CITY METRO NETWORK EXPANSION WITH REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper presents a method to solve the city metro network expansion problem using reinforcement learning (RL). In this method, we formulate the metro expansion as a process of sequential station selection, and design feasibility rules based on the selected station sequence to ensure the reasonable connection patterns of the metro line. Following this formulation, we train an actor critic model to design the next metro line. The actor is a seq2seq network with attention mechanism to generate the parameterized policy which is the probability distribution over feasible stations. The critic is used to estimate the expected reward, which is determined by the output station sequences generated by the actor during training, in order to reduce the training variance. The learning procedure only requires the reward calculation. Thus our general method can be extended to multi-factor cases easily. Considering origin-destination (OD) trips and social equity, we expand the current metro network in Xi'an, China, based on the real mobility information of 24,770,715 mobile phone users in the whole city. The results demonstrate the effectiveness of our method.

1 INTRODUCTION

City metro network plays an important role in public transportation system. With the development of city, new transportation demands have led to the expansion of metro network. The last few years have witnessed tremendous expansion of metro network (Sun et al., 2018). On the other hand, the expansion of the metro network in turn has a profound impact on the city. The expanded lines may change the mobility trend of city population. Most previous research focuses on the design of metro network from scratch (Gutiérrez-Jarpa et al., 2018; Laporte & Pascoal, 2015). However, in the subsequent construction process, the dynamic of the city has been different from that in the initial stage. The original scheme may not be suitable for the current situation. Therefore, it is more reasonable to gradually design new lines to expand the metro network according to current city dynamic.

Usually, transportation planning objectives are mobility-based, such as maximizing OD trips. As the society progressed, sustainability has increasingly become the demand of city development. The sustainability prompts governments to re-recognize the role of the transport system, and thereby influences their transport policy (Manaugh et al., 2015). One conception of sustainability, social equity, which can be measured by the distributable benefit accessibility (Behbahani et al., 2019), has been acknowledged important. There have been several real transportation plans considering social equity (Arsenio et al., 2016). Metro network, an important transportation system, has a great influence on social equity. Therefore, in this work, we consider both OD trips and social equity to expand city metro network.

However, the problem becomes difficult when the city becomes large. First, it is difficult to formulate the problem efficiently (Laporte & Mesa, 2015). Existing studies formulate the problem as nonlinear integer programming, and call for an exponential number of subtour elimination constraints to ensure the rationality of expanded metro lines (Gutiérrez-Jarpa et al., 2013; Wei et al., 2019), which hinders solving the problem efficiently.

Second, the huge solution space makes it difficult to find a good solution effectively. Exact methods are inapplicable for integer programming problems with large solution space (Farahani et al., 2013). One common method in existing studies is to limit the search space. Previous work (Gutiérrez-Jarpa

et al., 2013; Wei et al., 2019; Laporte & Pascoal, 2015; Gutiérrez-Jarpa et al., 2018) predefines corridors based on expert knowledge, and only consider to design metro lines in these corridors. Their results depend on expert guidance, and it is possible that the best solution is left out. Therefore, we call for a method which does not require expert knowledge.

Carefully handcrafted heuristics, embedded with problem-specific knowledge, are usually efficient for large search space (Dufourd et al., 1996). However, the factors considered in the expansion of metro line vary with different cities and stages. Once the objectives of metro expansion change, heuristic methods need to be revised. Rather than handcrafting different heuristics for different objectives, a general method is necessary.

In this paper, we propose a method without expert knowledge to solve the city metro network expansion problem using RL. We formulate the metro line expansion as a process of sequential station selection, an MDP, and design feasibility rules based on the selected station sequence to ensure the reasonable connection patterns of the metro line. This formulation efficiently characterize the expansion of the metro line, without heavy constraints like existing studies (Wei et al., 2019).

Following this formulation, we propose an actor critic model (Konda & Tsitsiklis, 2000) to generate the next metro line. The actor is based on the seq2seq model (Sutskever et al., 2014). In the actor, an encoder characterizes the timely metro station information in the expansion process. After encoding, an RNN decoder is used to characterize the sequence information of the selected stations. Moreover, we employ an attention layer to produce the probability distribution over feasible candidate stations. Only requiring the reward calculation, we train the model with the critic reducing training variance, in order to find the high-priority metro line following feasibility rules. The reward function takes the final output station sequence as input only, which is friendly to objective changing. Therefore, our model is general for different objectives. Without expert knowledge, the learning procedure drives the policy to keep track of the better solution during the search and to search for better solutions. Its natural exploration mechanism determines that RL is suitable for large scale solution space.

