Reinforcement Learning for Autonomous Driving using CAT Vehicle Testbed

John Nguyen[*], Hoang Huynh[†], Eric Av[‡], Rahul Bhadani[§], Tamal Bose[¶]

Abstract

We discuss a deep reinforcement learning implementation using the CAT Vehicle Testbed and discuss the merits of simulation-based deep reinforcement learning. After implementing a simple 3 layered neural network which learned using deep Q-learning, we found some challenges associated with ROS and Gazebo. In spite of these challenges, we demonstrate that our reinforcement learning architecture can teach a car to avoid obstacles. With these preliminary results, we discuss what new things we can do and how we can implement more advanced methods.

Introduction

Reinforcement learning is a type of artificial intelligence that rewards and penalizes an agent on what action it takes given its current state. The state describes the agent's environment and action is used to change the state of the environment. The agent is penalized for failing the task and is rewarded for making progress on the task. Deep reinforcement learning utilizes a neural network to compute the action which will give the greatest reward. Due to the generality of reinforcement learning, researchers have been trying to build robots to accomplish complex physical tasks, such as driving [1] [2] [3]. This approach gained notoriety after the introduction of ALVINN [4], the first seminal paper in deep reinforcement learning for autonomous driving. The authors trained ALVINN to achieve 90 % accuracy in predicting whether to turn left or right, given an image of a road. They trained ALVINN using computer-simulated pictures of roads. Though their methods are simple by modern standards, their results demonstrate the utility of deep reinforcement learning for autonomous driving. In this project, we attempt to train a similar driving agent, but with the use of simulations and laser sensors instead of simple images.

In reinforcement learning, the agent responds to its environment with a policy; a mapping from perceived states and possible actions. The reinforcement learning process refers to the agent trying to optimize its policy for maximum reward. Therefore, we try to balance exploration and exploitation. In general, the agent attempts to exploit its knowledge about the environment to get a maximum reward. On the other hand, the agent must balance exploitation with exploring and getting new knowledge. If an agent spends too much time exploring, the agent may not be able to optimize its policy. On the other hand, if the agent spends too much time exploiting, the agent may not recognize an entirely different action it could take which may lead to a better reward. To this end, we employ an epsilon decreasing strategy, where we have probability ϵ to explore and probability $1 - \epsilon$ to explore in each episode. The value of epsilon will change once we have explored enough of the environment.

In 2004, the DARPA Grand Challenge [5] was established, where teams were given sensors and instructed to train a car to drive through a desert near Barstow, California. Subsequent years had similar challenges.

[*] nguy2539@umn.edu, Corresponding Author, University of Minnesota
[†] s_hum@coloradocollege.edu, Colorado College
[‡] eav@zagmail.gonzaga.edu, Gonzaga University
[§] rahulbhadani@email.arizona.edu, The University of Arizona
[¶] tbose@email.arizona.edu, The University of Arizona

This challenge is often cited as igniting American interest in autonomous driving. In the past decade, a number of significant advancements in autonomous driving have been made [6] [7] [8]. This area of research has been popular due to commercial, military and civil interest. Some companies such as Tesla and Uber already have their fleets of autonomous vehicles. Unmanned drones are already used for military purposes. Public imagination is often captured by the perceived convenience of autonomous vehicles.

Training autonomous vehicles have infamously been difficult due to the difficulty in representing driving conditions. Many people have developed deep reinforcement learning methods for autonomous driving using images [4] [9] [10]. By attaching a camera to the car, the agent can base its actions off of the video frames captured by the camera. Unfortunately, this method relies on computer vision algorithms to transform an image into a sequence of numbers. In our project, we represent a state as two floating-point values: distance to the nearest object and angle to the nearest object. Therefore, we do not need to incorporate computer vision methods in our learning. Since we are focusing on a simple obstacle avoidance agent, we do not need the complex regression and classification algorithms used in more complex problems.

