A GRAPHIC ADVANTAGE

DMITRY KUDINOV EXPLAINS HOW USING A GPU, A NEURAL NETWORK AND MACHINE LEARNING CAN SIGNIFICANTLY SPEED UP THE PLANNING OF COMPLEX TRANSPORT ROUTES

Calculating travel times is a foundational element of transportation logistics, urban design, asset management, retail, and countless other areas in the public and private sectors. Building a route from one location to another is now possible using numerous services and applications quickly and for free.

But what if you need to build a route that's more complex than simply a line from point A to point B? One that has multiple, nonlinear stops, and which also factors in traffic as a variable? Things suddenly become trickier when you want to find the best sequence, along with expected arrival times. While this is a simple enough scenario to model, many large companies face the challenge of calculating travel times at a scale of thousands of stops, with millions of travel-time estimations.

At Esri, we employ traditional algorithms that have been used successfully to tackle these problems for many years. However, these are expensive to scale and often demand routine engineering efforts to meet complex modelling requirements. Looking for alternative approaches, we went to hardware manufacturer NVIDIA, which provided its graphics processing unit (GPU). We then used artificial intelligence (AI) and machine learning to train a neural network to create a prototype system to predict and visualise travel times for transport networks with a large number of complex, hard-to-model and hidden variables. Here's how we did it.

The experiment

Logistics companies schedule multiple stops for multiple vehicles simultaneously. Large companies, while doing next-day planning, schedule thousands of stops with hundreds of vehicles per day, treating this as a single optimisation problem. But besides the number and locations of stops, there are other aspects of transport that are hard to model, such as seasonal traffic patterns and individual route preferences. Although some of these factors are hard to formalise and even harder to represent with traditional algorithms, they are integral properties of modern transport and have already been captured and buried deep inside existing individual GPS tracks.

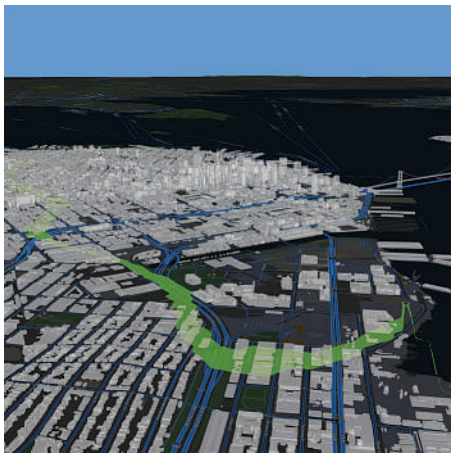

Figure 1. Boundary (in green) showing five-minute travel time from downtown San Francisco

While large-scale logistics challenges ask for high-throughput computations, more nuanced scenarios demand more flexibility without increasing the complexity and maintenance costs of the model. So, we decided to see if both requirements could be met with the help of machine learning. Using a simulated set of 300 million 'journeys' (GPS tracks represented only by two locations – departure points and destinations – as well as departure time and the number of minutes it took to travel) covering roads in the region of California and Nevada in the United States, we trained a neural network to predict travel times and record them on a ground truth transport graph. After being trained, the neural network could predict travel times – with measurable accuracy – between any two locations in California and Nevada, taking the departure time into account and effectively embedding the factor of road congestion into its function.

Once trained, the neural network produced predictions with enormous throughput: a single desktop machine with an NVIDIA GV100 GPU card can calculate more than 300,000 ETAs per second, which is at least two to three orders of magnitude faster than common traditional deterministic algorithms.

Logistics companies already use various algorithms to solve multivehicle scheduling problems, and at the core of most of them lies the so-called Origin-Destination (OD) Cost matrix, which needs to be calculated first and is filled with ETAs for any possible combination of two stops. For instance, if a company needs to make a net 1,000 stops with its trucks, the OD Cost matrix will calculate one million ETAs. The neural network can completely populate this matrix in only three seconds.

Additionally, it can successfully figure out accurate representations of road congestion patterns by using GPS tracks only, which proves its ability to learn hidden and deeply buried patterns. It also indicates a great potential for further fine-tuning with user-preferred routes or adapting to individual driving habits or commuter preferences.

The details

NVIDIA's GPU gave us the ability to train neural networks of a realistic size, and the various experimental travel times were shortened 20- to more than 50-fold, thanks to the massive parallelisation of matrix operations needed for training. This made the search for optimal neural network architecture and numerous hyperparameter values feasible and effective.

We then used TensorFlow and Keras libraries to build a dense, fully connected neural network – or multilayer perceptron (MLP) – with 16 hidden layers and a total of 10 million trainable parameters. To reduce the overfitting, we added a dropout node right before the output layer. The input was represented by normalised pairs of coordinates for departure and destination locations and departure time; the output was a single value showing the number of minutes it took to travel from point A to point B at a given time.

