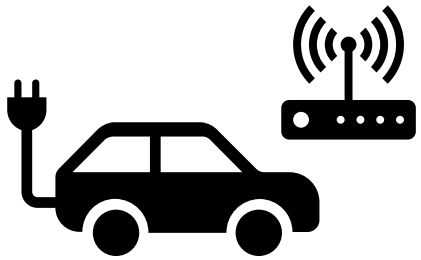
PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Dissipating Stop-And-Go Waves in Closed and Open Networks Via Deep Reinforcement Learning

Kreidieh, A. R., Wu, C., & Bayen, A. M. (2018)

Objetivo del trabajo



https://www.purosautos.com/noticias/en-que-ciudades-se-producen-los-peores-embotellamientos-del-mundo/

Generación de estrategias de control de tráfico mediante ***connected and automated vehicles*** (CAV's) en base a aprendizaje reforzado (RL), en redes abiertas (usando *Transfer Learning* a partir de redes cerradas).

Se hace énfasis en **la estabilización del tráfico (disminuir ondas *Stop-and-Go*)** considerando **una mezcla compuesta por CAV's y vehículos conducidos por humanos**.

Metodología

En términos generales, se aplica RL en conjunto con microsimuladores de tráfico. Los problemas de RL se estudian principalmente como un **Problema de Decisión de Markov (MDP)** con tiempo discreto, definido por:

$$(S, A, P, r, \rho_0, \gamma, T)$$

donde

S: Espacio de estados
A: Espacio de acciones
P: Probabilidad de transición
r: Función de recompensa
ρ0: Un distribución de estados inicial
γ: Factor de descuento (0,1]
T: Horizonte temporal

Para un **POMDP** (MDP parcialmente observable) se requiere adicionalmente:
Ω: Conjunto de observaciones
O: Distribución de probabilidad de observaciones.

Metodología

En el MDP, el **agente recibe inputs** desde el **entorno** e interactúa con éste llevando a cabo **acciones**, las que a su vez son definidas por una **política estocástica** parametrizada por **θ.**

El objetivo del agente es aprender un política óptima:

$$\theta^* := argmax_\theta \eta(\pi_\theta) \quad \text{donde} \quad \eta(\pi_\theta) = \sum_{i=0}^{T} \gamma^i r_i$$

Es la **recompensa esperada (descontada)** sobre una trayectoria **τ de pares de acciones y estados**, para todo el horizonte temporal.

En este trabajo, los parámetros de las políticas se definen iterativamente, mediante métodos de gradiente para políticas.

Metodología – Ondas Stop-and-Go

En el contexto microscópico de este trabajo (usando modelos de seguimiento vehicular), las dinámicas de cada vehículo α se describen mediante las ecuaciones diferenciales:

$$\frac{dh_\alpha}{dt} = v_l(t) - v_\alpha(t) \qquad \frac{dv_\alpha}{dt} = f(h_\alpha(t-\tau), v_\alpha(t-\sigma), v_l(t-\kappa))$$

Con:

h_α, v_α, v_l : Headway del vehículo, velocidad del vehículo y velocidad del líder, respectivamente

f : Ecuación de aceleración

τ, σ, κ : Desplazamientos temporales

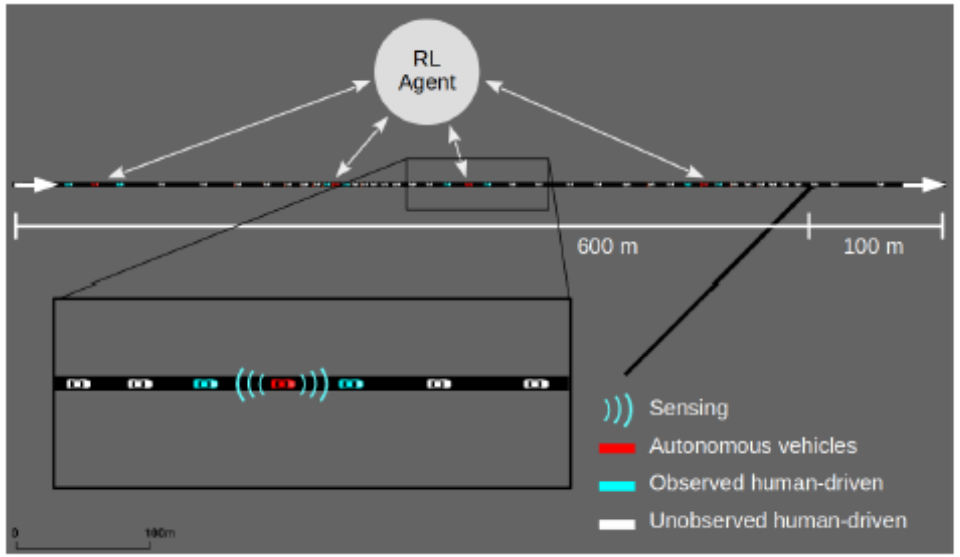
Estos **modelos forman la base para las transiciones** del MDP estudiado.