ROS and Gazebo

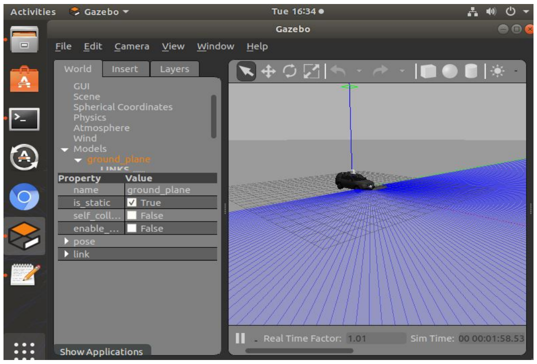
ROS (Robot Operating System) [11] is an open-source robotics software used to help integrate the various parts of a robot. A ROS program can be considered a graph of nodes, where each node represents a different process. In this project, we integrate the CAT Vehicle Testbed [12], which already utilizes ROS, with a new node that manages the deep reinforcement learning architecture. We use sensor data from the CAT Vehicle as input to our neural network, and the steering of the car as the output.

The CAT Vehicle Testbed is already compatible with the Gazebo simulator. Gazebo is a virtual environment that allows us to train and test the car in any virtual world. Since the Gazebo simulations are computationally intensive, it became necessary to restart Gazebo every training episode. Otherwise, the real-time factor of the simulation would become alarmingly low. Since ROS does not exactly work in real-time when low real-time factor decreases, we find our training to be much slower than how a real-life version of the simulation of the planned behavior.

ROS and Simulated World in Gazebo

Gazebo simulation environments can be opened through Ubuntu in various ways. The simulated environment in Gazebo is facilitated through a world file. A world file is essentially a file written in XML code that uses the ROS libraries to declare the environment variables for Gazebo to create in simulation.

We first started with a basic world which includes the URDF model of the CAT Vehicle as seen in Figure 1. This world file's code is the basis for the rest of the training the control behavior in our simulation.

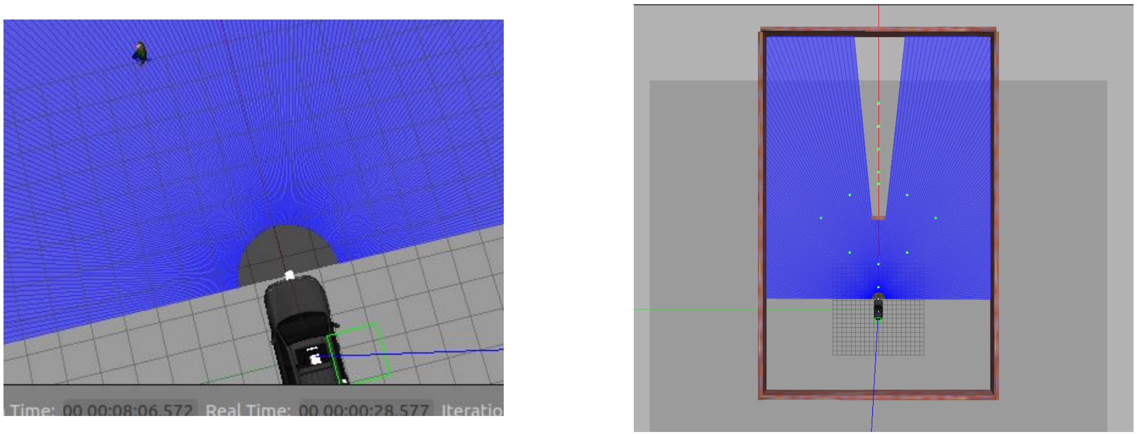


Property	Value
name	ground_plane
is_static	☑ True
self_coll...	□ False
enable_...	□ False

Real Time Factor: 1.01 Sim Time: 00 00:01:58.53

Figure 1: Basic world with CAT Vehicle

We created various other simulation scenario using world files to test basic obstacle detection and avoid-



(a)

(b)

Figure 2: (a). Human actor collision avoidance world (b). Waypoints worth extra reward points to encourage the CAT Vehicle to travel on them.

ance. The brick world, for example, 2b was created as a simple detection and avoidance simulation example. The CAT Vehicle is enclosed in a bounding box to help reset the simulation when running the deep Q-learning algorithm. We want the car to crash into the bounding wall to communicate to the algorithm that it should restart the Gazebo world and choose different values for the next run. Considering the amount of processing power required to run the simulation effectively, we needed to wait hours before the agent showed any tangible results. We also created a world that included a human actor to simulate a pedestrian walking along a crosswalk for the car to detect and avoid. This world includes the XML code for the actor instead of including it in the launch file. However, we ended up simplifying our use cases scenario considering the amount of time it takes to train the neural network. We chose to focus more on obstacle collision and avoidance instead and ended up with a world with given rewarded way-points for the CAT car to maneuver while avoiding a brick wall.

