Cellular Traffic Forecasting and Analysis with Efficiency Transformer

by Hexuan Weng

Thesis submitted in fulfilment of the requirements for
the degree of

Master of Analytics (Research)

under the supervision of Prof. Ling Chen, Dr. Yanbin Liu

University of Technology Sydney
Faculty of Engineering and Information Technology

December 2023

Certificate of Original Authorship

I, *Hexuan Weng*, declare that this thesis is submitted in fulfilment of the requirements for the award of *Master of Analytics (Research)*, in the *Faculty of Engineering and Information and Technology* at the *University of Technology Sydney*.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Hexuan Weng

Date: 22 December, 2023

Statement of Authorship Attribution

This thesis contains publications as follows,

Publication Weng, H., Liu, Y. and Chen, L., 2023, November. Spatial Bottleneck Transformer for Cellular Traffic Prediction in the Urban City. In Australasian Joint Conference on Artificial Intelligence (pp. 265-276). Singapore: Springer Nature Singapore.

Presented in Chapter 3.

Statement: I co-designed the technique with Yanbin Liu and Ling Chen. We proposed a novel model called *ST-InducedTrans*, to dynamically explore the large geographical correlations (spatial) and periodic variations (temporal). Specifically, a Spatial Bottleneck Transformer is devised to obtain spatial correlations from the most relevant grids in the geographical area, at the cost of linear complexity. Final manuscript was prepared with Dr. Yanbin Liu and Prof. Ling Chen.

Abstract

The purpose of forecasting is to be able to analyse unknown data and make strategic decisions in an accurate, robust and interpretable manner. Time series problems are an integral part of the forecasting problem and use statistics and machine learning modelling to analyse time series data to inform strategic decisions. This research focuses on new solutions to time series problems through novel deep learning models, mainly including time series prediction using mobile traffic data.

Cellular networks have witnessed an exponential growth in data traffic due to the proliferation of mobile devices and the increasing demand for high-bandwidth applications. Efficiently managing this burgeoning traffic has become a critical challenge for telecommunication providers. Forecasting and analyzing cellular traffic patterns are crucial for optimizing network performance, resource allocation, and ensuring a seamless user experience. In recent years, deep learning techniques, especially Transformers [45], have emerged as powerful tools for handling the complexity and dynamic nature of cellular traffic data.

This thesis presents a comprehensive review of statistical methods, machine learning-based methodologies and deep learning-based methodologies applied to cellular traffic forecasting and analysis. It provides an overview of traditional forecasting techniques and highlights the limitations that have prompted the adoption of deep learning models. Various deep learning architectures such as recurrent neural networks (RNNs) [47], long short-term memory networks (LSTMs) [25], convolutional neural networks (CNNs) [53], and hybrid models are examined in the context of their applications for traffic prediction and analysis. Furthermore, this thesis discusses the challenges and opportunities associated with employing deep learning in cellular traffic management. Issues such as data heterogeneity, scalability, interpretability, and real-time processing constraints are pointed out, along with potential solutions and future research directions.

There are two main work in our research. In the Chapter 3, we mainly focus on how to enhance the efficiency of the cellular traffic forecasting. Due to the widespread use of portable devices and the advancement of 5G technology, we have received a significant amount of mobile data, which requires prediction models for cellular traffic data. However, forecasting mobile traffic data efficiently is challenging due to the complex spatial and temporal correlations, especially when the mobile data comes from a large geographical area. To tackle this challenge, we propose a new model, called *ST-InducedTrans*, to dynamically explore the large geographical correlations (spatial) and periodic variations (temporal). Specifically, a Spatial Bottleneck Transformer is devised to obtain spatial correlations from the most relevant grids in the geographical area at the cost of linear complexity. For the temporal blocks, we embed the elaborately selected temporal clues into a temporal Transformer to offer useful temporal prompts for cellular prediction. Finally, several spatial and temporal blocks are effectively stitched into a whole model for complementary cellular traffic prediction. In Chapter 4, we add more cross-domain datasets into the cellular traffic data to improve prediction accuracy, including base station geographical location, POI distributions and social activity information. In both Chapter 3 and 4, We conducted comprehensive experiments on the public real-world cellular data from Telecom Italia, Milan [10]. Results show that our model outperforms the state-of-the-art methods on three metrics (MAE, NRMSE, and R^2) at the cost of lower time complexity.

Overall, this thesis consolidates the current state-of-the-art in utilizing deep learning, mainly Transformer, for cellular traffic forecasting and analysis, highlighting its potential to revolutionize how telecommunication networks anticipate and manage traffic demands.

Acknowledgements

I extend my heartfelt appreciation to everyone who contributed to the completion of this thesis.

Foremost, I am immensely grateful to my supervisor, **Prof. Ling Chen**, whose unwavering guidance and dedicated time throughout my *Master of Analytics (Research)* were invaluable. Her continuous support and insightful feedback were instrumental in bringing this thesis to fruition.

I am indebted to my co-supervisor, **Dr. Yanbin Liu**, whose constructive guidance and invaluable suggestions significantly shaped both my research and professional growth.

My heartfelt acknowledgments also go to **Changlu Chen, Chaoxi Niu, Dr. Yang Lin, Zihan Zhang, Brodie Skriveris** for their unwavering support and enlightening discussions.

To my family, whose unwavering support and understanding sustained me during the most challenging moments of my life, I am forever grateful.

Lastly, I extend my appreciation to the wider research and machine learning communities. Their invaluable research contributions and open-source resources have played a pivotal role in shaping and enriching the content of this thesis.

Contents

Chapter 1

Introduction

1.1 Introduction

The realm of time series forecasting dilemmas has garnered significant interest among researchers and scientists. The establishment of the International Association of Forecasters (IFF) occurred 25 years ago, specifically addressing the intricacies of time series forecasting [8]. Preceding 2005, a considerable volume of approximately 1000 papers focused on time series forecasting, leading to substantial advancements in research. Time series data typically exhibits distinct traits, showcasing trends that depict overarching patterns over time and seasonality, representing fluctuations in behavior within specific periods.

Cellular traffic data typically adhere to a time series format and exhibit distinct periodic patterns [45]. Over the last decade, the demand for 5G has surged, leading to a notable surge in mobile data traffic. Enhancement of cellular networks spans various aspects, including time, space, frequency, energy efficiency, and advanced signal processing techniques, offering avenues for eco-friendly initiatives and energy conservation. Accurate prediction in cellular prediction stands as a pivotal necessity.

Owing to its recurring patterns, temporal aspects play a crucial role in mobile traffic prediction. Previous studies have extensively employed Autoregressive Integrated Moving Average (ARIMA) and Recurrent Neural Networks (RNN) for forecasting cellular traffic. These models, designed for time series analysis, primarily emphasize the utilization of temporal elements [33]. ARIMA, a statistical technique, is unsuitable for nonlinear data, while RNNs face limitations in effectively handling extended time series information. The sequential nature of RNN operations results in prolonged processing times when managing extensive time series datasets.

In the spatial domain, Convolutional Neural Networks (CNNs) [53] and Graph Neural Networks (GNNs) [59] are commonly employed to capture spatial correlations. CNNs exploit diverse filter windows to capture spatial information, yet they are constrained by the distance spanned by geographical locations. For instance, long-range spatial correlations between distant grids may not be adequately captured. GNNs require an adjacency matrix and face privacy concerns when acquiring detailed data to construct this matrix [45]. Various statistical approaches, such as Pearson Correlation, are utilized to compute the adjacency matrix.

To circumvent the constraints posed by the limited receptive field in CNNs and the necessity for an adjacency matrix in GNNs, our study introduces an innovative model centered on the Transformer architecture. The Transformer model offers a broader receptive field and is adaptable to parallel computation. Our primary goal is to dynamically capture intricate spatial correlations, thereby enhancing the efficiency of the forecasting process. Additionally, our aim is to augment the precision of cellular traffic forecasting by leveraging this novel approach.

Expanding upon our objectives, we endeavor to explore the Transformer's capacity to capture nuanced spatial dependencies across cellular traffic data. By leveraging its inherent strengths in handling extensive contextual information, we seek to revolutionize the predictive accuracy of cellular traffic patterns. Furthermore, our research aims to validate the efficacy of this Transformer-based model by conducting comprehensive evaluations against existing state-of-the-art methods. Through these endeavors, we strive to not only enhance prediction accuracy but also pave the way for more efficient and adaptable forecasting frameworks in the realm of cellular traffic analysis.

1.2 Background

The realm of cellular traffic forecasting has garnered considerable interest recently, drawing attention from a multitude of researchers. The diligent efforts of scholars in this domain have culminated in the application of three primary methodologies to address time series challenges: statistical models, machine learning models, and deep learning models. A comprehensive exploration and detailed analysis of the literature review encompassing these methodologies will be presented in Chapter 2. Herein, we provide an overview of these three domains to set the stage for the ensuing comprehensive literature review.

Diving deeper into these methodologies, statistical models encompass a range of techniques leveraging mathematical formulations and historical data patterns to infer future trends. Machine learning models, on the other hand, harness algorithms and computational techniques to discern patterns and make predictions based on training data. Lastly, deep learning models, characterized by multi-layered neural networks, excel in capturing complex relationships within data, particularly in time series contexts. This section aims to provide a foundational understanding before the forthcoming literature review.

Statistical Models Cellular traffic forecasting, within the domain of statistical models, has undergone significant evolution over time. Early studies in time series analysis introduced the Exponential Smoothing method, marking a foundational step in addressing time series problems [12]. This method laid the groundwork for subsequent advancements, notably the Autoregressive Integrated Moving Average (ARIMA) model, which emerged as a prominent tool for time series forecasting [11].

Moreover, advancements building upon Snyder's work [60] led to the development of State Space Models within the context of cellular traffic prediction [21]. These models presented an advantageous solution, particularly for non-linear exponential smoothing methods [52]. However, one of the limitations of State Space Models in cellular traffic forecasting lies in their inability to capture intricate time series correlation patterns [42].

Machine Learning Models The escalating diversity and intricacy within mobile traffic data have imposed limitations on traditional statistical methods. Consequently, there has been a notable surge in research endeavors focused on leveraging machine learning techniques to enhance time series predictions, thereby overcoming these limitations and yielding more accurate forecasts [54]. Machine learning approaches enable the extraction of pertinent temporal and spatial insights from traffic data.

In the realm of machine learning algorithms applied to cellular traffic prediction, a variety of methodologies have been employed. Traditional machine learning models, particularly shallow neural networks featuring feed-forward multi-layer architectures, have demonstrated efficacy in model training and algorithm learning, resolving numerous time series challenges to a certain extent.

However, the exponential growth in mobile traffic data volume has resulted in escalating data sizes and heightened data dimensionality [22]. Consequently, traditional machine learning tools, including shallow neural networks, face limitations in handling these large-scale datasets. As a result, specific problems within cellular traffic prediction become exceedingly intricate or impractical for conventional machine learning models [76].

Deep Learning Models Numerous researchers have turned to deep learning methodologies to address the challenges posed by extensive time series problems, a response to the limitations encountered in traditional machine learning techniques. Within the realm of deep learning models, prominent architectures include Recurrent Neural Networks (RNN) [47], Convolutional Neural Networks (CNN) [53], and Transformer-based models [66].

RNN models excel in handling sequential data, notably time series information, and have been extensively researched for capturing temporal dependencies. However, they face challenges related to gradient instability, especially when predicting prolonged time series data.

CNNs, another prevalent deep learning architecture, effectively captures spatially relevant information through convolutional filters without relying on prior features. Zhang's study [77], for instance, employs LSTM to process temporal data and CNN to handle spatial information, subsequently merging the outcomes.

The Transformer model, a novel architecture reliant solely on attention mechanisms, offers enhanced parallelism and quicker convergence rates [66]. In time series problems, Transformers effectively gather relevant information using self-attention mechanisms. Notably, the Transformer-based approach has showcased promising outcomes in dealing with mobile traffic data.

While RNNs, CNNs, and encoder-decoder models remain foundational in sequence modeling, they tend to be more complex. In contrast, the Transformer model's reliance on attention mechanisms provides superior parallelism and has demonstrated noteworthy efficacy, particularly in achieving robust results within shorter timeframes. The success of Transformer-based models in mobile traffic data underscores their potential in addressing complex time series problems.

1.3 Research Aims and Outline

1.3.1 Research Aims and Plan

My primary research goal involves the utilization of authentic cellular traffic data to develop innovative and efficient deep learning models. Grounded in this research theme, I've structured my investigation into two key phases, each with its specific objectives.

Stage 1: explore spatial correlation dynamically and build spatial-temporal transformer for cellular traffic forecasting task. Accurate prediction outcomes hold significant potential for refining precision traffic engineering, optimizing demand-aware network resource allocation, and enhancing public transportation systems [3]. Prior research commonly relied on adjacent matrices to discern spatial correlations, an approach notorious for its computational intensity in calculating inter-grid correlations. To mitigate this challenge, we introduce a novel model named the *Spatial Bottleneck Transformer*. This innovative model aims to capture spatial dependencies efficiently, thereby enhancing prediction accuracy while drastically reducing computational complexity from quadratic to linear levels. The proposed Spatial Bottleneck Transformer can seamlessly integrate within temporal blocks, offering versatility in its application. In this stage of our research, our primary objective is to construct a transformer-based encoder and decoder structure explicitly tailored to elevate the precision of cellular traffic forecasting tasks. This strategic focus aligns with our goal of advancing predictive accuracy while addressing computational challenges associated with spatial correlation capture in cellular traffic modeling.

Research question: *Given real-world cellular traffic data in the urban city, how can we develop a novel model to achieve more accurate forecasting results and reduce time complexity considering more advanced spatial dependency?*

Stage 2: develop a deep learning model based on Transformer with cross-domain big data. Deep learning has demonstrated superior predictive capabilities in traffic modeling compared to traditional methodologies. However, the intricacies of cellular traffic forecasting extend beyond mere reliance on spatial and temporal factors, encompassing various external influences impacting traffic generation. Consequently, our current focus revolves around exploring the transformer attention model's applicability in incorporating both internal and external factors into the forecasting process. These factors include but are not limited to base stations, Points of Interest (POI), and social activities. In this stage of our research, we aim to leverage real-world industrial data to investigate the integration of cutting-edge deep learning techniques. The primary objective is to develop a comprehensive framework that accounts for both internal and external factors, harnessing the capabilities of the transformer attention model. By incorporating diverse external factors such as base stations, POI, and social activities, we endeavor to enhance the predictive accuracy and contextual relevance of cellular traffic forecasting models. This phase represents a strategic progression in our research trajectory, aiming to bridge the gap between theoretical advancements and practical applicability by utilizing real-world industrial data and state-of-the-art deep learning methodologies.

Research question: *Given real-world cellular traffic data with cross-domain big data, how can we leverage and transfer those external factors into the model and improve comparable cellular traffic forecasting accuracy?*

1.3.2 Report Outline

This thesis comprises publications and is structured as follows:

Chapter 1 sets the context by presenting the research background, delineating the research questions, and defining the objectives.

Chapter 2 delves into a comprehensive literature review on time series prediction within cellular traffic, encompassing successful methods in Statistical, Machine Learning, and Deep Learning domains. This chapter also highlights the research gap existing between prior studies and current challenges.

Chapter 3 focuses on advancing cellular traffic prediction through novel transformer-based models aimed at refining both efficiency and accuracy. The chapter introduces an innovative model designed to enhance forecasting outcomes by capturing spatial correlation information.

Chapter 4 introduces the integration of cross-domain big data as external information to augment the precision of cellular traffic prediction results. This chapter emphasizes the amalgamation of cross-domain data, such as base station, Points of Interest (POI), and social media information, with cellular traffic data as a primary focal point.

Chapter 5 concludes the thesis, summarizing our contributions to cellular traffic forecasting and proposing potential avenues for future research within this domain.