Based on real city-scale human mobility information of 24,770,715 mobile phone users obtained from a citywide 3G cellular network, we expand the current metro network in Xi'an, China. The results demonstrate the effectiveness of our method.

Our contributions are as follows:

1. We formulate the expansion of a metro line as a process of sequential stations selection, a Markov decision process (MDP). We design feasibility rules based on the selected station sequence to ensure the reasonable connection patterns of metro line, which is more efficient to formulate the problem than integer programming models.
2. We firstly propose a RL based method to solve the city metro network expansion problem. Without expert knowledge, our general method can be easily extended to the metro expansion considering multi-factors.
3. We incorporate social equity concerns into metro network expansion. Compared with the realistically planned lines, the results show the rationality of considering social equity in transportation planning.
4. We use real city-scale human mobility information to expand a metro network. The experimental results demonstrate the effectiveness of our method.

2 Related work

2.1 Metro network design

Metro network plays an important role in public transportation system. For metro network design, Laporte & Pascoal (2015) predefine a set of corridors, and propose path based algorithms. Considering the connection with existing lines, Wei et al. (2019) expand the metro network in predefined corridors. However, these methods depend on predefined corridors by planner, which may be interfered by human judgment. Gutiérrez-Jarpa et al. (2018) propose a greedy generation heuristic to select a set of corridors with higher passenger traffic from a set of candidate corridors. Enumerating all candidate corridors is impossible, which may leave out some good solutions. The corridors may need frequent adjustments, once the objectives change. In addition, these studies adopt non-linear

integer programming models to formulate the metro network design problem, and call for exponential number of constraints to ensure the rationality of expanded metro lines, which makes it difficult to solve the problem efficiently. To cope with these problems above, we propose a general method that is suitable for diverse objectives and does not require prior knowledge to expand city metro network.

2.2 REINFORCEMENT LEARNING

The strength of RL lies in its powerful decision-making ability. RL has made great progress in complicated tasks like playing Atari games (Mnih et al., 2013), robot training (Yang et al., 2017) and combinatorial optimization (Bello et al., 2016; Nazari et al., 2018). RL is proved efficient in problems with high dimensional search space, such as Go (Silver et al., 2016) and StarCraft (Vinyals et al., 2017). Meanwhile, classical combinatorial optimization problems like minimum cut (Li et al., 2018b), traveling salesman problem (Bello et al., 2016; Kool & Welling, 2018) and vehicle routing problem (Nazari et al., 2018) are solved by RL. Li et al. (2018a) points out that RL can find important factors which may be ignored by human. We believe that RL has the ablity to solve the metro expansion problem.

3 PROBLEM FORMULATION

In this paper, we design the next metro line to expand the current metro network in the target city. The metro line is determined by stations and line routing, and is allowed to connect with existing lines to form transfer stations.

For a target city, we divide it into $n \times n$ grids in a two dimensional space $\{g_i\}_{i=0}^{n^2-1}$. Each grid g_i is a square with a width of d, and its center is a candidate station s_i. We define the expansion of metro network on an undirected graph $\mathcal{G} = (\mathbb{S}, \mathbb{E})$, where $\mathbb{S} = \{s_0, s_1, ..., s_{n^2-1}\}$ contains all candidate stations and $\mathbb{E} = \{(s_i, s_j) \,|s_i, s_j \in \mathbb{S}\}$ contains all direct links between stations in existing lines. Each grid g_i is associated with a compound index of development D_i, and any two candidate stations s_i and s_j are associated with the two-way symmetrical OD trips $od_{i,j} (= od_{j,i})$ between them. We present the expanded metro line as an ordered station sequence $Z = (z_1, z_2, \ldots, z_T)$, $z_i \in \mathbb{S}$, and it should satisfy the following constraints.

- The consecutive stations must follow the minimum-maximum distance rules.
- The line routing should ensure reasonable connection patterns of stations, avoiding subtour and meandering line.
- The number of the stations is limited by N.
- The budget is limited by B.