ROS and Gazebo Launch Files

The ROS and Gazebo platform requires launch files to be executed in the command line for the simulation to appear in Gazebo. Launch files are essentially XML files that contain the initialization code for Gazebo which includes the world file path and any other paths desired. For our purposes, we also include the CAT Vehicle paths and topic paths to initialize the position and speed of our CAT Vehicle when Gazebo begins running the launch file.

This is not to confuse when first creating a world and robot directly from the start-up of Gazebo. Saving a world created in Gazebo saves an SDF format file which includes the XML code to interface with Gazebo packages but not necessarily ROS packages. Also, we need to specify paths and maneuver around logistical problems to add an SDF file to our launch file. For these reasons, launch files will only include URDF modeled files and XML coded world files.

Experiment Design

In this case, the neural network (NN) is our agent, and it controls the car. For this reason, we will describe the actions of the car and the neural network as the agent. We used Deep Q-Learning to balance exploration and exploitation. Our neural network consists of 3 fully connected layers: an input layer, a single processing layer, and the output layer. Our input layer takes in 2 arguments: the distance to the nearest object and the angle of the car to the nearest object. Due to sensor limitations, an object is only recognized if it is within 80 meters of the car. Both the input and processing layers have a ReLU activation function, and the output layer has a linear activation function. The output layer would provide action output from one of three actions: straight, turn left or turn right. The car was kept at a constant velocity of 2 m/s, and only the steering angle would change. Turning left would give the car -0.3 m/s of angular velocity. Turning right would give the car 0.3 m/s of angular velocity.

Our reinforcement learning system rewarded the car when the car was following a path or avoiding obstacles, while penalized the car for crashing into things. Each time-step, the neural network is given a state and returns one of the three actions described above. Each state consisted of an ordered pair (d, θ). d represents the distance from the car to the nearest object, and $d \in [4.5, 80]$. θ represents the angle of the nearest object with respect to the front of the car and $\theta \in [-\frac{\pi}{4}, \frac{\pi}{4}]$. Should the NN tell the car to go straight, we reward +2 points. If the car turns left or right then we reward +1 point. This is to incentivize the car to follow the path it is currently on. If the car drives within 5 meters of an object, we simulate a crash and penalize the car with a large negative reward. Furthermore, there are waypoints in the environment which form a path. If the car drives close enough to any of these waypoints, it is rewarded. If the car went to every waypoint, the simulation would end and the car would receive a large positive reward. Due to time constraints and debugging, we ran the experiment for 292 episodes, and at most 2000 timesteps per episode.

When we want to train a new neural network, we first initialize the network and set up the deep Q-learning parameters. Afterward, we start the episodes iteration. At the start of each episode, we launch the Gazebo simulation. After waiting a minute for everything to load properly, we take the initial observation and enter the timestep loop. At the start of each timestep, the car is fed the current state and determines an action. Car executes this action and takes the next state. After the timestep loop is finished, either because the car crashed or this episode reached the maximum number of timesteps, we save the neural network and close the Gazebo simulation. We need to save the neural network because in the event our Gazebo simulation does not load properly, the program will crash. Saving the network allows us to load the network using another program and continue training. After closing the simulation, we have our program wait for two minutes, so all ROS associated programs are closed before we restart ROS. If we do not wait for sufficient time, two instances of ROS will attempt to run at the same time, which will cause our program to crash. To train the car, we needed to incorporate loading and closing the simulation and saving the neural network each episode.

