Efficient Transformers for Financial Data

Final Project of
MODELS OF
SEQUENTIAL DATA

Skoltech

Skolkovo Institute of Science and Technology

TEAM 1
Vladimir Baikalov
Kovaleva Maria
Konstantin Shlychkov
Vo Ngoc Bich Uyen

Skoltech

Skolkovo Institute of Science and Technology

PLAN

1. Introduction
2. Full Attention Model
3. Performer Model
4. Informer Model
5. Comparison
6. Conclusion

Introduction

- Financial Data - information about receipts and expenditures on a bank card

  => **predict client's gender**

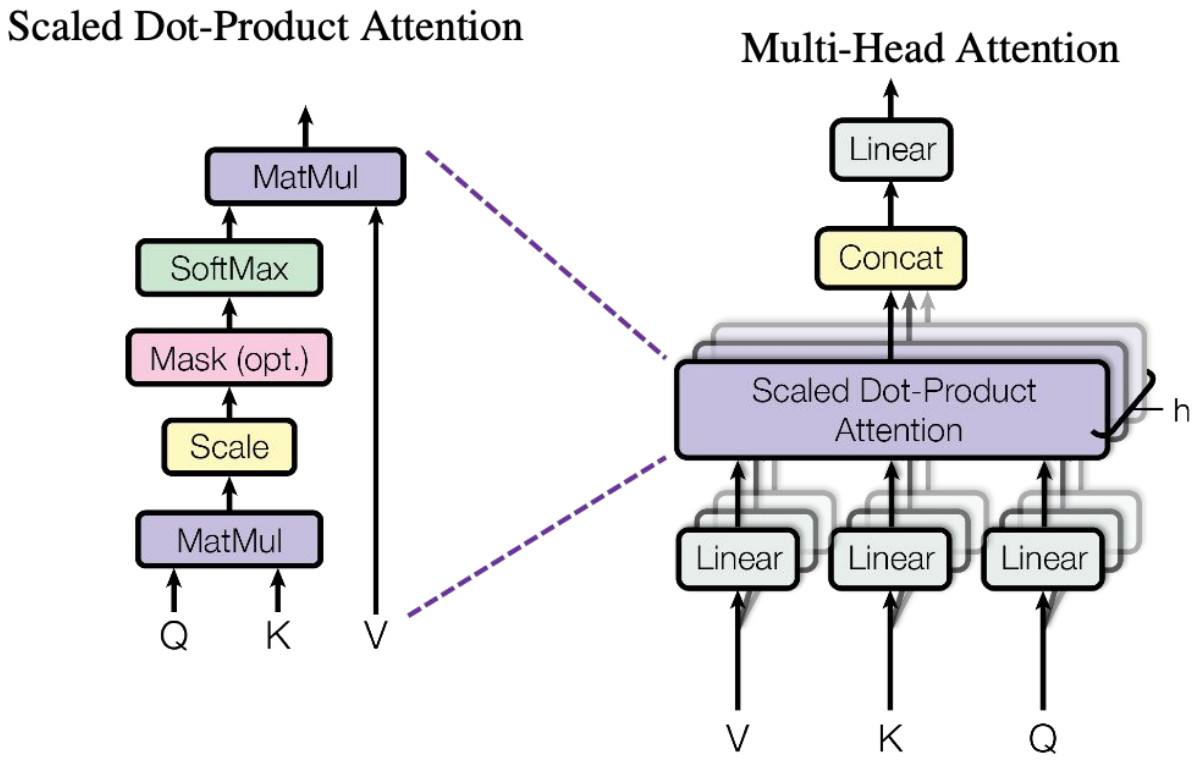
  by 3 models: **Baseline, Performer** & **Informer**

