GDAŃSK UNIVERSITY
OF TECHNOLOGY

Subject card

Subject name and code	Optimization of Structures & Calculations in Neural Networks, PG_00054195		
Field of study	Informatics		
Date of commencement of studies	February 2024	Academic year of realisation of subject	2024/2025
Education level	second-cycle studies	Subject group	Optional subject group Subject group related to scientific research in the field of study
Mode of study	Full-time studies	Mode of delivery	at the university
Year of study	2	Language of instruction	Polish
Semester of study	3	ECTS credits	3.0
Learning profile	general academic profile	Assessment form	assessment

Conducting unit	Department of Multimedia Systems -> Faculty of Electronics, Telecommunications and Informatics	
Name and surname of lecturer (lecturers)	Subject supervisor	dr hab. inż. Piotr Szczuko
	Teachers	dr hab. inż. Piotr Szczuko

Lesson types and methods of instruction	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	15.0	0.0	15.0	15.0	0.0	45
	E-learning hours included: 0.0						

Learning activity and number of study hours	Learning activity	Participation in didactic classes included in study plan	Participation in consultation hours	Self-study	SUM
	Number of study hours	45	2.0	28.0	75

Subject objectives	The goal is to present theory, practice and problems solving in a domain of models optimisation. Techniques for structure prunning, sparsing, architecture simplification, calculations accelerations are presented. Various approaches for effective training, robustness assurance, accuracy and precision for real-world applications, e.g. in case of limited resources or noisy data.

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K7_U42] can solve engineering and research problems including design, assessment and maintenance of information systems and applications, using experimental methods and management techniques	Student creates the project, using appropriate tools, justifies the choice. Student conducts experiments and examinations, measures model accuracy. Correctly formulates conclusions based on the measured characteristics.	[SU1] Assessment of task fulfilment [SU3] Assessment of ability to use knowledge gained from the subject
[K7_W43] Knows and understands, to an increased extent, the nformal, technical and social aspects of the operation of complex information systems in the information society and in the global information n infrastructure.	Student is able to apply tools and justify the need for optimization of processing and architectures in neural networks. Student knows how chosen methods influence accuracy and performance.	[SW3] Assessment of knowledge contained in written work and projects
[K7_W06] Knows and understands, to an increased extent, the basic processes taking place in the life cycle of devices, facilities and technical systems.	Student defines goals of the project, states conclusions. Student is able to correctly justify selection of methods and tools. Student knows and can comment on theoretical aspects of the task.	[SW3] Assessment of knowledge contained in written work and projects
[K7_W41] Knows and understands, to an increased extent, the standards, production methods, life cycle and development trends of software as well as information systems and applications.	Student knows typical methods for optimization of algorithms and architectures, can apply, justify their use, formulate conclusions, estimate and predict possible results. Knows the difference between various use-cases, centralised vs. distributed, edge processing vs. server processing.	[SW3] Assessment of knowledge contained in written work and projects
[K7_U07] can apply advanced methods of process and function support, specific to the field of study	Student created a machine learning model and optimized it with respect to the model goal, model structure. Student correctly used chosen library and programming language.	[SU1] Assessment of task fulfilment [SU4] Assessment of ability to use methods and tools

Subject contents

Neural model reduction, calculations accelerations.

Quantisation, sparsification, knowledge distillation.

Noisy labels training,

Network architectures search.

Self-supervised training, pre-training.

Models uncertainty estimation (calibration, test-time dropout, ensambling, Bayes networks)

Models robustness, adversarial techniques,

Hybrid models, weight-agnostic, capsule nets.

Prerequisites
and co-requisites

Assessment methods
and criteria

Subject passing criteria	Passing threshold	Percentage of the final grade
Project	51.0%	30.0%
Colloquy	51.0%	35.0%
Laboratory	51.0%	35.0%

Recommended reading	Basic literature	Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, Alexandra Peste, (2021) Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks. [2102.00554] (arxiv.org) Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. 2020. A Survey of Model Compression and Acceleration for Deep Neural Networks. (2020). arXiv:cs.LG/1710.09282 Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural Architecture Search: A Survey. (2019). arXiv:stat.ML/1808.05377 Manish Gupta and Puneet Agrawal. 2020. Compression of Deep Learning Models for Text: A Survey. (2020). arXiv:cs.CL/2008.05221 V. Sze, Y. Chen, T. Yang, and J. S. Emer. 2017. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proc. IEEE 105, 12 (2017), 22952329. https://doi.org/10.1109/JPROC.2017.2761740
	Supplementary literature	Tensorflow model optimization (2022) https://www.tensorflow.org/model_optimization Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey. (2020). arXiv:cs.LG/2009.06732
	eResources addresses	Adresy na platformie eNauczanie:
Example issues/ example questions/ tasks being completed	• Describe and comment on one of chosen methods for optimisation, justify its use: network pruning and sparse processing, quantisation, knowledge distillation. • Justify the need for for data sparsification and architecture sparsification, and benefits of those operations. • Describe how a training on noisy labels can be efficiently performed. • Describe how the self-supervision and pre-training work. What are the benefits of these procedures. • Describe methdos for calibration of neural networks, dropout and models ensambling.. • How to estimate the model robustness? • Give an example of weight-agnostic model, and application of capsule networks.	
Work placement	Not applicable	