Deep Q Learning Details

Neural Network Structure

Our neural network consists of three fully connected layers: an input layer, a single processing layer, and the output layer (Figure 3). Our input layer takes in two arguments: the distance to the nearest object and the angle of the car to the nearest object. Both the input and processing layers have a ReLU activation function, and the output layer has a linear activation function. The output layer would return predicted Q values of three actions: straight, turn left, and turn right. The loss function of the model is the mean squared error. We use Adam optimizer to update network weights.

Exploitation and Exploration

Exploration is the action of the CAT Vehicle to find information about the environment. Exploitation is the action of the CAT vehicle to maximize return based on known information. There is a trade-off between exploration and exploitation. If the CAT vehicle only exploits the environment, it may miss out opportunities to discover higher reward paths. However, if the vehicle always explores the environment, it cannot utilize known data to maximize return. Therefore, to balance the exploration and exploitation, we set the exploration rate to start at 1.0 and decay it by multiplying in 0.995 after each time step. This ensures that the CAT vehicle starts by exploring the environments and then exploits as it receives more information. [14]

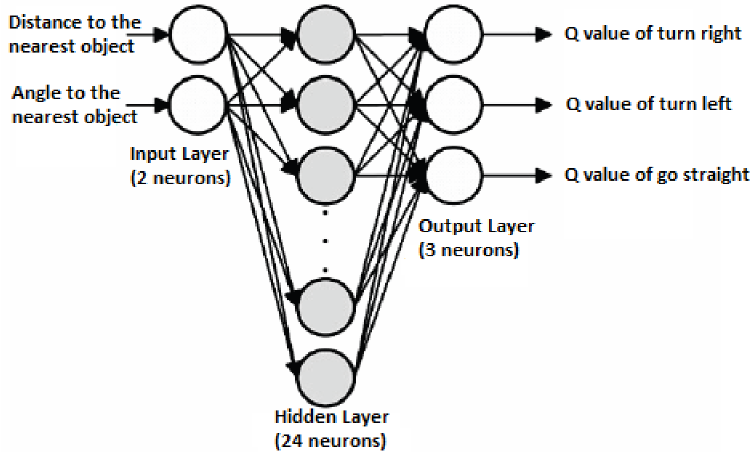


Figure 3: Representation of neural network used for training. Image modified from Cojbasic et. al [13]

Experience replay

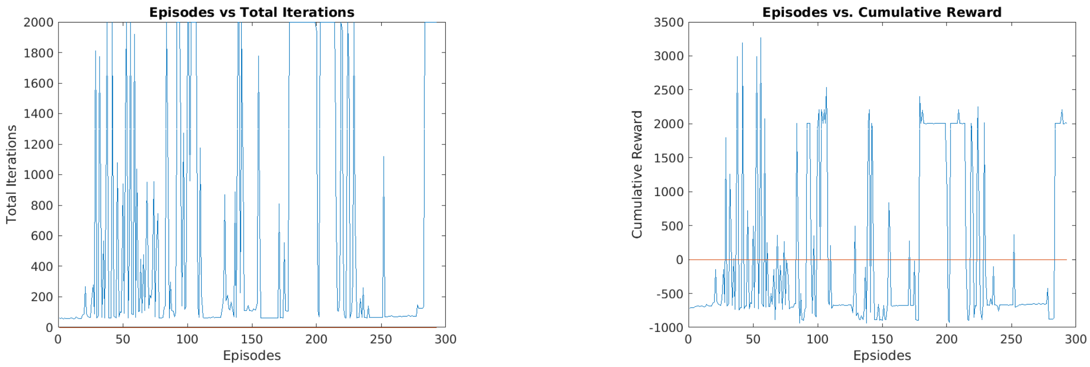
When the CAT vehicle sequentially obtains data from the simulation, the neuron network calculates the loss function based on consecutive samples, which are highly correlated and decrease learning efficiency. Therefore, we apply a technique called experience replay to increase learning process and minimize the undesirable temporal correlations [15]. A memory is created to store the sequential data. After each time step, we take a random batch of data in memory to train the model. Bellman optimality equation is used to update the Q values. We then calculate the loss based on the state and updated Q values.

