Synthetic 5G Traffic Generation: A Machine Learning Approach

Karsten Cedric van der Deijl[1] **Supervisors:** Nitinder Mohan[1], Marco Colocrese[1]

[1]Delft University of Technology

TU Delft

Introduction

- No public **5G** packet **datasets**; privacy/proprietary limits.
- Existing tools (e.g., TRex) ⇒ constant-rate, miss real timing.
- Existing work: flow-level, **not packet-level** [1].

Related Work

- Bytewise synthesis: **PAC-GAN**, **PacketCGAN** & **PcapGAN** ⇔ GANs.
- PAC-GAN: **novel** nibble encoding, high validity [2].
- Tabular ⇒ state-of-the-art choices: **TabularARGN**, **TVAE**, **CTGAN**, **REaLTabFormer**.
- TabularARGN: **best performance** on diverse datasets [3].
- Compare key models: PAC-GAN (novel direct) vs TabularARGN (top tabular) [3].

Research Question

*How can **machine learning techniques** be used to generate **synthetic 5G** network **traffic**? What ML techniques are most suitable for this task?*

1. *What are the existing ML-based methods for synthetic traffic generation?*
2. *How do the methods compare in terms of fidelity and ease of integration?*

Dataset and Preprocessing

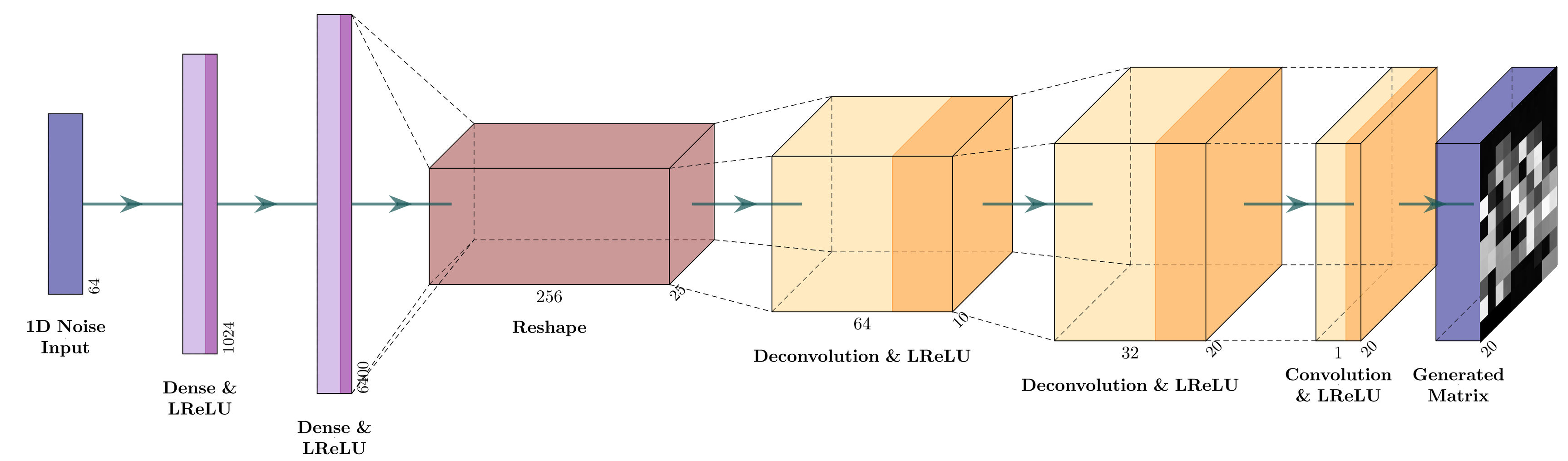
- **2.68M** packets from **simulation 5G network** [4].
- **90/10** train/test **split**: 2.41M/0.26M.
- Parsed with **Scapy** to table; omitted low-variance & variable length fields.

