A Deep Value-network Based Approach for Multi-Driver Order Dispatching

Xiaocheng Tang
DiDi Mountain View | AI Labs
Joint work with
Zhiwei (Tony) Qin, Fan Zhang, Zhaodong Wang, Zhe Xu, Yintai Ma, Hongtu Zhu & Jieping Ye

KDD Oral 2019
Anchorage, AK, USA

DiDi

DiDi is the world's leading mobile transportation platform

550+M riders
30+M work opportunities
10B rides per year

DiDi

Outline

Motivation

A Semi-MDP Formulation

Learning and Multi-Driver Planning

- State Representation
- Lipschitz Regularization
- Context Randomization
- Multi-City Transfer

Experiment Results

- Simulations using real-world data
- Online A/B tests

DiDi

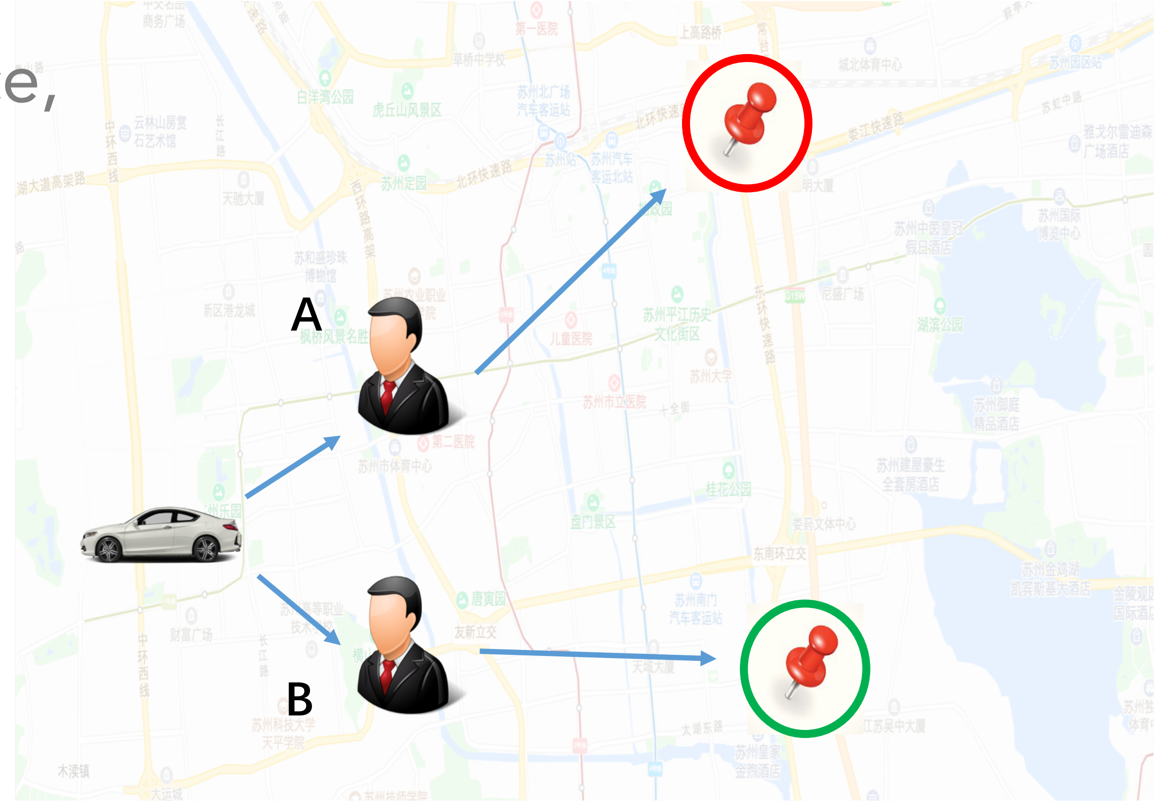
MOTIVATION

Same trip fee, pickup distance, passenger features, etc.

- Person A (-> hot)
- Person B (-> cold)

Which one to fulfill?

Reduce total idling time of the drivers!