Chapter 2

Literature Review

In Chapter 1, we provided an overview of the historical evolution of time series forecasting, delineating three primary domains: statistical methods, machine learning methods, and deep learning methods. Notably, deep learning methods are categorized within machine learning approaches. However, for the sake of our research focus on investigating and resolving time series forecasting challenges through deep learning frameworks, we have specifically delineated deep learning as a distinct segment within the literature review. Moving forward, the subsequent section will systematically introduce statistical methods and deep learning techniques.

2.1 Statistical Methods

Since the 1950s, statistical models have played a crucial role in time series forecasting tasks. Initially, these methods didn’t garner significant attention. However, with the increasing significance of time series problems, there emerged a necessity to comprehend the intrinsic relationships within time series data—such as trends, seasonality, and autocorrelation. This understanding was pivotal in achieving more precise predictions for future time series outcomes. This segment of the literature review will delve into key statistical methodologies that have substantially influenced subsequent research, notably Exponential Smoothing (ES) [12], Autogressive Integrated Moving Average (ARIMA) [11], and State Space Model (SSM) [37].

2.1.1 Exponential Smoothing

The Exponential Smoothing method, originating around 1950, represents one of the earliest time series techniques. Despite its inception in the 1950s, this method initially didn't attract considerable attention from statisticians [12][26][69]. First introduced by Brown (1959), it was later recognized as a variant of Moving Average, albeit lacking considerations for trend and seasonality. Its fundamental formula $S_t = a \cdot y_t + (1 - a)S_{t-1}$ involves S_t denoting the smoothed value at time t, y_t representing the actual value at time t, S_{t-1} indicating the smoothed value at time $t - 1$, and a as the smoothing constant, typically within the range of [0, 1].

Muth and John were among the pioneers formalizing the statistical underpinnings of Exponential Smoothing [50], while Pegels et al. highlighted the existence of seasonality and trend features in time series problems [16]. Gardner and Everette conducted a seminal review of Exponential Smoothing, classifying published articles, significantly boosting research interest in this method [7]. Snyder's proposal of an effective Exponential Smoothing model in spatial contexts laid a foundation for subsequent studies [60]. Taylor introduced a new statistical model rooted in Exponential Smoothing, building upon previous empirical findings [65]. Hyndman expanded on Taylor's approach, categorizing time series problems into distinctive trends and seasons to delineate crucial trend and seasonal characteristics [28]. While Exponential Smoothing models suit linear data with evident trends or seasonality, their performance tends to falter with non-linear data patterns.

2.1.2 AutoRegressive Integrated Moving Average

During the early phases of time series modeling, significant attention was dedicated to core assumptions and assessments, including stationarity and autocorrelation. These foundational principles notably contributed to shaping classical time series models like the Autoregression Method (AM) [75], Moving Average (MA) [75], and Autoregressive Integrated Moving Average (ARIMA) [11]. These statistical methodologies employ time as the independent variable and the sequence values as dependent variables, constructing regression models based on historical data. Specifically, autoregressive models and autoregressive moving average models are adept at capturing temporal structures inherent in the data [11].

Numerous studies [24, 63] have highlighted the advantages of integrating leading indicators into statistical models to bolster forecasting accuracy. Seasonal ARIMA adeptly manages multiple seasonal cycles within time series data [73]. Moreover, The ARIMAX model expands upon ARIMA by incorporating explanatory variables, overcoming ARIMA's limitation in capturing seasonal patterns.

Although initially tailored for linear time series, extensions of ARIMA have emerged to model non-linear and non-stationary time series. Miller and Williams introduced shrinkage estimators to enhance forecasting accuracy by integrating multiplicative seasonal factors into ARIMA [48]. Hyndman further explored diverse trend and seasonality relationships in ARIMA modeling [27], while Lee et al. combined ARIMA with genetic programming to model non-linear time series, leveraging their complementary characteristics [41]. ARIMA has used data from Chinese cities for cellular traffic forecasting in some studies [29], while SARIMA has also used accurate Ethiopian data for forecasting [31]. Regarding the base station problem, Zhang [2] proposed to use SARIMA for time series learning and proposed a new model to obtain the influence of the geographical location on the base station utilization for the clustering.

2.1.3 State Space Model

State Space Models (SSMs) constitute a diverse and influential family of classic statistical models widely utilized in time series analysis [37]. They offer a flexible and comprehensive framework for understanding structured temporal data. SSMs stand out for their ability to represent complex temporal patterns and model time series with well-understood structures. Their versatility allows a clear representation of underlying processes within the data. However, despite their rich modeling capabilities, SSMs face limitations. One primary constraint is their inability to effectively learn patterns from multiple related time series simultaneously. Additionally, designing SSMs often necessitates in-depth domain knowledge to appropriately structure and model the time series data.

Several studies demonstrate the diverse applications and adaptations of SSMs in time series analysis. For instance, Dong et al. proposed SSM techniques aimed at stationarizing time series before integrating them into linear models [19]. Douc et al. explored the realm of non-linear State Space Models to enhance time series forecasting accuracy [20]. Additionally, Ives and Dakos introduced innovative time-varying and threshold models using linear SSMs, capturing dynamic changes in non-linear time series [35].

While SSMs offer a versatile and powerful framework for modeling structured time series data, their limitations underscore the challenges in handling multiple related time series and the requisite domain expertise for effective model design. Understanding these strengths and limitations is crucial for utilizing SSMs effectively in time series analysis and addressing various complexities inherent in temporal data modeling.

2.2 Machine Learning Methods

Machine learning methods have played a pivotal role in the analysis and prediction of time series data, offering diverse tools to discern patterns and forecast future trends without relying on deep learning architectures. Traditional machine learning algorithms such as K Nearest Neighbors and Neural Networks have been extensively utilized in time series analysis.

2.2.1 K Nearest Neighbors

K Nearest Neighbors (KNN) is a well-established algorithm in time series forecasting that operates by identifying the top K nearest neighbors within the training dataset and predicting future values based on the average of these neighbors' future values. This method is relatively straightforward to implement and is particularly effective in capturing repetitive patterns frequently observed in time series data. One of the key reasons for the popularity of KNN in time series forecasting is its ability to identify similarities in historical patterns, enabling predictions based on the nearest neighbors' historical information [15].

In many instances, KNN is employed for making non-autoregressive forecasts, where predictions are generated without explicitly considering the temporal order or sequential dependencies in the data. This non-sequential approach is suited for scenarios where the focus is on capturing similarities and patterns across the historical dataset rather than sequential trends [15]. However, this non-autoregressive nature might limit its capability to capture complex temporal dependencies present in certain time series datasets.

To enhance KNN's forecasting accuracy, researchers have developed variations such as Pattern Sequence-based Forecasting (PSF). This variation, as mentioned by [9], aims to address some of the limitations of standard KNN by leveraging pattern sequences within the time series data to facilitate more precise predictions. For instance, Cai et al. [15] improved KNN's performance by incorporating spatiotemporal correlations, leading to enhanced accuracy in multi-step forecasting.

2.2.2 Neural Networks

Neural Networks (NNs) are computational models inspired by the workings of the human brain. They are built upon the backpropagation algorithm, a significant advancement introduced in the study of neural networks [40]. However, this algorithm has its limitations; it may converge to a local minimum and encounter issues like the exploding or vanishing gradient problem. A typical neural network comprises neurons organized into one or more hidden layers and an output layer. The network's weights and biases are initialized randomly and then adjusted through a learning algorithm to minimize error and optimize performance.

In the context of cellular traffic prediction, Neural Networks have been leveraged to capture complex patterns and dependencies within the data. Researchers have explored the use of NN architectures to model the dynamics of cellular traffic more effectively. For instance, Raza et al [57] introduced a framework for electricity load forecasting that combined the predictive outcomes of three distinct neural networks. They trained each individual model using a global particle swarm optimization method, aiming to enhance their overall forecasting performance.

The application of Neural Networks in cellular traffic prediction involves utilizing these architectures to interpret and forecast traffic patterns based on historical data. By employing diverse NN models and optimizing their parameters, researchers aim to achieve more accurate and robust predictions in cellular traffic forecasting tasks. Studies like that of Raza et al [57]. showcase efforts to harness multiple neural network models in tandem, demonstrating the potential of combined frameworks to improve predictive accuracy in forecasting cellular traffic trends.

2.3 Deep Learning Methods

In the age of extensive data, the analysis of time series has evolved as a pivotal component within the realm of AI technology. The fusion of time series analysis with deep learning models has led to the development of more sophisticated and advanced approaches. This section predominantly focuses on exploring previous research conducted on mobile traffic data utilizing the principles of deep learning theory.

2.3.1 Recurrent Neural networks

Recurrent Neural networks (RNNs) [47] Recurrent neural networks' powerful temporal modelling capabilities also play mobile traffic forecasting. RNN is a special type of artificial neural network adapted to work for time series data or data that involve sequences. And this model is essential in natural language processing and time series forecasting. However, RNNs may encounter gradient dispersion and has its limitation when dealing with long-term sequence data. For example, when RNNs deal with long-term sequence data, it can only obtain information from relatively recent sequences but does not have a memory function for earlier sequences, thus losing information. LSTM can be used to forecast long-term sequence data.

As to the mobile traffic data, Azari's research [61] mainly combination of ARIMA method and LSTM to improve the forecasting accuracy. In Kuber's research[34], LSTM is mainly used to build temporal dependent models. Regarding spatial components, other auxiliary information, such as the location of airports, banks, or restaurants, is primarily used for modelling. In the LSTM research, Azari[61] and Kuber[34] mentioned the benefits of better resource allocation through prediction.

2.3.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) [53] have gained extensive traction in the realm of computer vision. The grid-like structure of cellular traffic data, resembling images, has prompted prior investigations into merging convolutional neural networks with recurrent neural networks to forecast mobile traffic patterns. For instance, Zhang et al. [3] proposed a model leveraging Milan's data through a 3D-ConvNet coupled with LSTM for predictive analysis. This fusion of Convolutional Neural Network and Long Short-Term Memory effectively captures both temporal and spatial dependencies inherent in the data. Zhang's STCNet (Spatial-Temporal Cross-domain Neural Network) [76] incorporates metadata, such as time of day, day of the week, and additional relevant data, serving as input to a two-layer neural network comprised of LSTM. This innovative model integrates cross-domain data, enriching the information within the CNN structured with two layers. The resultant CNN-RNN model harmonizes intricate spatio-temporal traffic patterns in mobile traffic prediction, notably enhancing predictive accuracy.

The amalgamation of Convolutional Neural Networks and Long Short-Term Memory is particularly noteworthy in mobile traffic prediction due to their ability to effectively extract temporal and spatial dependencies. Zhang's STCNet [3] significantly leverages this hybrid model by incorporating diverse metadata as inputs to an LSTM-based neural network, enhancing the contextual richness of CNN-based models. By integrating cross-domain data into the CNN structure, Zhang's innovative approach advances the prediction accuracy of complex spatio-temporal traffic patterns in mobile networks. This CNN-RNN fusion not only captures the temporal and spatial aspects of traffic data but also underscores its potential to significantly enhance predictive performance in mobile traffic forecasting tasks.

2.3.3 Encoder-Decoder

The encoder and decoder functions within the framework are pivotal components that transform input data into a desired format and then decode it into the target format, respectively. The versatility of the encoder-decoder paradigm allows the hybridization of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) as both encoder and decoder components, offering flexible configurations adaptable to diverse tasks. For instance, Zhang introduced the Spatio-Temporal Neural Network (STN) model, employing a fusion of Convolutional Long Short-Term Memory (LSTM) and 3D Convolutional Neural Network within the encoder-decoder architecture [3]. This fusion captures temporal and spatial dependencies while efficiently processing mobile traffic data.

Further advancements in the encoder-decoder paradigm have led to innovative models like SpectraGAN proposed by Kai et al [6]. This model integrates an encoder and a generator, where the encoder assimilates contextual information, while the generator encapsulates the hidden context details and transforms them into the required format [6]. This distinctive architecture emphasizes the importance of context representation and translation within the encoder-decoder structure for various applications. Additionally, the Transformer architecture has recently gained attention for its parallel encoder-decoder structure. Liu et al. pioneered the application of Transformer architecture in cellular traffic prediction, demonstrating its effectiveness in modeling temporal and spatial dependencies within traffic data [45].

2.3.4 Graph Neural Network

Over the past decade, Graph Neural Networks (GNN) [59] have gained considerable traction, revolutionizing various domains. A pivotal advancement in the realm of graph convolutional networks emerged in 2013, marked by Bruna et al.'s foundational study [13], which introduced a variant of graph convolution grounded in spectral graph theory. This approach provided a basis for spatial-based graph convolutional networks by leveraging spectral methods capable of processing entire graphs simultaneously.

Within the realm of mobile traffic data analysis, Graph Neural Networks (GNNs) have emerged as a critical tool. Wang's study employed GNN-D (Graph Neural Network with Decomposed Cellular Traffic Model) to forecast metropolitan mobility data, incorporating diverse factors such as land use, population dynamics, holidays, and social activities [68]. Similarly, Lin utilized MPGAT (Multivariate and Propagation Graph Attention Network) to scrutinize changes in traffic flow induced by outdoor mobile traffic data, with a primary focus on examining road intersections within urban areas [5]. These graph neural network models have demonstrated their efficacy in accurately predicting movement patterns or areas with complex traffic dynamics, enabling better traffic control measures to mitigate accidents and congestion.

The evolution and widespread application of Graph Neural Networks (GNNs) have substantially reshaped the analysis and understanding of complex network structures like cellular traffic data. Pioneering studies such as Wang's utilization of GNN-D and Lin's implementation of MPGAT underscore the effectiveness of GNNs in capturing intricate relationships and patterns within mobile traffic data. These models, enriched with diverse parameters and features, exhibit promising potential in forecasting mobility and traffic flow dynamics within urban landscapes.

2.3.5 Transformer

Cellular traffic forecasting has seen remarkable advancements with the emergence of Transformer architectures. Initially introduced as a novel sequence model, Transformers revolutionized data processing in various domains, including cellular traffic prediction. Unlike traditional models, Transformers rely solely on attention mechanisms, dispensing with recurrent or convolutional layers, thereby enabling access to historical sequences regardless of temporal distance [66]. Recent studies have demonstrated the superiority of Transformer architectures over Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in modeling complex temporal dependencies and achieving improved performance in diverse tasks.

In the realm of cellular traffic prediction, several Transformer-based models have surfaced to address specific challenges and enhance forecasting accuracy. Li et al. proposed the LogSparse Transformer, which effectively resolves memory bottleneck issues encountered by the standard Transformer model by enabling selective attention to relevant parts of the past history [42]. Addressing spatio-temporal dependencies, Cai et al. introduced the Traffic Transformer, designed explicitly to model intricate spatio-temporal relationships at different scales for traffic forecasting tasks [14].

Furthermore, Lim et al. proposed the Temporal Fusion Transformer (TFT), a significant contribution enabling interpretable time series forecasting [44]. This model effectively captures temporal dynamics while ensuring interpretability, a crucial aspect in comprehending and explaining the predicted outcomes. Additionally, Rasul et al. introduced a Transformer-based model employing a conditioned normalizing flow, which effectively models multivariate temporal dynamics for forecasting tasks [56].

Moreover, Wu et al. devised the Adversarial Sparse Transformer (AST), incorporating adversarial training strategies to improve forecasting accuracy from a global perspective in a non-autoregressive manner [70]. These studies collectively highlight the versatility and adaptability of Transformer-based architectures in capturing intricate cellular traffic patterns and dynamics, indicating their potential to significantly advance the field of cellular traffic forecasting.

