Department of Computer Science

Constructing the Face of Network Data

Ertza Warraich, Muhammad Shahbaz

1. Problem Statement

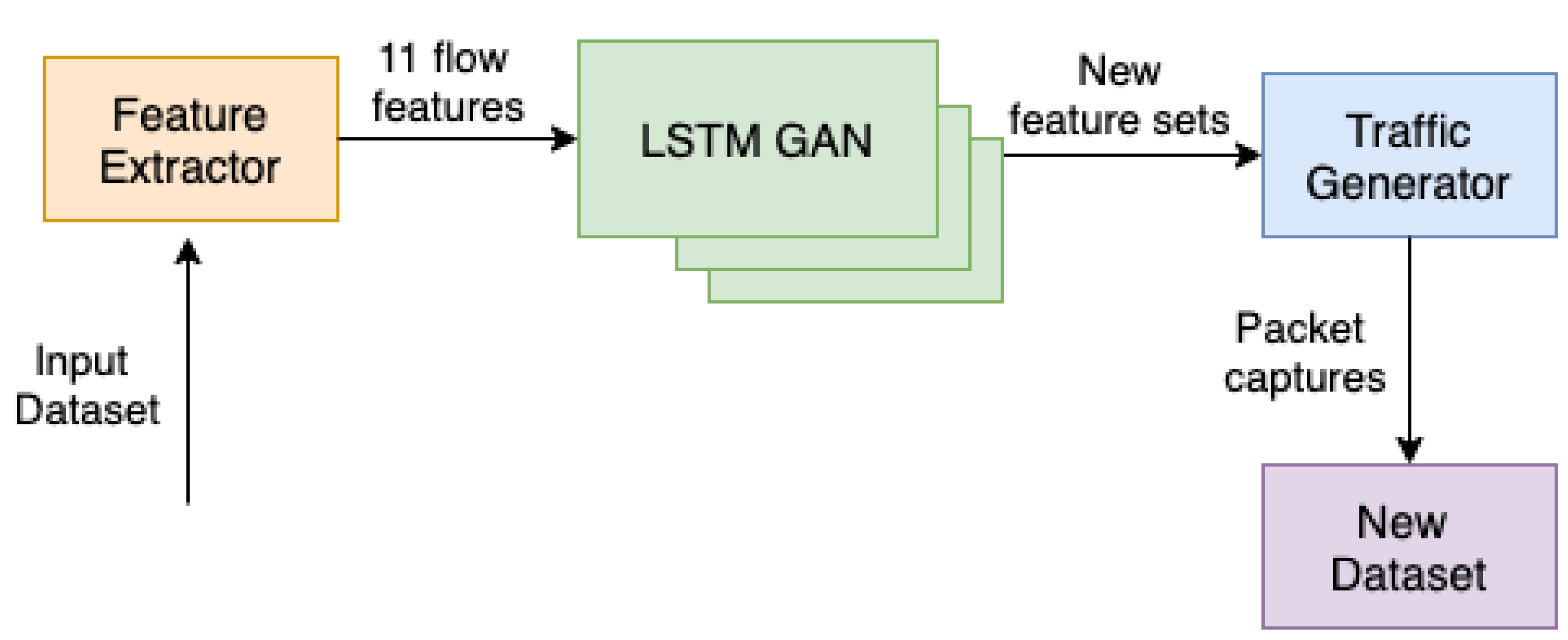
- Network **datasets** are **essential** for operating, managing, and **understanding** ever-changing **modern networks**. However, such **datasets are rare**:
  a) due to **privacy and sensitivity** concerns
  b) and **provisioned as feature sets** instead of packet traces

- Having **pre-selected features** severely **limits the application** of datasets as they can only be used in a subset of Machine Learning systems. But **feature extraction and selection** is also a formidable undertaking requiring **domain mastery** of the dataset.