Cold area

Hot area

DiDi

MOTIVATION

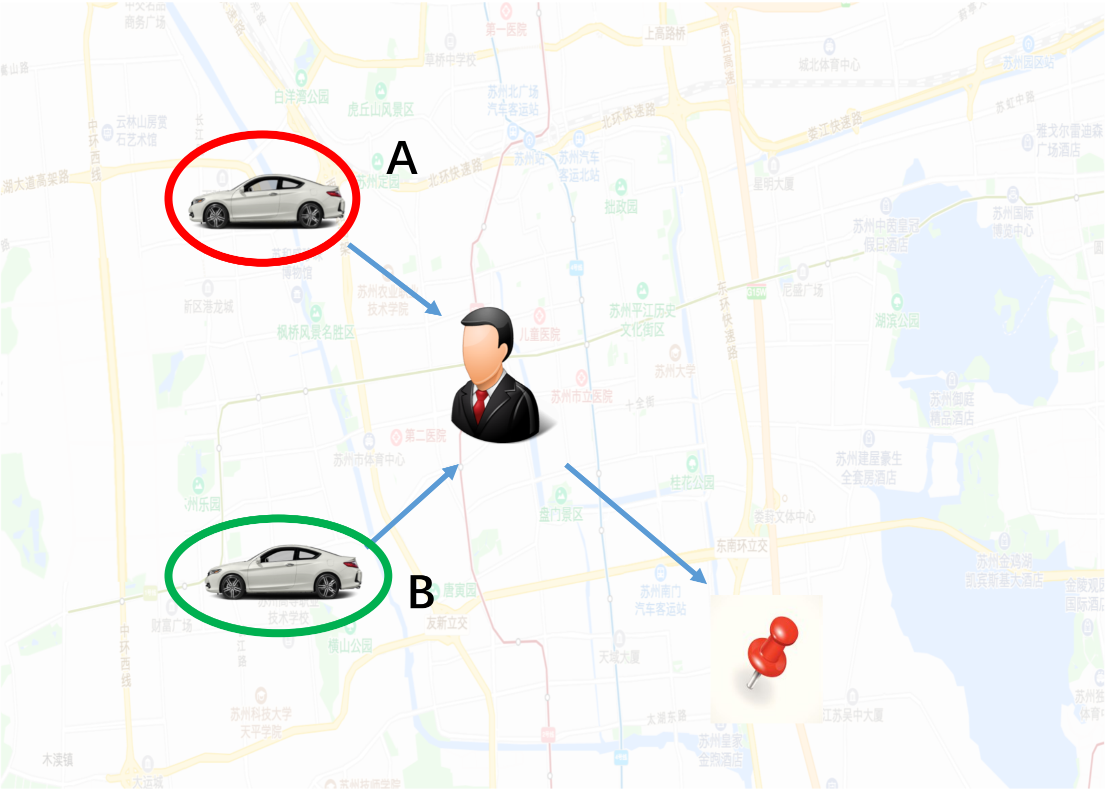
Same pickup distance,

driver features, etc.

- Driver A (hot)
- Driver B (cold)

Which one to dispatch?

Reduce total idling time of the drivers!

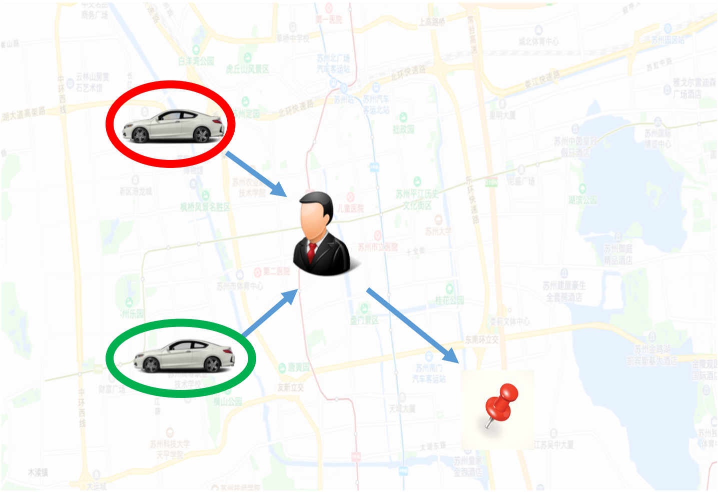


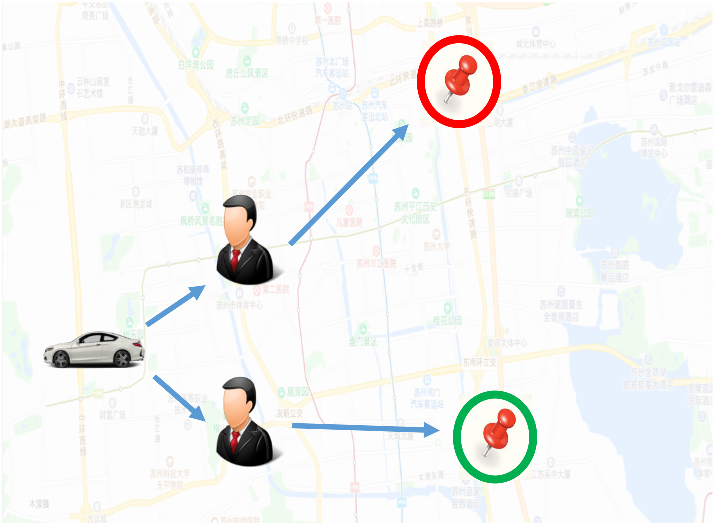
Cold area

Hot area

MOTIVATION

Cold area   Hot area





Reduce idling time → **Increase fulfillment** → **Increase driver income**

QUESTIONS

What defines a **hot**/**cold** area?

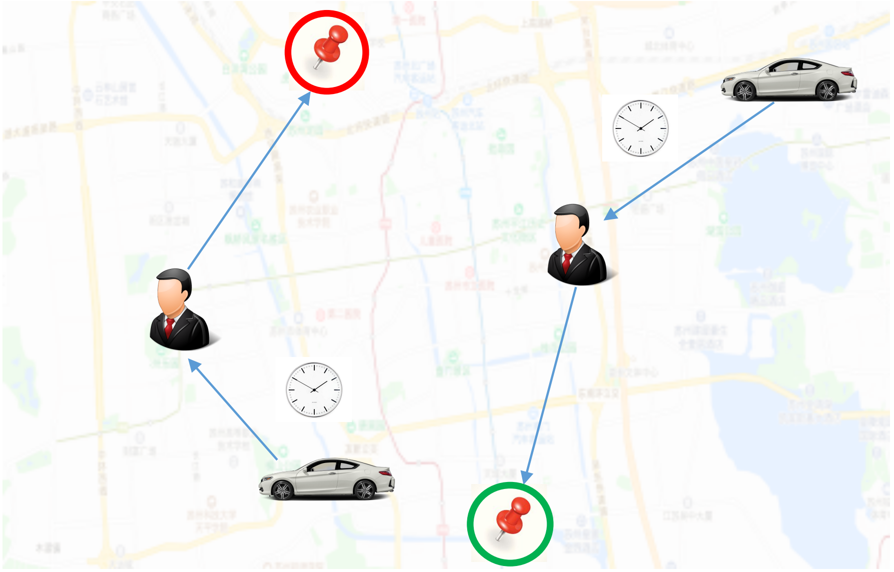
Why **reinforcement learning** (why not **supervised learning**)?