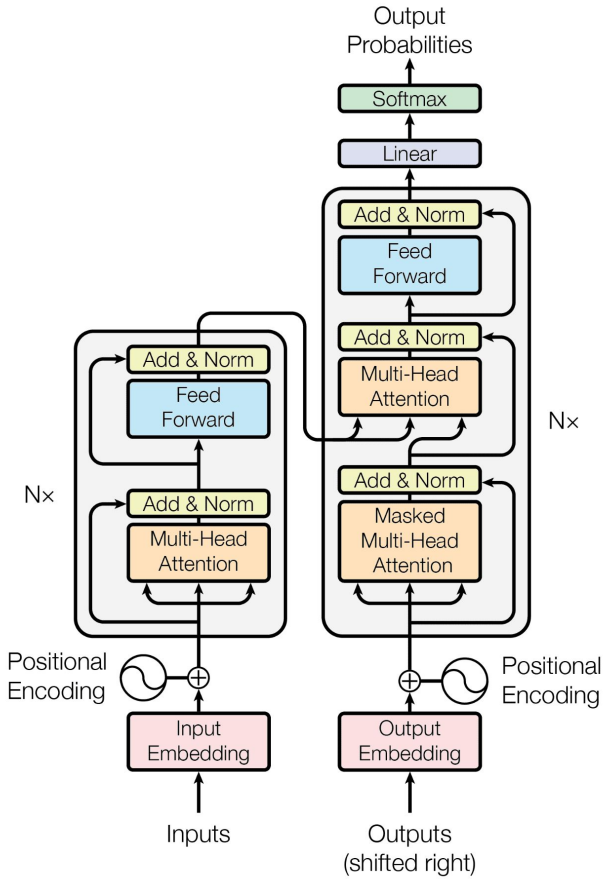
- Comparing the efficiency of them in terms of *memory* and *computation.*

Skoltech

Skolkovo Institute of Science and Technology

Full Attention Model





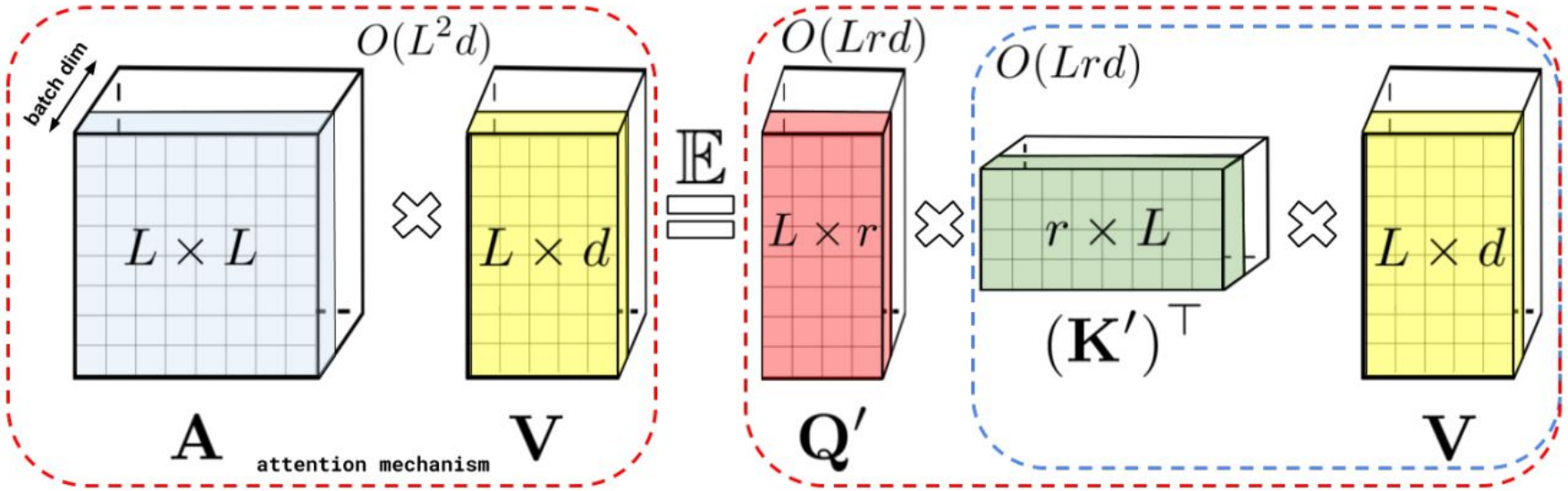
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Skoltech

Skolkovo Institute of Science and Technology

Performer Model



$$\phi(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}}(f_1(\omega_1^\top \mathbf{x}), ..., f_1(\omega_m^\top \mathbf{x}), ..., f_l(\omega_1^\top \mathbf{x}), ..., f_l(\omega_m^\top \mathbf{x})),$$