Satisfied by the new line Z, we denote the newly added OD trips as $R_{od}(Z)$ and the social equity indicator as $R_{ac}(Z)$, which aims to maximize the total benefits of the society. More details about these two are in Appendix A. The objective of the new metro line is to maximize the weighted sum of OD trips and social equity

$$\omega\left(Z\right) = \alpha_1 \times R_{od}(Z) + \alpha_2 \times R_{ac}(Z) \tag{1}$$

where α_1 and α_2 are the weights of added OD trips and social equity, and $\alpha_1 + \alpha_2 = 1$.

4 METHOD

4.1 RL FORMULATION

We focus on the city metro network expansion problem. We denote the station selected at step t as z_t, and the selected station sequence until step t as $Z_t = \{z_1, z_2, \ldots, z_t\}$. Given the graph $\mathcal{G}$, we aim to learn a parameterized policy $P\left(Z|\mathcal{G}\right)$ to maximize ω.

$$P\left(Z|\mathcal{G}\right) = \prod_{t=1}^{T} P\left(z_t|\mathcal{G}, Z_{t-1}\right), \tag{2}$$

and

$$M_t = M(Z_t) \tag{3}$$

Equation (3) characterizes the update of feasibility rules. We use a binary vector M_t to indicate whether each station can be selected in next step. According to the rules, feasible stations are set to 1, oherwise 0 (see Section 4.2 for more details). Important RL elements are listed as follows.

Environment. The environment correspond to the city in which the metro is expanded. The RL agent will build the station in order, completing the metro network expansion.

State. For step t, the selected station sequence Z_t is the current state. After the agent chooses the next station z_{t+1}, the state will be updated with Z_{t+1}.

Action. We define the station z_t which the agent chooses to build next step as the action. Once the action is decided, the current state will be updated.

Reward. Since RL are set to maximize the reward, we define the value of objective ω as reward. The reward function is calculated when the metro expansion is done.

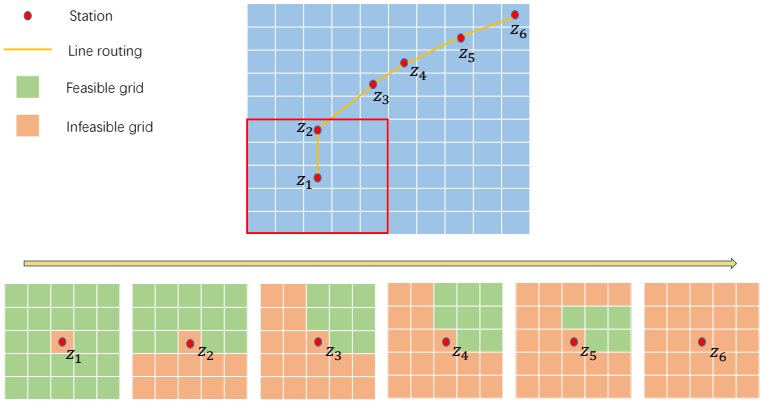


Figure 1: An example of metro line expansion.

4.2 Feasibility Rules

In this section, we present the feasibility rules (3) to satisfy the constraints in Section 3. The RL agent can only select stations in a certain filter to ensure the minimum-maximum distance rule. The filter is a square shape with $m \times m$ grids. To ensure reasonable connection patterns of stations, we design the action direction rules based on history actions to determine the candidate action in the next step (see Appendix B for more details). During the expansion, the expanded metro line is allowed to connect with existing lines to form transfer stations, but it cannot coincide with the existing lines. Per unit length of metro line and per station consume a certain cost. Once the number of selected stations reaches the upper limit N, or the cost of construction exceeds the budget B, the expansion process is terminated. Based on these rules, the agent dynamically acquires the next optional stations, which avoids exponential constraints in existing studies (Gutiérrez-Jarpa et al., 2013; Wei et al., 2019). An example is shown in Figure 1.

4.3 Network Architecture

As depicted in Figure 2, the neural network serially chooses the next station, with continuous interaction with the environment. We use the seq2seq model with attention mechanism to construct the neural network (Sutskever et al. (2014); Vaswani et al. (2017)).