DiDi

A SEMI-MDP FORMULATION

State, s:=(l,μ,**u**) is the

- geo-coordinates (l) of the driver
- the raw time stamp (μ)
- the contextual feature vector (**u**), e.g. the supply-demand conditions, driver service statics, etc.



Option, o the k-step transition of the driver

Reward, R is the total fee collected for the trip

- a function of s and o

Policy, π(o|s) is a function that

- maps a state s to a distribution over the action space (stochastic policy) or a particular action (deterministic policy)

DiDi

A SEMI-MDP FORMULATION

State value function, V(s): expected cumulative reward that.

- the driver will gain till the end of an episode if he/she starts at state s and follows a policy π

$$V^{\pi}(s) := E\{\sum_{i=t+1}^{T} \gamma^{i-t-1} r_i | s_t = s\}$$

- Similar to standard MDPs, we can write **Bellman equations** for general policies and options given one-step transition (s_t, R_t, s_{t+k})

$$V^{\kappa+1}(s_t) \leftarrow \frac{R_t(\gamma^{k_t} - 1)}{k_t(\gamma - 1)} + \gamma^{k_t} V^{\kappa}(s_{t+k_t}).$$

A SEMI-MDP FORMULATION

State value function, V(s): expected cumulative reward that.

- the driver will gain till the end of an episode if he/she starts at state s and follows a policy π

$$V^{\pi}(s) := E\{\sum_{i=t+1}^{T} \gamma^{i-t-1} r_i | s_t = s\}$$

- Similar to standard MDPs, we can write **Bellman equations** for general policies and options given one-step transition (s_t, R_t, s_{t+k})

Smooth version of reward clipping

$$V^{\kappa+1}(s_t) \leftarrow \frac{R_t(\gamma^{k_t} - 1)}{k_t(\gamma - 1)} + \gamma^{k_t} V^{\kappa}(s_{t+k_t}).$$

Use of a secondary neural network to ensure training stability

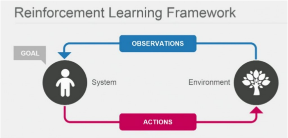
Parameterized by a neural network

Training target

DiDi

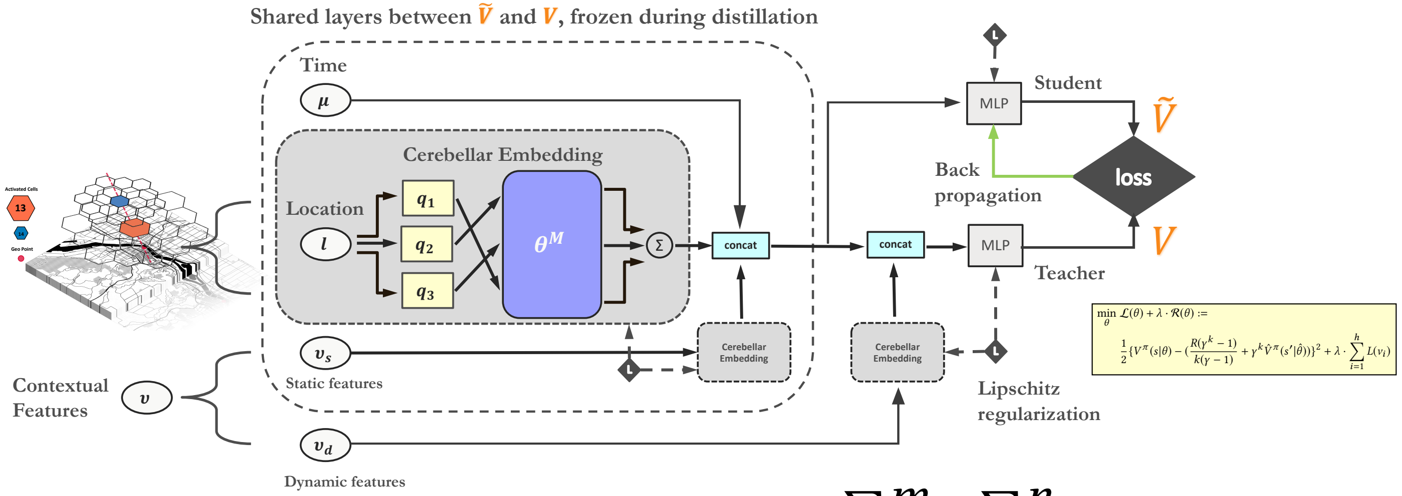
LEARNING AND PLANNING

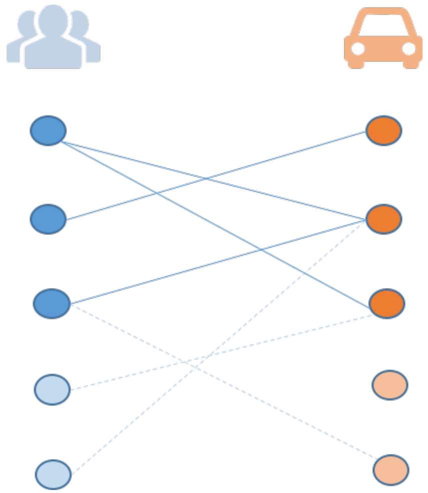
Deep Reinforcement Learning



+

Bipartite Graph Matching





$$\max_{x \in C} \sum_{i=1}^{m} \sum_{j=1}^{n} \rho_{ij} x_{ij}$$

Objective:

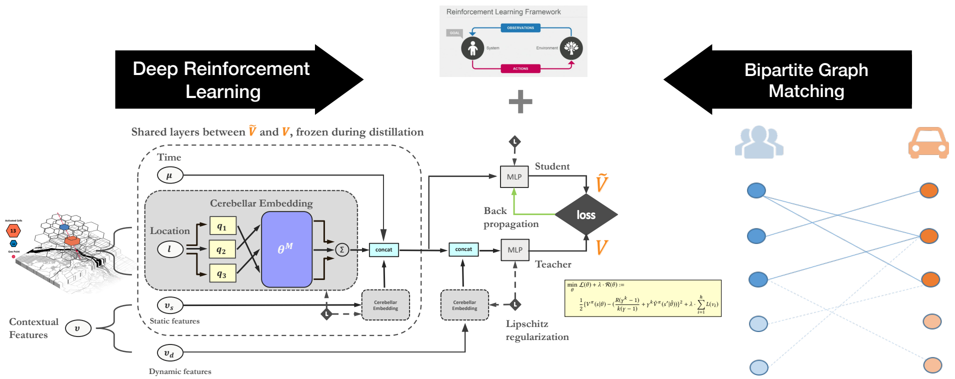
maximize the total utilities of the assignments where the utility scores are computed as the Temporal Difference error between order's destination state and driver 's current state, e.g.,

Spatiotemporal optimality!

$$\rho_{ij} = R_{ij} \frac{(\gamma^{k_{ij}} - 1)}{k_{ij}(\gamma - 1)} + \underline{\gamma^{k_{ij}} V(s_j)} - \underline{V(s_i)} + \Omega \cdot U_{ij}$$

DiDi

LEARNING AND PLANNING



Policy Evaluation

Learning ⇄ **Planning**

Policy Improvement

- Planning using the new value network, which is fitted against data generated by the old value network
- Learning needs to strike a balance between fitting the target while avoiding divergence from the previous value network, e.g., on-policy methods like PPO, TRPO, etc.
- Significant improvement is obtained by iterating between online planning and offline learning

DiDi

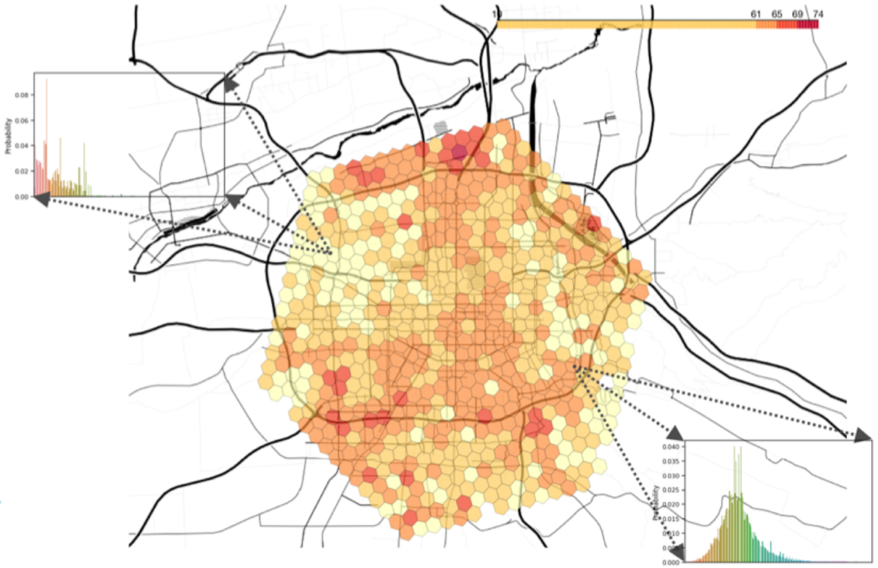
ANSWERS

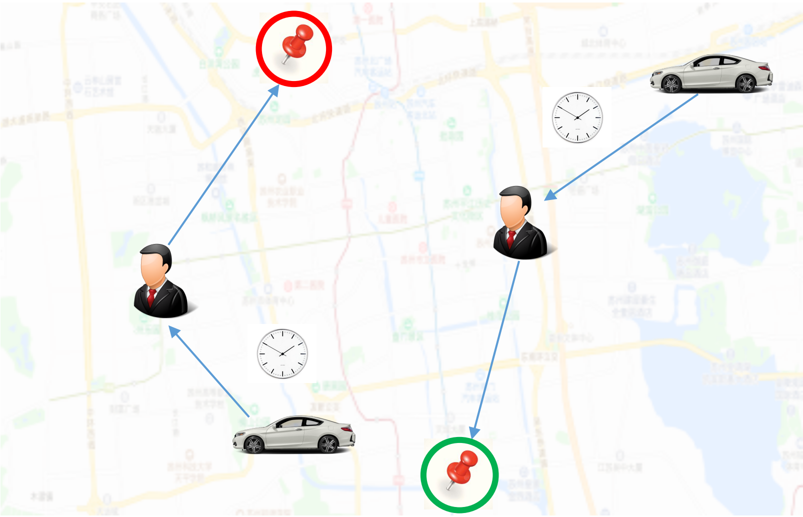
What defines a hot/cold area?

- The expectation of a driver's earning potential till the end of a day, e.g., long-term value

Why we care about long term value?

