Fingerprinting Vehicles Using CAN Bus Data Snapshots

David R Crow, 2d Lt, USAF

July 10, 2019

1 Problem Domain

In an environment of ever-increasing numbers of malicious actors, the Air Force requires a reliable intrusion detection system (IDS) in each of its vital systems. My thesis research project hopes to determine whether inherent causal relationships between various functions or metrics in a car can be identified and then leveraged as an IDS. This project aims to provide the foundation for such a system.

To do so, we will attempt to identify which of nine distinct vehicles generated a given segment of controller area network (CAN) bus data. This is a multiclass classification problem which asks the following question: does a given vehicle generate data with some characteristic unique to that vehicle? In other words, does a given vehicle leave identifiable fingerprints on its data?

A demonstrated ability to identify a vehicle using only its CAN bus data likely implies that unique relationships are present (and evident) in the data. Although success in this domain will not necessarily alter the course of my thesis research, failure may indicate that the causal relationships necessary for an IDS are simply not present.

2 Deep Learning Task

One-shot learning is a deep learning (DL) fingerprinting method that may prove useful for this project. Says Andriy Burkov: "In one-shot learning, typically applied in face recognition, we want to build a model that can recognize that two photos of the same person represent that same person. If we present to the model two photos of two different people, we expect the model to recognize that the two people are different" [1]. By treating relatively small segments of CAN bus data like photographs, we can build a *vehicle* recognition model.

One type of neural network, the siamese neural network (SNN), allows for effective one-shot learning without requiring exceptionally large numbers of training examples. With an SNN, we do not need to split the data for a given car into a large number of tiny snapshots; instead, we can utilize relatively large snapshots that are much more likely to capture potential vehicle-specific fingerprints. Ideally, an SNN trained on these snapshots can then effectively classify data segments.

3 Data

The dataset comes from Oak Ridge National Laboratory (ORNL) in the form of a `.db` file. In total, the lab shared over three gigabytes of temporally-organized CAN bus data recorded during 34 captures of nine different vehicles. The database details each vehicle's *make*, *model*, *year*, *electrification_level*, and fuel types; this information is shown in Table 1.

Some vehicles, like the Subaru Outback, only have one capture; others, like the Ford Fusion and the Nissan Leaf, have several (the Leaf has the most captures with 14). Each capture contains metadata which includes a *vehicle_id* (values in 1-9), a *capture_id* (values in 1-32, 46, 52), a *timestamp* (the Unix time at the start of the capture), a *tag* (e.g., cmax_diagnostics, f150_drive, or leaf_eco), a *diagnostics* flag (a boolean which represents whether or not the capture is for diagnostic purposes), and a *description* (e.g., "eco mode from calhouns-oak-ridge to Bobby home" or "driving around ORNL campus while injecting diagnostics on loop").

Table 1: The ORNL Dataset's Vehicle Metadata

vehicle_id	make	model	year	electrification_level	fuel_type_primary	fuel_type_secondary
1	Toyota	Tacoma	2008	NULL	Gasoline	NULL
2	Toyota	Corolla	2009	NULL	Gasoline	NULL
3	Nissan	Leaf	2011	BEV (battery electric vehicle)	Electric	NULL
4	Ford	C-Max	2013	Plug-in hybrid	Electric	Gasoline
5	Chevrolet	Volt	2015	Plug-in hybrid	Electric	Gasoline
6	Ford	F-150	2014	NULL	Gasoline	NULL
7	Ford	Fusion	2016	Plug-in hybrid	Electric	Gasoline
8	Subaru	WRX	2017	NULL	Gasoline	NULL
9	Subaru	Outback	2009	NULL	NULL	NULL

Table 2: Percentage of Messages Belonging to Each Vehicle in the ORNL Dataset

Vehicle ID	Messages	Proportion
1	640,591	1.57 %
2	1,044,769	2.57 %
3	22,526,385	55.30 %
4	1,897,692	4.66 %
5	7,076,525	17.37 %
6	1,222,985	3.00 %
7	4,570,278	11.22 %
8	1,080,978	2.65 %
9	671,279	1.65 %

Additionally, each capture contains between 37,333 and 7,076,525 messages; the mean and median values over all captures are 1,197,984.76 and 1,062,873.50, respectively. In total, the dataset includes 40,731,482 CAN bus messages. Each of these messages includes a *capture_id*, a *timestamp*, an *arbitration_id* (which "identifies the message and indicates the message's priority" [2]), a *dlc* (or data length code, which simply counts the number of data bytes), and the hexademical *data* itself.

The wrangling process for these data is straightforward. DB Browser for SQLite, a program that allows one to view and modify `.db` files, enables structured query language (SQL) queries and comma-separated values (CSV) file output. One can use an SQL query to export exactly those fields one desires to a CSV file, which one can then read into a Pandas dataframe for DL purposes.

4 Truth Data and Performance

The nine vehicles detailed in Table 1 constitute the nine classes for this classification task. Every message in the dataset has an associated capture, and every capture has an associated vehicle; thus, the ORNL dataset is fully-labeled. As Table 2 shows, the messages are not evenly distributed over the nine classes. However, many DL techniques consider imbalanced datasets, and so training an effective model is still possible.

It is relatively easy to measure the performance of a sufficiently-trained model. For a multiclass classification problem, one-versus-all confusion matrices often prove useful. The F-measure for each confusion matrix (as detailed in [3]) highlights the model's performance over the nine classes. Specifically, it indicates whether a vehicle's unique fingerprint – if such a fingerprint even exists – is present in a snapshot of that vehicle's CAN bus data. If necessary, we can also micro- or macro-average these scores to obtain a single performance metric for the model.

5 Research Support

This project supports my own research. We aim to determine whether forecasting methods like empirical domain modeling (EDM) and machine learning can be used as an effective IDS. To do so, we hope to identify significant causal relationships; if we can use (say) EDM to demonstrate that Value A always precedes Value B, then an observed Value A or Value B – but not both – could indicate faulty equipment or a malicious actor in the system. We can devise more complex relationships, but the theory is essentially the same.

In this context, this project will be useful. By employing modern DL techniques, we hope to determine whether CAN bus data can uniquely identify a vehicle. If so, it's likely that causal relationships unique to a given vehicle are present in the vehicle's data; it may be possible to then leverage these relationships into an IDS. However, if instead this project shows that CAN bus data cannot fingerprint a given vehicle, attempts to construct a viable IDS for said vehicle using assumed relationships in the CAN bus data may prove infeasible.

References

[1] A. Burkov, *The Hundred-Page Machine Learning Book*. 2019.

[2] National Instruments, "Controller Area Network (CAN) Overview." https://www.ni.com/en-us/innovations/white-papers/06/controller-area-network–can–overview.html, 2019. [Online; accessed 8-July-2019].

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.