At step t, the encoder which includes a single 1-dimensional CNN layer is used to embed the information of candidate stations into a d-dimensional space. During the encoding step, we consider several information for each grid $\{g_i\}_{i=1}^{n^2-1}$, including coordinates and whether it is selected as a station by the RL agent. The information after embedding can be denoted as $X^t = \{X_i^t\}_{i=1}^{n^2-1}$.

The decoder consists of an RNN layer and is used to output the current hidden state h^t. It takes the embedded information of the last selected station z_t and memory hidden state as input. After

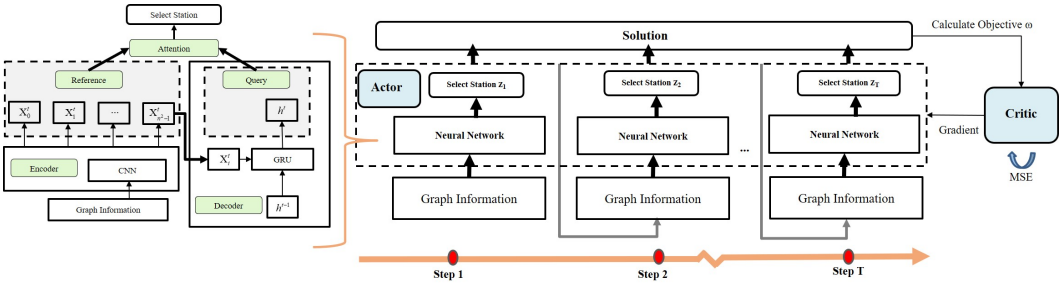


Figure 2: Network architecture

decoding, the attention mechanism is employed. The attention mechanism takes the encoder output X_t as reference and the hidden state h^t as query to generate the probability distribution which indicates the probability of each candidate station. In the attention mechanism, the probability is calculated as

$$q_i^t = v_a^T \tanh(W_a[X_i^t; h^t]),\ a^t = \text{softmax}(q^t) \tag{4}$$

$$c^t = \sum_i a_i^t X_i^t \tag{5}$$

$$P(z_{t+1}|\mathcal{G}, Z_t) = \text{softmax}(v_c^T \tanh(W_c[X_i^t; c^t]) \oplus H \cdot M_t) \tag{6}$$

where a^t is the relevant parameter, and c^t is the context parameter. W_a, W_c, v_a^T and v_c^T are training parameters. $[\ldots;\ldots]$ represents the element-wise concatenation operator and $\oplus$ represents the element-wise sum operator. H is a huge constant and M_t reflects the feasibility rules.

Once the station is selected, the current hidden state and the embedded information of the selected station will be the input of next decoder step. The process of choosing stations repeats until any termination condition is reached. Detailed processing procedure is presented in Algorithm 1.

Algorithm 1 Processing Procedure

Input Graph $\mathcal{G}$

for t $= 0, ..., \mathrm{T} - 1$ **do**
  Update the feasible stations M_t by Equation (3).
  Compute embeddings X^t of the current state by encoder.
  Compute the query h^t according to last hidden state h^{t-1} and the embedding of last selected station X_i^t by decoder (Initialize h^0 when $t = 0$).
  Compute the attention value a^t by Equation (4), and the probability $P(z_{t+1}|\mathcal{G}, Z_t)$ by Equation (6).
  Select Station z_{t+1} with the probability $P(z_{t+1}|\mathcal{G}, Z_t)$.
end for
return Solution $= Z_T$

4.4 Training

Taking learning methods into consideration, due to the lack of labels, the metro expansion problem cannot be tackled with supervised learning. The quality of supervised learning model is closely related to the quality of labels which are hard to obtain.

In contrast, RL provides the method for training the network. By setting the network parameters as θ, given the graph $\mathcal{G}$, the training objective can be denoted as

$$J(\theta|\mathcal{G}) = \mathbb{E}_{Z\sim p(\cdot|\theta)}\omega(Z|\mathcal{G}) \tag{7}$$

Parameters are trained with the policy gradient algorithm (Williams (1992)).

$$\nabla J(\theta|\mathcal{G}) = \mathbb{E}_{Z\sim p(\cdot|\theta)}[(\omega(Z|\mathcal{G}) - b(\mathcal{G}))\nabla_\theta \log p_\theta(Z|\mathcal{G})] \tag{8}$$

where $b(\mathcal{G})$ represents the baseline for reducing the training variance.