- This is a sequence decision problem
- The dispatching action is temporally extended



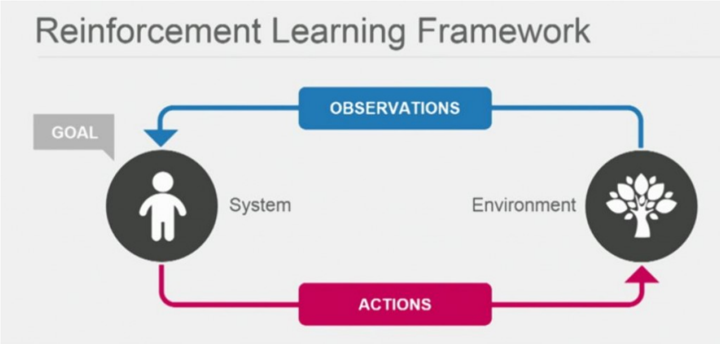


DiDi

ANSWERS

Why reinforcement learning (why not supervised learning)?

- The value network is obtained from fitting the driver's historical income (target)
- The "target" changes as soon as a new value network is deployed in the environment
- Learning involves the balance between fitting the target while avoiding divergence from the previous value network, e.g., on-policy methods
- Hard to do off-policy + importance sampling since we act by solving a combinatorial problem instead of according to a probability distribution



Why is this important?

- Significant improvement by online + offline iterations
- No "labeling" cost
- No "investment budget" or "subsidizing" cost
- The system automatically improves itself (reinforcement)

DiDi

QUESTIONS

How to learn a **good** **value network** for dispatching?

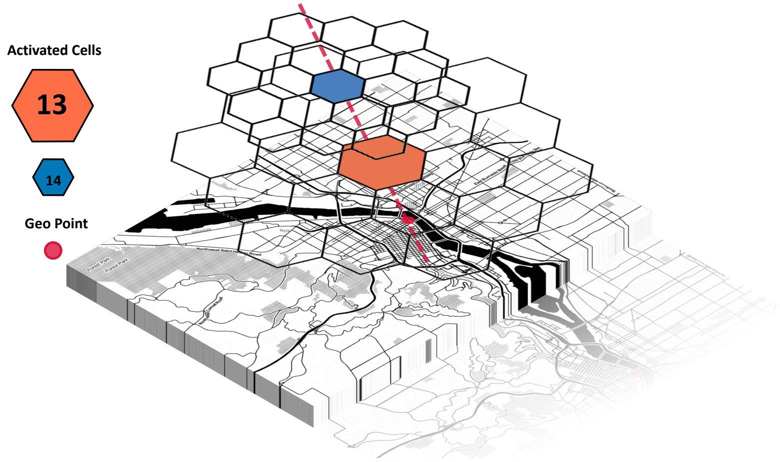
LEARNING AND PLANNING

State representation

Lipschitz regularization

Context randomization

Multi-city transfer

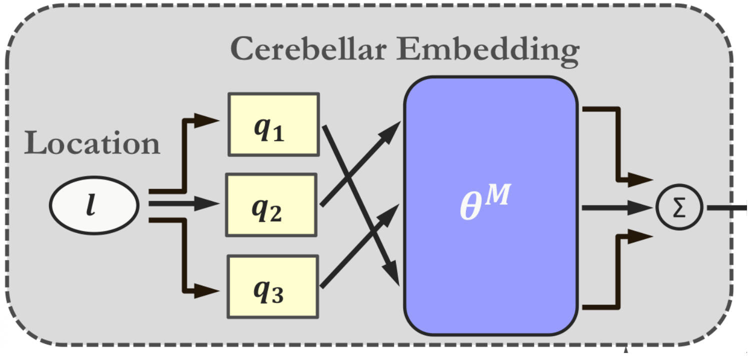


- **Memory-based Neural Network**

  Distributed representation

- **Hierarchical Hexagon Tiling System**

  To capture unique properties of specific streets, neighborhoods, and cities, we let the model learn a hierarchy of representations for areas of different size, with the precise location represented in the model by the sum of the embeddings of its location at various scales.



LEARNING AND PLANNING

State representation

Lipschitz regularization

Context randomization

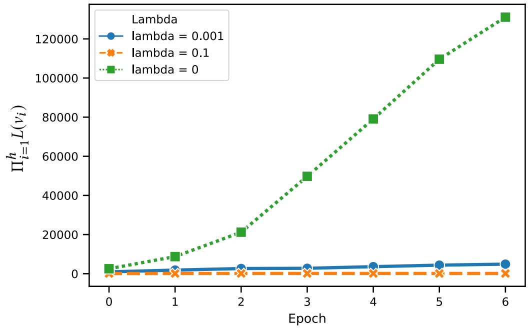
Multi-city transfer

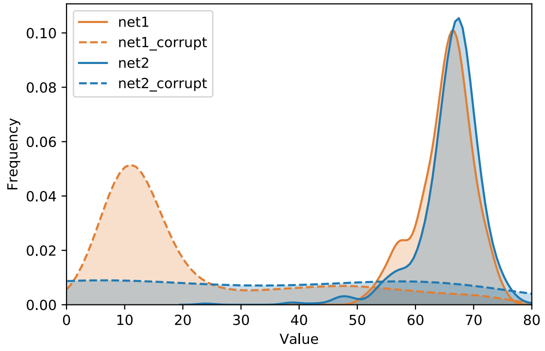
- **Lipschitz value function**

  The variation of the function w.r.t. a change in its input is bounded by the **Lipschitz constant**

- **Regularize this constant during training**

  To induce a smoother value estimations and to stabilize the **nonlinear Bellman update** (replacing the target network introduced by the original DQN paper [Mnih et al., 2015]). We find that this improves learning dynamics and policy convergence.





$$V^{\kappa+1}(s_t) \leftarrow \frac{R_t(\gamma^{k_t}-1)}{k_t(\gamma-1)} + \gamma^{k_t} V^{\kappa}(s_{t+k_t}).$$

$$V^{\kappa+2}(s_t) \leftarrow \frac{R_t(\gamma^{k_t}-1)}{k_t(\gamma-1)} + \gamma^{k_t} V^{\kappa+1}(s_{t+k_t}).$$

Bad value becomes target resulting in more bad values!