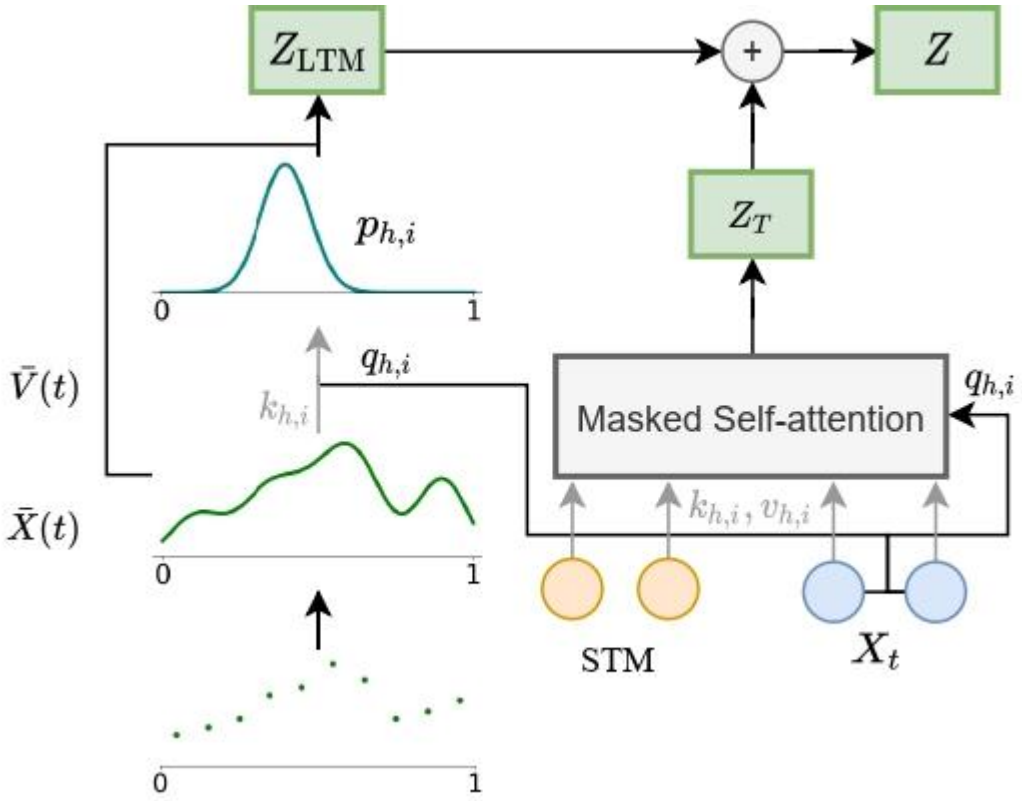
- Based on FAVOR+
- New space complexity O(Lr + Ld + rd)
- New time complexity O(Lrd)
  where L – sequence length, r – latent dimension size, d – embedding size

Skoltech
Skolkovo Institute of Science and Technology

Informer Model

Main points:

- Continuous state representation:
  - convert input sequence in the continuous signal as linear combination of N radial basis functions via regression
  - add attention: also continuous via probability distribution computing by neural network
- Advantages:
  - Complexity will be linear O(LN) rather than quadratic $O(L^2)$
  - “Constant length” N which can be smaller than original L