During training, the well known actor critic algorithm is used. For each step, the actor chooses the next station to build according to the network in Section 4.3. The critic is set to calculate the baseline in Equation (8). The network architecture is simple for critic, containing a single Dense and ReLu layer after the original information is embedded with CNN. Detailed algorithm is shown in Algorithm 2.

Algorithm 2 Actor Critic Training

Require: Batch size B, Training epoch E, Step T
Initialize actor parameters θ
Initialize critic parameters θ_c
for $\mathrm{epoch} = 1, ..., \mathrm{E}$ **do**
  Reset city metro network $\mathcal{G}$
  for $\mathrm{t} = 1, ..., \mathrm{T}$ **do**
    Select the next station by actor, $z_t \leftarrow \mathrm{SampleSolution}\,(p(\cdot|\theta))$
    Update feasibility rules $M_t = M(Z_t)$
  end for
  Find a solution Z^i for each batch, where $i \in \{1, \ldots, B\}$
  Calculate baseline $b(\mathcal{G})$ by critic
  $\nabla J(\theta|\mathcal{G}) \leftarrow \frac{1}{B}\sum_{i=1}^{B}(\omega(Z^i|\mathcal{G}) - b(\mathcal{G}))\nabla_\theta \log p_\theta\left(Z^i|\mathcal{G}\right)$
  Update the parameters of actor $\theta \leftarrow \mathrm{Adam}(\theta, \nabla J(\theta|\mathcal{G}))$
  $\nabla L_c \leftarrow \frac{1}{B}\sum_{i=1}^{B}(\omega(Z^i|\mathcal{G}) - b(\mathcal{G}))^2$
  Update the parameters of critic $\theta_c \leftarrow \mathrm{Adam}(\theta_c, \nabla L_c)$
end for

5 EXPERIMENTS

We conduct a case study to investigate the behavior of our method. The data and parameters are presented in Section 5.1. We conduct our method to expand the metro network in Section 5.2. Compared with baseline in Section 5.3, we demonstrate that our method is not in need of expert knowledge. We sequentially expand several metro lines in Section 5.4.

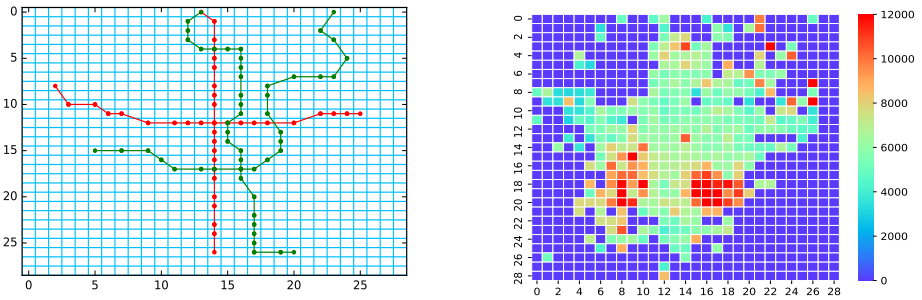
5.1 DATA AND PARAMETERS

The research is conducted based on the metro network in Xi'an, Shaanxi Province, China. Its first line started operation on September 16, 2011, and four lines are in operation by September 2019. Our experimental data, coming from a citywide 3G cellular data network, records the mobility information of 24,770,715 mobile phone users from October 1, 2015 to October 31, 2015. Figure 3(a) shows the metro network with 4 lines, where the red lines represent the existing 2 lines before October, 2015; the green lines represent the subsequently opened 2 lines; the dots represent stations.

Through data analysis referring to Yin et al. (2017), we set each grid as a square with 1000 meters width and divide the study area into 29×29 grids. Correspondingly, we set the size of filter in Section 4.2 as 5×5 according to Laporte & Pascoal (2015). Subsequently, we calculate the OD trips between any two grids. Figure 3(b) presents the distribution of house prices in 2015. The house price of grid g_i is used to characterize its index of development D_i, which is applied to the calculation of social equity (see Appendix A for more details).

