ANITA

Anonymous big data A project funded by FFG

Implementations of generative deep neural network architectures

Deliverable D5.1

Author(s): Michael Platzer, Klaudius Kalcher

Reviewer(s): Peter Eigenschink

Document version: 0.2
Date: 07.06.2021

WU Wirtschaftsuniversität Wien Vienna University of Economics and Business
MOSTLY·AI
George Labs
STATISTIK AUSTRIA Die Informationsmanager
FFG Promoting Innovation.
Federal Ministry Republic of Austria Climate Action, Environment, Energy, Mobility, Innovation and Technology
Programme **"ICT of the Future"** - the initiative of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK)

Disclaimer

This deliverable describes the work and findings of the AI-Based Privacy-Preserving Big Data Sharing for Market Research (Anonymous Big Data (ANITA)) project.

The authors of this document have made every effort to ensure that its content was accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this deliverable are responsible for any possible errors or omissions as well as for any results and actions that might occur as a result of using the content of this document.

Table of contents

1 Summary

With the results of WP4's simulation studies showing that existing GAN-based and VAE-based architectures struggle in retaining information even for non-sequential mixed-type datasets, the emphasis of WP5 shifted towards three alternative architectures: Transformers, LSTMs with Attention, Temporal Convolutions. We implemented reference implementations of these selected generative deep neural network architectures and performed hyper parameter explorations on top of Virtual Data Lab (see D4.1.).

Please note, that key findings, the actual implementations, and their corresponding research reports are project internal, and not contained within this document. This document only contains brief introductions to the model classes, as well as selected snippets from the internal research reports.

2 Researched Architectures

2.1 Transformers

Transformer is a machine learning model introduced in “Attention is all you need” (Vaswani et al., 2017), a paper that became popular due to its successful applications within natural language processing. The Transformer architecture has some key advantages compared to the sequence modeling techniques like LSTMs (Hochreiter 1997). These main advantages are (Vaswani et al. 2017):

1. It can learn long-range dependencies – all historic prior events are equidistant to predicted current event
2. It can be trained significantly faster – layer outputs can be calculated in parallel, instead of a series like with an RNN

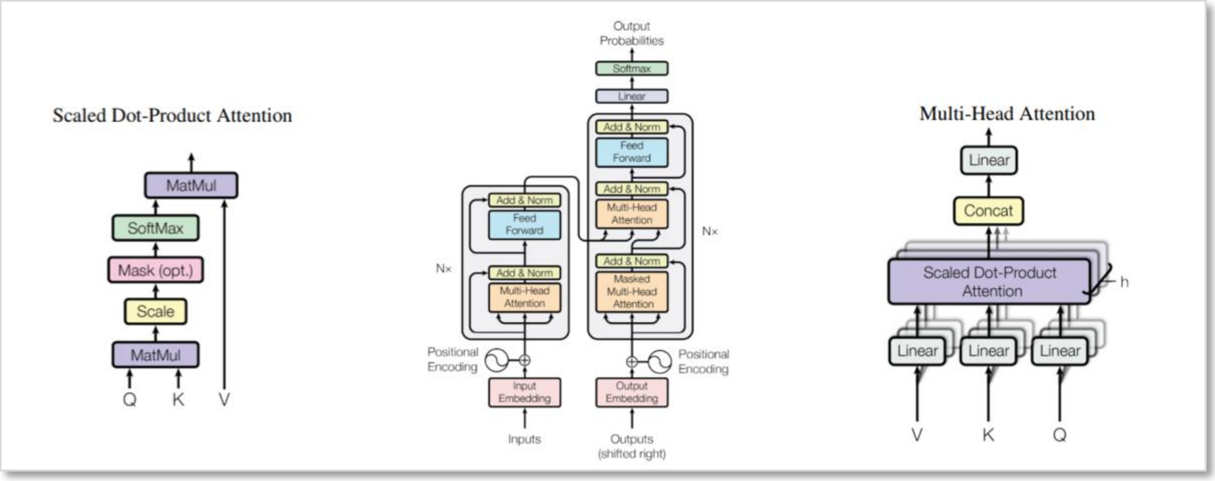


Figure 1 Key Components of Transformer Model Architecture (Vaswani et al, 2017)