Chapter 3

Spatial Bottleneck Transformer for Cellular Traffic Prediction in the Urban City

Due to the widespread use of portable devices and the advancement of 5G technology, we have received a significant amount of mobile data, which requires prediction models for cellular traffic data. However, accurately forecasting mobile traffic data is challenging due to the complex spatial and temporal correlations, especially when the mobile data comes from a large geographical area. To tackle this challenge, we propose a new model, called *ST-InducedTrans*, to dynamically explore the large geographical correlations (spatial) and periodic variations (temporal). Specifically, a Spatial Bottleneck Transformer is devised to obtain spatial correlations from the most relevant grids in the geographical area, at the cost of linear complexity. For the temporal blocks, we embed the elaborately selected temporal clues into a temporal Transformer to offer useful temporal prompts for cellular prediction. Finally, several spatial and temporal blocks are effectively stitched into a whole model for complementary cellular traffic prediction. We conducted comprehensive experiments on the public real-world cellular data from Milan [10]. Results show that our model outperforms the state-of-the-art methods on three metrics (MAE, NRMSE, and R^2) at the cost of lower time complexity.

3.1 Introduction

In recent years, cellular traffic prediction has become a prominent area of focus due to the advancement of 5G technology. With the development of portable devices and the internet, mobile phones become an essential part of our daily life. Since cellular traffic data normally contains spatial and temporal information as well as their interactions, efficiently identifying the complex spatial and temporal dependence is crucial for accurate prediction. Previous studies have employed statistical methods, machine learning methods and deep learning methods to improve both the accuracy and efficiency of cellular traffic prediction.

Statistical methods extract specific correlation measures from the individual time series. For example, Zhao et al. [80] used Anselin Local Moran's I statistic measure, while Zhang et al. [77] and Liu et al. [45] used the Pearson correlation coefficient to manually get the spatial correlation between the target grid and its neighboring cells. These methods used pre-defined statistics and shallow models, so they struggled to capture the complex non-linear spatial-temporal correlations in real-world cellular traffic data [60, 11].

Machine learning methods utilize various traditional algorithms such as Gradient Boosting Decision Tree (GBDT) [38], Gaussian mixture model (GMM) [32], and support vector regression [58] to improve the prediction performance. However, similar to statistical methods, they are also constrained by shallow models.

Deep learning methods leverage the representation capability of modern deep neural networks to solve the above problems, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Graph Neural Networks (GNNs), and Transformers. Qiu et al. [55] proposed a model using RNNs in the temporal block in their research. However, its performance deteriorates quickly in predicting long-range time-series data. CNNs [39] are used to extract the spatial correlation in cellular traffic prediction. For example, in [77], cellular data of the whole geographical area is treated as an image that can be processed by CNNs to obtain spatial correlations. The disadvantage of CNNs is that the local receptive fields can only model adjacent spatial information within a small region range. GNNs are employed to model the spatial correlations by a graph structure. For example, CNN&GNN model was proposed in [80] to process spatial features in cellular traffic prediction. But its graph structure requires an adjacent matrix with fine-grained data which is difficult to obtain due to privacy issues.

The emergence of Transformer advanced research in various fields such as Natural

Language Processing (NLP) and computer vision. Hence, it can also serve as a good network architecture for mobile traffic prediction. On the one hand, Transformer can resolve the long-range prediction inability of RNNs by escaping the gradient vanishing problem. On the other hand, Transformer has a global receptive field using the self-attention mechanism, which relaxes the local limitations of CNNs and is free from the adjacent matrix of GNNs. Therefore, the vanilla Transformer has been used in [45] to improve the performance of cellular traffic prediction. Since the complexity of the vanilla Transformer is quadratic w.r.t the input length, modeling the spatial grids of real-world cellular traffic (e.g., 100×100 in Milan) is computationally expensive and infeasible. In this context, a novel, dynamic and more efficient model variant is required to reduce the heavy computation load of Transformer.

Motivated by this, we propose a new model, called **ST-InducedTrans**, to efficiently capture the complex spatial-temporal dependencies of a large mobile grid for accurate cellular traffic prediction. The whole model is composed of several spatial and temporal blocks. Each spatial block is a well-devised Spatial Bottleneck Transformer, which introduces a smaller-size query (i.e. **inducing point**) to only focus on the K most relevant spatial correlations and recovers the input length by an original-size query. This design significantly reduces the Transformer complexity and also manages to automatically and dynamically select the most relevant region correlations, even for far-away regions. Each temporal block is a vanilla Transformer augmented with elaborately selected temporal clues (e.g., day of the week and holidays). These clues offer important prompts for extracting certain cellular traffic patterns, e.g., the difference between weekdays and weekends. The overall structure is built by stitching the spatial and temporal blocks with several fusion layers. We verify the effectiveness of our model design on a widely-used cellular prediction benchmark: Milan. The state-of-the-art comparison, parameter analysis and visualization corroborate the superiority and efficiency of our model design.

The contributions of this paper are summarized as follows:

- We propose ST-InduscedTrans, a novel and efficient model for cellular traffic prediction on a real-world large grid.
- We design a Spatial Bottleneck Transformer in ST-InducedTrans to capture the spatial dependencies, which can improve the prediction accuracy and reduce the time complexity from quadratic to linear.
- Informative temporal clues (e.g., day of the week and holidays) are embedded in the temporal Transformer to provide useful temporal prompts beneficial for cellular prediction.
- Comprehensive experiments are conducted on the real-world benchmark dataset Milan, which verifies the superiority and efficiency of our method.

3.2 Related Work

Amidst the era of big data, time series analysis has evolved as a subset of AI technology, drawing significant interest from researchers toward cellular traffic forecasting. Many scholars have engaged statistical methods and traditional machine learning techniques for cellular traffic data prediction. Moreover, the fusion of cellular traffic challenges with advanced deep learning models has led to the development of more sophisticated predictive frameworks. This section primarily focuses on presenting the relevant literature within the domain of mobile traffic prediction.

3.2.1 Statistical Methods

Several noteworthy statistical models have found application in cellular traffic forecasting, notably including Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing (ES), and the Holt-Winters (HW) model. ARIMA, characterized by its autoregressive, difference, and moving average components, has been utilized by Guo et al. [29] in predicting cellular traffic data from various Chinese cities.

In addressing the base station problem, Zhang et al. [2] employed Seasonal Autoregressive Integrated Moving Average (SARIMA) for time series learning. Their work introduced a novel model aimed at understanding the geographical impact on base station utilization through clustering techniques. Furthermore, Snyder [60] contributed significantly by proposing an effective Exponential Smoothing model within spatial modules, laying the foundation for subsequent research endeavors.

Earlier approaches predominantly relied on the empirical properties of these research methodologies. Taylor et al. [65] proposed a statistical model based on Exponential Smoothing, while Hyndman et al. [28] extended this approach by segmenting time series problems into five trends and three seasons. This segmentation elucidates the trend and seasonal characteristics of time series problems. Although these statistical methods require less computation, they primarily rely on linear relationships between inputs and outputs, which poses challenges in identifying nonlinear relationships.

Combining statistical and machine learning methodologies, Azari's research [61] integrates the ARIMA method with Long Short-Term Memory (LSTMs) to unveil nonlinear relationships, showcasing a hybrid approach that merges the strengths of both statistical and machine learning techniques.

3.2.2 Machine Learning Methods

LightGNM, a novel tree-based model, was introduced to expedite the training process of traditional Gradient Boosting Decision Trees (GBDT) in cellular traffic forecasting, as outlined by Ke et al. [38]. This model found application within the final prediction model, enhancing its efficiency. In a related domain, Zhang et al. [32] proposed a Gaussian Mixture Model (GMM) aimed at minimizing power consumption for unmanned aerial vehicle transmission and mobility. Their model significantly reduced power requirements for downlink transmission and mobility processes.

Additionally, Support Vector Regression (SVR) [58] emerged as another method employed in predicting cellular traffic data. However, compared to deep learning models, most machine learning models in cellular traffic forecasting are relatively shallow.

One of the primary limitations of these methods is their oversight of spatial-temporal data correlations and challenges in scalability when dealing with high-dimensional datasets. While machine learning models generally outperform statistical methods, only a few of these models have demonstrated superiority over deep learning models in cellular traffic forecasting [36]. This highlights the ongoing pursuit for methodologies that effectively capture spatial-temporal correlations and scalability in high-dimensional data, while also emphasizing the comparative advantages and limitations of different predictive approaches in this domain.

3.2.3 Deep Learning Methods

Numerous attempts have been made to address challenges in cellular traffic forecasting using deep learning methodologies. Recurrent Neural Networks (RNNs), as highlighted by Medsker et al. [47], have been instrumental in uncovering temporal variations within cellular traffic prediction, particularly within temporal blocks. Although Qiu et al. [55] employed RNNs in the temporal block of their model, they encountered limitations regarding accuracy in predicting extended time series data.

Kuber et al. [34] primarily relied on Long Short-Term Memory (LSTMs) to construct temporal-dependent models in their research. Concurrently, they incorporated auxiliary information (such as the geographical locations of airports, banks, or restaurants) to enhance spatial components. Moreover, both Azari [61] and Kuber [34] highlighted the advantageous implications of improved resource allocation via prediction models.

However, the sequential nature of RNNs and LSTMs poses challenges in efficiently processing input sequences, resulting in slower training processes. Given that cellular traffic data can be likened to image-based grids, researchers have combined Convolutional Neural Networks (CNNs) with RNNs to predict mobile traffic data. While CNNs, as emphasized by Kim et al. [39], excel in handling spatial components, they often struggle to capture distant spatial information beyond adjacent regions. Additionally, Graph

Convolutional Networks (GCNs) and Graph Neural Networks (GNNs) have gained traction in processing spatial-related information in recent studies.

Notably, Zhang et al. [3] proposed a ConvLSTM model, employing LSTMs to handle long temporal information while utilizing CNNs to explore spatial dependencies. Nonetheless, limitations persist, particularly in accurately describing dependencies among adjacent grids.

Incorporating the encoder-decoder paradigm, researchers have hybridized CNNs and RNNs, leveraging their strengths for diverse tasks. For instance, Zhang et al. [3] introduced the Spatio-Temporal Neural Network (STN) within an encoder-decoder framework, combining Convolutional Long Short-Term Memory and 3D Convolutional Neural Network. Similarly, Xu and Kai [6] proposed SpectraGAN, integrating an encoder and a generator for contextual and hidden information processing, respectively.

The advent of Transformer architectures has sparked interest in cellular traffic prediction, pioneered by Liu et al. [45]. Transformers, renowned for their parallel encoder-decoder structure and success in Natural Language Processing (NLP), have shown promise in various domains. Liu et al. [45] devised novel transformer modules, STB and TTB, specifically tailored for cellular traffic prediction. However, the computational complexity of basic transformers, operating at $O(n^2)$ for a 100x100 grid, poses significant computational challenges, demanding the development of more efficient models to alleviate computational burdens.

3.3 Problem Formulation

We conduct cellular traffic research on the large-scale, real-world public telecommunication dataset from a well-known telecommunication provider: Telecom Italia. The dataset, collected from Milan, divides the large geographical areas into a $H \times W$ grid. Various cellular traffic activities are recorded, including Received SMS, Sent SMS, Incoming Call, Outgoing Call and Internet usage.

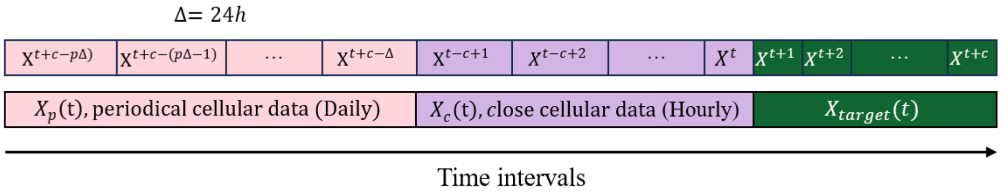


Figure 3.1 : $X_p(t)$, $X_c(t)$, $X_{target}(t)$ at the $X(t)$ interval.

In this paper, we deal with the cellular traffic forecasting problem. In particular, we study the behavior of Call data prediction. Cellular traffic prediction is performed on the $H \times W$ grid representing the whole geographical location. Each cell of the grid reflects the cellular traffic in a certain location. As a time-series prediction problem, we make use of two types of historical data: *close* neighboring cellular data and *periodical* cellular data. Denote the current time step of traffic data as *t* in Fig. 3.1; we want to predict future data $X_{target}(t) \in \mathbb{R}^{N \times c}$ of the future c steps, where N is the number of grids. Then, we define close neighboring data $X_c(t) \in \mathbb{R}^{N \times c}$ as the previous c steps just before future data on all N locations. We define periodic data $X_p(t) \in \mathbb{R}^{N \times p \times c}$ as historical data having the same multiplicative intervals relative to future steps.

To summarize, the cellular traffic forecasting problem is defined as follows: $\boldsymbol{Given}$ cellular traffic data from N geographical grids: (1) Close neighboring cellular data $X_c(t) = (X^{t-c+1}, X^{t-c+2}, \ldots, X^t) \in \mathbb{R}^{N \times c}$, (2) Periodical cellular data $X_p(t) = (X^{t+1-p\Delta}, \ldots, X^{t+c-p\Delta}, \ldots, X^{t+1-\Delta}, \ldots, X^{t+c-\Delta}) \in \mathbb{R}^{N \times p \times c}$, $\boldsymbol{Forecast}$ future traffic data $X_{target}(t) = (X^{t+1}, X^{t+2}, \ldots, X^{t+c}) \in \mathbb{R}^{N \times c}$.

3.4 Methodology

In recent years, the revolutionary improvements that Transformer has brought to computer vision and Natural Language Processing (NLP) have attracted much attention from the academic community. Some work, such as [45], has used the vanilla Transformer to predict cellular traffic and has achieved certain improvements. For real-world cellular traffic prediction problems, data usually comes from a large geographical area (e.g.,

100×100 in Milan). In this case, the adoption of Transformer is computationally heavy and infeasible due to its quadratic time complexity. We propose a novel *ST-InducedTrans* model for cellular traffic prediction, which contains a well-devised *Spatial Bottleneck Transformer* to obtain the spatial correlations from all grids by inducing points (i.e., bottleneck). Moreover, we embed informative temporal clues in the temporal Transformer to augment the temporal pattern extraction.

However, due to the large time complexity of Transformer in terms of large geographical grids, most previous studies have adopted different statistical methods to obtain the most relevant spatial correlations, such as Anselin Local Moran's I Statistic Measure [77] and Pearson Correlation Coefficient [45]. The main problem with these statistical approaches to select grids is that when there is a large number of data, the computational cost of calculating the correlation between different grids can be expensive. Therefore, this paper proposes a Spatial Bottleneck Transformer to obtain spatial information on the top K grids with higher similarity among all grids by inducing points (i.e., bottleneck). This inducing point can be learned continuously in the transformer.

The main problem with these statistical approaches is that when the amount of data is large, the computational cost of calculating the correlation between different grids can be expensive. Therefore, this paper proposes a Spatial Bottleneck Transformer to obtain spatial information on the top K grids with higher similarity among all grids by inducing points. This inducing point can be learned continuously in the transformer.