5.2 METRO EXPANSION RESULTS

We conduct our method to design the next metro line according to the objective (1). The results are shown in Figure 4, and corresponding indicators are presented in Table 1. With the decrease of α_1, the expanded line shifts from satisfying more OD trips in Figure 4(a) to achieving higher social equity indicator in Figure 4(c). Among them, the expanded line in Figure 4(c) passes through

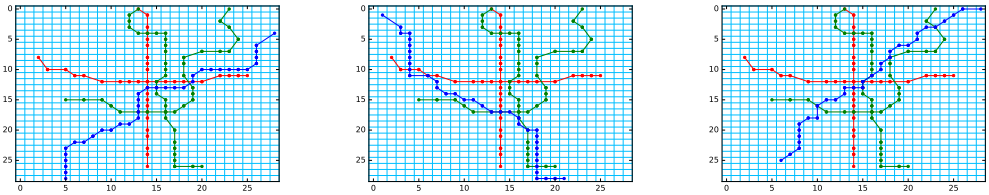


(a) Metro network. (b) House price (RMB).

Figure 3: The current city operational status.

the grids with high development level according to Figure 3(b), which intuitively demonstrates the effectiveness of our method when social equity is considered only. While considering OD trips and social equity as equally important, the expanded line satisfies two factors in a balanced way. Its shape is like a partial combination of the 2 subsequently opened lines after October, 2015.

Different objectives lead to different metro network. The objectives of metro expansion vary with different cities and stages, which may cause existing methods, whether heuristics or predefined corridors, to be revised. By changing the reward function, our method can be easily extended to different objectives, without problem-specific knowledge. Therefore, our method is general, and more suitable for metro expansion.



(a) $\alpha_1 = 1, \alpha_2 = 0$ (b) $\alpha_1 = 0.5, \alpha_2 = 0.5$ (c) $\alpha_1 = 0, \alpha_2 = 1$

Figure 4: The next expanded metro line with higher priority. The blue lines are our expanding lines.

Table 1: The newly satisfied OD trips and social equity

	R_{od}	R_{ac}	ω
$\alpha_1, \alpha_2 = 1, 0$	48.17	34.60	48.17
$\alpha_1, \alpha_2 = 0.5, 0.5$	26.79	33.18	29.98
$\alpha_1, \alpha_2 = 0, 1$	27.38	36.64	36.64

5.3 Comparisons with baseline

We compare our method against the baseline proposed in Wei et al. (2019). As for the baseline, we predefine corridors and end nodes as Wei et al. (2019) did. We conduct the baseline to expand metro line only in these corridors, with the end nodes as endpoints of metro lines.

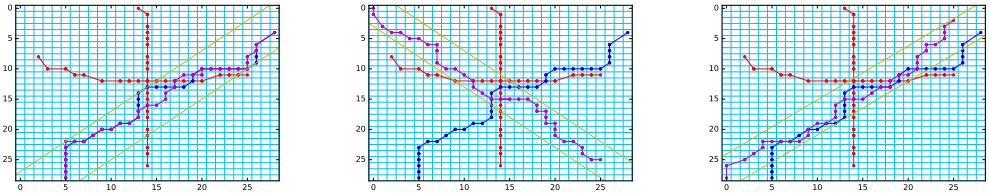
Specifically, we firstly predefine corridor 1-3 in Table 2, which respectively contain our expanded lines with different objectives in Section 5.2. Taking the first predefined corridor in Figure 5(a) as an example, the area between the two yellow lines is the predefined corridor. After the corridor is determined, the optimization is conducted by Gurobi with the same objective function (maximizing OD trips). The result of corridor 1 is slightly better than ours according to Table 2. The same results appear in corridor 2 and 3 for the corresponding objectives.

Then we randomly predefine the fourth and fifth corridors, which are shown in Figure 5(b) and Figure 5(c). Similarly, we conduct the baseline in these two corridors with different objectives respectively. Corresponding results are presented in Table 2. Compared with our method, all the expanded lines in the fourth and fifth corridors show bad performance.

Table 2: Comparisons with baseline (Wei et al., 2019).

