Predicting the Vitality of Stores Along the Street Based on Business Type Sequence via Recurrent Neural Network

Zidong Liu[1], Yan Li[2], and Xiao Xiao[3(✉)]

[1] The University of Texas at Austin, 310 Inner Campus Drive, Austin TX 78712, USA
zidong.liu.22@utexas.edu
[2] University of Sydney, Camperdown, NSW 2006, Australia
yali3816@uni.sydney.edu.au
[3] Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, TO, Italy
xiao.xiao@polito.it

Abstract. The rational planning of store types and locations to maximize street vitality is essential in real estate planning. Traditional business planning relies heavily on the subjective experience of developers. Currently, developers have access to low-resolution urban data to support their decision making, and researchers have done much image-based machine learning research from the scale of urban texture. However, there is still a lack of research on the functional layout with shop-level accuracy. This paper uses a sequence-based neural network (RNN) to explore the relationship between the sequence of store types along a street and its commercial vitality. Currently, the use of RNNs in the architectural and urban fields is very rare. We use customer review data of 80streets from O2O platforms to represent the store vitality degree. In the machine learning model, the input is the sequence of store types on the street, and the output is the corresponding sequence of business vitality indexes. After training and evaluation, the model was shown to have acceptable accuracy. We further combined this evaluation model with a genetic algorithm to develop a business planning optimization tool to maximize the overall street business value, thus guiding real estate business planning at a high resolution.

Keywords: Machine learning · Big data analysis · Vitality prediction · Recurrent neural network · Genetic algorithm

1 Introduction

Business management experience shows that the order in which stores are located on the street has a significant effect on business. Stores at the street corner tend to have higher popularity and therefore higher rents. People without specific goals are more likely to shop at the first supermarket they see. Store's neighbours may also have a complex impact on its operation, depending on the types of stores. For example, two supermarkets that are located close to each other may have a vicious competition, but

for McDonald's and KFC, putting them together may help enhance their visibility on the street. A bank placed next to a luxury store may help to increase the sales of the luxury store.

1.1 Problem Statement

The current business planning model still relies heavily on the subjective experience of real estate developers, which leads to the uncertainty in planning results and adversely affects the profitability of businesses. There has been increasingly data analysis such as customer base analysis and regional vitality analysis to support low-resolution issues like the proportion of store types [5]. However, the resolution of these data is still not sufficient to guide shop-level planning. There is still a lack of research for more precise planning such as the location sequence of business types along a street. Therefore, this paper has a strong practical research significance.

1.2 Literature Review

There have been many studies on region vitality through machine learning. However, most of them are image-based (GAN-based) and do not achieve store-level accuracy. Among these studies, GAN models are predominant. A study transforms the citizens' cycling route data into an urban heat map to represent community vitality and explores its relationship with urban fabric [8]. Similar approaches can be used to predict other urban metrics, such as urban crime rate [3] and commercial value [7]. However, due to the limitation of computation and data resolution, the generated results always have ambiguous areas. This is the reason why some studies have attempted to vectorize images before performing machine learning [9].

In this study, we choose RNN as the basic neural network model. RNN is based on sequential data, widely used in natural language processing, advertising recommendations and so on. Compared with other models, RNN's features are highly compatible with our research object and goal. Here are the reasons:

1. RNN uses sequential data as input and output.
2. In RNN models, the order of data has a decisive influence on the results.
3. The input and output in the RNN training set can be of different lengths.

Among the sparse RNN-based studies in the architectural and urban fields, there is one relevant to the topic on business optimization [4]. Using the behaviour of pedestrian inside a mall as data, the researchers trained a behavioural predictor that can infer the pedestrian's walking direction. This model in turn guides the design of the mall, leading to higher commercial value on the pedestrian's expected route. In addition, some researchers have tried to use RNNs from the perspective of software operation.

1.3 Project Goal

The paper aims to explore the relationship between the order of store business types along the street and their commercial vitality by a sequence-based neural network (RNN). The

machine learning model simulates the behaviour of people walking down the street and passing through stores. In the model, the input is the sequence of store types and the output is the sequence of vitality indexes. After training, this machine learning model can predict the vitality of each store, thus guiding real estate business planning at a high resolution.