LEARNING AND PLANNING

State representation

Lipschitz regularization

Context randomization

Multi-city transfer

- **Historical trajectory augmentation**

  During training we augment each historical driver trajectory with contextual features extracted from the production logging system

- **Build noise and variance into training**

  It is common to notice a ±30 minutes shift of the rush hour peak and the real-time statistics. Also the logging system often comes with scheduling bias.

- **Hierarchical range query**

  Instead of matching with the exact spatiotemporal status, we implement the procedure such that it allows the specification of a range for a given query and returns all features within that range throughout the history.

DiDi

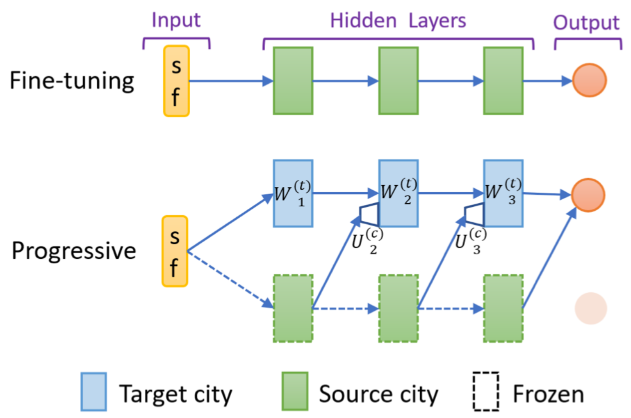
LEARNING AND PLANNING

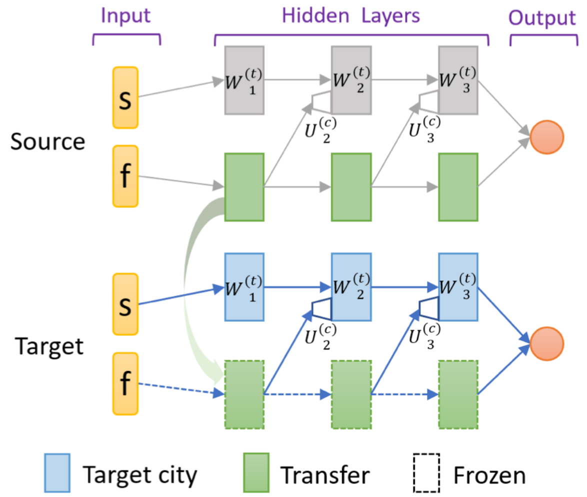
State representation

Lipschitz regularization

Context randomization

Multi-city transfer





▶ **Correlated-feature progressive transfer**

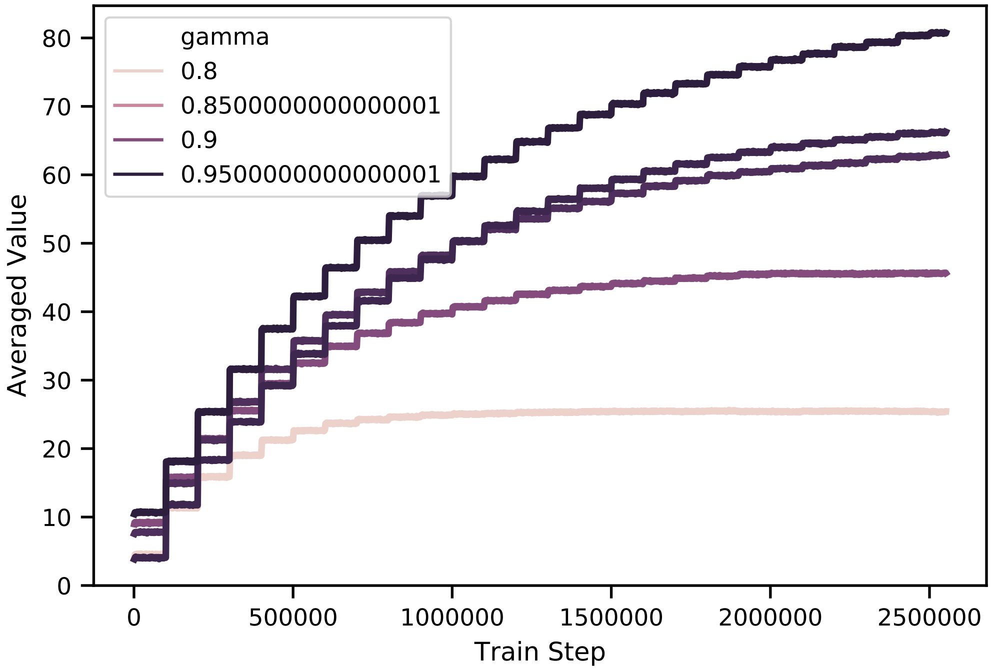
Instead of using a fully-connected network which takes all state elements as an entirety during training, we build and train a parallel progressive structure with two separate input groups.