TabularARGN

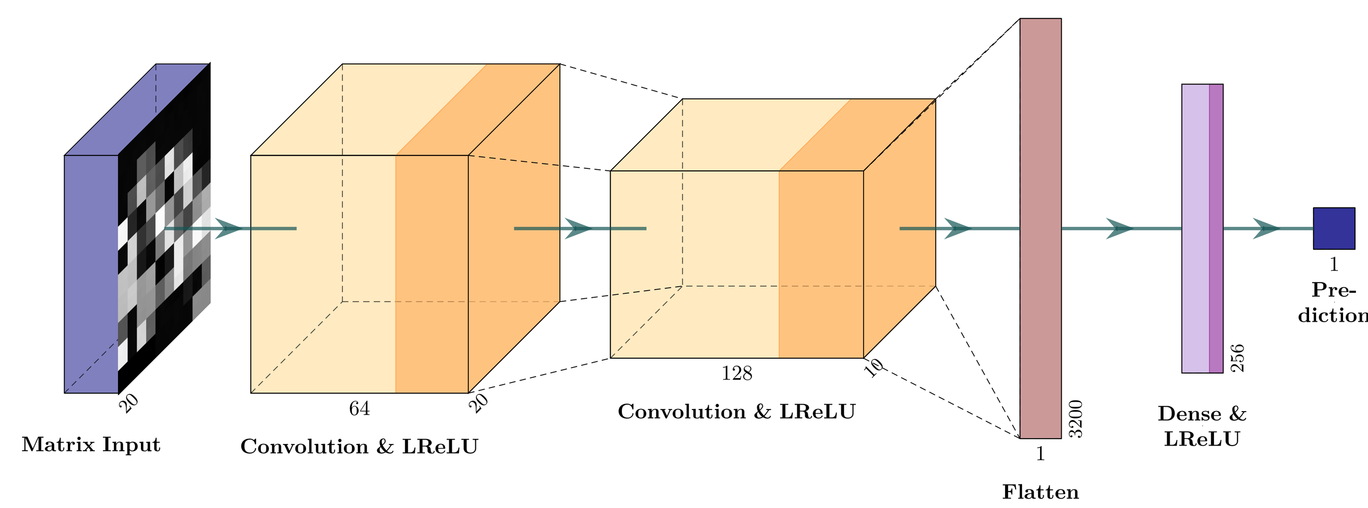
- **Auto-regressive** NN for tables.
- Predicts fields from each other ⇒ capture inter-column relations.
- Implementation from authors [3].

PAC-GAN

- **Convolutional GAN**, nibble-matrix packet encoding.
- Adversarial training with Wasserstein loss.
- TensorFlow implementation, adapted from original; focus on **headers** + **timedelta**.



(a) Generator Architecture



(b) Discriminator Architecture

Figure 1. PAC-GAN Model Architecture
Created using PlotNeuralNet [5]

Evaluation Metrics

Validity

Fractional **Protocol-Compliance**

Marginal Distribution

- **Univariate**: distribution match
- **JSD**: similar category distribution.
- **EMD**: similar numeric value distributions.

Joint Distribution

- **Bivariate**: pairwise field relationships preservation.
- **Coverage**: measures missing real patterns
- **Precision**: measures sample realism
- **Recall**: measures captured diversity
- **Density**: measures clustering in real regions

Results

Metric	PAC-GAN	TabularARGN
Validity ↑	96.89%	100.00%
EMD ↓	0.0191	0.0121
JSD ↓	0.2281	0.2404
Univariate ↑	0.8215	0.7069
Bi-Variate ↑	0.6651	0.4814
Coverage ↑	0.1786	0.0015
Recall ↑	0.9843	0.9986
Density ↑	0.2316	0.0027
Precision ↑	0.5318	0.0073

(a) Evaluation Results

	PAC-GAN	TabularARGN
# Parameters	7.87M	0.47M
Size on Disk	30.09MiB	5.52 MiB
Generation time for 500K	2.84s	3.38s
Training time	~5hrs	~14min

(b) Model Complexity

Table 1. Evaluation Results (1a) & Model Complexities (1b), for PAC-GAN & TabularARGN

- Both models show **good validity**; TabularARGN nearly guarantees protocol adherence.
- Both **capture marginal distributions** well; Slightly better PAC-GAN.
- PAC-GAN balances joint metrics ⇒ **diverse and realistic** samples.
- PAC-GAN captures **inter-field dependencies**.
- TabularARGN misses inter-field dependencies ⇒ **less realistic** outputs.

Contributions

- Developed and applied expansive **evaluation framework**.
- Compared **Tabular** vs **direct-GAN** approach.
- Introduced inter-packet **timing** modeling.

Conclusion and Discussion

- Deep generative models: **high-fidelity**, protocol-valid 5G headers.
- PAC-GAN: **best** trade-off, **realistic** & **valid**.
- Future: incorporate **more models** and protocol **layers**, explore **privacy preservation.**

References

[1] Adrien Schoen, Gregory Blanc, Pierre-François Gimenez, Yufei Han, Frédéric Majorczyk, and Ludovic Me. A Tale of Two Methods: Unveiling the Limitations of GAN and the Rise of Bayesian Networks for Synthetic Network Traffic Generation. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 273–286, July 2024. ISSN: 2768-0657.

[2] Adriel Cheng. PAC-GAN: Packet Generation of Network Traffic using Generative Adversarial Networks. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0728–0734, October 2019. ISSN: 2644-3163.

[3] Paul Tiwald, Ivona Krchova, Andrey Sidorenko, Mariana Vargas Vieyra, Mario Scriminaci, and Michael Platzer. TabularARGN: A Flexible and Efficient Auto-Regressive Framework for Generating High-Fidelity Synthetic Data, February 2025. arXiv:2501.12012 [cs].

[4] Cooper Coldwell, Denver Conger, Edward Goodell, Brendan Jacobson, Bryton Petersen, Damon Spencer, Matthew Anderson, and Matthew Sgambati. Machine Learning 5G Attack Detection in Programmable Logic. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pages 1365–1370, 2022.

[5] Haris Iqbal. HarisIqbal88/PlotNeuralNet v1.0.0, December 2018.

K.C.vanderDeijl@student.tudelft.nl