2   Methodology

The research process is divided into three parts: data collection, model training and model evaluation. We collected data of stores along the streets from O2O platforms including Gaode Map, Meituan and Dianping and transformed these data into sequences that can represent the types of stores and their sales status. After that, the sequence data are entered into the seq2seq model and trained in the LSTM layers. Then the model outputs the sequence of letters that can represent the vitality level. Finally, we use Cross Entropy Loss Function and the prediction accuracy function to evaluate the effectiveness of this prediction model (Fig. 1).

After obtaining the prediction model, a street outside the training set is used to verify the effectiveness of the model. Furthermore, we can combine this prediction model with a genetic algorithm to develop a business planning optimization tool: it automatically gives the best ranking order based on the input store types to maximize the business value of the whole street.
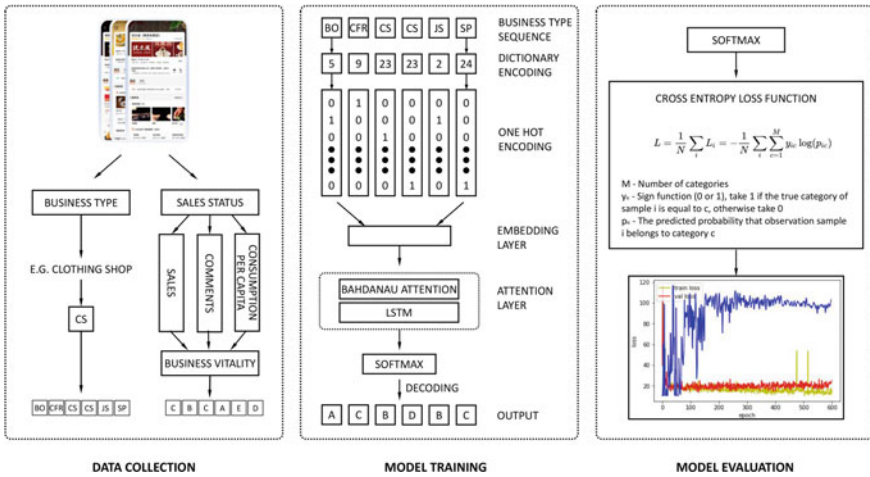

Fig. 1. Research framework

2.1   Data Collection

We selected 80 streets, 1261 stores, and 29 store types from 8 representative cities in China from O2O platforms (Fig. 2). As the main O2O platforms vary from city to city and

different merchants on the same street might choose different platforms, it is necessary to collate data from multiple mainstream platforms. In this research, the commercial data was comprehensively collected on Meituan, Dianping and Gaode Map. In this way, we collect as complete data as possible for every store on each of the 80 streets. Regarding a tiny number of shops with missing data, we take the average of the nearby shops of the same type as a replacement. O2O platforms provide a variety of information: shop type, number of reviews, per capita spending. There is also information on sales volume (some semi-annual, some monthly).

Fig. 2. POI data statistics

2.2 Data Processing

Quantitative assessment of business vitality is very complex since no platform provides direct information on the sales of every shop in the street. Based on the assumption that all shops have the same review rate, we can use the number of reviews multiplied by the per capita spend to estimate the sales of each shop. However, after research, we found that the type of shop significantly impacts the number of reviews. For example, milk tea shops and fast-food restaurants tend to have very high review rates. In contrast, some support facilities such as banks and bicycle repair points have low review rates though their existence can have a significant impact on the surrounding stores.

