Network-aware Multi-agent Reinforcement Learning for Adaptive Navigation of Vehicles in a Dynamic Road Network

M.A.Sc. Thesis of Fazel Arasteh
York University, Toronto, Canada
Traffic Congestion

++ Trip Time
++ Air Pollution
Driver Frustration

Rush Hours
Traffic Incidents
Road Maintenance
Weather Condition
Expensive Solution: Construct More Roads
Economical Solution: Algorithmic Solution
Base Method: Shortest Path First Algorithm (SPF)
Static Network: Constant Travel Times

Optimal
Dynamic Network: Changing Travel Times

Traffic Jam

SPF

Travel Time Prediction
Limitation 1: Inaccurate Long-term Travel Time Predictions
Limitation 2: Greedy SPF! No Collaboration!
Vehicle Navigation Problem

Fleet of Vehicles

Traffic Jam

Collaboration

Improve Travel Time
Problem Definition
Vehicle Navigation Problem

Given:

Road network of the controlled area:

\[ W=\{R,I\} \]

State of road network \( W \) at time \( t \) which is the expected travel time in every road at time \( t \):

\[ S_w^t \]

\[ E(\text{Travel-Time}(r)) \mid r \in R \]

A set of origin-destination trips with time-label \( \tau \) indicating when the trip starts:

\[ \text{Trips}=\{(o, d, \tau) \mid o \in R, d \in I \} \]
Vehicle Navigation Problem

Task:
Generate a path for each trip:

\[ \text{Path(trip)} \mid \text{trip} \in \text{Trips} \]

Objective:
Minimizing the average travel time for all the trips:

\[ \tau' = \text{finish time of trip } (o, d, \tau) \]
\[ \text{Travel-Time(trip)} = \tau' - \tau \]
\[ \text{AVTT} = \frac{\sum \text{Travel-Time (trip)}/|\text{Trips}|}{\text{trip} \in \text{Trips}} \]

Minimize AVTT
Method: GNN + RL
Background
Reinforcement Learning

- State $s \in S$
- Take action $a \in A$
- Get reward $r$
- New state $s' \in S$
Q-learning, DQN
Graph Neural Networks
Graph Attention Networks
Methodology
Vehicles as Agents

Too Many Agents!

Complex Environment!
Packet Routing Problem

Constant # Agents

Simple Environment
Road Network - IP Network Analogy
MARL Formulation

0 1 0 0

Agents → Intersection
State → Agent ID+ $S_w + D$
Action → Next Road
Next State → Next Agent ID+ $S_w + D$
Reward → $-\Delta T$
Closed Controlled Area
Challenge 1: Huge Irrelevant Network State

Travel Time is a function of the Network State (Traffic State).

Most information is irrelevant:
- Global Information → Complex Model
- Local Information
- Neighborhood Info?

Neighborhood Info?
Solution: Graph Attention Networks

Neighborhood Captured
Locality of Access: Intuition for Routing
Exception: Disconnected Near Intersections!
Challenge 2: Intersection IDs

<table>
<thead>
<tr>
<th>Normalized Coordinates</th>
<th>One-hot Encoding</th>
</tr>
</thead>
<tbody>
<tr>
<td>x=-0.12</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

Preserves Locality of Access

- 2 dimensional data

Hard for NN to separate → No Convergence

Easy for NN to separate

How to have both?
Solution: Space Filtering (e.g. Z-Order)

<table>
<thead>
<tr>
<th>x:</th>
<th>000</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
</tr>
</thead>
<tbody>
<tr>
<td>y:</td>
<td>000</td>
<td>001</td>
<td>100</td>
<td>101</td>
<td>200</td>
<td>201</td>
<td>300</td>
<td>301</td>
</tr>
<tr>
<td>000</td>
<td>00000 00001</td>
<td>00100 00101</td>
<td>01000 01001</td>
<td>01010 01011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>001</td>
<td>00010 00011</td>
<td>00110 00111</td>
<td>01010 01011</td>
<td>01011 01011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>010</td>
<td>01000 01001</td>
<td>01100 01101</td>
<td>01110 01111</td>
<td>01111 01111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>011</td>
<td>01100 01101</td>
<td>01110 01111</td>
<td>01111 01111</td>
<td>01111 01111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10000 10001</td>
<td>10010 10011</td>
<td>10100 10101</td>
<td>10101 10101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>10010 10011</td>
<td>10110 10111</td>
<td>11010 11011</td>
<td>11011 11011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>10100 10101</td>
<td>10110 10111</td>
<td>11100 11101</td>
<td>11110 11110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>10110 10111</td>
<td>11100 11101</td>
<td>11110 11111</td>
<td>11111 11111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Solution: Space Filtering (e.g. Z-Order)