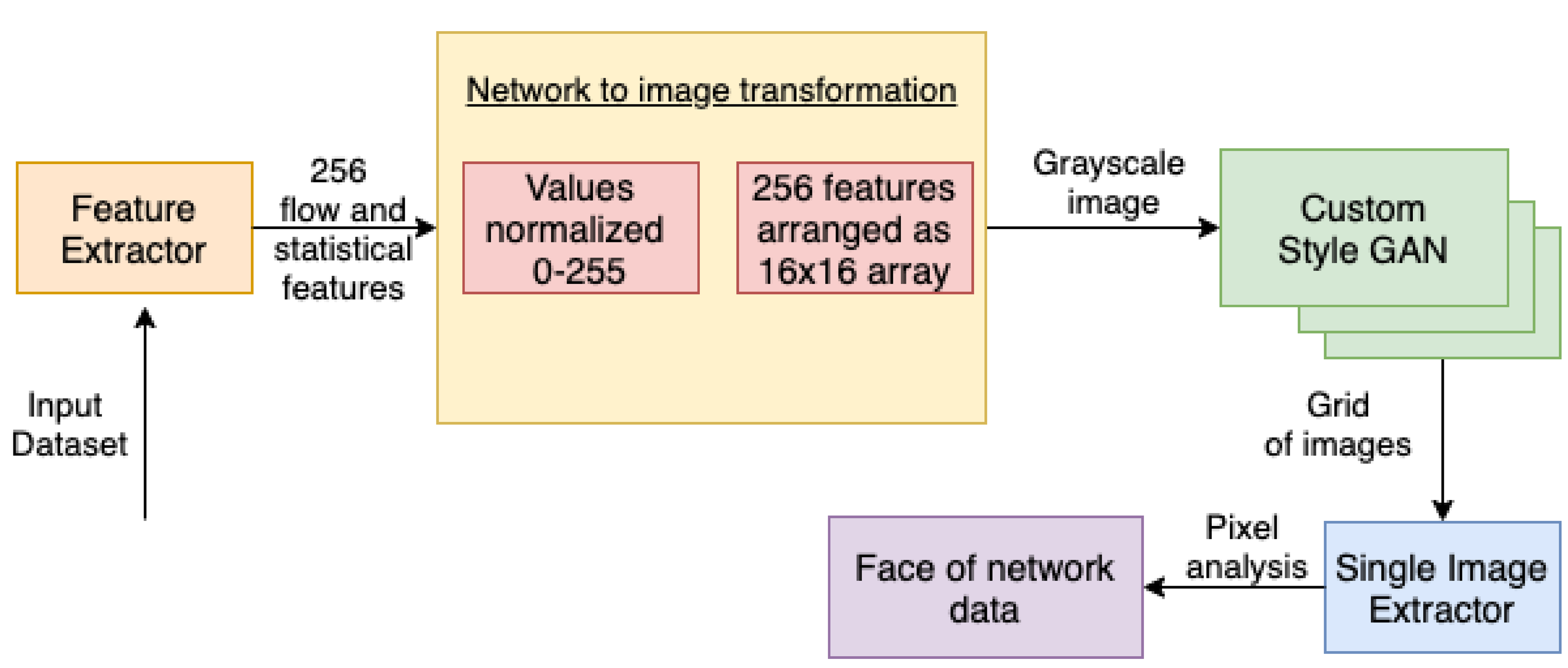
2. Approach

- We present a **GAN based approach** for the two problems:
  a) generate new and timely datasets (packet traces) automatically
  b) and **find the key features** to create face of network data

Dataset is generated in the following way:
1. Features extraction from input dataset
2. Feed the features to our **LSTM GAN** framework
3. GAN learns and **generates new feature sets**
4. Traffic w/ **random bits** is generated using those features
5. New dataset in packet capture format is output



To construct the face and find the most important features:
1. Extract **256 features** from input network dataset and transform them to an **image representation**
2. The features are fed to our custom **StyleGAN**
3. It outputs a **grid of newly generated images**
4. The most prominent pixels in the generated images identify what are the most important features of that dataset



3. Experimental Setup

- We test the quality of newly generated datasets by **applying our framework** to a **well-known problem of censorship circumvention** and **traffic classification.**
  1. We use a **Skype dataset** to train our GAN
  2. New Skype dataset is output from our framework
  3. The new dataset is passed through state-of-the-art **Skype traffic classifiers** and results are evaluated

Model	Training Time	Accuracy	F1-Score
Logistic Regression	9.3 seconds	99.32%	1.00 0.99
Multi-Layer Perceptron	3.16 minutes	99.82%	1.00 1.00
K-Nearest Neighbors	1.58 minutes	99.95%	1.00 1.00
Decision Tree	4.1 seconds	99.96%	1.00 1.00
AdaBoost	0.58 minutes	99.96%	1.00 1.00
Random Forest	4.7 seconds	99.99%	1.00 1.00

