Rumburská 260/15
190 00 Praha 9, Praha
+420 777 162 444
vladyorsh@gmail.com

Uladzislau Yorsh

Education

2026 (est.) **Charles University in Prague**, PhD.

2022 **Czech technical university in Prague**, Ing, with honors.

2020 **Czech technical university in Prague**, Bc.

Experience

Dec 2023–current **Data Scientist Contractor**, Cerebri AI.
- Designing new training methods for Large Language Models in the Text2SQL domain with a purpose to reduce the training and inference token count.

Apr 2021–Oct 2022 **Research Assistant**, The BigCode project.
- Proposed, implemented and evaluated two architectures for processing sequential inputs with $\mathcal{O}(n)$ complexity w.r.t. a sequence length.

May 2021–Oct 2021 **Data Science Intern**, Recombee s.r.o.
- Proposed and developed a new model for the next basket prediction task, which improved the IoU score from 0.185 to 0.203.
  - 8M interactions dataset; Tensorflow implementation.
  - Used a LSTM which incorporated a suitable "last matters more" inductive bias.

Mar 2021–May 2021 **Research Assistant**, Inference Technologies.
- Proposed and implemented an unsupervised classification algorithm of wafer bin map defects, which improved the Cohen kappa score from 0.76 to 0.81
  - Used an autoencoder to embed WBMs into a latent space, a denoising variant to make a more robust embedding for classes with defects looking similar to noise.
  - Additionally experimented with several variational and adversarial autoencoder variants.

Nov 2022–Dec 2022 **Visiting Researcher**, TU Dortmund.
- Proposed and implemented an algorithm for an automatic extension of existing ontologies with concepts mined from text based on metric learning.
- Used fastText for candidate mining and CharBERT for predicting links between a candidate and present concepts.

Publications

ICLR 2024 Tiny Paper **On Difficulties of Attention Factorization through Shared Memory**.
- Explored the problems which raise on application of the models utilizing memory to factorize attention to reduce its complexity (e.g. Luna or Set Transformer), and demonstrated that the full potential of the models is not used.
- Proposed architecture changes which led to significant performance improvements on all the considered tasks.

ICANN 2022 **Linear Self-Attention Approximation via Trainable Feedforward Kernel**.
- Proposed and implemented a new attention mechanism with a linear complexity w.r.t. an input sequence length.
- Evaluated the model on the LRA benchmark and beaten most of the baseline models.

ITAT 2022 **Text-to-Ontology Mapping via Natural Language Processing Models**.
- Explored the possibilities of an automatic assignment of an ontology to a text document.
- No other similar works were known at the moment of the publication.

Other Projects

Aug 2021–Feb 2022 **SimpleTRON: Simple Transformer with O(N) Complexity**.
- Proposed and implemented a new attention mechanism with a linear complexity w.r.t. an input sequence length.
- Evaluated the model on the LRA subset and outperformed all other models on the considered tasks at the moment of publication.

Feb 2023–Jun 2023 **Shared Task on Automatic Minuting**.
- Adopted one of the sub-quadratic Transformer variants for summarization of long texts.
- Proposed the cross-attention mechanism for the encoder-/decoder-only architecture.

Skills

Languages: Python, C, C++, Scala, Java, R

Domains: Machine Learning, Computer Vision, Signal Processing, Natural Language Processing, Data Preprocessing, Theoretical Informatics, Statistics

Technologies: PyTorch, Tensorflow, Keras, JAX, Flax, Ludwig, MATLAB, SQL, Docker, Apache Cassandra, Elasticsearch, MongoDB, Hadoop, SPARK, git, REST

Communication: English (B2), Czech (B2), German (A2), Russian (native)

Research Interests

Handling very long sequences is a challenging machine learning task I tackle. The goal is to make neural networks to be able to scale along the sequence length up to hundreds of thousands or potentially millions of tokens. As this task is highly hardware demanding, I am also keen on developing parameter-efficient models that are able to reach state-of-the-art performance using less amount of compute, and put an additional effort into learning theory to support my findings.