Transformer Models for Generating Networked Time-Series Data

Yusuf Elnady

Transformer Models for Generating Networked Time-Series Data

Yusuf Elnady

1. Intro
2. Goal
3. Approach
4. Evaluation Plan

Intro

- **Limited data access** is a longstanding barrier to data-driven research.
- Data-driven techniques are central to networking and systems research.
  - allows network operators and system designers to take new data-driven management decisions.
- In practice, the benefits of data-driven research are restricted to **those who possess data**.
- Even when **collaborating stakeholders** have plenty to gain, they are reluctant to share datasets for fear of revealing business secrets and/or violating user privacy.
  - Example: an Internet Service Provider may need **workload-specific optimizations** from an **equipment vendor**.

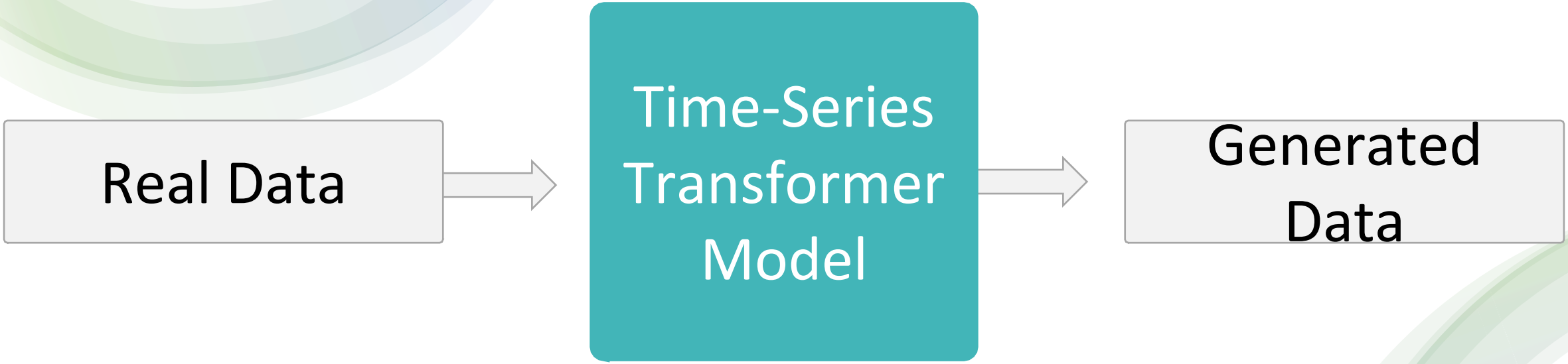
Intro

- **Why to have a generative framework for networked data?**
  - Such a **framework** can **enhance the potential of data-driven techniques** by making it easier to obtain and share data.
  - By just obtaining a **few records of new attack** (small dataset), you can generate a lot of similar datapoints to increase the dataset size.
  - Large number of samples is always useful in data analytics and downstream tasks.

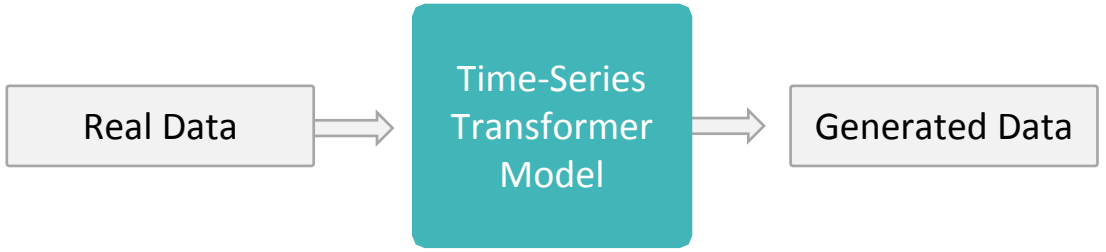
Goal

- My focus is on synthesizing **multi-dimensional time series measurements (X)** associated with **multi-dimensional metadata (Classes).**
  - Common in **networking** and **systems applications.**
- My goal is generating new samples that can outperform baseline algorithms on downstream tasks.
- Example of network and systems datasets
  - Cluster requests (e.g., Google Cluster Usage Trace)
    - The logs contains 1) **measurements** of task resource usage, and 2) **the exit code** of each task.
  - Web sessions (e.g., WWT)
  - Bandwidth measurements (e.g., FCC MBA)

Approach

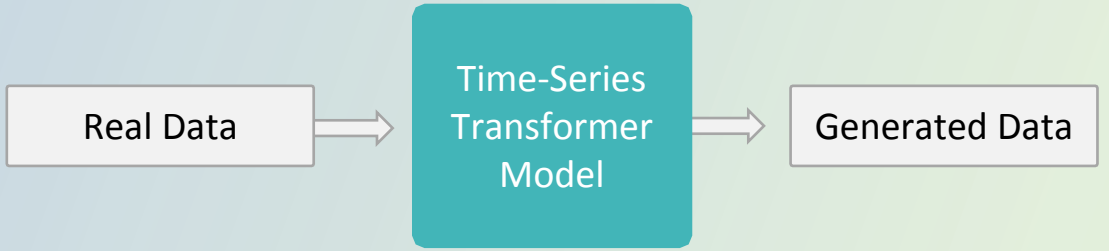


Approach



- **Why do I need to use transformer models?**
  - It is the state-of-the-art model for processing sequential data.
  - It solves the problem of capturing long-term temporal correlations.
  - It uses ***Attention*** to boost the speed.
    - No RNN cells anymore.
    - It outperforms the previous king: RNNs, LSTMs, and GRUs.
  - They don't require sequential data to be processed in order.
    - You can run the code on multi cores of the GPU.
    - It facilitates training on large datasets.
- The challenge is applying transformer models on real-valued (time-series) data.

Evaluation Plan



Training Phase

Generated Data → Predictor (MLP, Naïve Bayes, Logistic Regression)

Evaluation Phase

Real Data → Trained Predictor (MLP, Naïve Bayes, Logistic Regression) → Predicted Class

Predicted Class → **Accuracy** ← Actual Class

References

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. (2017). Attention Is All You Need.
- Charles Reiss, John Wilkes, and Joseph L Hellerstein. 2011. Google cluster-usage traces: format+ schema. Google Inc., White Paper (2011), 1–14.
- Federal Communications Commission. 2018. Raw Data - Measuring Broadband America - Seventh Report. (2018). https://www.fcc.gov/reportsresearch/ reports/measuring-broadband-america/raw-data-measuring-broadband-america-seventh.
- Google. 2018. Web Traffic Time Series Forecasting. (2018). https://www.kaggle.com/c/web-traffic-time-series-forecasting.
- Lin, Z., Jain, A., Wang, C., Fanti, G., & Sekar, V. (2020). Using GANs for Sharing Networked Time Series Data. Proceedings of the ACM Internet Measurement Conference.

Thanks for Listening!

Any Questions?