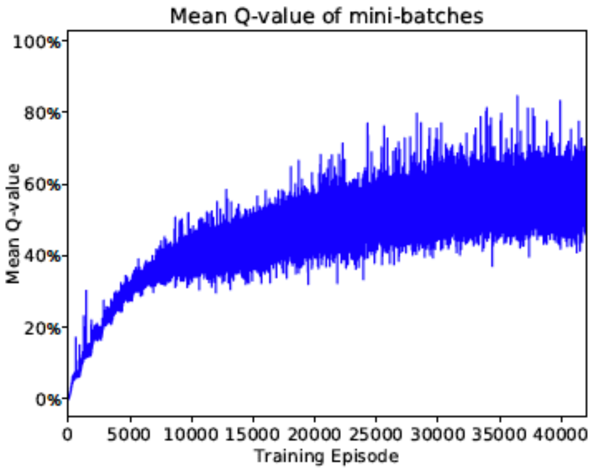
DiDi

EXPERIMENT RESULTS

Training curve

- Better dynamics and convergence compared to DQN





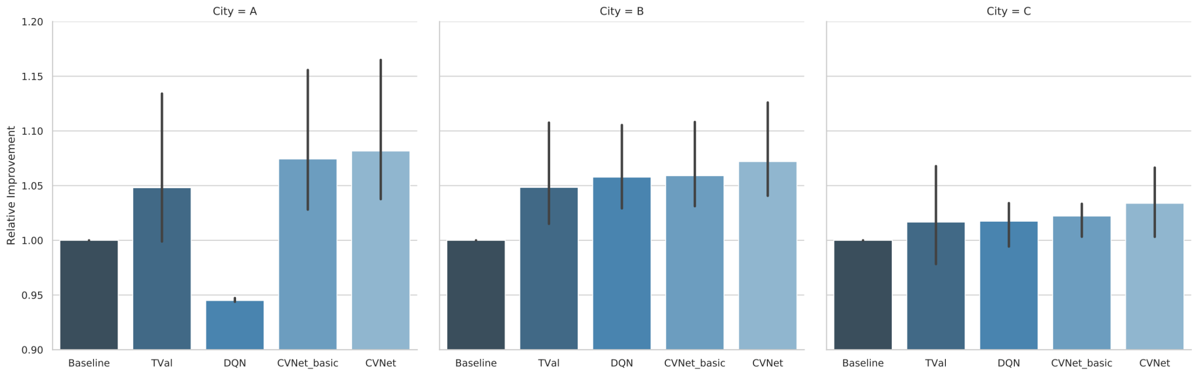
(a) DQN training

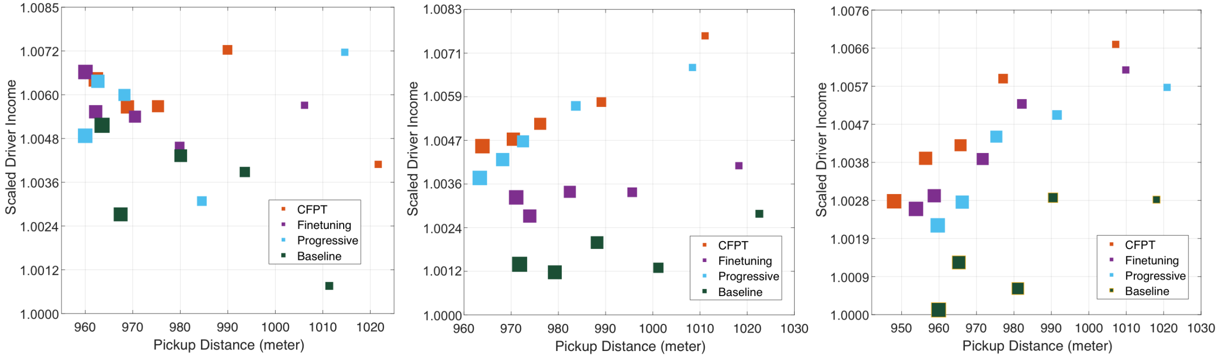
DiDi

EXPERIMENT RESULTS

Simulations with real-world data

- (Top) CVNet achieves an average improvements (across days) from 3% to 8%.
- (Bottom) Compare transfer methods (from city A to B, C and D) with baselines.