Method	$\alpha_1 = 1$			$\alpha_1 = 0.5$			$\alpha_1 = 0$		
	R_{od}	R_{ac}	ω	R_{od}	R_{ac}	ω	R_{od}	R_{ac}	ω
Corridor 1	48.33	-	48.33	-	-	-	-	-	-
Corridor 2	-	-	-	27.53	33.93	30.73	-	-	-
Corridor 3	-	-	-	-	-	-	-	37.14	37.14
Corridor 4	23.97	-	23.97	16.70	22.68	19.59	-	26.87	26.87
Corridor 5	28.63	-	28.63	27.52	21.21	24.37	-	21.95	21.95
Our model	48.17	-	48.17	25.03	31.63	28.33	-	36.64	36.64

It can be vividly shown that the baseline method relys heavily on expert guidance. Once the corridor contains several good solutions, optimization tools such as Gurobi perform well in metro expansion problem. While the conditions is far more complicated in real world, causing it difficult to find the appropriate corridors. The fourth and fifth corridors indicate the bad results of choosing the unsuitable area. However, regardless of expert knowledge, RL shows the ability to find a good solution. The results are close to those conducted by Gurobi, showing the effiveness of our method.



(a) The first predefined corridor. (b) The fourth predefined corridor. (c) The fifth predefined corridor.

Figure 5: Comparisons with the expanded lines in predefined corridors. The area between two yellow lines is the predefined corridors. The blue (violet) lines are the expanded lines with our (baseline) method, considering only OD trips.

5.4 Expansion of multiple metro lines

Considering OD trips and social equity as equally important, we sequentially design the second metro line with the expanded line in Figure 4(b) as the existing line. The expanded line presented in violet is shown in Figure 6. Its lower left part is similar to the manually designed sixth metro line in reality. In this way, we design the metro lines sequentially and gradually expand the metro network.

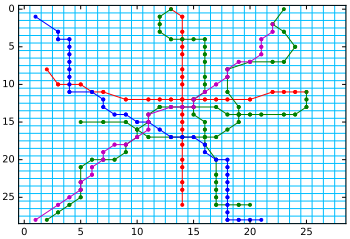


Figure 6: Multiple expanded metro lines.

6 Conclusion

This paper presents a method to solve the city metro network expansion problem using RL. By formulating the metro line expansion as a process of sequential station selection, we train a seq2seq model with attention mechanism to generate a parameterized policy. Without expert knowledge, the parameterized policy generates the next metro line. Our method is general for different objectives, thus it is suitable for the expansion of metro network with multi-factor objectives.

REFERENCES

Elisabete Arsenio, Karel Martens, and Floridea Di Ciommo. Sustainable urban mobility plans: Bridging climate change and equity targets? *Research in Transportation Economics*, 55:30–39, 2016.

Hamid Behbahani, Sobhan Nazari, Masood Jafari Kang, and Todd Litman. A conceptual framework to formulate transportation network design problem considering social equity criteria. *Transportation Research Part A: Policy and Practice*, 125:171–183, 2019.

Irwan Bello, Hieu Pham, Quoc V Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Hélène Dufourd, Michel Gendreau, and Gilbert Laporte. Locating a transit line using tabu search. *Location Science*, 4(1-2):1–19, 1996.

Reza Zanjirani Farahani, Elnaz Miandoabchi, Wai Yuen Szeto, and Hannaneh Rashidi. A review of urban transportation network design problems. *European Journal of Operational Research*, 229 (2):281–302, 2013.

Gabriel Gutiérrez-Jarpa, Carlos Obreque, Gilbert Laporte, and Vladimir Marianov. Rapid transit network design for optimal cost and origin–destination demand capture. *Computers & Operations Research*, 40(12):3000–3009, 2013.

Gabriel Gutiérrez-Jarpa, Gilbert Laporte, and Vladimir Marianov. Corridor-based metro network design with travel flow capture. *Computers & Operations Research*, 89:58–67, 2018.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pp. 1008–1014, 2000.

WWM Kool and M Welling. Attention solves your tsp. *arXiv preprint arXiv:1803.08475*, 2018.

Michael Kuntz and Marco Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. *International Journal of Geographical Information Science*, 28 (9):1904–1921, 2014.

Gilbert Laporte and Juan A Mesa. The design of rapid transit networks. In *Location science*, pp. 581–594. Springer, 2015.

Gilbert Laporte and Marta MB Pascoal. Path based algorithms for metro network design. *Computers & Operations Research*, 62:78–94, 2015.

Yexin Li, Yu Zheng, and Qiang Yang. Dynamic bike reposition: A spatio-temporal reinforcement learning approach. 2018a.

Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Combinatorial optimization with graph convolutional networks and guided tree search. *CoRR*, abs/1810.10659, 2018b.