Skoltech

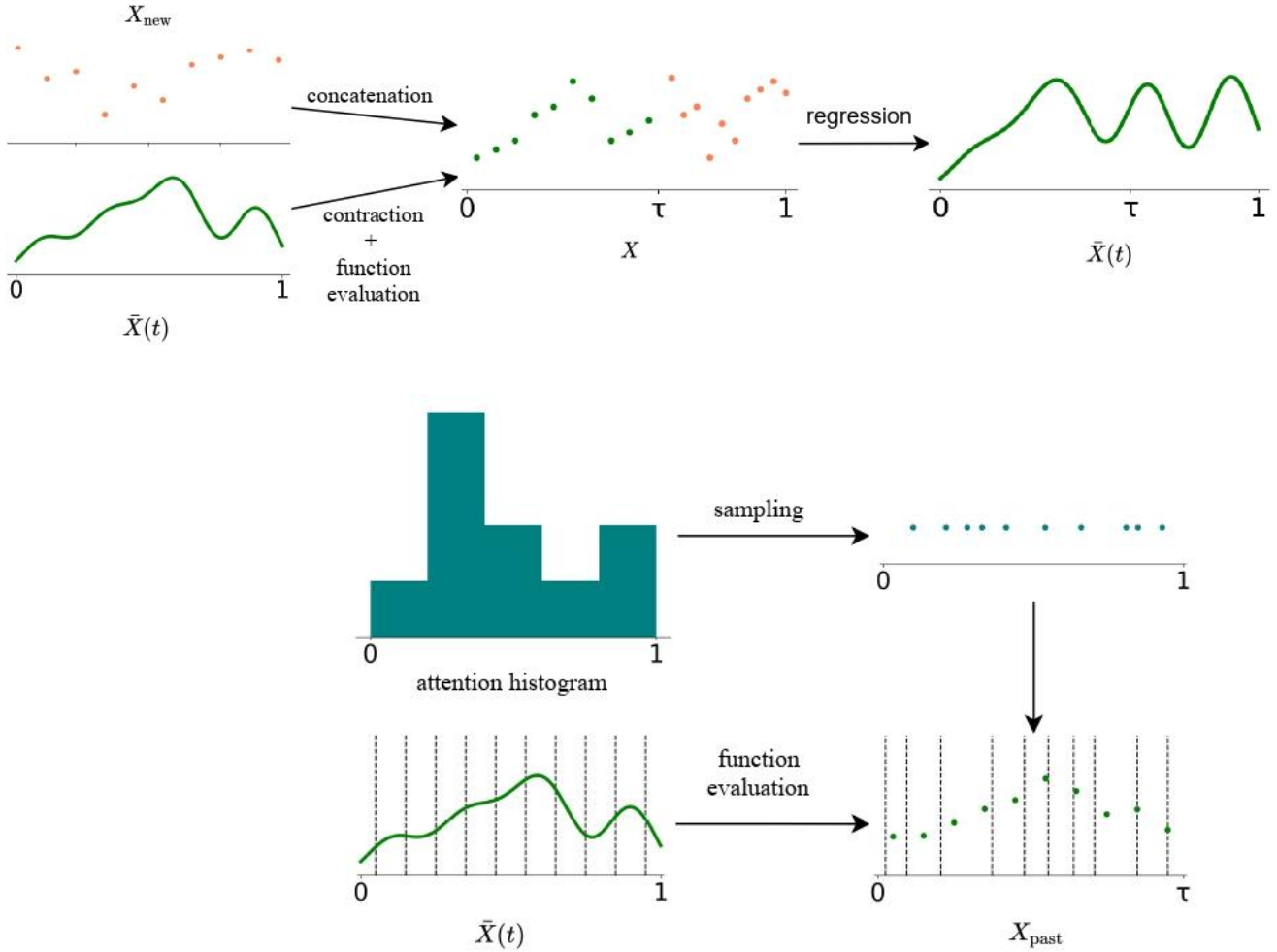
Skolkovo Institute of Science and Technology

Informer Model

Skoltech
Skolkovo Institute of Science and Technology

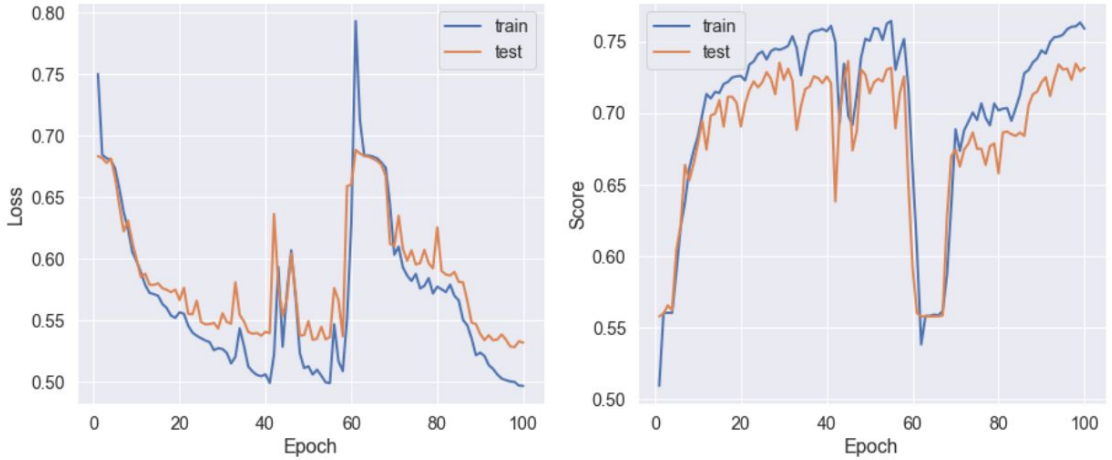
Additional points:

- unbounded memory: we can add points from previous continuous space representation
- sticky memories: sample according to previous attention

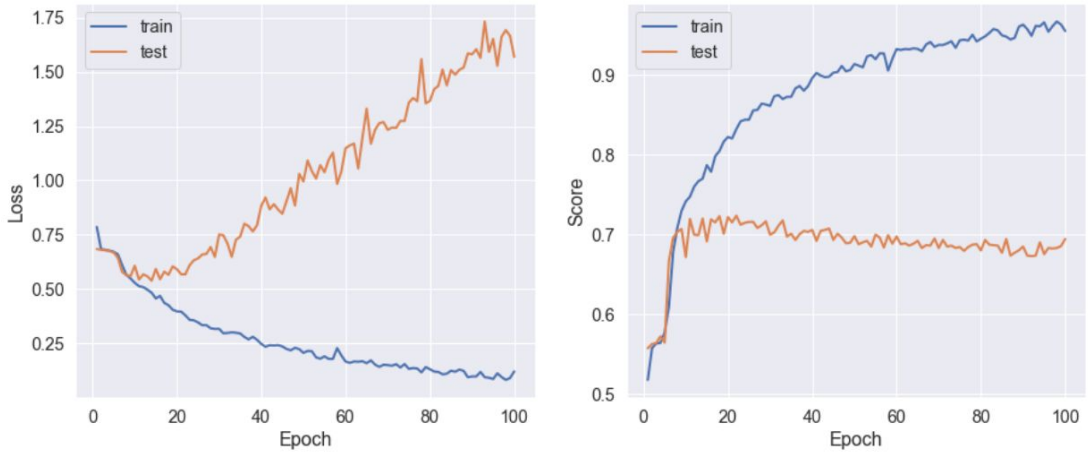


Results of training

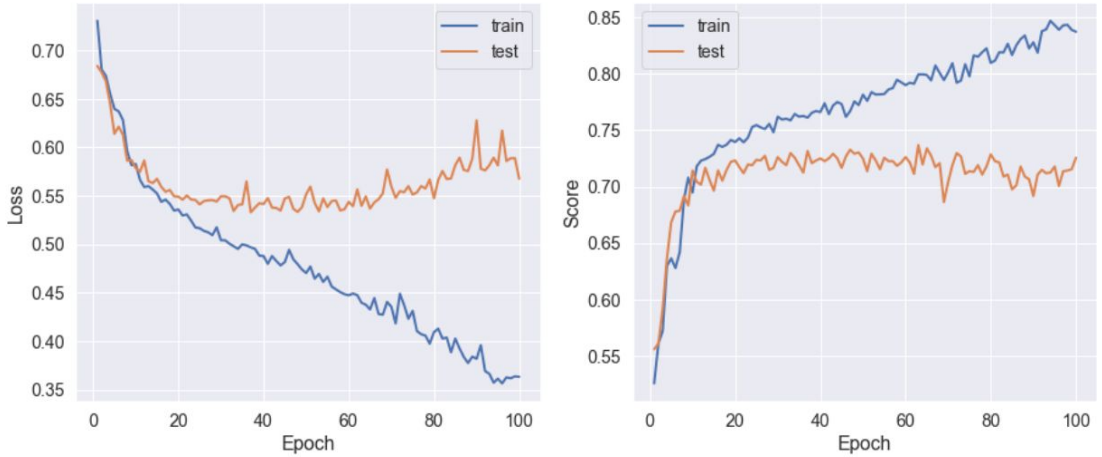
All the models achieve at least 70% accuracy score on the test data