For benchmarking, and hyper parameter explorations we leveraged the Virtual Data Lab. The following chart depicts a selected sample result across the four datasets of Virtual Data Lab, for a range of hyper parameter settings. As can be seen, the performance with respect to our introduced consistency metrics (see D4.2.) varies significantly across the settings, showing the sensitivity of the transformer architecture with respect to datasets and hyper parameters.

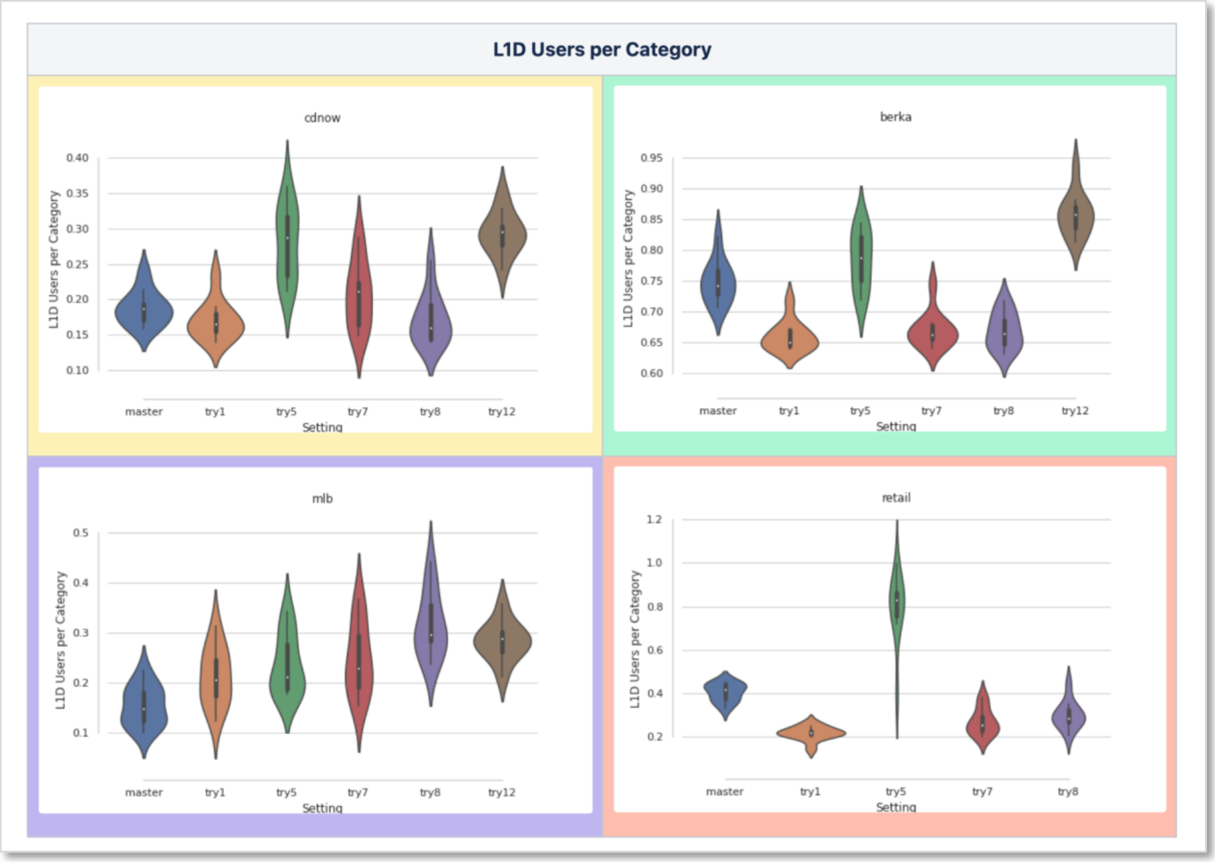
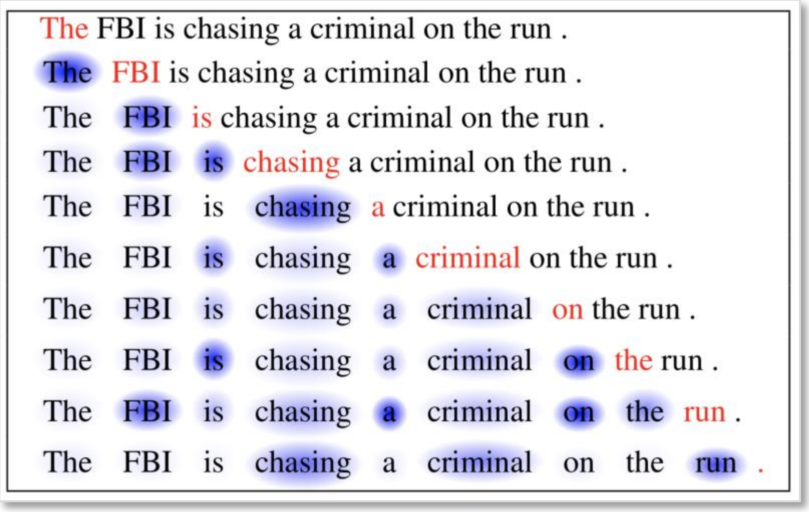


