What are the main points when writing about custom dataset?

* Why I make a custom dataset ?
* How the dataset was made?

Custom Dataset

The CIC-IDS-2017 dataset is a widely used dataset in the field of network intrusion detection system (NIDS). Current models that run on the CIC-IDS-2017 dataset achieve very good results. These good results are due to spliting of the dataset into a training set and test set. Normally, this division prevents the model from seeing instances of the test set in advance, thereby evaluating the generalization of the model. However, in the context of NIDS (for example KDD99, NSLKDD, CIC-IDS-2017 or most recently TON\_IOT\_Net), these datasets are all formed in a simulated environment, with machines running attack using scripts. I make two assumptions that:

* Models trained in a simulated environment will only classify well in that simulated environment.
* Models learn the properties of the attack script, not the essence of the attack type. In the case of using different attack tool, the model will not be able to identify it.

How dataset was made ?

There are three step in making this dataset.

First I simulate a DoS attack by using wrk tool. This tool is an HTTP benchmark tool, working by send stream of HTTP request to the server. Then I capture the pcap file in webserver using tcpdump. Then I pass the pcap file through CICFlowMeter. This is the same tool used for generating well-knowned CIC-IDS-2017 dataset. I mark all network flow with Destination IP and Source IP equal to IP of attack machine. I perform 2 attack scenario, Single Machine Attack And Multiple Machine Attack. More detail of how dataset was made can be found at <https://github.com/NguyenQuangMinh0504/IDS-Dataset>