In order to provide a more objective assessment of the commercial viability of shops, a relative quantity approach is applied here. For these 1261 shops, we compare the number of reviews multiplied by the value of per capita consumption within each type of shop,

and then classify their relative vitality into five classes: ABCDE. For example, there are 75 pastry shops, so we rank their vitality, then the top 10 are ranked A, 11–25 are ranked B, and so on (Fig. 3). For those supporting facilities with few reviews like banks, we unify their vitality value C. After calculating the vitality values of the stores in these 80 streets, we can get some interesting statistical conclusions. Shanghai, Nanjing, Wuhan and Suzhou have higher average store vitality than Kunming and Changsha, which is in line with daily experience: store vitality is positively correlated with the economic development of a city (Fig. 4).
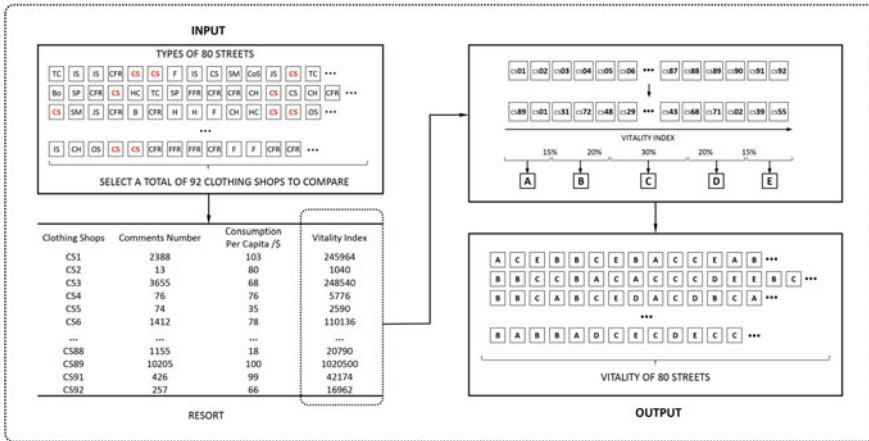
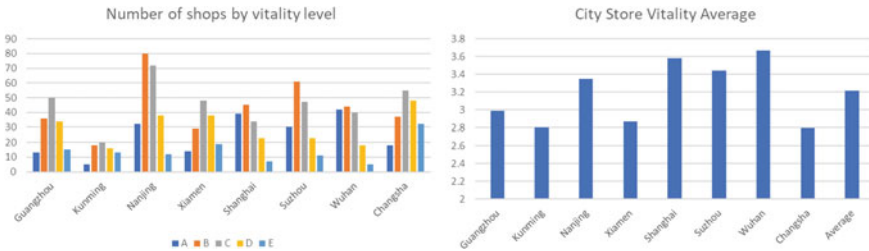**Fig. 3.** Translating business data into relative vitality values

Fig. 4. Comparison of city store vitality

2.3 Training Set Expansion

The machine learning model simulates the behaviour of people walking down the street and passing through stores that is a one-way experience. However, since both ends of the street can be the starting points, the sequences can all be trained in reverse, so the dataset was expanded from 80 streets to 160. To expand the sample size further, we extracted all the subsequences whose length are greater than five from the beginning of these

160 sequences (Fig. 5). This is reasonable because we may not go through the whole street in daily shopping but finish shopping after passing several stores. By this method, we obtained a total of 1820 sequential data. This method of expanding the database is inspired by the research of Weixin Huang's team on the modelling operation, in which they also applied a similar subsequence approach [2].
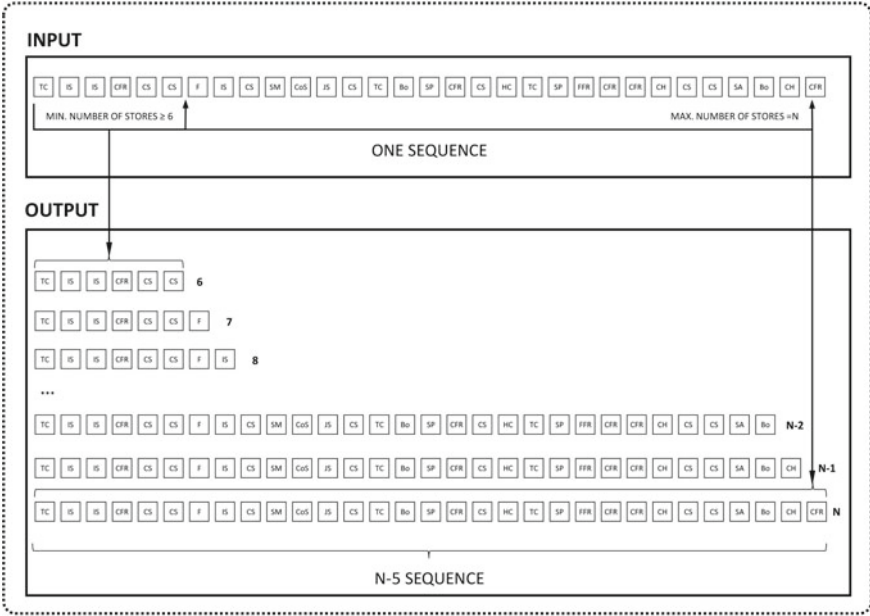