Figure 2 Sample results from internal research report. Lower L1D scores are better.

2.2 LSTM with Global Self-Attention

A second model architecture investigated was a LSTM (Hochreiter 1997) enhanced by global self-attention (Bahdanau et al. 2015). LSTMs are a popular type of Recurrent Neural Network architecture used for sequences. Internally, sequences are processed by record by record, and the output of an LSTM layer is the last hidden state. This can be expanded by applying an attention layer on top of all hidden states, which allows the model to selectively attend, i.e., to amplify or soften the effect of each record in the hidden state.



The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

The FBI is chasing a criminal on the run .

Figure 3 Illustrative example for self-attention mechanism. Prediction task is to determine the next word (highlighted in red). Attention allows us to selectively place emphasis on preceding words (highlighted in blue). [Cheng et al 2016]

Self-attention is selected because we want to attend to previous time steps. Global attention is selected because we want to attend to all parts of the sequence. And dot product is chosen as alignment function, because of the promising results combined with global attention reported in Luong et al 2015.

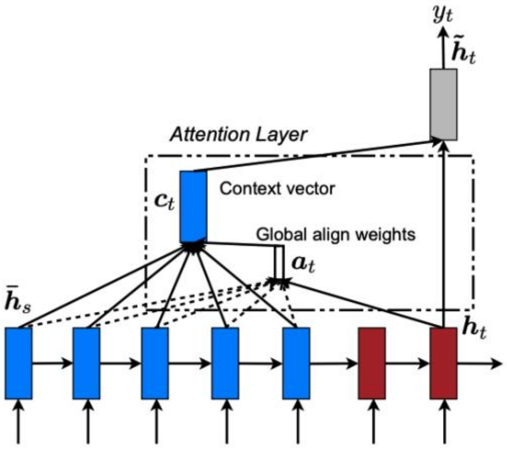


Figure 4 Architecture of a global attention layer (Luong et al 2015)

With the reference implementation in place, we benchmarked it on top of Virtual Data Lab to check for any improvements across metrics, across datasets against a vanilla LSTM implementation.