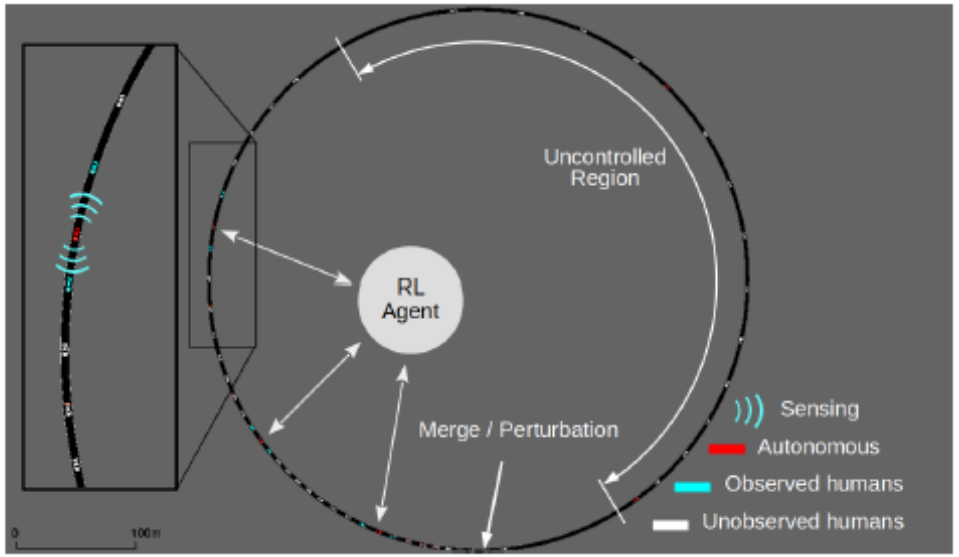
Metodología – Ondas Stop-and-Go

En **equilibrio**, todos los vehículos se mueven a **velocidades constantes y mantienen headways constantes**. En la práctica, esto no se da, originándose perturbaciones (olas) que se **propagan aguas arriba**, que pueden ocasionar que el tráfico se detenga por completo.

Metodología – Escenarios

Se considera una red de autopista de una pista (700 metros, 2000 veh/hr) con una rampa de acceso (100 veh/h), que crea perturbaciones. Una cantidad variable de CAV's están presentes en el flujo (de 0% a 10%).





Para el *Transfer – Learning*, se comienza por definir un anillo cerrado de 1400 metros, con 50 vehículos en total. Para simular la rampa, se introducen perturbaciones en un punto. Los CAV's se entrenan inicialmente en el anillo.

Metodología – Modelo de seguimiento vehicular y CAV's

Para modelar a los **conductores humanos** en el simulador, se utiliza el **Intelligent Driver Model (IDM)**, cuya función de aceleración es la siguiente:

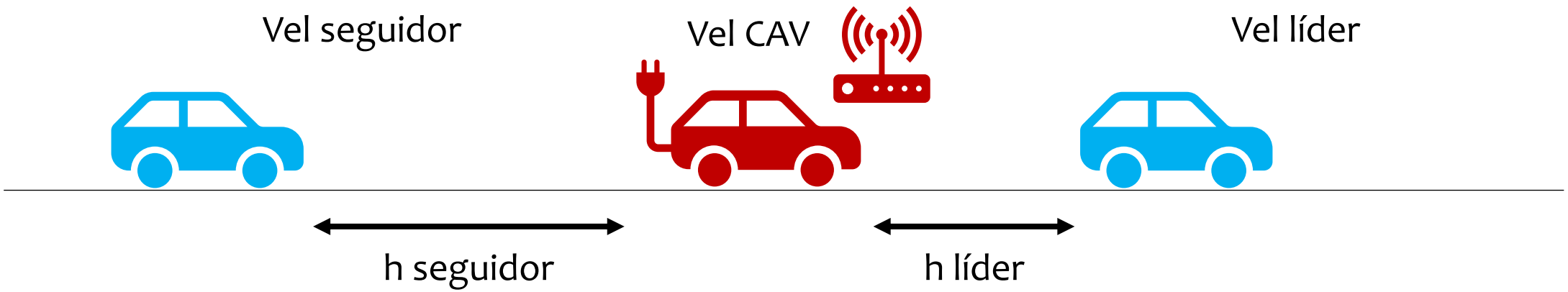
$$\dot{v} = a\left[1 - \left(\frac{v}{v_0}\right)^{\delta} - \left(\frac{s^*(v,\Delta v)}{s}\right)^2\right]$$

$$s^*(v,\Delta v) = s_0 + max(0, vT + \frac{v\Delta v}{2\sqrt{ab}})$$

Por otro lado, se reemplazan conductores humanos por CAV's de acuerdo a la proporción deseada de estos últimos.

Metodología – Observaciones y acciones

El **espacio de observaciones** del agente consiste en características observables locales, tales como la velocidad y headway del vehículo siguiente y anterior al CAV, además de la velocidad de este último.



El **espacio de acciones** consiste en aceleraciones (dentro de un margen) para cada CAV.

Metodología – Función de Recompensa

La función de **recompensa** elegida promueve **altas velocidades para el sistema**. Siendo $v_i(t)$ y $h_i(t)$ la velocidad y el headway (temporal) del vehículo i en el paso t, respectivamente, la función de recompensa se define como:

$$r = \underbrace{\|v_{des}\| - \|v_{des} - v(t))\|}_{\text{Busca la proximidad del sistema a una velocidad deseada } v_{des}\text{, manteniendo penalización por término de simulación anticipada}} - \underbrace{\alpha \sum_{i \epsilon (CAV)} max\left[h_{max} - h_i(t), 0\right)]}_{\text{Penalidad usada para identificar características locales de congestión (bajos headways)}}$$

Metodología – Simulaciones

Los experimentos se implementan en Flow, un entorno computacional para ejecutar RL profundo en microsimuladores de tráfico.

Las simulaciones se ejecutan en SUMO. Se utilizan pasos de **0,2 segundos**, y una **duración de 3600 segundos**. Al agente se le entregan actualizaciones del estado y genera **nuevas acciones en incrementos de 1 segundo**, repitiendo estas acciones durante los próximos **5 pasos consecutivos**.