Fig. 5. Sub-sequences generation for the training set expansion

2.4 Machine Learning

Machine training is based on the Seq2Seq attention model (Fig. 1). Data set is divided into the training set, validation set and test set according to the ratio of 7:2:1. We evaluate the effectiveness of this model by two functions: Cross Entropy Loss Function (Eq. 1) and the Prediction Accuracy Function (Eq. 2). The Prediction Accuracy Function is formulated by the specific issue of this paper. The difference between the predicted value and the target value varies depending on the predicted value (Table 1). The accuracy of random guess is the sum of all the values in Table 3 divided by 25 equals 46.56%.

$$L = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i L_i \sum_{c=1}^{M} y_{ic} \log(P_{ic}) \tag{1}$$

Table 1. Accuracy calculation table

Accuracy		Target vitality				
		A	B	C	D	E
Perdicted vitality	A	1	0.66	0	0	0
	B	0.75	1	0.5	0.33	0.25
	C	0.5	0.66	1	0.66	0.5
	D	0.25	0.33	0.5	1	0.75
	E	0	0	0	0	1

M: Number of categories y_{ic}: Sign function (0 or 1)

P_{ic}: The predicted probability that ith item belongs to category c

$$P = \frac{1}{n_t} \sum_{i=0}^{\min{(n_t, n_p)}} 1 - \frac{\Delta r_i}{\max(R - r_t, r_t - 1)} \times 100\%, \quad \Delta r_i = |r_{ip} - r_{it}| \quad (2)$$

R: Range of vitality level

n_t: Target sequence length n_p: Predicted sequence length

r_{it}: Target vitality of the i th term r_{ip}: Predicted vitality of the i th term

The training results after 600 epochs with 15 batches per epoch are shown in Fig. 6. The training effect is good. The model never enters the overfitting state since the training loss curve and the validation loss curve remain stable and the accuracy curve keeps increasing.
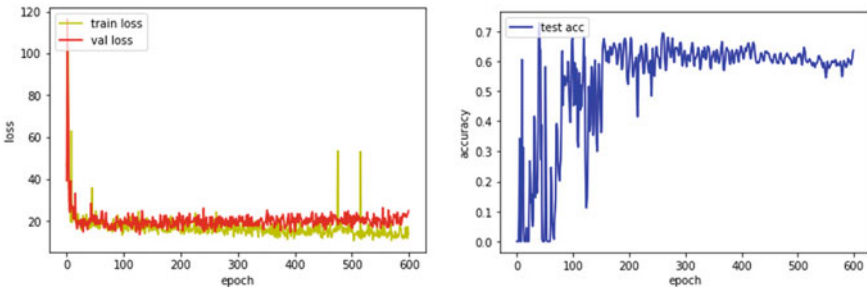