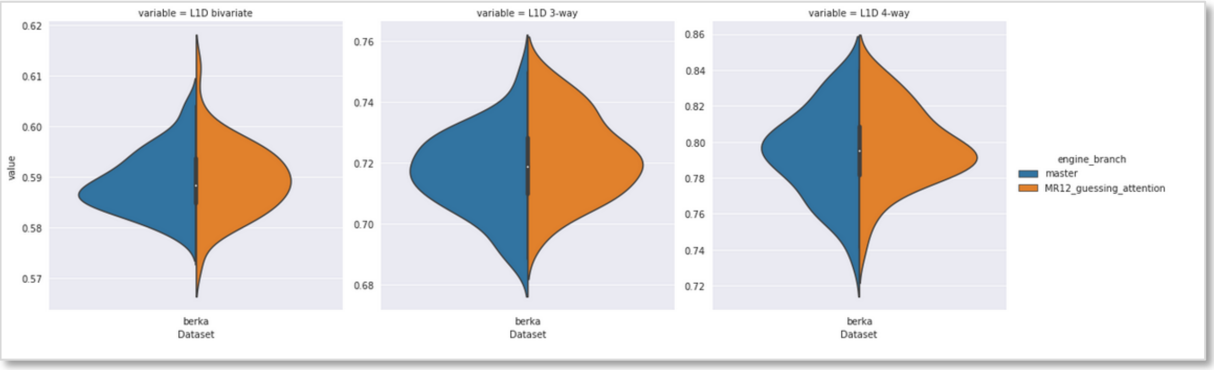


Figure 5 Sample results from internal research report. Lower L1D scores are better.

2.3 Temporal Convolutions

Thirdly, temporal convolutional networks (Bai, Kolter & Koltun 2018) were implemented, and compared to base LSTM models for sequence modeling. These consist of dilated, causal 1D convolutional layers, that have equal input and output lengths.

Advantages

- Support long as well as short term memory – increase field of perception
- Training is easily parallelizable
- Model can outperform LSTMs and RNNs in certain tasks

Disadvantages

- Inferences requires the whole sequence, which can result in memory becoming the bottleneck
- Different domains have different requirements on the history length, thus require adapting hyper parameters of receptive field.

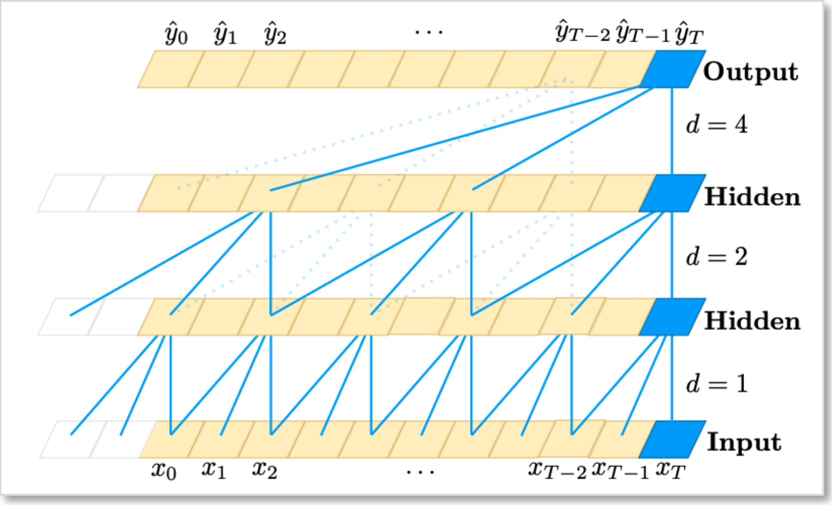

$\hat{y}_0$ $\hat{y}_1$ $\hat{y}_2$ . . . $\hat{y}_{T-2}$ $\hat{y}_{T-1}$ $\hat{y}_T$
Output
$d = 4$
Hidden
$d = 2$
Hidden
$d = 1$
Input
x_0 x_1 x_2 . . . x_{T-2} x_{T-1} x_T

Figure 5 Model architecture of Temporal Convolutional Network

The following chart depicts a selected accuracy metric across the four datasets of Virtual Data Lab across ten independent runs, for a set of 7 hyper parameter settings (c1…c7).