Preserved Locality

Separable by

Log (N) dimensions
Model Architecture

Network State

Routing Query (Destination IDs)

Routing Response

GAT Layer

Q-network

Router Agents

LOS (- \Delta T)
Reward Function Justification: End2End travel time prediction

\[ Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha (R_{t+1} + \gamma \max_{a \in A_{t+1}} Q(S_{t+1}, a) - Q(S_t, A_t)) \]

- \( \gamma = 1, \alpha = 1 \)

- \( S_{t+N} \) terminal

- \( r = -\Delta T \)

\[ Q(S_t, A_t) \leftarrow R_{t+1} + \max_{a \in A_{t+1}} Q(S_{t+1}, a) \]

\[ Q(S_t, A_t) \leftarrow R_{t+1} + R_{t+2} + \ldots + R_{t+N} \]

\[ Q(S_t, A_t) \leftarrow -\Delta T_1 - \Delta T_2 - \ldots - \Delta T_N = -\text{Travel Time} \]
Experimental Evaluation
Datasets: Grid Network
Datasets: Downtown Toronto

- Bloor st
- Bathurst st
- Don Valley H.W.
- Gardiner Exp H.W.
Datasets: Traffic Demand

- Uniform Demand
- Biased Demand
Dynamics Simulation

Traffic Jam

traffic-state-change-period
congestion-epsilon
congestion-speed-factor
traffic-state-change-period
<table>
<thead>
<tr>
<th>Base Lines, Algorithm Versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel Time Shortest Path First (SPF)</td>
</tr>
<tr>
<td>Travel Time Shortest Path First with Rerouting (SPFWR)</td>
</tr>
<tr>
<td>Q-routing</td>
</tr>
<tr>
<td>AN(0hop), AN(1hop), AN(2hop)</td>
</tr>
</tbody>
</table>
Evaluation Metric

- Average Travel Time (AVTT)
- Routing Success (RS)
Performance Evaluation (Online Training)

Toronto Average Travel Time

- AN(0hop)
- AN(1hop)
- AN(2hop)
- Q-routing

Toronto Average Travel Time (Last 100 Episodes)

Toronto Routing Success
Performance Evaluation (Online Training)

5x6 Average Travel Time

5x6 Average Travel Time (Last 100 Episodes)

5x6 Routing Success
Performance Evaluation (Offline Testing Settings)

- 2000 Uniform Trips
- 200 Biased Trips

Routing Success = 2200
## Performance Evaluation (Offline Testing Results)

<table>
<thead>
<tr>
<th></th>
<th>5x6 Network</th>
<th>D.T Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPF</td>
<td>173.4</td>
<td>551.7</td>
</tr>
<tr>
<td>SPFWR</td>
<td>205.1</td>
<td>475.6</td>
</tr>
<tr>
<td>QR</td>
<td>159.6</td>
<td>∞</td>
</tr>
<tr>
<td>AN(0hop)</td>
<td>143.7</td>
<td>477.6</td>
</tr>
<tr>
<td>AN(1hop)</td>
<td>138.4</td>
<td>476.4</td>
</tr>
<tr>
<td>AN(2hop)</td>
<td>145.4</td>
<td>479.3</td>
</tr>
</tbody>
</table>
Locality of Access Evaluation
Attention Evaluation
Conclusion, Limitations & Future Work
Conclusion
Limitations & Future Work

Limitations:
1. Reliability
2. Scalability
3. Network State Capturing

Future Work:
1. Shared Policies
2. Hierarchical Routing
3. Traffic Signal Control
Thank You
Collaborative Policies
Reward Function Justification: End2End travel time prediction

\[ Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha (R_{t+1} + \gamma \max_{a \in A_{t+1}} Q(S_{t+1}, a) - Q(S_t, A_t)) \]

\[ \gamma = 1, \alpha = 1 \]

- \( Q(S_t, A_t) \leftarrow R_{t+1} + \max_{a \in A_{t+1}} Q(S_{t+1}, a) \)

- \( Q(S_t, A_t) \leftarrow R_{t+1} + R_{t+2} + \ldots + R_{t+N} \)

- \( r = -\Delta T \)

- \( Q(S_t, A_t) \leftarrow -\Delta T_1 - \Delta T_2 - \ldots - \Delta T_N = \text{Travel Time} \)