Fig. 6. Training results

3   Case Study

We chose Gungyuan West Street in Nanjing, outside the training set, to apply our trained evaluation model. Gongyuan West Street is in the historical centre area of Nanjing, with a wide variety of businesses and high popularity. The commercial situation of the site is shown in Fig. 7.
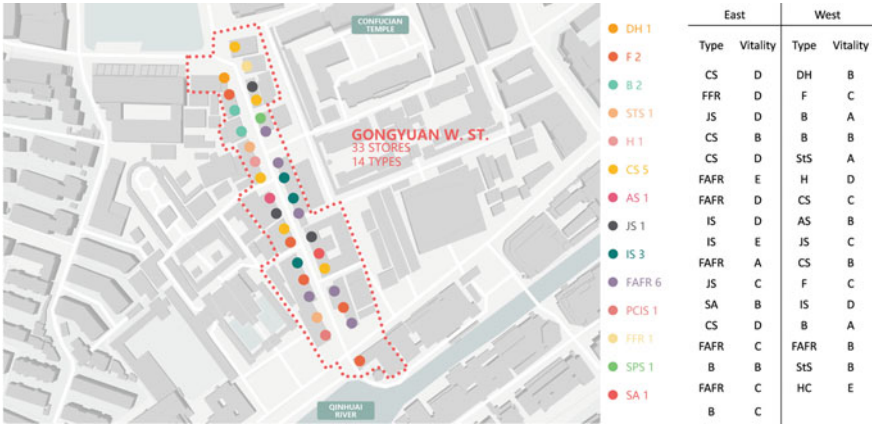

| | East | | West | |
	Type	Vitality	Type	Vitality
	CS	D	DH	B
	FFR	D	F	C
	JS	D	B	A
	CS	B	B	B
	CS	D	StS	A
	FAFR	E	H	D
	FAFR	D	CS	C
	IS	D	AS	B
	IS	E	JS	C
	FAFR	A	CS	B
	JS	C	F	C
	SA	B	IS	D
	CS	D	B	A
	FAFR	C	FAFR	B
	B	B	StS	B
	FAFR	C	HC	E
	B	C		

Legend: DH 1, F 2, B 2, STS 1, H 1, CS 5, AS 1, JS 1, IS 3, FAFR 6, PCIS 1, FFR 1, SPS 1, SA 1

GONGYUAN W. ST. 33 STORES 14 TYPES

Fig. 7. Vitality of stores in Gongyuan West Street

The types of stores in West Street were input into the trained model, and the output vitality prediction was "b c b c b c b c b c b c b c b c b c", with an accuracy of 77% according to Formula (2) (Table 1). Experiment 1 adds a movie theatre at the beginning of the street, and the model had a higher expectation of street vitality (Fig. 8). Experiment 2 arranges the same kinds of stores together. The model also has a higher expectation of the overall vitality of the street (Fig. 9 and Table 2).

```
testSeq = ["TC","DH","F","B","B","StS","H","CS","AS","JS","CS","F","IS","B","FAFR","StS","HC"]
testSeq_str=" ".join(str(i) for i in testSeq).lower()
print(m.infer(testSeq_str)[0])
```
```
[191]  ✓  0.7s
...    b c b b c b b c b b c b b c b c b
```

Fig. 8. Vitality of stores in Gongyuan West Street after Experiment 1

3.1   Vitality Optimization Based on Genetic Algorithm

Further, we combined this evaluation model with a genetic algorithm to develop a reference tool that can provide suggestions for optimizing the location of stores. The vitality levels correspond to specific numbers: A scores 5, B scores 4, C scores 3, D scores 2,

```
testSeq = ["DH","CS","CS","FAFR","F","F","StS","StS","B","B","H","AS","JS","IS","B","HC"]
testSeq_str=" ".join(str(i) for i in testSeq).lower()
print(m.infer(testSeq_str)[0])
```
[226] ✓ 0.6s

... b c b c b c b b c b c b a a a a

Fig. 9. Vitality of stores in Gongyuan West Street after Experiment 2

Table 2. Expected street vitality after location adjustment

Target	B	C	A	B	A	D	C	B	C	B	C	D	A	B	B	E	
Prediction	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	
Accuracy	100%	100%	80%	75%	80%	75%	66%	75%	66%	75%	66%	75%	80%	75%	100%	50%	77%

and E scores 1. The genetic algorithm takes the total score of vitality as the optimization target. At each iteration, the genetic algorithm randomly swaps two store locations. Through continuous iterations, the genetic algorithm then gives the optimal solution of this prediction model.