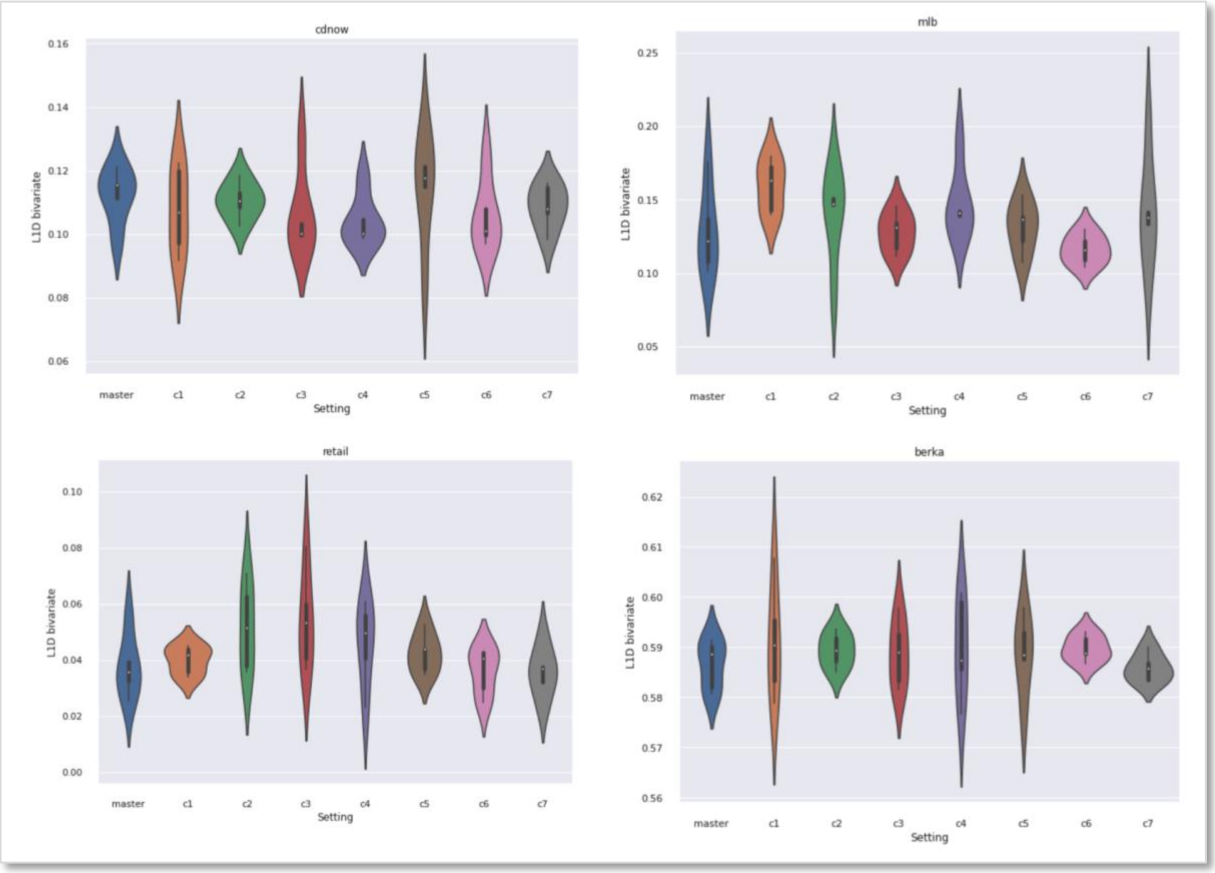


Figure 6 Sample results from internal research report. Lower L1D scores are better.