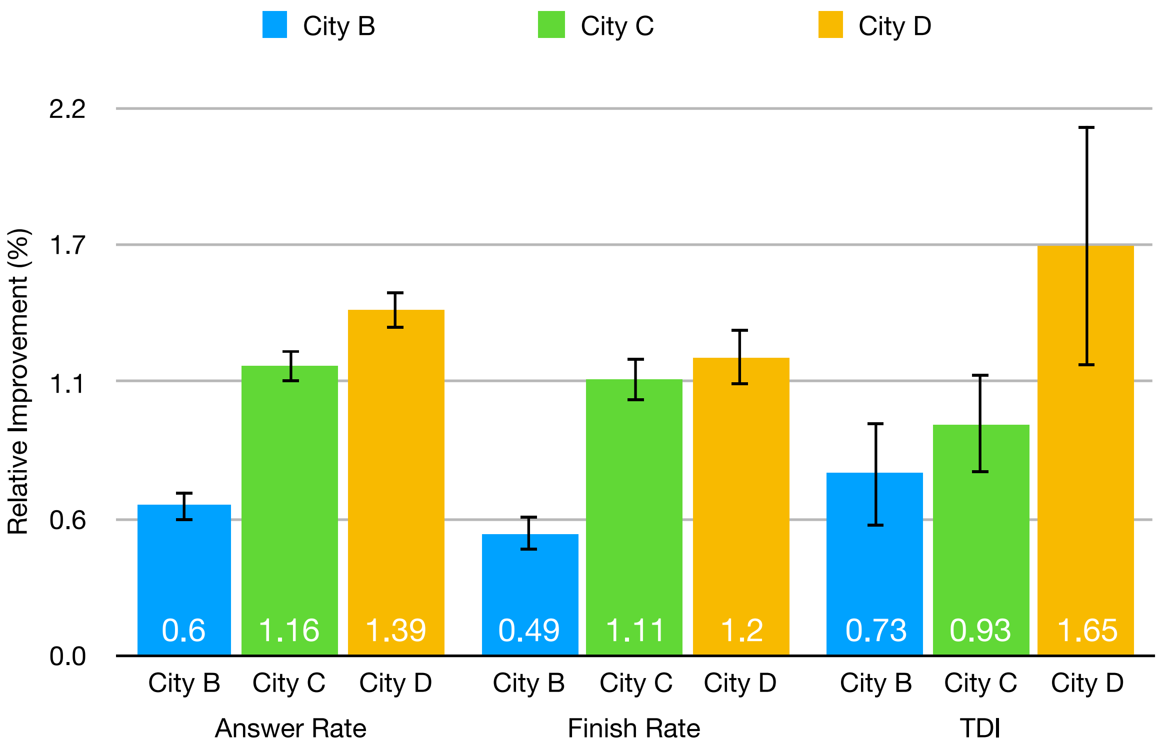


(a) City B (b) City C (c) City D

EXPERIMENT RESULTS

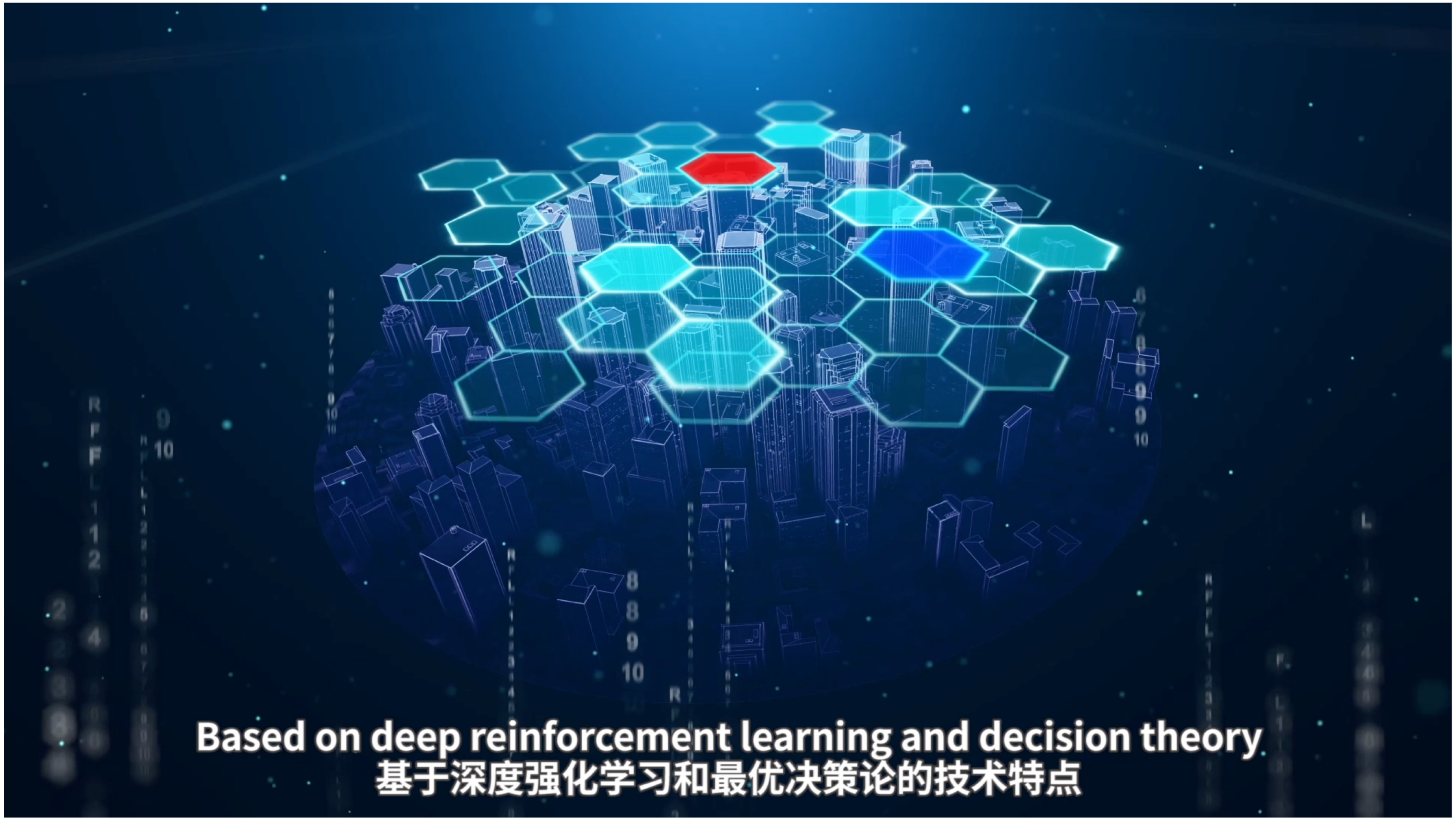
Online A/B tests

- We conduct large scale online A/B tests, which demonstrate that the proposed method achieves significant improvement on both total driver income and user experience related metrics



DiDi

SUMMARY



DiDi

Thank You.

Xiaocheng Tang, Ph.D. 唐小程
Staff Research Scientist

AI Labs @ DiDi Chuxing
Mountain View, CA, USA
Email: xiaochengtang@didiglobal.com
Linkedin: https://www.linkedin.com/in/xiaochengt/

DiDi