After hundreds of iterations, the system did find a solution with a high vitality index: "CS F STS CS AS JS IS STS B HC B DH B H F FAFR". The vitality prediction for this sequence is: "B C B A A A A A A A A A A A A A A A A A" with a score of 76. Figure 10 records an evolutionary process.
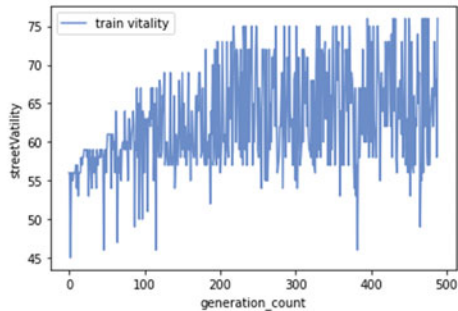

Fig. 10. Overall street vitality optimization by genetic algorithm

4   Conclusion

This paper presents a method that uses machine learning to predict commercial vitality along streets and provide optimization advice. This study has important practical value for high-precision business planning. Although there have been many machine learning studies based on urban texture images, few studies are accurate to the prediction of vitality of stores. Compared with previous studies, this study creatively interpreted people's walking and shopping behaviour in the street as a linear sequence. It converted POI data collected from the O2O platform into a sequence format to train the RNN model.

In the future, this study still has much room for improvement. The accuracy of the current model is still not high enough. In the data collection stage, a larger data set is needed in the future. Since the information accuracy requirement is very high (relative location of each store), the automatic POI data collection method based on geographic coordinates is not applicable. Currently, we use manual methods to collect data one by one along the street. In the future, however, automated data collection algorithms will have to be developed to replace the current manual methods to remarkably expand the scale of the training set. In the data processing stage, there are many noise points in the data set due to many factors affecting the vitality of the real-world stores. In the future, homogenized data algorithms will be used to eliminate the effect of noise [6]. In the model training phase, we will use more RNN models such as Transformer, GRU, BiLSTM to compare which model is more suitable for this research in the future.

References

1. Gao W et al (2021) A data structure for studying 3D modeling design behavior based on event logs. Autom Constr 132(103967):103967. https://doi.org/10.1016/j.autcon.2021.103967
2. Gao W et al (2022) Command prediction based on early 3D modeling design logs by deep neural networks. Autom Constr 133(104026):104026. https://doi.org/10.1016/j.autcon.2021.104026
3. He J, Zheng H (2021) Prediction of crime rate in urban neighborhoods based on machine learning. Eng Appl Artif Intell 106(104460):104460. https://doi.org/10.1016/j.engappai.2021.104460
4. Karoji G, Hotta K, Hotta A, Ikeda Y (2019) Pedestrian dynamic behaviour modeling. In: Proceedings of the 24th international conference on computer-aided architectural design research in Asia: intelligent and informed, CAADRIA 2019, pp 281–290
5. Schlegel A, Birkel HS, Hartmann E (2021) Enabling integrated business planning through big data analytics: a case study on sales and operations planning. Int J Phys Distrib Logist Manag 51(6):607–633. https://doi.org/10.1108/ijpdlm-05-2019-0156
6. Sheng Q et al (2018) The application of space syntax modeling in data-based urban design—an example of Chaoyang square renewal in Jilin city. Lands Archit Front 6(2):102. https://doi.org/10.15302/j-laf-20180211
7. Shou X, Chen P, Zheng H (2021) Predicting the heat map of street vendors from pedestrian flow through machine learning. In: Proceedings of the 26th international conference on computer-aided architectural design research in Asia: projections, CAADRIA 2021, pp 569–578
8. Sun YJ, Jiang L, Zheng H (2021) A machine learning method of predicting behavior vitality via urban forms. In: Proceedings of the 40th international conference on computer aided design in architecture: distributed proximities, ACADIA 2021, pp 160–168

9. Xia X, Tong Z (2020) A machine learning-based method for predicting urban land use. In: Proceedings of the 25th international conference on computer-aided architectural design research in Asia: anthropocene, CAADRIA 2021, pp 21–30