3.4.1 Spatial Bottleneck Transformer

Spatial Bottleneck Transformer (SBT) is the major contribution in this paper, which eliminates the quadratic scaling problem of all-to-all attention of a vanilla Transformer and decouples the network depth from the input's size, allowing us to construct very deep models. Concretely, SBT is a Transformer-based neural network architecture. Unlike the vanilla Transformer, each SBT block is composed of two Multihead Attention Blocks (MABs), as shown in Fig. 3.2. We design a bottleneck structure for the two MABs. The

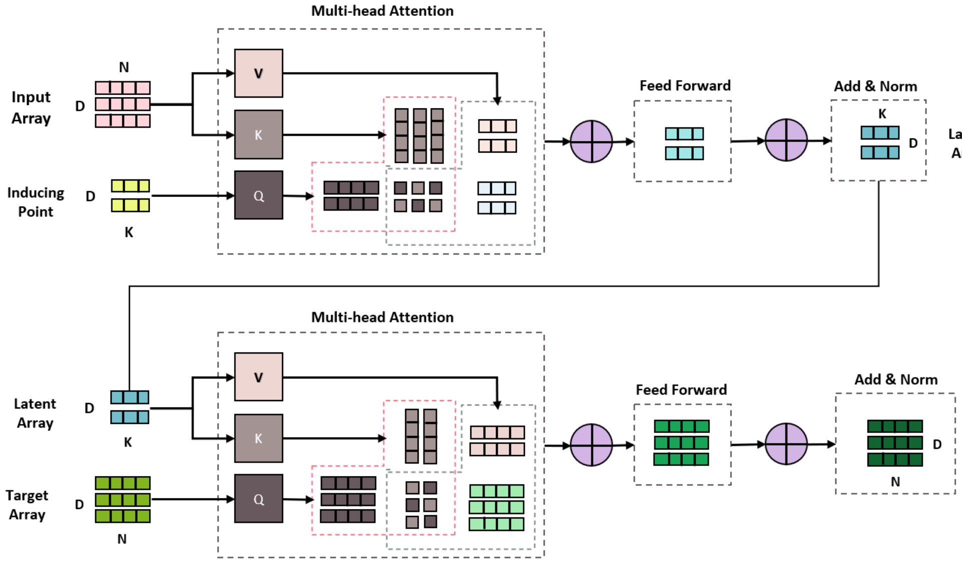


Figure 3.2 : Spatial Bottleneck Transformer

first MAB (*encoder*) takes as input a small-size query to extract the K most relevant features with the cross-attention mechanism. Then, the second MAB (*decoder*) retrieves the spatial correlations from the extracted K relevant features with another cross-attention.

The bottleneck refers to mapping from N input elements to a much smaller $K(K \ll N)$ features and then recovering the original length size N. There are several advantages of the bottleneck design over the vanilla transformer design: (1) the computation complexity is significantly reduced with this design; (2) the bottleneck design has the feature selection effect by only selecting K relevant features. This is consistent with cellular traffic properties that only a small number of locations sharing similar patterns are highly related in cellular traffic prediction; (3) K can be adjusted to reflect different circumstances, such as different cities and different grid-scale grained.

The main components of the SBT are as follows.

Multi-Head Attention. This module is similar to the Multi-Head Attention in the vanilla Transformer. It splits the vectors in the input sequence into multiple subsets and then performs a self-attentive computation on the vectors in each subset. With an input tensor of shape (N, D), the Multi-Head Attention outputs a tensor of the same shape (N, D), but embeds the correlations among all input elements inside. Here, N refers to the number of input elements, and D stands for the feature dimension. Formally, the module can be denoted below:

$$\begin{aligned} \mathrm{Attention}(Q, K, V) &= \text{softmax}\left(QK^{\mathrm{T}}\right) V, \\ \mathrm{Multihead}(Q, K, V) &= \mathrm{concat}\left(O_1, \cdots, O_h\right) W^O, \\ \text{where } O_j &= \mathrm{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \end{aligned} \tag{3.1}$$

Here, $W^O, QW_i^Q, KW_i^K, VW_i^V$ are the relevant fully connected layer weights.

Feed-Forward Network The feed-forward neural network has two fully connected layers for enhanced representation. For each input vector, the Feed-Forward Network maps it to a higher dimensional vector, which is then passed through a ReLU activation function and mapped back again to the original dimensional vector. This operation allows each element to capture more complex interactions after computation.

Layer Normalization Layer Normalization normalises each sub-layers output to improve the model's stability and generalisation. Instead of using the mean and variance of the entire set of inputs, it uses the mean and variance of the subset corresponding to each sample.

Spatial Bottleneck Transformer (SBT). SBT model comprises an Encoder and a Decoder. Unlike the vanilla Transformer, we adopt cross-attention mechanisms in the Multi-Head Attention module instead of self-attention. Specifically, in the *Encoder*, we introduce the **inducing point** $I \in \mathbb{R}^{K \times D}$ as the query probe to extract K most relevant features from the original input sequence $X \in \mathbb{R}^{N \times D}$ ($N = H \times W$ in a cellular traffic grid). Since $K \ll N$ in our design, the Encoder has the advantages of information compression and feature extraction, which only keeps the most related features beneficial for future cellular traffic prediction. I serves as a query (Q), and X serves as key (K) and value (V) in Eq. 3.1. Taking the Feed-Forward Network and layer normalization into consideration, we can represent the structure as:

$$\begin{aligned} \mathrm{CrossAttention}(I, X) &= \mathrm{LayerNorm}(H + \mathrm{FeedForward}(H)), \\ \text{where } H &= \mathrm{LayerNorm}(X + \mathrm{Multihead}(I, X, X))\,. \end{aligned} \tag{3.2}$$

Similarly, the *Decoder* adopts a cross-attention structure but designs different query, key, and value. We take the original input sequence X as query (Q), and the output from the Encoder as key (K) and value (V). This way, the output of Decoder has the same sequence length as the input.

The overall structure of the SBT can be represented as:

$$\begin{aligned} \mathrm{SBT}_K(X) &= \mathrm{CrossAttention}(X, H) \in \mathbb{R}^{N \times D}\,, \\ \text{where } H &= \mathrm{CrossAttention}(I, X) \in \mathbb{R}^{K \times D}\,. \end{aligned} \tag{3.3}$$

Here, K stands for the bottleneck size for choosing only K relevant features.

3.4.2 ST-InducedTran Model

We now describe the detailed structure of the ST-InducedTrans model, as shown in Fig. 3.3. Overall, ST-InducedTrans includes the temporal clues, spatial block and temporal block, devised to capture the spatial correlations and temporal patterns, respectively.

Temporal Clues. Before going into the details of the model architecture, we first describe how to leverage indicative temporal clues to help cellular traffic prediction. Our observation is that the cellular patterns vary with certain temporal events, e.g. weekdays and weekends exhibit different patterns in using mobile devices, thus leading to different prediction modes. Similarly, holidays are an important factor. Bearing this in mind, we define the one-hot indicative clues 'day-of-week' (7-dimension) and 'holidays' (1-dimension). Then, we concatenate them into the temporal clues $X_{tc} \in \mathbb{R}^{n \times (7+1)}$. With an additional projection, the clues can be mapped to $X_{tc} \in \mathbb{R}^{N \times D}$ (D is the same as all the Transformers) and used as temporal clues in the later Transformers.

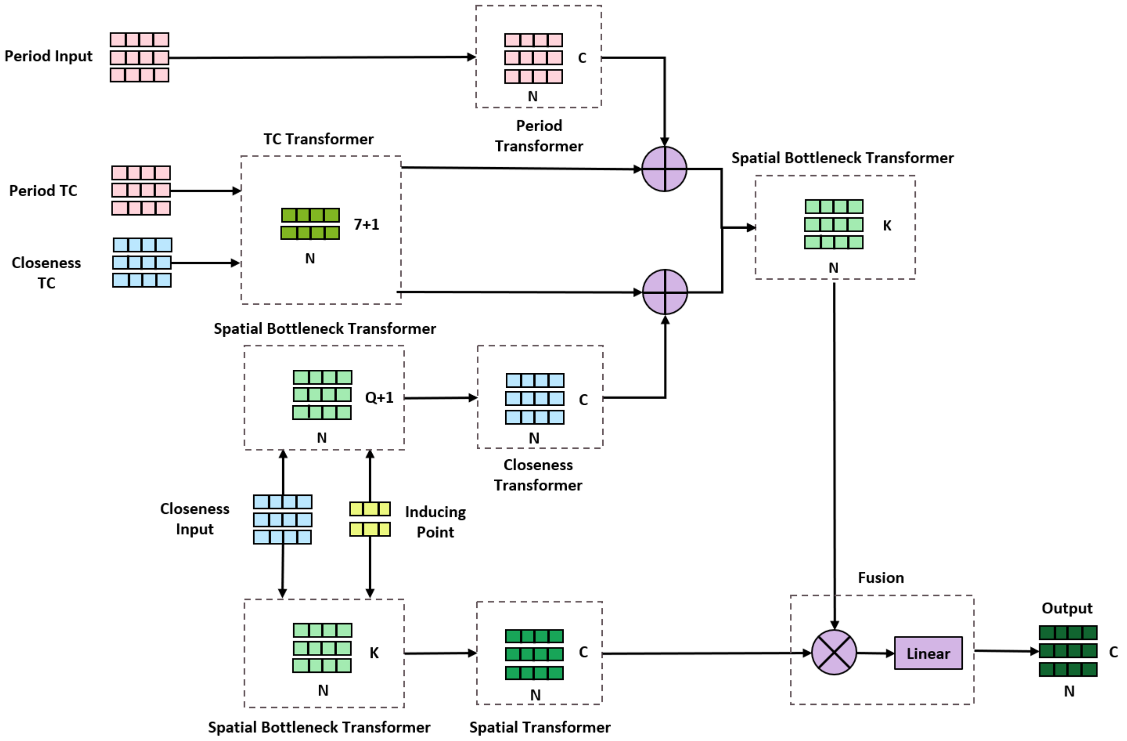


Figure 3.3 : ST-InducedTran Model

Spatial Block. There are two types of Transformer in the Spatial Block: Spatial Bottleneck Transformer and Spatial Transformer.

Spatial Bottleneck Transformer is designed to extract the possible correlations among all locations in the cellular traffic grid. Since in a real-world grid, not all grid locations are prediction relevant, the bottleneck structure can efficiently and effectively select only the relevant spatial locations helpful for cellular traffic prediction. As mentioned before, the parameter K of the inducing points in the bottleneck can be chosen to reflect the real-world scenarios, such as cities, grid scale, etc.

Spatial Transformer The input of this block is $X_c \in \mathbb{R}^{N\times c}$. The top K similar elements are selected and concatenated to $X_s \in \mathbb{R}^{N\times K\times c}$ via Spatial Bottleneck Transformer, which is a learning-based feature selecting module. The Spatial Transformer takes the vanilla Transformer structure and contains an encoder and a decoder. X_s is the encoder input, X_c is the decoder input, and the final output is $X_{sp} \in \mathbb{R}^{N\times 1\times c}$.

Temporal Block. There are also two types of Transformer in the Temporal Block: Closeness Transformer and Period Transformer. These two types process the close neighboring data $X_c \in \mathbb{R}^{N \times c}$ and the periodic data $X_p \in \mathbb{R}^{N \times p \times c}$ as described in Section 3.

Closeness Transformer We first augment the close neighboring data in the spatial context. The Q most related spatial locations are selected with the Spatial Bottleneck Transformer and concatenate with the grid's own data as $X_c^{aug} = \mathbb{R}^{N \times (Q+1) \times c}$. Then X_c^{aug} is input to the encoder of the Closeness Transformer. The period data X_p is averaged to get X_p^{avg} and input to the decoder of the Closeness Transformer. Closeness Transformer mainly processes the close data while it also refers to the period average for relevant information.

Period Transformer The period data X_p is directly input to the encoder. And the close data X_c is input to the decoder as reference information. Since the period data already contains various period patterns, there is no need to apply extra augmentations. Period Transformer processes the period data while it also refers to the close data for relevant information.

Stitching and Fusion. After getting all the modules and components, we stitch them in an overall model ST-InducedTrans, as shown in Fig. 3.3. Specifically, the model input includes, period data, close data, and temporal clues. The temporal block and the Spacial block are put in a parallel layout to process the data. Finally, the Fusion module is proposed (containing several MLP layers) to combine both the temporal and spatial processed information for better prediction.

In this experiment, the fusion method is relatively simple: the results of the bimodal modes are concatenated together, the dimension becomes $2 \times D$, and then the dimension is reduced from $2 \times D$ to D by a linear layer.

3.5 Experiments

3.5.1 Dataset

We use the common benchmark dataset Milan [10], whose data is collected from Telecom Italia. It contains aggregated cellular traffic, including SMS, call service and internet in the city of Milan. The dataset contains several traffic activities in Milan, which consists of 1000 mesh overlays. The time duration for data collection is from 2013/11/01 00:00 to 2014/01/01 23:00 for two months (62 days with around 300 million records), with an interval of 10 minutes. We further aggregated the data into an interval of 60 minutes (1 hour). The geographical area of Milan is divided into a size of $H \times W$ grids, where H and W refer to the number of rows and columns and $H = W = 100$ (i.e. 10,000 grid cells). Each grid has an approximate area of 235×235 square meters. Following previous studies, we take the center 20×20 grid from the whole 100×100 grid, mainly to predict the cellular traffic in urban areas. In this paper, we use call data, and our approach also fits the other data sets including SMS and Internet usage.

3.5.2 Baseline

We compared the proposed ST-IndusiveTran Model with several classical baseline methods that are widely used in time series prediction

- Historical Average & ARIMA [11]: Traditional statistical method that uses one day of mobile phone traffic data from historical data to predict future traffic.
- STDenseNet [77]: This model's primary objective is to concurrently capture both spatial and temporal correlations within traffic across distinct cells. Incoming and outgoing traffic, depicted as a two-channel tensor matrix resembling an image, is employed at each instance. Utilizing a sliding window approach, training and test datasets are generated. Fusion techniques are employed to account for both proximity-based (closeness dependence) and time-based (period dependence) rela-

tionships within the model.

- STACN(w/o E) [80]: This model proposes a spatial-temporal attention-convolution network (STACN) that considers the dynamic spatiotemporal correlation of cellular traffic. A GCN+CNN convolutional module is investigated to capture mobile traffic's spatial and temporal characteristics through spatial and temporal attention.
- ConvLSTM [76]: STCNet (Spatial-Temporal Cross-domain Neural Network) can efficiently capture complex patterns in grid data. The model actively models three cross-domain datasets to capture the external factors influencing traffic generation. At the same time, the model proposes a clustering algorithm to divide urban areas into different groups, thus designing a continuous inter-intentional migration learning strategy. Finally, cross-domain data modelling is used to detect real data information.
- ST-Tran [45]: This model is the first to implement all modules in the transformer. It is of pioneering importance. In this model design, the authors propose a Temporal Transformer Block and a Spatial Transformer Block, and finally, fuse the results of the two modules.

3.5.3 Implementation Details

Hardware

All experiments were performed on a GeForce RTX 6000 with 24G RAM.

Optimizer

In this study, we experimented with two optimization algorithms, Adam and AdamW, and throughout our trials, we observed superior performance with Adam. Adam represents a departure from classical stochastic gradient descent methods. Unlike traditional stochastic gradient descent, where the learning rate remains constant during training, Adam offers the advantage of individualized learning rates for each parameter. This feature proves beneficial in handling sparse gradients, ensuring improved performance. Additionally, Adam adjusts the learning rate for each parameter by computing the weighted averages of recent gradients, enabling robust performance even in the presence of noisy problems. Notably, Adam's optimizer demonstrates reduced sensitivity to varying learning rates.

In this experiment, adam's β parameter was initialised to 0.98.

Regularization

For each sublayer, including the multihead attention mechanism and the fully connected layer, a dropout was used before entering the residual component and the layer norm, where dropout rate = 0.1, indicating a 10% setting of the output to 0.1. The dropout rate = 0.1 was also used after the positional encoding.

Hyperparameters

During this experiment, the transformers underwent training iterations with N=6. Notably, the dimensionality (d_{model}) varied across the transformer models: d_{model} equaled 64 in the spatial transformer, 256 in the closeness transformer, 128 in the period transformer, and similarly, 128 in the generic transformer. The d_{model} parameter signifies the representation size of vectors assigned to tokens upon their entry into a specific transformer module.