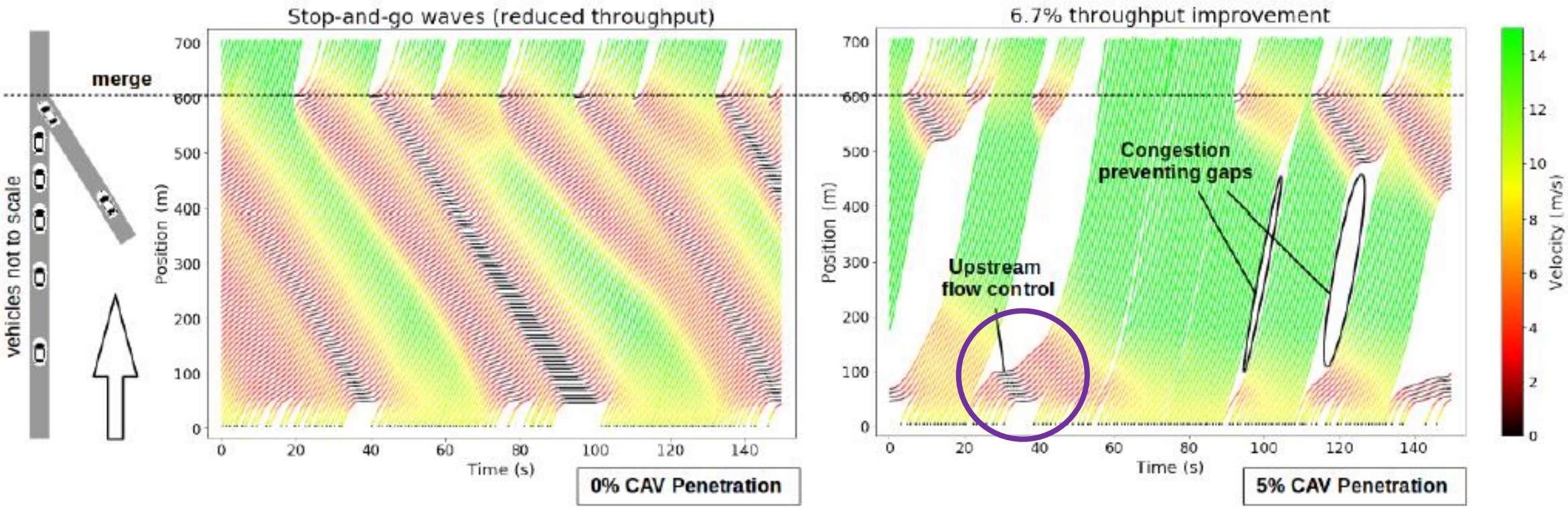
El método de gradiente para política utilizado para aprender la política de control es el Trust Region Policy Optimization (TRPO). Para la mayoría de los experimentos, se utiliza una política MLP Gaussiana Diagonal y una tanh como no linealidad.

Resultados

En base a pruebas considerando una presencia de **hasta un 10% de CAV's**, se obtienen los siguientes resultados:

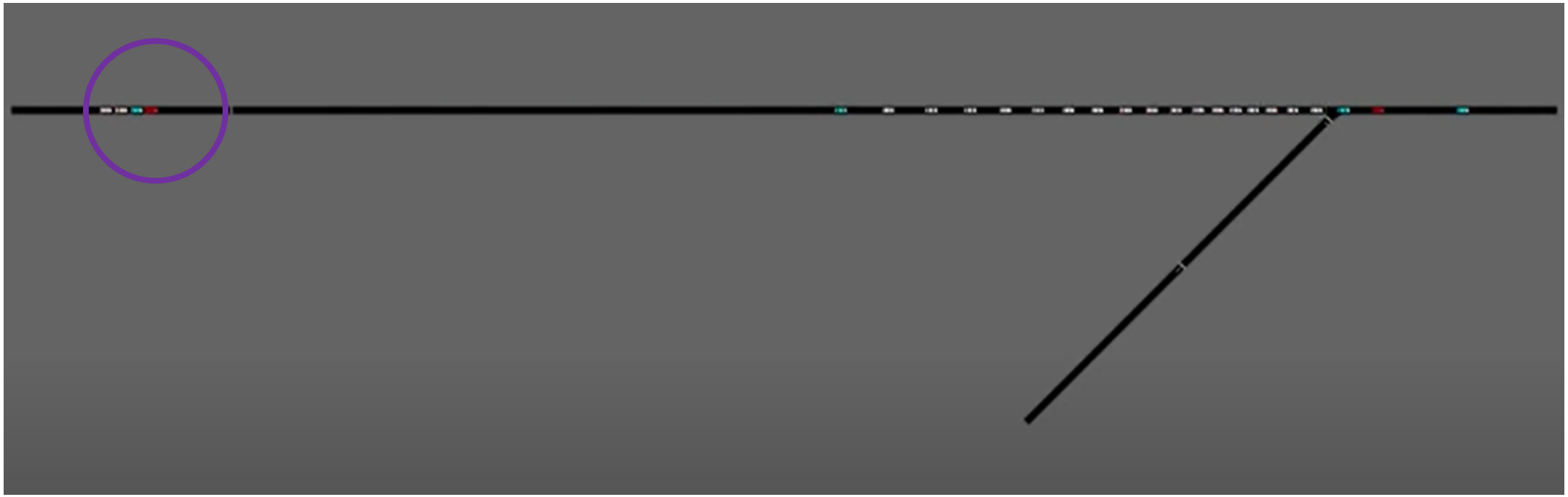
- Una presencia de 2,5% de CAV's ya reduce la frecuencia y magnitud de olas Stop-and-Go.
- Disipación casi total de las olas de Stop-and-Go, en una vía recta y red abierta.
- Velocidades se duplican.
- Rendimiento de la red mejora un 13%.
- Es posible transferir los resultados de redes cerradas (anillo) a la red abierta.

