cells are only about 2 hours.

The WAVENET model got a better result compared use the other linear models.



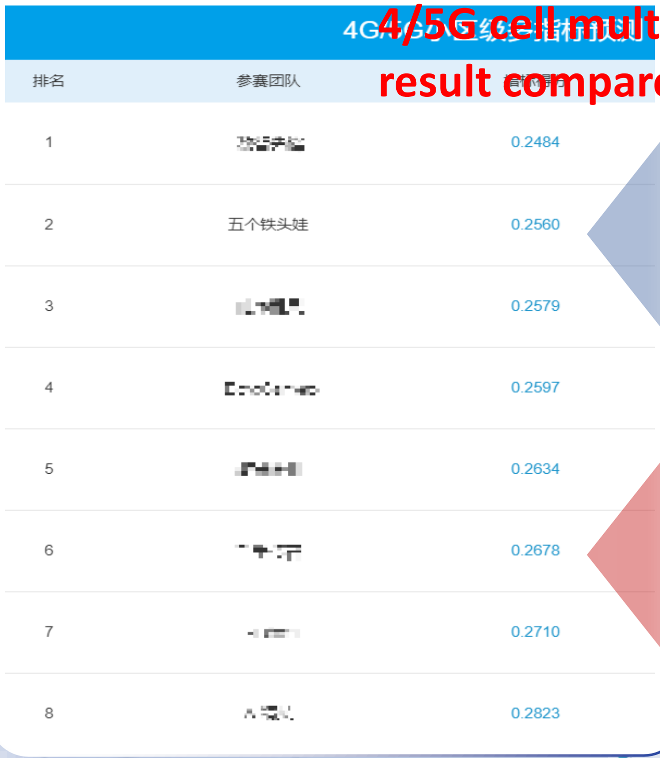
In the program ,4 cities, 6 indicators in all cells, to generate 48 reports respectively, requiring 4 cores 32g + 1 V100 training and predictive sharing for 5 hours, and the prediction results are 0.2560-0.2565.



Linear model training and prediction takes 2 hours of CPU

Linear model score 0.2665

4/5G cell multi-index Prediction result compared with other models



WAVENET model 0.2560

Only use the linear model 0.2665 to directly enter the preliminaries 6

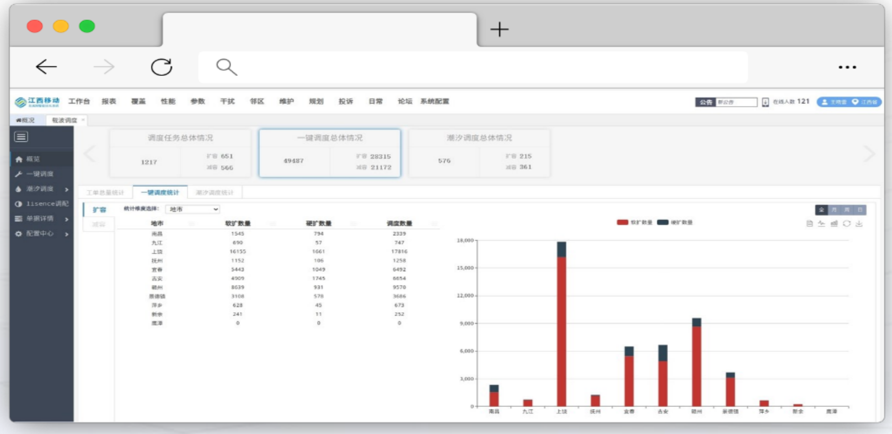
Value benefits

- China Mobile uses multi-index Predictions to promote the implementation of **AUTONOMOUS NETWORKS** in wireless network capacity analysis, and extremely pushes to achieve the full realization of L4 network autonomous driving capabilities in 2025.
- In the intelligent module of network capacity analysis, most of the provinces have implemented related applications such as intelligent power saving and automatic carrier dispatching functions, making outstanding contributions in terms of carbon peaking, cost saving and efficiency improvement.

China Mobile Intelligent Power Saving Platform

- ✓ Combine this project with China mobile's Jiutian Planform, created China Mobile's AI Power Saving Platform in March 2021 and achieved O&M automatically.
- ✓ By predicting low-traffic areas and time periods for shutdown operations, has saved about 30 million CNY in electricity bills in province, which is equivalent to reducing carbon dioxide emissions.
- ✓ This platform has improved the efficiency about 10% and been awarded for national innovation and state prize number more than 10.