Moreover, a polynomial learning rate decay strategy was adopted for this experiment. The initial learning rate commenced at 0.001, while the training batch size for the entire model was set to 16, conducted over a span of 500 training rounds. These specifications were crucial elements in fine-tuning the performance of the transformer models in the experimental setup.

Table 3.1 : Input of Different Transformers.

Transformer	d_{model}	encoder input	decoder input
spatial	64	$X_s(t)$	$X_c(t)$
closeness	256	$X_c(t)$	$X_p(t)$
period	128	$X_p(t)$	$X_c(t)$

3.5.4 Evaluation Metrics

It is essential to evaluate the model's accuracy to describe its performance in the forecasting task. Evaluation metrics change according to the problem type. In this research, we use MAE (Mean Absolute Error), $RMSE$ (Root Mean Squared Error), and R^2 (Coefficient of determination).

Given $y_{\text{true}} \in \mathbb{R}^{N \times c}$ to be the ground truth mobile traffic at time step t, and $y^t_{\text{pred}} \in \mathbb{R}^{N \times c}$ to be the predicted cellular traffic values, and T to be all the time samples predicted, then, the metrics are defined as follows:

MAE quantifies the average absolute deviation between the original and predicted values, calculated by averaging their differences across the dataset. A lower error value closer to zero indicates a more favorable outcome. The formula for RMSE is outlined below.

$$MSE = \frac{1}{T} \sum_{t=1}^{T} \left(y^t_{\text{true}} - y^t_{\text{pred}} \right) \tag{3.4}$$

Root Mean Square Error (RMSE) measures the error rate using the square root of Mean Square Error. Mean Square Error (MSE) evaluates the prediction error of the model, ranging from $[0, +\infty]$. A lower error value closer to zero indicates better performance. The formula for RMSE is provided. Normalized root mean error (NRMSE) is an adaptation of RMSE that normalizes the error relative to the observed range of the variable.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(y_{\text{true}}^{t} - y_{\text{pred}}^{t}\right)} \tag{3.5}$$

R^2 denotes the coefficient indicating the goodness of fit between predicted and original values, with a range of [0, 1]. A higher R^2 value signifies better model performance.

$$R^2 = 1 - \frac{\sum_{t=1}^{T}\left(y_{\text{true}}^{t} - y_{\text{pred}}^{t}\right)^2}{\sum_{t=1}^{T}\left(y_{\text{ave}}^{t} - y_{\text{pred}}^{t}\right)^2} \tag{3.6}$$

3.6 Results & Discussion

3.6.1 Time Complexity Analysis

One major contribution of this research is to reduce the quadratic scaling problem of Transformer to a linear complexity problem. Denote N as the total number of grid cells (i.e. $H \times W$) we have, and we want to predict through top K spatially relevant data from N grid cells with Spatial Bottleneck Transformer. Our design reduces the number of calculation operations from N^2 of the vanilla Transformer to NK ($K \ll N$). This means the complexity reduces from quadratic to linear. Tab. 3.2 presents the time complexity comparison. Here D is the dimension of the features in the intermediate layers. The significant reduction in time complexity makes the computation faster and possible to predict cellular traffic data in a larger-scale geographical grid.

Table 3.2 : Complexity of vanilla Transformer and Spatial Bottleneck Transformer.

Model Name	Complexity Per Layer
Vanilla Transformer	$O(N^2D)$
Spatial Bottleneck Transformer	$O(KND)$

Table 3.3 : Comparison with the state-of-the-art methods.

Methods	MAE	NRMSE	R^2
HA	18.7226	0.9687	0.4419
ARIMA	17.1895	0.8813	0.6564
LSTM	13.9438	0.6079	0.7802
STDenseNet	12.3168	0.6442	0.7842
STACN (w/o E)	12.6450	0.6210	0.7207
ConvLSTM	11.2308	0.5652	0.8097
ST-Trans	10.0244	0.5388	0.8273
ST-InducedTrans (w/o Fusion)	**9.7215**	0.5319	0.8317
ST-InducedTrans	9.7273	**0.5035**	**0.8493**

3.6.2 Comparison with the State-of-the-art Methods

In Tab. 3.3, we compare our method with various baselines, including the state-of-the-art method ST-Trans. Our model variant "w/o Fusion" means the Spatial Bottleneck Transformer is removed before the Fusion module (Fig. 3.3 left-right). Our method outperforms all the baselines with a large margin w.r.t. three metrics. Note that ST-Tran employed vanilla Transformer architecture. Our method is not only more efficient in time complexity but also much better in performance. The comparison with the "w/o Fusion" variant shows that the Spatial Bottleneck Transformer is effective and generally applicable in our model.

3.6.3 Parameter Analysis and Visualization

Parameter K. For too small and too large K, the performance deteriorates quickly, while in a reasonable range (15 to 20), the performance is stable. The best K is 20 according to the MAE measure, which is only 5% compared with the overall 400 grid cells.

In this module, the choice of K is significant. K can be selected in the range of

$[0, 400]$; when K is chosen as 0, the grid selection function is removed. To enhance the prediction performance, we experimented with the value of K between 0 and 50; the MAE value was the lowest for $K = 20$ when fusion was not implemented.

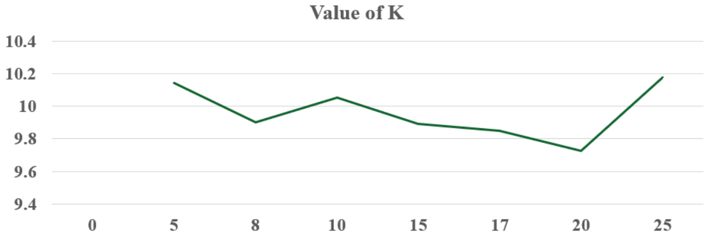


Figure 3.4 : Performance with respect to varying values of K

Prediction Visualization. We show the comparison between the predicted values and the ground truth of the 224th grid in Fig. 3.5 and Fig. 3.6 for the model “w/o Fusion” and our final model. Both models have a relatively good prediction trend. However, our model performs much better in detail, such as the time range between 75 and 100, verifying the effectiveness of the Spatial Bottleneck Transformer in the Fusion module.

3.7 Conclusion

This paper proposes a new ST-InducedTrans model to improve forecasting accuracy and computation complexity. We devise a new spatial transformer block, namely Spatial Bottleneck Transformer (SBT), to explore the spatial dependencies. In SBT design, we introduce a learnable inducing point to reduce the algorithmic complexity of the transformer from quadratic to $O(nm)$ linear. Moreover, we explore the informative temporal clues and incorporate them into the temporal embedding of the Transformer. Also, information on temporal embedding is added, including information on holidays and days of the week. Extensive experiments are conducted on the real-world cellular traffic dataset, which corroborates the efficiency and effectiveness of our ST-InducedTrans model. This novel ST-InducedTran model is applied to real-world public mobile traffic data. This model reduces the algorithm’s complexity and improves the accuracy of the prediction.

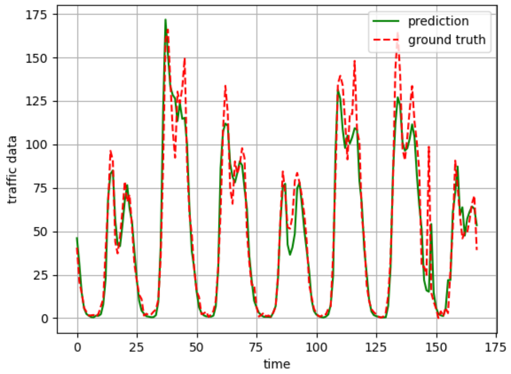


Figure 3.5 : Predicted values without Fusion

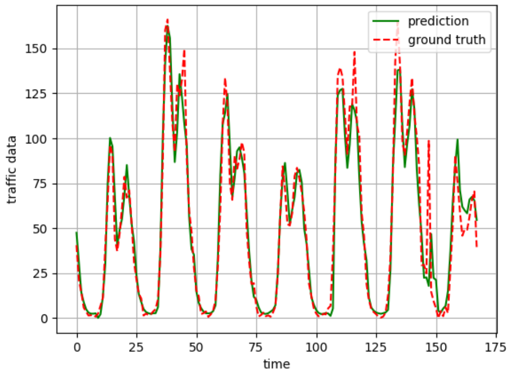


Figure 3.6 : Predicted values with Fusion

In future work, we can add more external information, including POIs, to improve the accuracy of the prediction.

Chapter 4

Transformer-Based Cellular Traffic Prediction across Diverse Domains of Big Data Sources

4.1 Introduction

Time series forecasting refers to predicting future trends and patterns from historical data, which has been used in many areas, including earthquake prediction, stock market prediction, electricity usage and vehicular traffic forecasting. Cellular traffic prediction has also become a heated research topic in time series forecasting, which attracts more researchers to improve accurate prediction of cellular traffic. Accurately predicting mobile traffic is of vital importance for both *operators* and *application developers*. For operators, accurate forecasting can help them plan and optimize network layouts, improving network quality and user experience. Moreover, the analysis of user behavior and demand can be used to predict future cellular traffic demand, thus helping operators to develop traffic packages and improve user satisfaction. For application developers, accurate prediction can help them tailor the performance and experience of their applications, thereby improving user satisfaction and the effectiveness of advertising. In this context, mobile traffic forecasting has become a hot research topic in the past ten years and gained much research attention.

In the previous chapter, we proposed a novel deep learning model, called ST-InducedTrans, to eliminates the quadratic scaling problem of all-to-all attention of a vanilla Transformer and decouples the network depth from the input's size, allowing us to construct very deep models. However, in order to obtain more accurate prediction results, more factors should be taken into consideration, which is a challenging task to combine extra spatial information with the cellular traffic itself. This combination will make our dataset more complicated.

Previous studies have done some research on the problem of mobile traffic prediction. A number of researchers have used statistical methods, machine learning methods and deep learning methods to improve cellular traffic prediction. Many statistical-based approaches are used to predict future mobile phone traffic, but most statistical methods are shallow models whose performance degenerates considerably over time[11]. Statistical methods operate on individual time series while ignoring spatial correlation. It is also hard to capture the complex non-linear spatial-temporal correlations in real-world cellular traffic data[60]. In previous research, Recurrent Neural Networks (RNNs)[47], Long Short-Term Memory (LSTMs)[23], Convolutional Neural Networks (CNNs)[39], Graph Convolutional Networks (GCNs)[79], and Graph Neural Networks (GNNs) are proposed to process grid-based data.

Transformer is an encoder-decoder structure and has achieved good results in Natural Language Processing (NLP) [66]. In recent years, Transformer has achieved good results in computer vision, video processing and other fields. Transformer can dynamically aggregate the most relevant features, have a larger receptive field and conduct parallel computation, which can somehow solve the problems mentioned in the previous methods. Some research [45] has used the vanilla transformer to predict cellular traffic, and has achieved certain improvements. However, since the data we want to predict is a 100×100 grid, the algorithm complexity of the vanilla transformer is $O(n^2)$. Calculating the cellular traffic of the entire area is very computationally expensive, and we need to create novel, dynamic and more efficient models to reduce the huge amount of calculation.

Most of researchers mainly consider the cellular traffic dataset itself as the main data source to improve the prediction accuracy and external information is rarely taken into account. However, external factors should be taken into considerations, for example, their user movements and visits. Some external information, including base stations data, POIs distribution and social activities information have certain level influence on cellular traffic generation [30], [72]. For example, people will visit the shopping malls more frequently compared to those in the rural areas. The number of the base station, the distribution of POI and the number of social activities for each cell will be vary based on the location of the area. Meanwhile, we have to consider the functionality of each zone when we predict our cellular traffic data. For instance, we can divide any large geographical areas into different functional zones based, including residential zones, commercial zones, industrial zones, recreational and open space, etc.

To effectively enhance mobile prediction results over expansive regions and capture the interplay between spatial and temporal data correlations, we incorporate external factors into our approach. Employing advanced clustering techniques, we segment areas into distinct functional zones. Building upon our prior work on the ST-InducedTrans model detailed in the previous section, a key focus of this chapter lies in integrating additional factors to adeptly capture correlations among various grid points. This integration constitutes a significant contribution. Drawing insights from a spectrum of extensive data sources beyond cellular network data, including GPS data and social media trends, holds promise in enriching predictive models. By amalgamating these diverse datasets, we aim to augment forecast accuracy and granularity across larger geographical extents. This multidimensional approach enables us to enhance the predictive capabilities and finer spatial resolution of our forecasts.

The primary objective of this paper centers on refining the precision of mobile traffic prediction by leveraging our innovative model introduced in the preceding chapter. This paper encapsulates several key contributions, outlined as follows:

- We present a novel model tailored explicitly for mobile traffic prediction based on ST-InducedTrans. Building upon our previous research, this model incorporates refined methodologies and advanced algorithms to bolster prediction accuracy.
- We introduce a comprehensive approach that integrates diverse factors beyond traditional data sources. This includes incorporating external data such as GPS information, social media trends, and other relevant datasets to enrich the predictive capabilities of our model.
- Our chapter delves into an in-depth exploration of spatial-temporal correlations

within mobile traffic data. By leveraging sophisticated techniques, we aim to capture and harness these correlations more effectively to improve prediction accuracy.

These contributions collectively represent our concerted effort to elevate the accuracy and applicability of mobile traffic prediction, paving the way for more effective and reliable forecasting methodologies. The rest of this paper is organized as follows. Section 2 introduces some related work. Section 3 describes the problem formulation and preliminaries. Section 4 presents the proposed novel model architecture with cross domain datasets, including base station information, Points of Interest (POIs) distribution and social media activities. Section 5 presents the performance comparison between the proposed network and a baseline scheme, followed by Section 6 to show the result and some discussions. Conclusions are drawn in Section 7.

4.2 Related Work

This section will presents a range of models focusing on different aspects of spatial and temporal dependencies in cellular traffic forecasting. Spatial models like CNNs, GNNs, GCNs and variants address different trade-offs between preserving topology and extracting spatial features.Temporal models, including RNNs, LSTM, GRU and Transformers, offer varying capabilities in handling temporal dynamics, long-range dependencies, and computational efficiency. In conclusion, this section highlights the diverse methodologies employed in cellular traffic forecasting, each with its strengths and limitations concerning spatial and temporal dependencies, paving the way for further advancements and integrated models in this field.

4.2.1 Spatial Dependencies

ARIMA [49] & Bayesian Networks [67] exhibit advantages in handling highly nonlinear traffic patterns, providing a baseline for understanding temporal trends and patterns in cellular traffic data. They are useful for capturing nonlinear relationships within the data.

Convolutional Neural Networks (CNNs) adopted in [46] [1] [78] extract spatial features by converting traffic networks into regular grids. However, they might lose crucial topology information present in irregular traffic networks.

Graph Neural Networks (GNNs) [59] and Graph Convolutional Networks (GCNs) [18] generalize deep learning to non-Euclidean domains, exploring inherent traffic topology and preserving the graph structure. Models like STGCN [74] and DCRNN [43] use spectral and diffusion graph convolutions on directed graphs, respectively, to capture spatial dependencies. Graph Attention Networks (GATs) summarize geo-graph features using meta-learners [4], while Graph WaveNet captures hidden spatial patterns but with fixed spatial dependencies after training. Graph WaveNet with Dilation Convolution improves accuracy by learning hidden spatial patterns, and these models have limitations in scalability for long input sequences and capturing long-range dependencies affected by deeper layers [71].

Spatial-Temporal Transformer Networks efficiently model dynamic directed spatial dependencies in high-dimensional latent subspaces, not relying on predefined graph structures, and can adapt to varying spatial relationships.

4.2.2 Temporal Dependencies

Recurrent Neural Networks (RNNs), especially Gated Recurrent Units (GRU) [17] and Long-Short Term Memory (LSTM) networks, handle temporal dependencies and long-range relationships in traffic data. However, they suffer from gradient issues during training and might be computationally intensive for long sequences.