pic.1.Performer

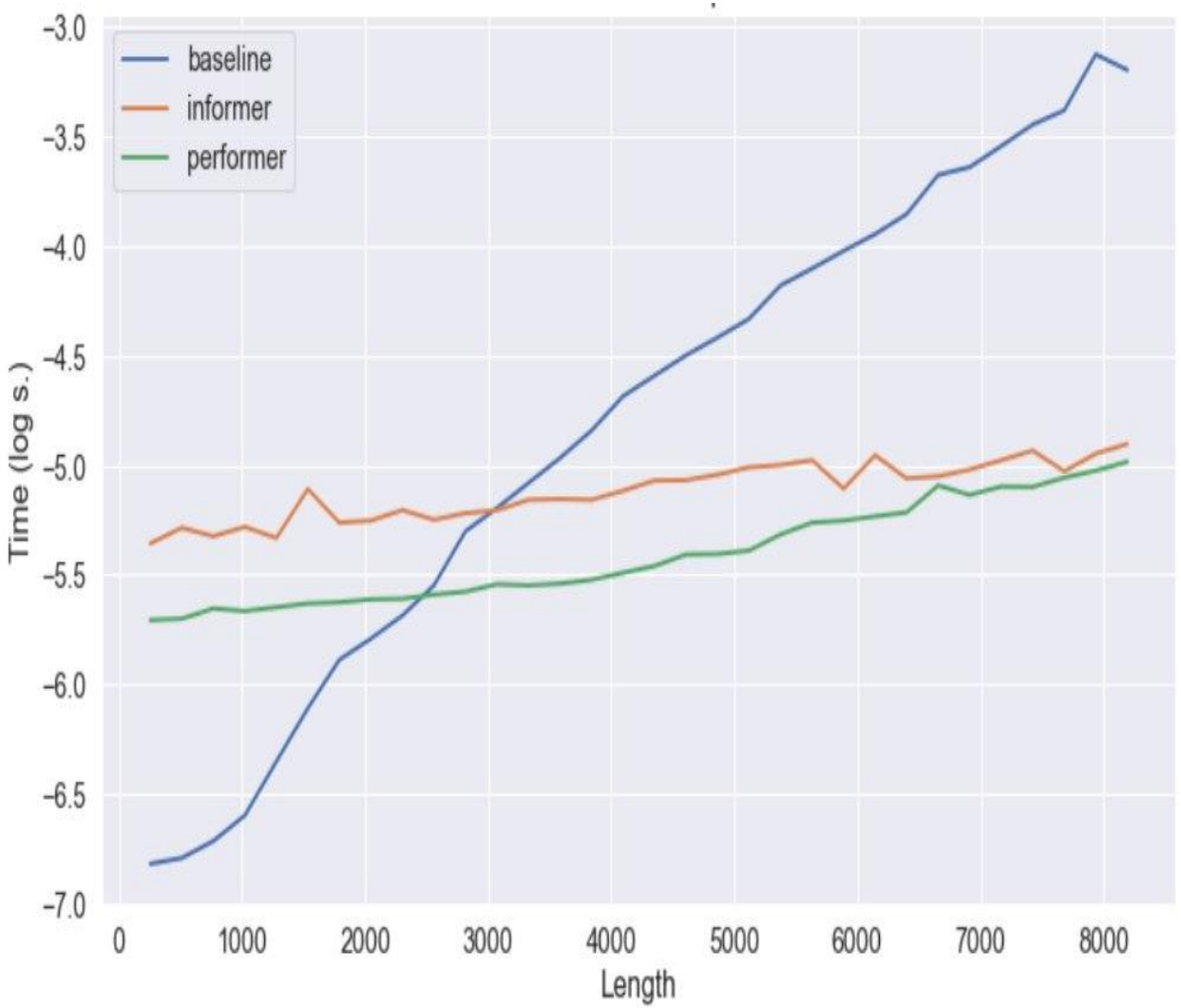


pic.2.Baseline

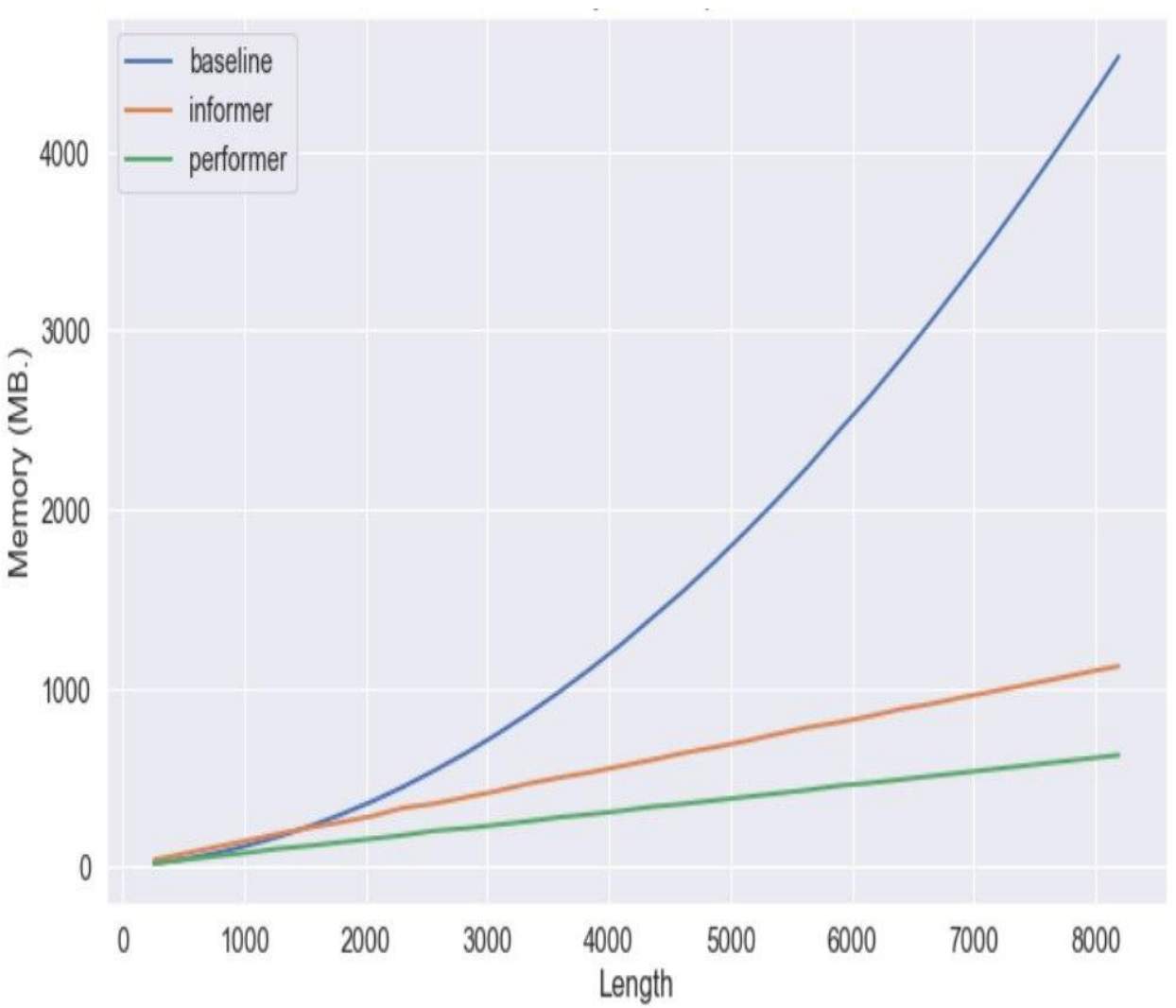


pic.3. Informer

Inference Time Comparison



Memory Consumption



Conclusion

Skoltech
Skolkovo Institute of Science and Technology

- **Vanilla transformers** - a sequence transduction model used in encoder-decoder architectures with multi-headed self-attention;
- **Informer** - a transformer extended with an unbounded long-term memory;
- **Performer** - a transformer relied on Fast Attention Via positive Orthogonal Random features mechanism.

Experiments on transaction data to determine a customer's gender

=> the Performer and Informer model <[memory and temporal complexity]< baseline model

Thx

for your attention

remember attention is all you need