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.

The dataset was constructed through a multi-stage process involving attack simulation, traffic capture, and flow generation. Initially, A DoS attack was simulated using the wrk tool, a benchmark tool that generates high volumes of HTTP requests to a target server. Concurrently, tcpdump command was employed to capture the network traffic generated during the attack. This captured traffic, stored in a pcap file, encapsulated a comprehensive representation of the attack dynamics. Subsequently, the captured pcap file was processed through CICFlowMeter, a tool renowned for its utilzation in the creation of CIC-IDS-2017 dataset. CICFlowMeter meticulosly parsed the network traffic and generated a comprehensive set of network flows, providing a granular view of the communication patterns within the captured data. To facilitate the subsequent classification tasks, a labeling process was undertaken. Network flows exhibiting Destination IP or Source IP matching the IP address of the attack machine were identified and labed as attack traffic. This labeling strategy enabled a clear distinction between legitmate network flows and those associated with the simulated DoS attack. For a more comprehensive exploration of the dataset creation methodology, refer to [1].

For dataset created with single machine attack, I use a simple ANN model with 3 layers. The whole dataset has 2529 records, in which 1007 records are normal flow.

<https://github.com/NguyenQuangMinh0504/IDS-Dataset>