Transformers excel in sequence learning by employing parallelizable self-attention mechanisms [66]. They effectively capture long-range time-varying dependencies in input sequences of varying lengths, offering adaptability and efficiency in modeling temporal dynamics.

Zhang et al. [76] proposed STCNet (Spatial-Temporal Cross-domain Neural Network), which used metadata, including time of the day, day of the week, and other data, as the input of the two-layer neural network of LSTMs, and used cross-domain data as the information in CNNs and employed the CNN-RNN model to combine the complex spatio-temporal traffic variability in mobile traffic prediction through temporal and spatial perspectives.

In chapter 3, we proposed a novel model, called ST-InducedTrans [62], to attempt to bridge spatial and temporal dependencies efficiently, offering promising directions for future research in cellular traffic forecasting.

4.3 Problem Formulation

Our research focuses on analyzing cellular traffic using a large-scale dataset obtained from Telecom Italia in Milan. This dataset divides the geographical area into a grid of $H \times W$ dimensions and records various cellular activities like Received SMS, Sent SMS, Incoming and Outgoing Calls, and Internet usage.

Our specific focus is on forecasting cellular traffic, particularly in predicting Call data. The prediction task involves the entire $H \times W$ grid, where each grid cell represents the cellular traffic in a specific location. To approach this time-series prediction problem, we utilize two types of historical data: neighboring cellular data and periodic cellular data.

Neighboring cellular data $X_c(t) \in \mathbb{R}^{N \times c \times x}$ refers to the preceding c time steps leading up to the future data point $X_{target}(t) \in \mathbb{R}^{N \times c}$ for all N locations, considering x external factors. On the other hand, periodic cellular data $X_p(t) \in \mathbb{R}^{N \times c \times p \times x}$ comprises historical data at regular intervals relative to the future steps, covering the same multiplicative intervals with x external information.

In essence, the challenge of cellular traffic forecasting involves leveraging the preceding c steps of cellular data in close proximity to the future prediction and historical data occurring at regular intervals to predict the future c steps of cellular traffic data across N geographical grids.

4.4 Methodology

4.4.1 Structure of Transformers

Transformer is a sequence-to-sequence model based on a self-attention mechanism, originally proposed by Ashish Vaswani et al. in 2017 [66], and is widely used in natural language processing (NLP), performing particularly well in machine translation tasks. Transformer has had the same impact on NLP as CNN has on computer vision, providing a framework for researchers in computer vision to avoid tedious additional steps, such as feature engineering. The emergence of the Transformer model allows for better results in most domains.

Transformer is also a type of encoder and decoder model. Transformer used in this study is the standard Transformer architecture, illustrated in Fig. 3.6, including encoder, decoder, self-attention, multi-head attention, embedding and positional encoding.

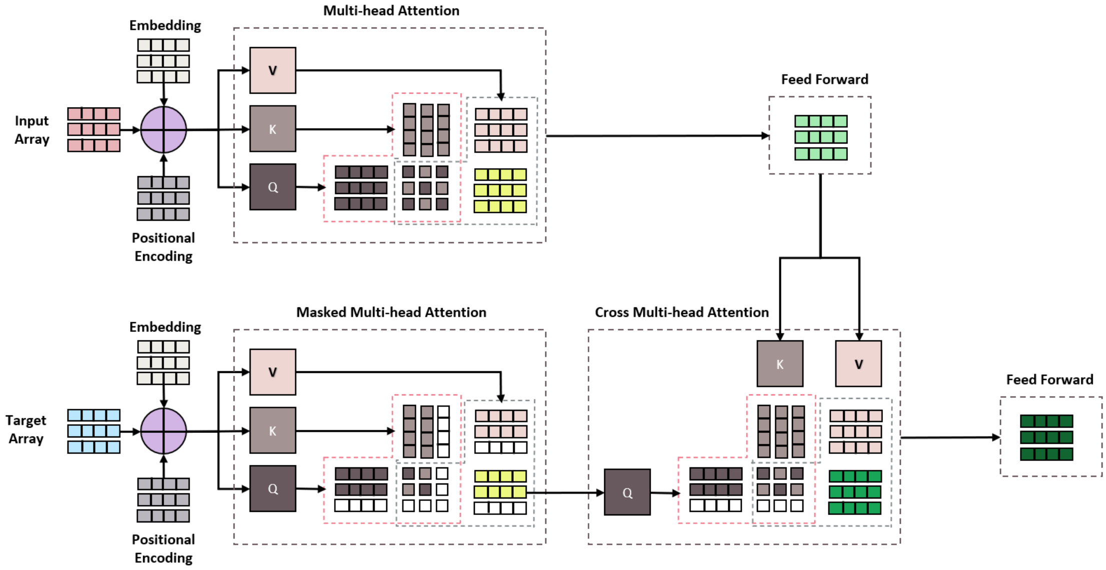


Figure 4.1 : Structure of basic transformer

Encoder The encoder converts the input sequence into a contextual vector. It consists of several identical layers stacked on each other, each consisting of two sub-layers: a multi-headed self-attention mechanism and a position-wise fully connected feed-forward neural network. The multi-headed self-attention mechanism is used to calculate the relationships between different positions in the input sequence and is a form of auto-regression. The fully connected feed-forward neural network is used to enhance the representation, in which Layer Normalization is used to obtain information about the data. Unlike Batch Normalization, which calculates the mean and variance for a mini-batch, Layer Normalization obtains the mean and variance for each sample. This has the advantage that the mean and variance are not affected when targeting a particularly long sequence. Each sub-layer is linked with residuals using layer normalisation. The output of each sublayer is:

$$Output = LayerNorm(x + Sublayer(x)) \tag{4.1}$$

Decoder The decoder is similar to the encoder in converting a context vector into an output sequence. It also consists of a stack of identical layers, each consisting of three sub-layers: a masked multi-headed self-attention mechanism, a multi-headed cross-attention mechanism and a fully connected feed-forward neural network. The multi-headed self-attention mechanism and the fully-connected feed-forward neural network serve the same purpose as in Encoder. The multi-headed encoder-decoder attention mechanism fuses the relationship between the context vector and the input sequence to generate the output sequence better. The main purpose of the mask used in the multi-headed self-attention mechanism is that information after t should not be seen at time t during decoding. The self-attention mechanism with a mask does not see information after time t. The masking effect is achieved by turning all information after time t into 0, thus ensuring that the training and prediction results are consistent.

Self-Attention The attention function is a function query, key, and value that maps a query and some key-value pairs into an output. The dimension of the output is the same as the dimension of value. The self-attention mechanism is the core part of the Transformer and is used to compute the relationships between different positions in the input sequence.

In the self-attention mechanism, each position in the input sequence is computed with all other positions, resulting in a weighted sum vector. The weights are derived by computing the similarity between the query and the other positions. This calculation allows the model to capture long-distance dependencies in the input sequence.

In this self-attention, we use dot-product attention, which is easier to implement than addictive attention. Query and key do dot product. The larger the dot product, the higher the similarity. The result of the inner product is then divided by $\sqrt{d_k}$ and softmax to find the weight, which is a non-negative weight, ranging from 0 to 1. When querying information with a query, the information distribution is obtained by multiplying the query and key and then multiplying by the information value to obtain the final attention based on the query. The formula of self attention is shown as follows.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{4.2}$$

Multi-head Attention Similar to the filter of CNN, the Multi-Head Self-Attention mechanism is an extension of the Self-Attention mechanism, which can learn different representations in different heads and improve the model's generalisation ability. Specifically, Multi-Head Attention splits the input sequence into multiple sub-sequences, performs the self-attention computation on each sub-sequence separately, concatenates the computation results and obtains the final representation by a linear layer. Instead of making a single attention function, the multi-Head Attention mechanism projects the entire queries, keys, and values into a d-dimensional h^{th} order attention function. Each function's output is concatenated and then projected back to obtain the final output. The reason for doing a multi-head attention mechanism is that there are not many parameters that can be learned in dot-product attention, and sometimes more parameters that can be learned are needed to learn different patterns. Query, key, and values are first projected to a lower dimension through a linear layer. The projection weights are learnable, split into h parts, in the hope that different projection weights can be learnt by splitting into h parts, in the hope that different projection methods can be learnt so that different patterns can be matched in the projected metric space. Finally, the h parts are concatenated to make a single projection, similar to convolutional neural networks with multiple output channels.

$$\begin{aligned} \operatorname{MultiHead}^{\mathrm{Mu}}(Q, K, V) &= \operatorname{Concat}\left(\operatorname{head}_1, \ldots, \operatorname{head}_{\mathrm{h}}\right) W^O \\ \text{where head } &= \operatorname{Attention}\left(Q W_i^Q, K W_i^K, V W_i^V\right) \end{aligned} \tag{4.3}$$

Embedding Embedding is mapping the input into a vector and learning a vector of length d to represent it for any sequence. An embedding is required in both Encoder and Decoder to convert each sample into a vector of equal d_{model} length.

Positional Encoding In the attention mechanism, the sequence is not sequential. The output is a weighted sum of the values. The weight is the similarity between the query and the key, computed independently of the information in the sequence. The order changes, but the output does not change. In the RNN model, the RNN uses the output of the previous output as the input to the next computation. This in itself is a sequence with sequential information in it. Positional Encoding is a method of encoding positional information into the input sequence based on some fixed function (3.5) that encodes the position into a fixed length vector, which is then added to the input vector, thus obtaining a representation with position information. It is mainly calculated using different periods using the sin and cos methods. The vector of length d_{model} records information about the current position i, which is then added to the embedding information of length d_{model}. This completes the process of adding the position information to the input.

$$\begin{aligned} PE_{(pos, 2i)} &= \sin\left(pos / 10000^{2i / d_{\text{model}}}\right) \\ PE_{(pos, 2i+1)} &= \cos\left(pos / 10000^{2i / d_{\text{model}}}\right) \end{aligned} \tag{4.4}$$

Feed Forward The fully connected feed-forward neural network acts once on each sample, and the transformer has two linear layers put together. The process of transforming the vector dimension from d_{model1} to d_{model2} and then projecting it back to d_{model1}.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{4.5}$$

4.4.2 ST-InducedTrans Model with Cross-Domain Datasets

The ST-InducedTrans model structure is outlined in Fig. 4.2. In essence, it comprises three key components: temporal clues, spatial block, and temporal block as we designed in the previous chapter. These components are designed to capture temporal patterns and spatial correlations. In this chapter, we not only incorporated supplementary data that includes spatial information, as well as we simplified the spatial block in the original the ST-InducedTrans model.

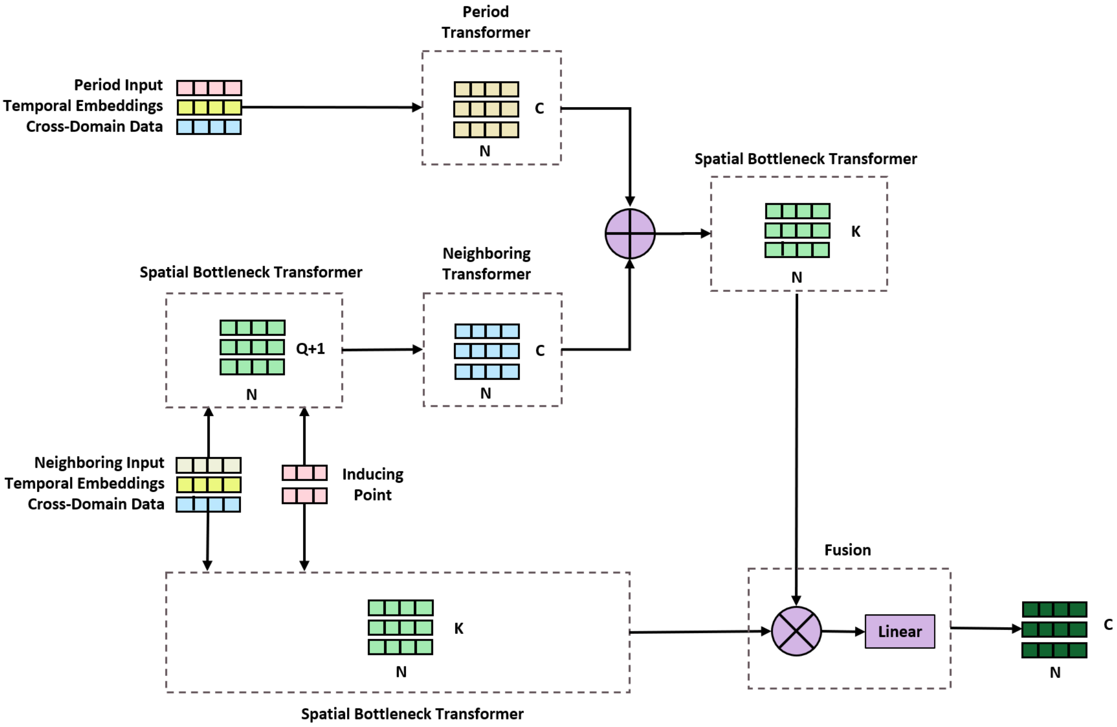


Figure 4.2 : ST-InducedTran Model with Cross-Domain Datasets

Temporal Embeddings In Chapter3, The ST-InducedTrans model incorporates indicative temporal embeddings, such as day-of-week and holiday indicators, recognizing their impact on cellular traffic patterns, and utilizes a concatenated representation that is projected to enable their utilization as temporal embeddings within Transformers. In this Chapter, the ST-InducedTrans model expands its indicative temporal embeddings to encompass additional temporal information, including day-of-the-week represented as numbers from 0 to 6 starting from Sunday, hour-of-day as values ranging from 0 to 23, binary indicators for weekdays (1 for weekdays, 0 for non-weekdays), and another binary indicator for weekends and holidays. Recognizing the similar traffic patterns on weekends and holidays, the model assigns a value of 1 when the current date falls on either a weekend or a holiday. For example, at the t-step of 05:00:00 11/05/2013, the respective temporal embeddings are 'day-of-week' (2), 'hour-of-day' (5), 'is-weekdays' (1), and 'is-weekend-holiday' (0). These temporal features are merged with both neighboring and periodic cellular data before being fed into the temporal block and spatial block. The combined feature vector forms the temporal embeddings $X_{te} \in \mathbb{R}^{n \times 4}$. Through an additional projection, these embeddings can be transformed to $X_{te} \in \mathbb{R}^{N \times D}$ (where D remains consistent across all Transformers) and employed as temporal embeddings in subsequent Transformers.

Cross-domain Data To represent the concatenation of different cross-domain datasets (BSs, POIs, and Twitter activities) with the original cellular traffic data, we denote the combined result as X_{Cross}, using the $\oplus$ symbol to denote the concatenation operation. This operation signifies the merging of the external spatial information with the original cellular traffic data. Detailed explanations regarding these cross-domain datasets will be provided in the upcoming section.

$$X_{\text{Cross}} = X_{\text{BS}} \oplus X_{\text{POI}} \oplus X_{\text{Social}} \tag{4.6}$$

Spatial Block

The *Spatial Bottleneck Transformer* is designed to uncover potential correlations among various locations within the cellular traffic grid. Recognizing that not all grid locations are pertinent for prediction in real-world scenarios, this bottleneck structure efficiently identifies and selects spatial locations crucial for accurate cellular traffic prediction.

Previously mentioned, the parameter K determining the inducing points within the bottleneck can be adjusted to reflect practical scenarios, such as city dimensions or grid-scale considerations. In contrast to the detailed Spatial Bottleneck Transformer discussed earlier, we've streamlined the spatial block. In this adaptation, the Transformer, functioning as an encoder-decoder model, incorporates the Spatial Bottleneck Transformer module, substituting the original encoder while retaining the decoder in its original form.