China Mobile Carrier Automatic Scheduling System

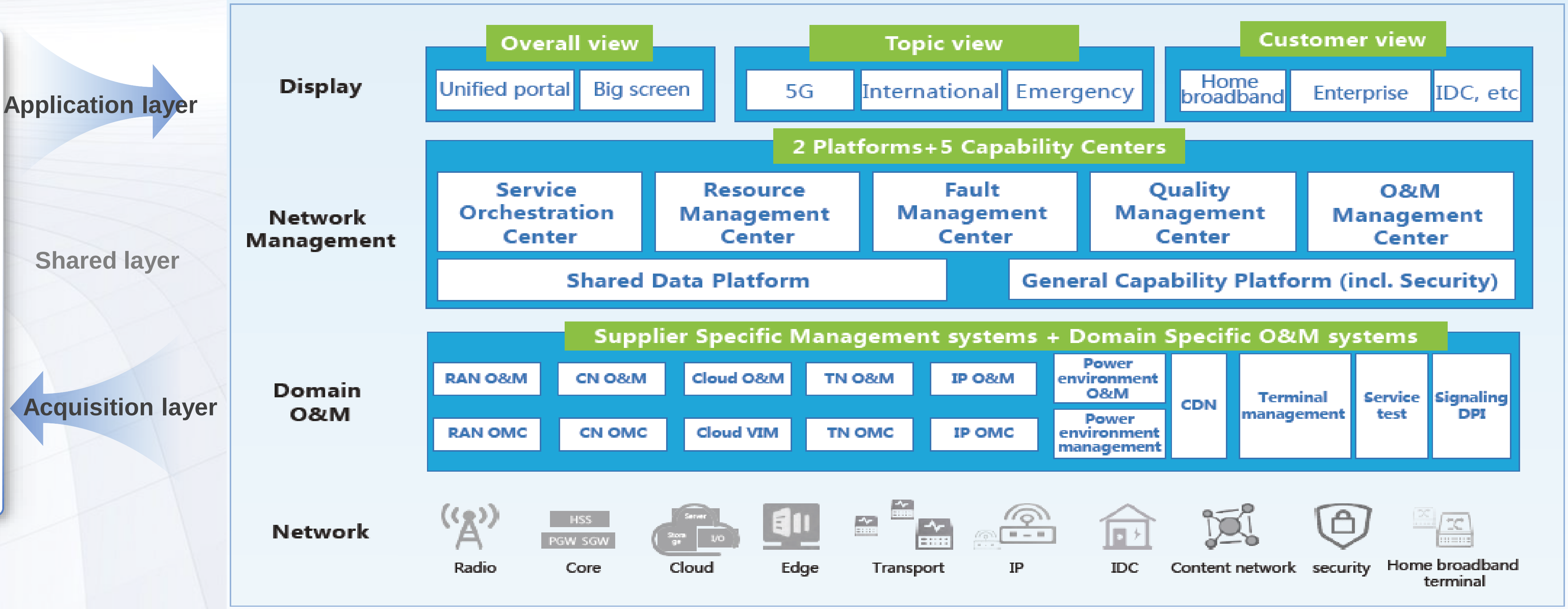


- ✓ Our province relies on the development of its own big data platform, which has been launched in June 2020.
- ✓ By predicting the dismantling of low-traffic cells and time periods, and the expansion of high-load cells, our province has automatically used platform capability for 49,000 times, which is equivalent to saving 9.8 million yuan in labor/license fee.

Application Prospect Strengthen the Capability of China Mobile's Autonomous Networks

- China Mobile's Autonomous Networks are based on the "2+5+N" system in terms of wireless network operation and maintenance. On the one hand, 2 core platforms and 5 capability centers are built to pull through the O&M process, complete the automation configuration activation, consolidate the data base, strengthen the end-to-end services guarantee, and achieve the reuse and sharing of capabilities.
- Among the core modules of "2+5+N", the intelligence of multi-index Prediction are playing important role in the further strengthened such as artificial intelligence, big data application in the network.

THE 4G/5G cell-level multi-index prediction module obtains the original data from the "2" platform, and provides the necessary support for the six capability centers' core modules for autonomous driving



THANK YOU

Ironhead-5

China Mobile Jiangxi Ltd.