Kevin Manaugh, Madhav G Badami, and Ahmed M El-Geneidy. Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in north america. *Transport policy*, 37:167–176, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

MohammadReza Nazari, Afshin Oroojlooy, Lawrence Snyder, and Martin Takac. Reinforcement learning for solving the vehicle routing problem. In *Advances in Neural Information Processing Systems*, pp. 9861–9871, 2018.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Yanshuo Sun, Paul Schonfeld, and Qianwen Guo. Optimal extension of rail transit lines. *International Journal of Sustainable Transportation*, 12(10):753–769, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

Yi Wei, Jian Gang Jin, Jingfeng Yang, and Linjun Lu. Strategic network expansion of urban rapid transit systems: A bi-objective programming model. *Computer-Aided Civil and Infrastructure Engineering*, 34(5):431–443, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Zhaoyang Yang, Kathryn E Merrick, Hussein A Abbass, and Lianwen Jin. Multi-task deep reinforcement learning for continuous action control. In *IJCAI*, pp. 3301–3307, 2017.

Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. Depicting urban boundaries from a mobility network of spatial interactions: a case study of great britain with geo-located twitter data. *International Journal of Geographical Information Science*, 31(7):1293–1313, 2017.

A ADDED OD TRIPS AND SOCIAL EQUITY INDICATOR

A.1 ADDED OD TRIPS

In our study, the expanded line Z is allowed to connect with existing metro lines to form transfer stations. Thus, our $R_{od}(Z)$ contains not only the OD trips between the stations on Z, but also contains the OD trips between the stations on Z and the stations on existing lines through transfer stations.

A.2 SOCIAL EQUITY INDICATOR

In this paper, the social equity indicator is calculated as the same in Behbahani et al. (2019). The expansion of metro has an impact on the accessibility of each area. We calculate the accessibility Ac_i of area i as

$$Ac_i = \sum_j D_j F\left(c_{ij}\right) \tag{9}$$

where

c_{ij}: the travel cost between i and j

$F\left(c_{ij}\right)$ is defined as

$$F\left(c_{ij}\right) = F\left(t_{ij}\right) = e^{-\beta t_{ij}} \tag{10}$$

where t_{ij} is travel time and β is an adjustment parameter.

D_j is the compound index of development and defined as

$$D_j = \sum_k w_k d_{j,k} \tag{11}$$

where $d_{j,k}$ is an economic variable and w_k is the weight.

In our study, we condider that t_{ij} is proportional to distance between i and j. Kuntz & Helbich (2014) consider the house price as a socioeconomic variable, thus we take the house price of grid g_i as its index of development D_j.

Under a utilitarianism theory, the social equity indicator $R_{ac}(Z)$ is defined as the total added benefits $\sum_i Ac_i$. We prefer the metro line satisfying greater social equity indicator $R_{ac}(Z)$.

B ACTION DIRECTION RULES

In practice, the metro line should avoid subtour and squiggly line. Based on history action direction, we design agent action direction rules to ensure the rationality of expanded metro lines, shown in Figure 7.

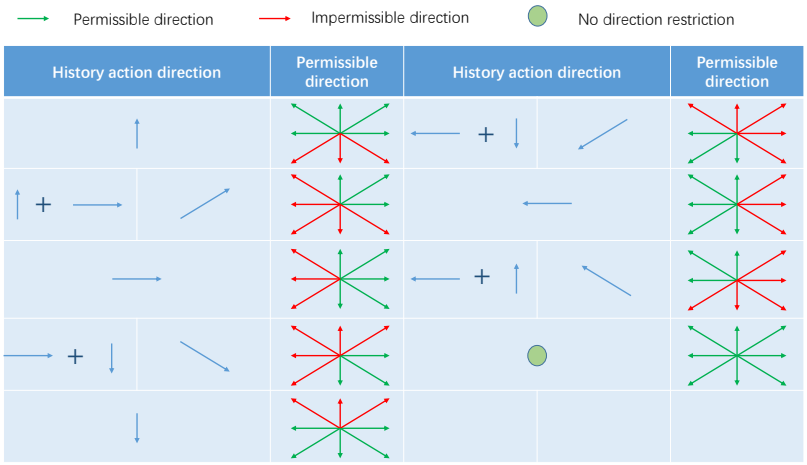


Figure 7: Action direction rules.