Temporal Block Within the Temporal Block, two distinct Transformers are employed: the Closeness Transformer and the Period Transformer. These specialized Transformers handle the processing of close neighboring data $X_c(t) \in \mathbb{R}^{N\times c\times x}$ and periodic data $X_p(t) \in \mathbb{R}^{N\times c\times p\times x}$ as detailed in Section 3.

Closeness Transformer: Initially, the close neighboring data is enriched within the spatial context. Through the Spatial Bottleneck Transformer, the Q most relevant spatial locations are selected and concatenated with the grid's native data, forming $X_c^{aug} = \mathbb{R}^{N\times(Q+1)\times c}$. This augmented data is then input into the encoder of the Closeness Transformer. Simultaneously, the periodic data X_p is averaged to create X_p^{avg}, which is fed into the decoder of the Closeness Transformer. The Closeness Transformer primarily handles close data while also referencing the averaged period data for pertinent insights.

Period Transformer: In contrast, the Period Transformer directly receives the period data X_p as input to its encoder. On the other hand, the close data X_c serves as reference information and is input into the decoder. As the period data inherently contains diverse period patterns, no additional augmentations are necessary. The Period Transformer focuses on processing the period data while also utilizing close data for relevant contextual information.

Fusion Once all the individual modules and components are obtained, they are assembled into the comprehensive ST-InducedTrans model (Fig. 4.2). The model's input comprises period data, close data, and temporal clues. Notably, the temporal block and the Spatial block are arranged in a parallel layout to concurrently process this input data.

To enhance predictive capabilities, a Fusion module, comprising several Multilayer Perceptron (MLP) layers, is introduced. This Fusion module serves the purpose of amalgamating the processed information from both the temporal and spatial pathways, aiming for improved prediction outcomes.

4.5 Experiments

4.5.1 Dataset

Cellular Traffic Dataset and Key Observations

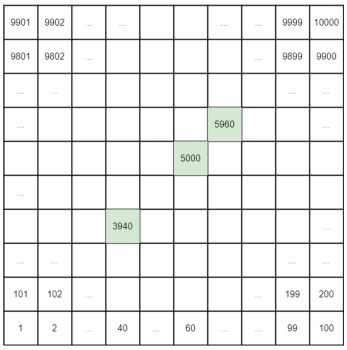
This research uses a large-scale real-world public telecommunication dataset from a large telecommunication provider in Italy. They offer two Italian areas: the city of Milan and the Province of Trentino. The dataset contains several traffic activities in Milan and Trentino for two months. [10] In our research, we will use the area of Milan, which consists of 1,000 mesh overlays. The spatial call aggregated detail information within grids using the following formula:

$$S_i(t) = \sum_{v \in C_{\text{map}}} R_v(t) \frac{A_{v \cap i}}{A_v} \tag{4.7}$$

The geographical area of Milan is divided into a size of $H \times W$ grids. H and W refer to the grid's number of rows and columns. In the dataset of 'city of Milan', $H = W = 100$, and the whole city is divided into 10000 grids.

Telecom Italia provider provides Call Detail Records. A Radio Base Station is provided when a user engages in an interaction, including receiving a phone call, making a phone call, receiving an SMS, sending an SMS and using the Internet. Therefore the data includes five components: Received SMS, Sent SMS, Incoming Call, Outgoing Call and Internet. Two-month data containing 300 million records is collected from 2013/11/01 00:00 to 2014/01/01 23:00 (62 days), with an interval of 10 minutes. We further aggregated the data into an interval of 60 minutes (1 hour). In the spatial dimension, the whole city is divided into 100×100 areas and each grid had an approximate size of 235×235

square meters [10].



9901	9902	...	...			...	...	9999	10000
9801	9802	...					...	9899	9900
...	...							...	...
...						5960			...
					5000				
...									
			3940						...
...	...							...	...
101	102	...					...	199	200
1	2	...	40	...	60	...	...	99	100

Figure 4.3 : Milan Grid.

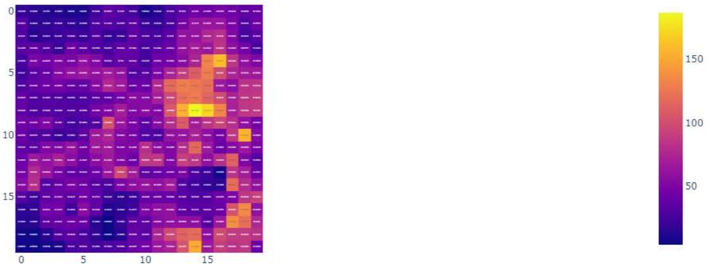


Figure 4.4 : 18:00-19:00 Network Activities

Following are some key observations:

Internet We have selected several Cells to visualise the network traffic data. We have selected Cell3940 as x, Cell5000 as y and Cell5960 as z. we can see that the daily network traffic activity shows a clear periodicity, and more periodicity can be considered in the temporal dimension. Also, in selecting different regional Cells, the variation in the central area is more pronounced and fluctuates more; for example, Cell1 is not in the central area and has less variation in magnitude.

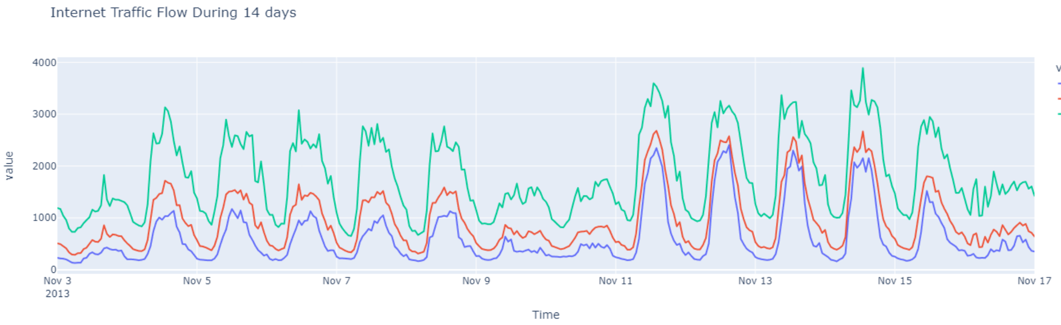


Figure 4.5 : Internet Traffic visualization.

SMS & Call Unlike Internet traffic, SMS and Call traffic patterns are significantly different on weekdays and weekends. In particular, in the central zone, the peak number of incoming SMS is lower on weekends than on weekdays. The change in the float was not as pronounced in non-central areas, and the number of incoming SMS messages was overall smaller in non-central areas than in central areas.

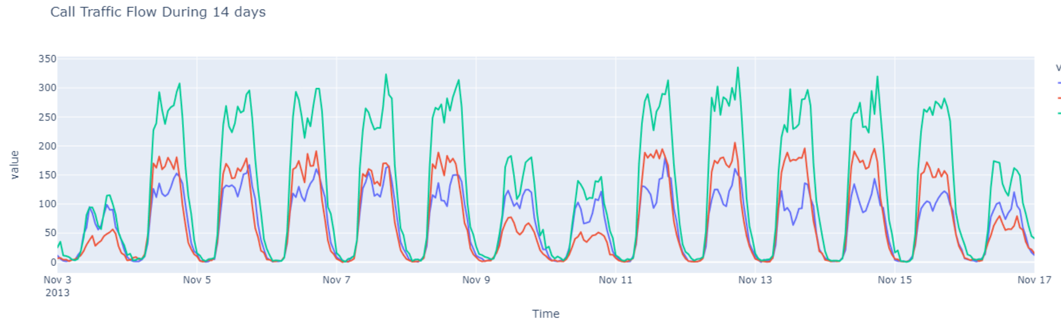


Figure 4.6 : Call in Traffic for two weeks.

Cross-Domain Datasets

Most of the current researchers focus on the spatio-temporal factors, which is the cellular traffic data itself. However, in order to improve the accuracy of the cellular traffic data, the importance of the external factors cannot be ignored. In our research, we consider

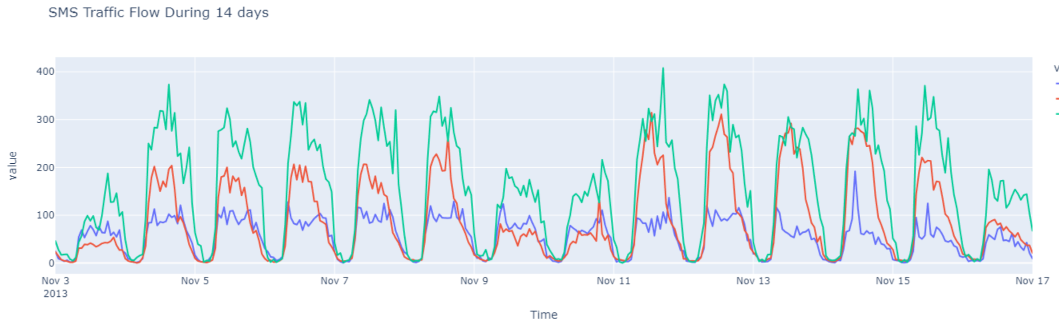


Figure 4.7 : SMS in Traffic for two weeks.

three different external data, which are the number of BSs information, POIs distribution information as well as social activities of each cell.

Base Stations information sourced from Open-CellID, a collaborative project collecting GPS positions of cell towers and their corresponding location area identity [51], offers valuable data including the longitude, latitude, and mobile country code. Utilizing the longitude and latitude data of each base station, we integrate the count of base stations with the original Milan cellular traffic dataset. This integration allows the collection of the number of base stations for each cell, denoted as $X_{\mathrm{BS}}^{(h,w)}$. The base station matrix can be visually represented as the collection of X_{BS} values for each cell, and the base station matrix can be shown as

$$\mathbf{X}_{\mathrm{BS}} = \begin{bmatrix} x_{\mathrm{BS}}^{(1,1)} & x_{\mathrm{BS}}^{(1,2)} & \cdots & x_{\mathrm{BS}}^{(1,W)} \\ x_{\mathrm{BS}}^{(2,1)} & x_{\mathrm{BS}}^{(2,2)} & \cdots & x_{\mathrm{BS}}^{(2,W)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{BS}}^{(H,1)} & x_{\mathrm{BS}}^{(H,2)} & \cdots & x_{\mathrm{BS}}^{(H,W)} \end{bmatrix} \tag{4.8}$$

We use Google Places API [64] to obtained POIs information. POIs distribution information includes subway stations, stores, churches, cafe, restaurants and so on. In the Table 4.1, there are 13 different POIs included. Not all of the POIs have a huge influence on our cellular traffic prediction. Therefore, we only collect cafes, bars and restaurants, and then combine other POIs into a different features. Each POI can be denoted as $X_{\mathrm{POI}}^{(h,w)}$, and the POI matrix can be represented as

$$\mathbf{X}_{\mathrm{POI}} = \begin{bmatrix} x_{\mathrm{POI}}^{(1,1)} & x_{\mathrm{POI}}^{(1,2)} & \cdots & x_{\mathrm{POI}}^{(1,W)} \\ x_{\mathrm{POI}}^{(2,1)} & x_{\mathrm{POI}}^{(2,2)} & \cdots & x_{\mathrm{POI}}^{(2,W)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{POI}}^{(H,1)} & x_{\mathrm{POI}}^{(H,2)} & \cdots & x_{\mathrm{POI}}^{(H,W)} \end{bmatrix} \tag{4.9}$$

Table 4.1 : Detailed Cross-Domain Datasets

Dataset	Type	Number
Cellular Traffic	SMS / Call / Internet	around 300 million
POI	Subway station	104658
	Store	19748
	Restaurant	4666
	Bar	3192
	Lodging	2922
	Hospital	1585
	School	1284
	Cafe	995
	Bank	882
	Park	765
	Church	512
	Parking	392
	Library	188
Base Stations	GSM / CDMA / LTE	69909
Social Activities	Twitter	269290

Social activity information, which represents user movements and user usage for certain area, is also included as part of our cross-domain dataset. Twitter information is the main social activity data we obtained from Dandelion API, including the location and keywords. For each cell, we use $X_{\mathrm{Social}}^{(h,w)}$ to represent the number of Twitter activities of cell (h, w). The social media matrix can be expressed as

$$\mathbf{X}_{\text{Social}} = \begin{bmatrix} x_{\text{Social}}^{(1,1)} & x_{\text{Social}}^{(1,2)} & \cdots & x_{\text{Social}}^{(1,W)} \\ x_{\text{Social}}^{(2,1)} & x_{\text{Social}}^{(2,2)} & \cdots & x_{\text{Social}}^{(2,W)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\text{Social}}^{(H,1)} & x_{\text{Social}}^{(H,2)} & \cdots & x_{\text{Social}}^{(H,W)} \end{bmatrix} \tag{4.10}$$

These cross-domain datasets are static datsets, because they do not change over time. Once they are obtained, they can be fed into our model for training. In this section, we use heatmaps with 5 layers to represent the original cellular traffic, grids in $H \times W$ format, Base Stations distribution, Social Activities (Twitter), and Points of Interest Distribution.

4.5.2 Baselines

In this section, our main objective is to contrast the refined ST-InducedTrans model with the original innovative model introduced in Chapter 3. This comparison is especially pertinent as we have already evaluated several classical and state-of-the-art methods. Additionally, the ST-Trans model proposed by Liu et al. (2021) [45] will also be included in this comparison analysis.

- **ST-Trans** [45]: The novel aspect of this cellular prediction model lies in its exclusive use of Transformer-based modules throughout. It intricately crafts separate Transformer-based branches for spatial and temporal processing, amalgamating the acquired information from these branches at the final stage. This model's design entails a Temporal Transformer Block and a Spatial Transformer Block, with the eventual fusion of the outcomes from these two distinct modules.

- **ST-InducedTrans**: In ST-InducedTrans, we have engineered a Spatial Bottleneck Transformer to capture spatial dependencies, aiming to enhance prediction accuracy while simultaneously reducing time complexity. This innovation transitions the computational complexity from quadratic to linear, leading to more efficient processing while maintaining or improving predictive performance.

4.5.3 Evaluation metrics

In our comprehensive model evaluation, we employed three extensively used metrics in time-series and cellular traffic prediction:

- MAE (Mean Absolute Error)
- $NRMSE$ (Normalized Root Mean Squared Error)
- R^2 (Coefficient of determination)

For both MAE and $NRMSE$, lower values indicate better results, with proximity to zero signifying improved accuracy. Conversely, a higher value for R^2 signifies better model performance.

4.6 Results and Discussions

Our primary contribution in this research involves augmenting the accuracy of cellular traffic prediction by integrating additional external information. We introduced an enhanced model, named ST-InducedTransPlus, which builds upon the ST-InducedTrans framework. Notably, our model streamlines the spatial block in comparison to ST-InducedTrans. In our design, we simplified the conventional spatial transformer within the spatial block and adjusted certain parameters within the spatial bottleneck transformer to enhance efficiency. Notably, our spatial bottleneck transformer comprises both encoders and decoders, allowing for more efficient processing while preserving predictive accuracy.

In Table 4.2, our experiments involved various combinations, incorporating temporal embeddings and additional spatial information encompassing Base Station distributions, Point of Interests distributions, and social activities. These experiments were conducted across two primary models: ST-Trans [45] and ST-InducedTransPlus. The findings from our experiments indicate the substantial impact of X_{Cross} on improving cellular traffic prediction accuracy. Notably, the accuracy improvements were observed in both the ST-Trans and ST-InducedTransPlus models. The most promising outcomes were observed

Table 4.2 : Different External Combination in Transformers

Combination	MAE	NRMSE	R^2
ST-Trans	10.0244	0.5388	0.8273
ST-Trans + X_{TE}	9.91146	0.54215	0.82519
ST-Trans + X_{Cross}	9.84599	**0.51274**	0.83197
ST-Trans + X_{TE} + X_{Cross}	9.88978	0.53153	**0.84364**
ST-InducedTransPlus	9.89902	0.52308	0.83727
ST-InducedTransPlus + X_{TE}	9.94422	0.53959	0.82683
ST-InducedTransPlus + X_{Cross}	9.79733	0.52778	0.83433
ST-InducedTransPlus + X_{TE} + X_{Cross}	**9.76610**	**0.51406**	**0.84283**

in the streamlined ST-InducedTrans model, particularly when incorporating temporal embeddings and cross-domain datasets.