Training and Results

We trained the car using a simple Gazebo world simulation. The car was enclosed by brick walls with a smaller brick wall (Figure 2). In our simulation, the car is the only moving object and can only drive forward, so the car cannot go in reverse. Due to the size of the car model and position of the laser scan sensor, the minimum distance the sensor can detect is 4.7 meters. For this reason, we determine the car has crashed if the sensor detects an object that is within 5 meters of the car. If the car crashes, the episode ends and the car receives a large penalty. If the car went within 1 meter of a waypoint, the car would receive a moderate reward. If the car went to 10/13 waypoints, the episode will end and the car will receive a large reward. We limit each episode to running at most 2000 timesteps. If the car does not crash for all 2000 timesteps nor did it go to 10/13 waypoints, the simulation will end and the car will receive neither a large penalty nor a large reward.

For each timestep, the car can drive right, straight or left. If the car does this without crashing, it receives a small reward. We debated on whether to penalize the car for driving in each timestep or to reward it. In theory, we should penalize the car for taking an action so it would be incentivized to follow the path. In early training, we noticed a negative reward for taking an action would cause the car to crash itself upon spawning. This is likely because the path of waypoints is not dense enough, so the car does not follow its path. So assuming the car does not follow the path and crashes, it will receive a more negative reward the longer it drives. Therefore if the car crashes faster, it will receive a less negative reward. To avoid this situation, we decided to reward the car for driving. Later in training, we realized the car was only following the path to an extent and began circling the obstacle. Since the car was avoiding obstacles, the car gained a small reward for each timestep. So instead of following the path, the car simply drove around the obstacle and avoiding crashes every timestep to maximize its reward. We plot the reward and episode length vs episodes in Figure 4b.

Notice in Figure 4 that the reward per episode does not appear to follow a trend. This is likely due to some



(a)

(b)

Figure 4: (a). The number of timesteps of each episode. (b). The cumulative reward at the end of each episode. The differences in (a) and (b) are due to the vehicle traveling to waypoints during the episode, resulting in a greater reward.

error in the way we saved and loaded the neural network. During training, the program had crashed several times. When loading the neural network, we were unable to recover some parameters of the parameters of our learning algorithm, such as the discounted epsilon. In particular, we were unable to recover the estimated discounted rewards for the current action and state. Due to this, the learning process was difficult. To alleviate these issues, we need methods to guarantee some parts of our program finish loading before others start. Due to the nature of real-time simulations and crashing, our neural network takes about 40 hours to train, assuming it does not crash. The problem of crashing and the time needed to train present interesting problems we will need to explore at a later date.

Conclusion

In this paper, we introduced a reinforcement learning architecture compatible with the CAT Vehicle Testbed. This enabled us to train the car in a simple Gazebo world we created. This example demonstrates how we can now use simulated worlds with complex physics engines for deep reinforcement learning. Training autonomous vehicles via simulations have been historically difficult due to needing to simulated the physics of actually driving. Since the CAT Vehicle already uses ROS and Gazebo, we can control the car in a simulation for its training. We created a deep reinforcement learning architecture for autonomous driving which can be trained using widely used open-source software with a physics engine.

Further exploration of this topic is necessary. Now that we have built the reinforcement learning framework for the CAT Vehicle Testbed, we can begin to incorporate more complex learning methods and test on how these methods affect deep reinforcement learning for autonomous driving. In particular, we intend to implement meta-learning methods to attempt to train the car more quickly. Meta-reinforcement learning attempts to teach a neural network the ability to abstract problems. For instance, we managed to teach a neural network to control a car to avoid a stationary obstacle by circling the obstacle. In further work, we would like to train the network to also avoid moving obstacles, such as a person. Unfortunately, using our current architecture, we would need to train the car for this specific scenario. Meta-reinforcement learning abstracts stationary and mobile obstacles avoidance into simple obstacle avoidance. Ideally, we would want to place the car in a never before seen obstacle course with stationary and mobile obstacles and have the car avoid all obstacles and reach its destination with minimal training. Due to the challenge of training autonomous vehicles, it is worthwhile to attempt to maximize desired policies while minimizing training.