Classifiers' accuracy on actual Skype dataset

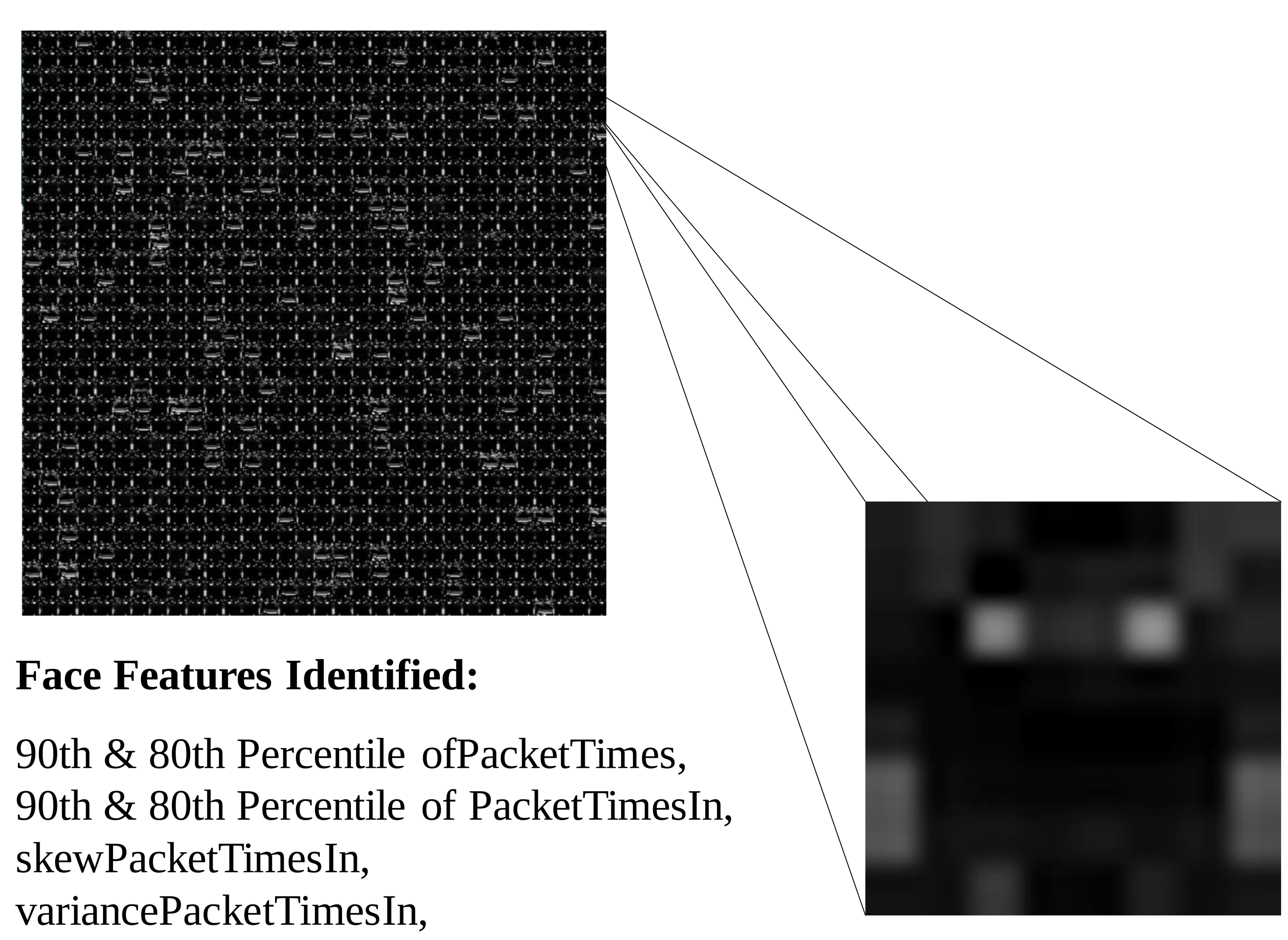
4. Evaluation

- Almost all the **traffic passes through as Skype** in our experimental setup

Model	Classified as Skype	Classified as Other
Logistic Regression	100.0000%	0%
Multi-Layer Perceptron	100.0000%	0%
K-Nearest Neighbors	99.0396%	0.9604%
Decision Tree	99.1597%	0.8403%
Random Forest	100.0000%	0%
AdaBoost	100.0000%	0%

Classifying our framework's traffic

- We **enhance and extrapolate a singular image** from the grid generated by our StyleGAN to take a closer look at the **face of network data** and **highlight the key features** which make a network dataset unique from other datasets.



Face Features Identified:

90th & 80th Percentile ofPacketTimes,
90th & 80th Percentile of PacketTimesIn,
skewPacketTimesIn,
variancePacketTimesIn,
skewPacketTimesOut,
variancePacketTimesOut