4.7 Conclusion

In conclusion, the enhancements introduced in this chapter significantly elevate the predictive capabilities of the ST-InducedTrans model in cellular traffic data analysis. By integrating temporal embeddings and cross-domain datasets comprising diverse static information gathered from various APIs, this study achieves notable improvements in prediction accuracy. The strategic concatenation of these datasets and embeddings with the cellular traffic data, coupled with the streamlined approach to spatial block simplification, demonstrates a substantial positive impact on prediction performance. Through rigorous experimentation utilizing public real-world datasets and comprehensive assessments involving varied combinations and baseline models, this work substantiates the effectiveness of these enhancements in advancing the predictive power of the ST-InducedTrans model, presenting a promising avenue for further advancements in spatiotemporal data analysis and forecasting.

Chapter 5

Conclusion

5.1 Summary

Cellular traffic forecasting is of significant importance in the telecommunications industry, including network optimization, capacity planning, resource management, service quality, and energy preparation. Accurate cellular traffic prediction can assist in efficiently allocating resources, reducing congestion, and enhancing overall network performance. Predicting cellular traffic aids in capacity planning for network infrastructure, which enables telecom operators to anticipate future demands and make necessary upgrades or expansions to handle increased traffic. During natural disasters or unexpected events, accurate forecasts assist in managing network loads, prioritizing services, and ensuring communication continuity.

Research on cellular traffic forecasting has seen significant advancements in utilizing various methodologies, including statistical methods, machine learning, deep learning, including transformer-based approaches.

Traditional statistical models like ARIMA (AutoRegressive Integrated Moving Average), Exponential Smoothing, and Seasonal Decomposition have been historically used for cellular traffic forecasting. While these methods provide a good baseline, they might struggle to capture complex nonlinear patterns and relationships present in modern cellular networks.

Machine Learning techniques have gained popularity due to their ability to capture nonlinear relationships and handle large volumes of data. Algorithms such as Support Vector Machines (SVM), Random Forests, Gradient Boosting, and Regression models have been applied to predict cellular traffic patterns. Feature engineering and selection are crucial in enhancing the performance of these models.

Deep Learning methods, particularly recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs), have shown promise in modeling temporal dependencies and sequential patterns in cellular traffic data. LSTMs, specifically, are adept at capturing long-term dependencies and have been used to forecast network traffic with varying success rates. Transformer architectures, such as the Transformer model originally introduced for natural language processing (NLP), have recently been adapted for time series forecasting tasks. Variants like the Temporal Fusion Transformer (TFT) and other transformer-based models have shown remarkable capabilities in capturing complex temporal relationships, handling multiple seasonalities, and achieving state-of-the-art performance in various time series forecasting tasks, including cellular traffic prediction.

The core contribution of this thesis is to propose a novel model called *ST-InducedTrans*, which is a transformer-based model to improve the efficiency of the calculation whereas improve the accuracy of the cellular traffic results. Based on the existing model, we also add extra external factors into the cellular traffic data to improve the accuracy of the prediction results.

In Chapter 3, we introduced the novel concept of *Spatial Bottleneck Transformer* on how to eliminate the quadratic scaling problem of all-to-all attention of a vanilla Transformer and decouple the network depth from the input's size, allowing us to construct very deep models. In our model, there are two main blocks: Temporal Blocks, where we embed the elaborately selected temporal clues into a temporal Transformer to offer useful temporal prompts for cellular prediction; Spatial Blocks, which are designed to extract the possible correlations among all locations in the cellular traffic grid. Our novel model has reduced the quadratic scaling problem of Transformer to a linear complexity problem.

Based on the novel model proposed in Chapter 3, the main target of Chapter 4 is to improve the accuracy of cellular traffic prediction by adding extra cross domain datasets, including base station information, POIs distribution and social activities data. Meanwhile, we also consider temporal meta data as part of the prediction model to improve the prediction accuracy. In Chapter 4, we consider the effectiveness of the cross domain datasets into two different models: Vanila Transformer and ST-InducedTrans. Inspired by the success of ST-InducedTrans model, combining extra cross domain dataset and meta data can provide more dimensions to improve the accuracy of the datasets.

The effectiveness of the proposed model and result has been published in [62].

5.2 Challenges and Future Work

For the future work, the research presented in this thesis can be extended in the following directions.

Currently, all the work we have conducted is based on the city area of Milan rather than the whole area of the geographical area. In Chapter 3, ST-InducedTrans model has improved the efficiency of predicting cellular traffic data. The future of cellular traffic forecasting aims to extend its scope beyond city-wide areas to predict and manage traffic on a large geographical scale, potentially covering regions. Future research will likely focus on developing models capable of making macroscopic predictions that cover larger geographical regions. This involves understanding and forecasting traffic patterns acrossed diverse terrains, population densities and regional variations.

In Chapter 4, we considered kinds of cross-domain datasets to enhance the accuracy and granularity of forecasts for larger geographical areas and some other sources of big data can be included, such as satellite imagery, weather patterns, and economic indicators. Integrating satellite data and remote sensing technologies could provide valuable insights into geographical features, infrastructure, and population density, aiding in the development of more accuerate predictive models for larger areas.

To predict larger geographical areas, hierarchical forecasting models might be developed. These models could use a multi-level approach, where predictions are made at various levels of granularity, such as country-wide, regional, and local levels, allowing for a comprehensive understanding of traffic dynamics. As the scale of data collection and analysis increases, ensuring privacy and addressing ethical concerns related to the collection and use of extensive geo-location data will be crucial aspects of future research.

In conclusion, future research in cellular traffic forecasting for larger geographical areas will focus on leveraging diverse datasets, and advanced modeling techniques to develop robust predictive models capable of handling the complexities of vast regions.

References

[1] *Deep spatio-temporal residual networks for citywide crowd flows prediction*, vol. 31, no. 1, 2017.

[2] *Traffic prediction based power saving in cellular networks: A machine learning method*, 2017.

[3] *Long-term mobile traffic forecasting using deep spatio-temporal neural networks*, 2018.

[4] *Urban traffic prediction from spatio-temporal data using deep meta learning*, 2019.

[5] *Multivariate and propagation graph attention network for spatial-temporal prediction with outdoor cellular traffic*, 2021.

[6] *Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data*, 2021.

[7] E. , “Exponential smoothing: The state of the art,” *Journal of forecasting*, vol. 4, no. 1, p. 1–28, 1985.

[8] J. and R. J. Hyndman, “25 years of time series forecasting,” *International journal of forecasting*, vol. 22, no. 3, p. 443–473, 2006.

[9] F. M. Alvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar, “Energy time series forecasting based on pattern sequence similarity,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, p. 1230–1243, 2010.

[10] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, “A multi-source dataset of urban life in

the city of milan and the province of trentino," *Scientific data*, vol. 2, no. 1, p. 1–15, 2015.

[11] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control.* John Wiley Sons, 2015.

[12] R. G. Brown, "Statistical forecasting for inventory control," 1959.

[13] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," arXiv preprint arXiv:1312.6203, 2013.

[14] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu, "Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting," *Transactions in GIS*, vol. 24, no. 3, p. 736–755, 2020.

[15] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 62, p. 21–34, 2016.

[16] P. C. Carl, "Exponential forecasting: Some new variations," Management Science, p. 311–315, 1969.

[17] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[18] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[19] Z. Dong, D. Yang, T. Reindl, and W. M. Walsh, "Short-term solar irradiance forecasting using exponential smoothing state space model," *Energy*, vol. 55, p. 1104–1113, 2013.

[20] R. Douc, E. Moulines, and D. Stoffer, *Nonlinear time series: Theory, methods and applications with R examples*. CRC press, 2014.

[21] J. Durbin and S. J. Koopman, *Time series analysis by state space methods*. OUP Oxford, 2012, vol. 38.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[23] A. Graves, “Long short-term memory,” Supervised sequence labelling with recurrent neural networks, p. 37–45, 2012.

[24] J. D. Hamilton and G. Perez-Quiros, “What do the leading indicators lead?” Journal of Business, p. 27–49, 1996.

[25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[26] C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages,” *International journal of forecasting*, vol. 20, no. 1, p. 5–10, 2004.

[27] R. J. Hyndman, “The interaction between trend and seasonality,” *International Journal of Forecasting*, vol. 20, no. 4, p. 561–563, 2004.

[28] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods,” *International Journal of forecasting*, vol. 18, no. 3, p. 439–454, 2002.

[29] *Traffic forecasting for mobile networks with multiplicative seasonal ARIMA models*. IEEE, 2009.

[30] *Understanding traffic dynamics in cellular data networks*. IEEE, 2011.

[31] *Mobile data traffic forecasting in UMTS networks based on SARIMA model: The case of Addis Ababa, Ethiopia*. IEEE, 2017.

[32] *Machine learning for predictive on-demand deployment of UAVs for wireless communications*. IEEE, 2018.

[33] *Graph attention spatial-temporal network for deep learning based mobile traffic prediction*. IEEE, 2019.

[34] *Traffic prediction by augmenting cellular data with non-cellular attributes*. IEEE, 2021.

[35] A. R. Ives and V. Dakos, "Detecting dynamical changes in nonlinear time series using locally linear state-space models," *Ecosphere*, vol. 3, no. 6, p. 1–15, 2012.

[36] W. Jiang, "Cellular traffic prediction with machine learning: A survey," *Expert Systems with Applications*, vol. 201, p. 117163, 2022.

[37] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[38] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[39] P. Kim and P. Kim, "Convolutional neural network," MATLAB deep learning: with machine learning, neural networks and artificial intelligence, p. 121–147, 2017.

[40] C.-N. Ko and C.-M. Lee, "Short-term load forecasting using svr (support vector regression)-based radial basis function neural network with dual extended kalman filter," *Energy*, vol. 49, p. 413–422, 2013.

[41] Y.-S. Lee and L.-I. Tong, "Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming," *Knowledge-Based Systems*, vol. 24, no. 1, p. 66–72, 2011.

[42] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in neural information processing systems*, vol. 32, 2019.

[43] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," arXiv preprint arXiv:1707.01926, 2017.

[44] B. Lim, S. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, p. 1748–1764, 2021.

[45] Q. Liu, J. Li, and Z. Lu, “St-tran: Spatial-temporal transformer for cellular traffic prediction,” *IEEE Communications Letters*, vol. 25, no. 10, p. 3325–3329, 2021.

[46] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, “Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction,” *Sensors*, vol. 17, no. 4, p. 818, 2017.

[47] L. R. Medsker and L. Jain, “Recurrent neural networks,” *Design and Applications*, vol. 5, p. 64–67, 2001.

[48] D. M. Miller and D. Williams, “Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy,” *International Journal of Forecasting*, vol. 19, no. 4, p. 669–684, 2003.

[49] W. Min and L. Wynter, “Real-time road traffic prediction with spatio-temporal correlations,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, p. 606–616, 2011.

[50] J. F. Muth, “Optimal properties of exponentially weighted forecasts,” *Journal of the american statistical association*, vol. 55, no. 290, p. 299–306, 1960.

[51] N. OpenCellID, “The world’s largest open database of cell towers,” 2022.

[52] J. K. Ord, A. B. Koehler, and R. D. Snyder, “Estimation and prediction for a class of dynamic nonlinear statistical models,” *Journal of the American Statistical Association*, vol. 92, no. 440, p. 1621–1629, 1997.

[53] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” arXiv preprint arXiv:1511.08458, 2015.

[54] M. Paolini and S. Fili, “Mastering analytics: How to benefit from big data and network complexity: An analyst report,” *RCR Wireless News*, vol. 20, no. 80, p. 120, 2017.

[55] C. Qiu, Y. Zhang, Z. Feng, P. Zhang, and S. Cui, “Spatio-temporal wireless traffic prediction with recurrent neural network,” *IEEE Wireless Communications Letters*, vol. 7, no. 4, p. 554–557, 2018.

[56] K. Rasul, A.-S. Sheikh, I. Schuster, U. Bergmann, and R. Vollgraf, “Multivariate probabilistic time series forecasting via conditioned normalizing flows,” arXiv preprint arXiv:2002.06103, 2020.

[57] Q. Raza, N. Mithulananthan, J. Li, and K. Y. Lee, “Multivariate ensemble forecast framework for demand prediction of anomalous days,” arXiv e-prints, p. arXiv–1811, 2018.

[58] N. I. Sapankevych and R. Sankar, “Time series prediction using support vector machines: a survey,” *IEEE computational intelligence magazine*, vol. 4, no. 2, p. 24–38, 2009.

[59] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE transactions on neural networks*, vol. 20, no. 1, p. 61–80, 2008.

[60] R. Snyder, “Recursive estimation of dynamic linear models,” Journal of the Royal Statistical Society. Series B (Methodological), p. 272–276, 1985.

[61] *Cellular traffic prediction and classification: A comparative evaluation of LSTM and ARIMA*. Springer, 2019.

[62] *Spatial bottleneck transformer for cellular traffic prediction in the urban city*. Springer, 2023.

[63] J. H. Stock and M. W. Watson, “Forecasting inflation,” *Journal of monetary economics*, vol. 44, no. 2, p. 293–335, 1999.

[64] G. Svennerberg, *Beginning google maps API 3*. Apress, 2010.

[65] J. W. Taylor, “Exponential smoothing with a damped multiplicative trend,” *International journal of Forecasting*, vol. 19, no. 4, p. 715–725, 2003.

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[67] J. Wang, W. Deng, and Y. Guo, “New bayesian combination method for short-term traffic flow forecasting,” *Transportation Research Part C: Emerging Technologies*, vol. 43, p. 79–94, 2014.

[68] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, and C. Peng, “Spatio-temporal analysis and prediction of cellular traffic in metropolis,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, p. 2190–2202, 2018.

[69] P. R. Winters, “Forecasting sales by exponentially weighted moving averages,” *Management science*, vol. 6, no. 3, p. 324–342, 1960.

[70] S. Wu, X. Xiao, Q. Ding, P. Zhao, Y. Wei, and J. Huang, “Adversarial sparse transformer for time series forecasting,” *Advances in neural information processing systems*, vol. 33, p. 17105–17115, 2020.

[71] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, “Graph wavenet for deep spatial-temporal graph modeling,” arXiv preprint arXiv:1906.00121, 2019.

[72] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” *IEEE/ACM transactions on networking*, vol. 25, no. 2, p. 1147–1161, 2016.

[73] R. A. Yaffee and M. McGee, *An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®*. Elsevier, 2000.

[74] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” arXiv preprint arXiv:1709.04875, 2017.

[75] U. Yule, “On a method of investigating periodicities in disturbed series, with special reference to wolfer’s sunspot numbers,” *Philosophical Transactions of the Royal Society of London Series A*, vol. 226, p. 267–298, 1927.

[76] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Communications surveys tutorials*, vol. 21, no. 3, p. 2224–2287, 2019.

[77] C. Zhang, H. Zhang, D. Yuan, and M. Zhang, "Citywide cellular traffic prediction based on densely connected convolutional neural networks," *IEEE Communications Letters*, vol. 22, no. 8, p. 1656–1659, 2018.

[78] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 3, p. 468–478, 2019.

[79] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, no. 1, p. 1–23, 2019.

[80] N. Zhao, Z. Ye, Y. Pei, Y.-C. Liang, and D. Niyato, "Spatial-temporal attention-convolution network for citywide cellular traffic prediction," *IEEE Communications Letters*, vol. 24, no. 11, p. 2532–2536, 2020.