We used mean squared error (MSE) as the loss function, and the Adamax optimiser with the initial learning rate of 1e-3. Training was performed for a total of 4,000 epochs on consecutive subsets

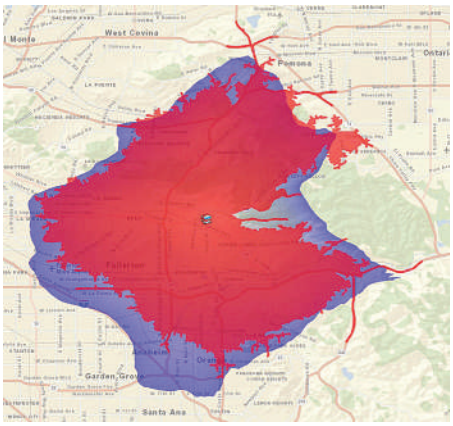

Figure 3. Predicted isochrone (blue) versus ground truth service area polygon (red). This mapped the accuracy of the predictive model

of 20 million journeys, simulating online learning. By the end of training, the MSE value on the validation set was about 13.5.

But can the neural network be usable at this point? While an MSE of 13.5 translates into a 3.7-minute standard deviation of predicted values from the ground truth, the routes in the California-Nevada region may differ significantly in size. So, a 3.7-minute difference may be acceptable for an hour-long route, but it would be excessively inaccurate for shorter routes.

HAVING A TOOL THAT ALLOWS YOU TO SEE EXACTLY WHERE AND WHEN GOODS AND CUSTOMERS WILL ARRIVE IS A CRITICAL ADVANTAGE

For a more precise metric of prediction accuracy, we constructed a chart that shows the variation of prediction accuracy as a function of route length. This provides a complete picture of where the accuracy is acceptable and where it is not. By analysing this chart and adjusting the route lengths' distribution in the training set accordingly, we can vary the resultant prediction accuracy without increasing the number of trainable parameters of the neural network.

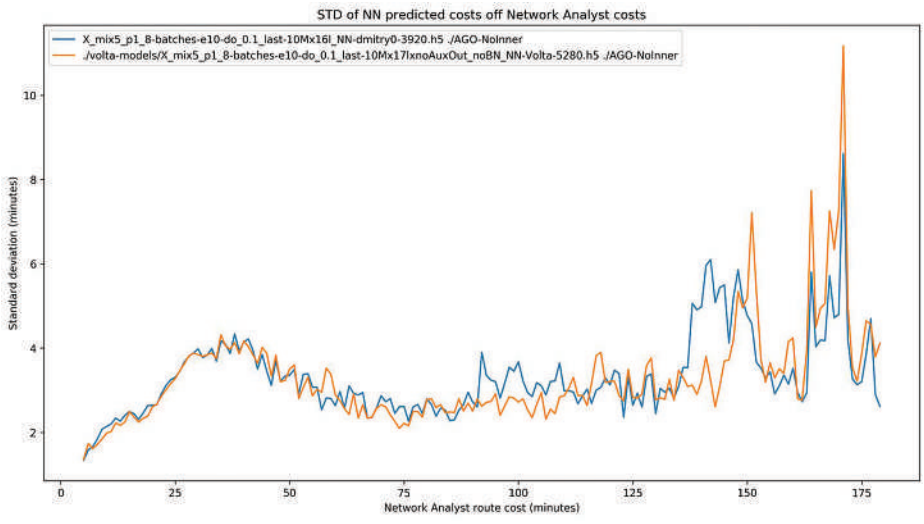

Figure 2. Variation of prediction accuracy as a function of route-length with two models compared on one chart

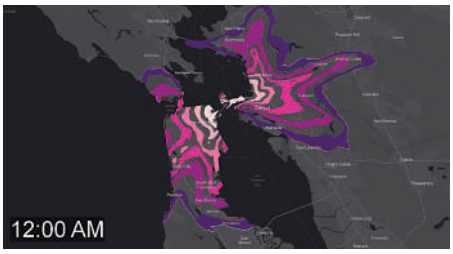

Figure 4. Road congestion patterns captured by the neural network during 'training'. The rings expand and contract depending on the time of day traffic was recorded

Another great tool we built for evaluating prediction accuracy is a Web Map Server (WMS) REST service endpoint wrapping our trained neural network. This service returns a geographically bound PNG image containing a travel time surface where every pixel is coloured corresponding to the time it takes to reach it from the central pixel. Once constrained by a maximum travel time value, the surface looks like an isochrone polygon, which, if overlaid on a traditional service area polygon on a map, allows a visual comparison of neural network predictions and the ground truth transport graph. Such isochrones represent reachability zones, which also change over time – shrinking during business hours and expanding back to full size during the night – and offer a great way to see how neural networks can perceive road congestion.

The road ahead

This ETA algorithm, combined with mapping and analytics, has many real-world applications beyond logistics. The ability to accurately predict arrival times and represent these values on a map is useful for any organisation that must factor in time and location as crucial parts of its business models.

For instance, this technology can be used by banks to perform ATM and branch location accessibility analyses. Or now that fast-food restaurants are offering services where customers can order ahead of time via a mobile app, they can ensure that the food and drinks are ready and fresh at exactly the time customers arrive at the restaurant location. With brick-and-mortar businesses racing to meet the expectations of consumers in a space where demand can be met at breakneck speed, especially by online retailers, having a tool that allows you to see exactly where and when goods and customers will arrive is a critical advantage.

Dmitry Kudinov is Senior Data Scientist at Esri (www.esri.com)