Resultados



Comparación de trayectorias con una presencia de 0% y 5% de CAV's.

Resultados

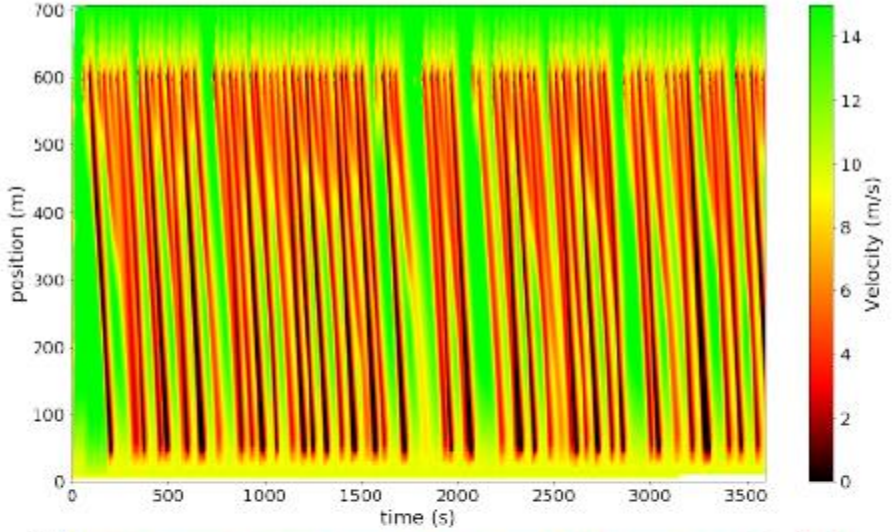


“Gating” del CAV al resto del flujo.

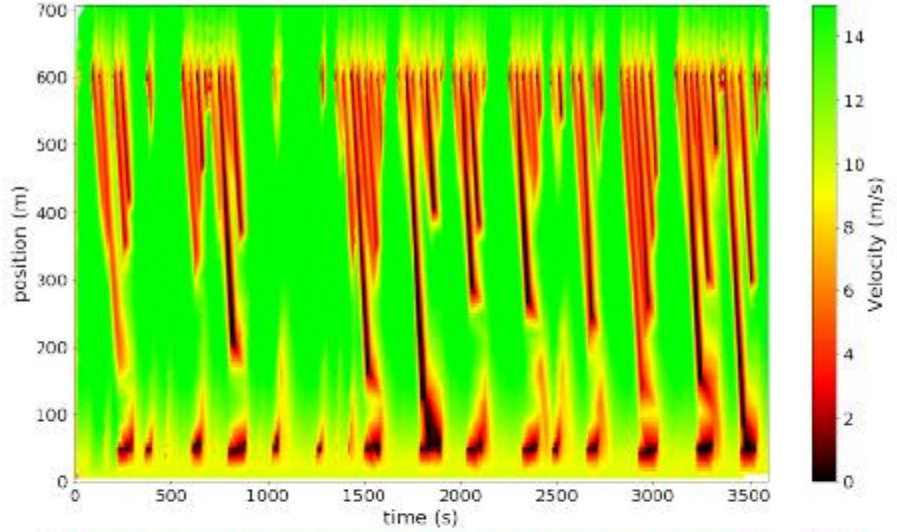
Resultados

Velocidades espacio-temporales para distintos % de CAV's.