The foremost in our method is how the simulation crashes when parts of the program do not load in order. Instead of relying on sleep statements and hoping parts of the program load in time, it would be best to have the program wait until necessary parts of the code have loaded before continuing to execute. Since we can utilize Gazebo simulations, the car can be trained in many other worlds and situations. One idea is to simulate highway driving scenarios. Some work has already been done on this topic [16], but was implemented on a scaled city. Once the car is sufficiently trained, implementing the policy with the CAT Vehicle on an actual highway. Another avenue to explore is creating a simulated city in Gazebo and training the car in urban driving. We have introduced a reinforcement learning architecture compatible with Gazebo and ROS, so we can train the car in any environment as long as we can create the environment.

Acknowledgment

This work was supported by the National Science Foundation as a part of the CAT Vehicle Research Experience for Undergraduates (REU) program under the award 1659428. In addition to that, the authors would like to thank Nancy Emptage for logistics related to the CAT Vehicle REU program.

References

[1] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars.," *CoRR*, vol. abs/1604.07316, 2016.

[2] U. Muller, J. Ben, E. Cosatto, B. Flepp, and Y. L. Cun, "Off-road obstacle avoidance through end-to-end learning," in *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf, and J. C. Platt, eds.), pp. 739–746, MIT Press, 2006.

[3] J. Koutník, G. Cuccu, J. Schmidhuber, and F. Gomez, "Evolving large-scale neural networks for vision-based reinforcement learning," in *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation*, GECCO '13, (New York, NY, USA), pp. 1061–1068, ACM, 2013.

[4] D. A. Pomerleau, "Advances in neural information processing systems 1," ch. ALVINN: An Autonomous Land Vehicle in a Neural Network, pp. 305–313, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989.

[5] R. Behringer, S. Sundareswaran, B. Gregory, R. Elsley, B. Addison, W. Guthmiller, R. Daily, and D. Bevly, "The darpa grand challenge - development of an autonomous vehicle," in *IEEE Intelligent Vehicles Symposium, 2004*, pp. 226–231, June 2004.

[6] E. Santana and G. Hotz, "Learning a driving simulator," *ArXiv*, vol. abs/1608.01230, 2016.

[7] A. Lavin and S. Gray, "Fast algorithms for convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4013–4021, 2015.

[8] M. Wulfmeier, D. Z. Wang, and I. Posner, "Watch this: Scalable cost-function learning for path planning in urban environments," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2089–2095, 2016.

[9] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuoglu, U. Muller, and Y. LeCun, "Learning long-range vision for autonomous off-road driving," *J. Field Robot.*, vol. 26, pp. 120–144, Feb. 2009.

[10] A. Giusti, J. Guzzi, D. Ciresan, F.-L. He, J. P. Rodriguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, D. Scaramuzza, and L. Gambardella, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, 2016.

[11] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[12] R. K. Bhadani, J. Sprinkle, and M. Bunting, "The cat vehicle testbed: A simulator with hardware in the loop for autonomous vehicle applications," *arXiv preprint arXiv:1804.04347*, 2018.

[13] Z. Cojbasic, V. Nikoli, E. Petrovic, V. Pavlovi, M. Tomi, I. Pavlovi, and I. iri, "A real time neural network based finite element analysis of shell structure," *Facta universitatis series Mechanical Engineering*, vol. 12, pp. 149–155, 06 2014.

[14] J. G. March, "Exploration and exploitation in organizational learning," *Organization Science*, vol. 2, no. 1, pp. 71–87, 1991.

[15] R. Liu and J. Zou, "The effects of memory replay in reinforcement learning," *CoRR*, vol. abs/1710.06574, 2017.

[16] K. Jang, E. Vinitsky, B. Chalaki, B. Remer, L. Beaver, A. A. Malikopoulos, and A. Bayen, "Simulation to scaled city: zero-shot policy transfer for traffic control via autonomous vehicles," in *Proceedings of the 10th ACM/IEEE International Conference on Cyber-Physical Systems*, pp. 291–300, ACM, 2019.