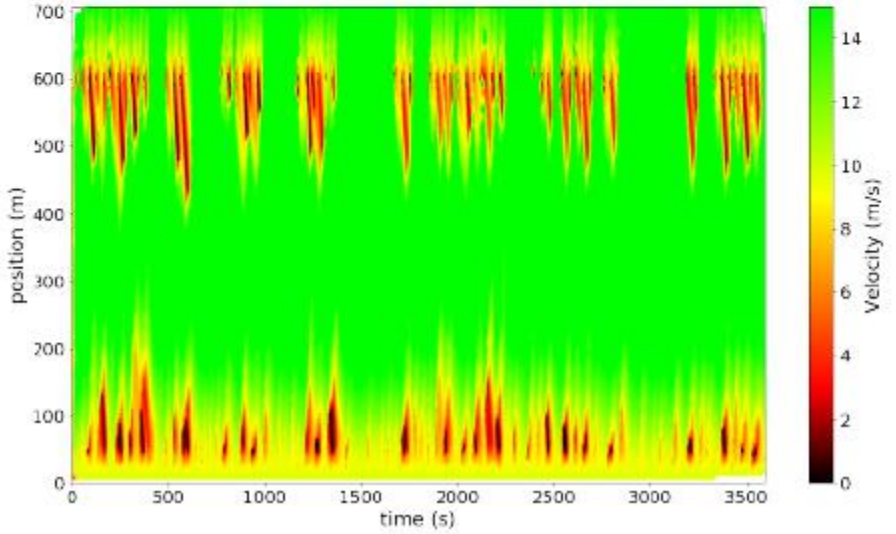
0%



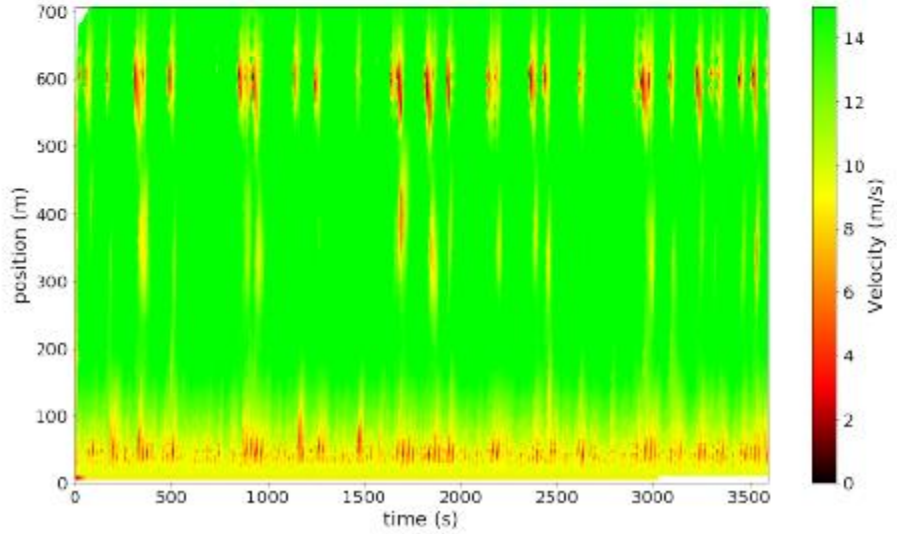
2,5%



5%

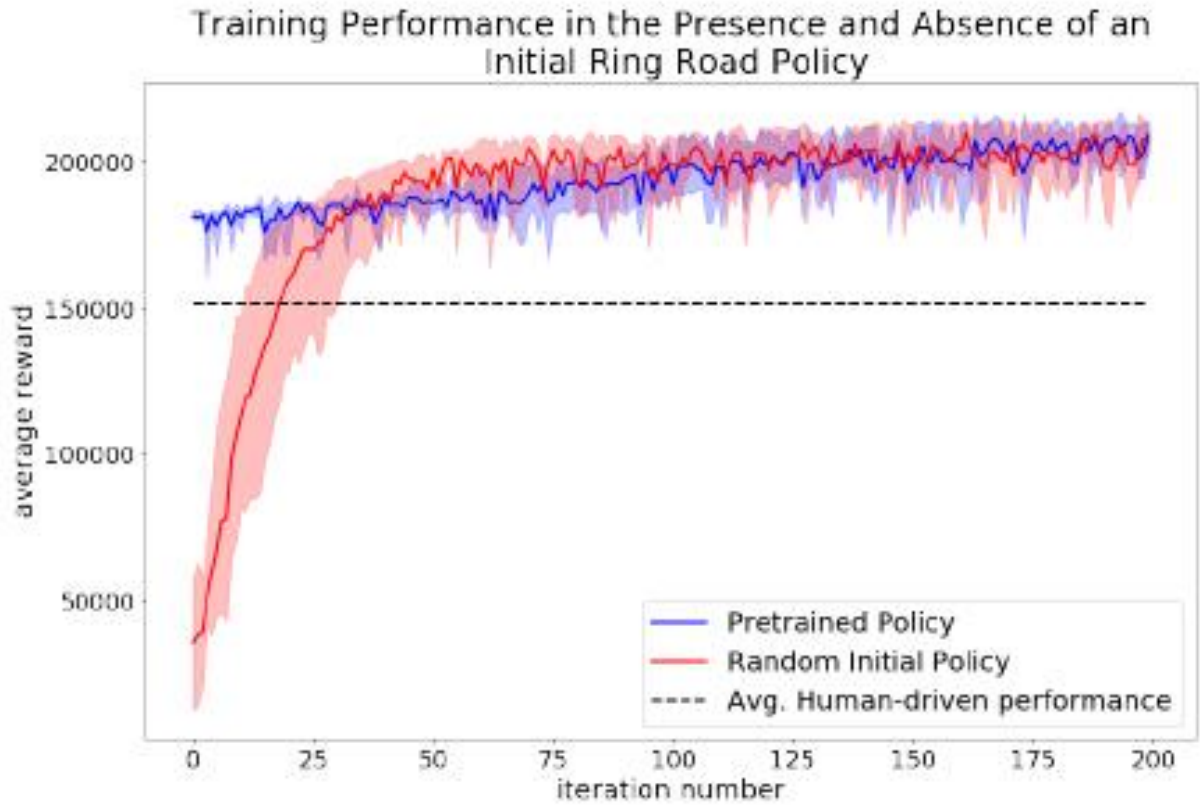


10%



Resultados

Comparación de la recompensa obtenida al usar modelo base y *Transfer – Learning*.



Discusión – Contribuciones Principales

https://www.lohud.com/story/news/transit/2018/11/26/ramp-meter-traffic-signals-287/2069633002/

Mediante la metodología planteada, se obtienen **incrementos notorios en el rendimiento de la red**, medido en cantidad de vehículos que salen de ella. Esta **mejora crece** a medida que **sube el número de CAV'S**, los que terminan actuando parecido a un *ramp-metering*.

Se demuestra que es posible realizar *Transfer – Learning* a partir de una red cerrada, para mejorar el entrenamiento y resultados en una red abierta.

Mediante **inversiones moderadas en CAV's** podrían obtenerse **grandes beneficios** en términos **sociales** (ahorros de tiempo de viaje, combustible, etc.).

Discusión – Comentarios

Es posible que los **parámetros** que se utilizan para el **IDM** puedan alterar los resultados obtenidos, si bien los autores mencionan que fueron calibrados de acuerdo a un flujo de autopista. Podría ser útil analizar **qué tan robusta es la metodología**, modificando los parámetros o el modelo de seguimiento vehicular.

$$\dot{v} = a\left[1 - \left(\frac{v}{v_0}\right)^{\delta} - \left(\frac{s^*(v, \Delta v)}{s}\right)^2\right]$$

Ecuación de aceleración del IDM

Discusión – Comentarios

Pensando en la aplicación de la metodología a la realidad, se podría analizar qué sucede si en vez de permitir una detención total del CAV, se trabaja con **velocidades mínimas**.

Realizar análisis para el caso de experimentos con **dos pistas**.

Otros elementos a analizar comprenden las posibles **fluctuaciones del flujo de entrada** y el **funcionamiento más realista de la rampa**, pues esta última funciona como fuente de perturbaciones más que un flujo realista.

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Dissipating Stop-And-Go Waves in Closed and Open Networks Via Deep Reinforcement Learning

Kreidieh, A. R., Wu, C., & Bayen, A. M. (2018)