Augmentation-aware Self-supervised Learning with Conditioned Projector

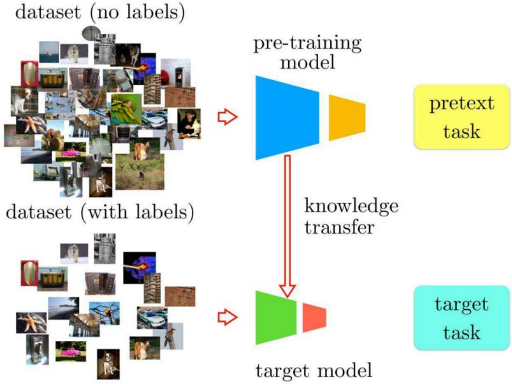
Marcin Przewięźlikowski, Mateusz Pyla, Bartosz Zieliński,
Bartłomiej Twardowski, Jacek Tabor, Marek Śmieja

ML in PL 2024

Three pillars of AI success

Problems with data

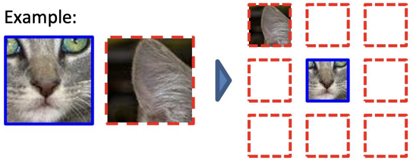
Self-supervised learning



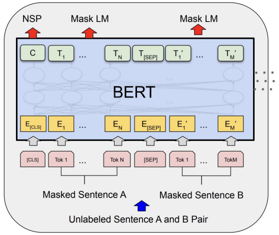
"the dark matter of intelligence" - Yann LeCun

What can serve as a pretext task?

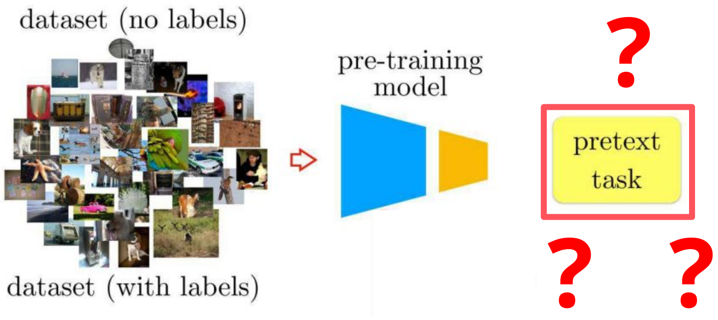
And how they did it before 2020s



Context prediction



Masked modeling
(i.e. BERT)



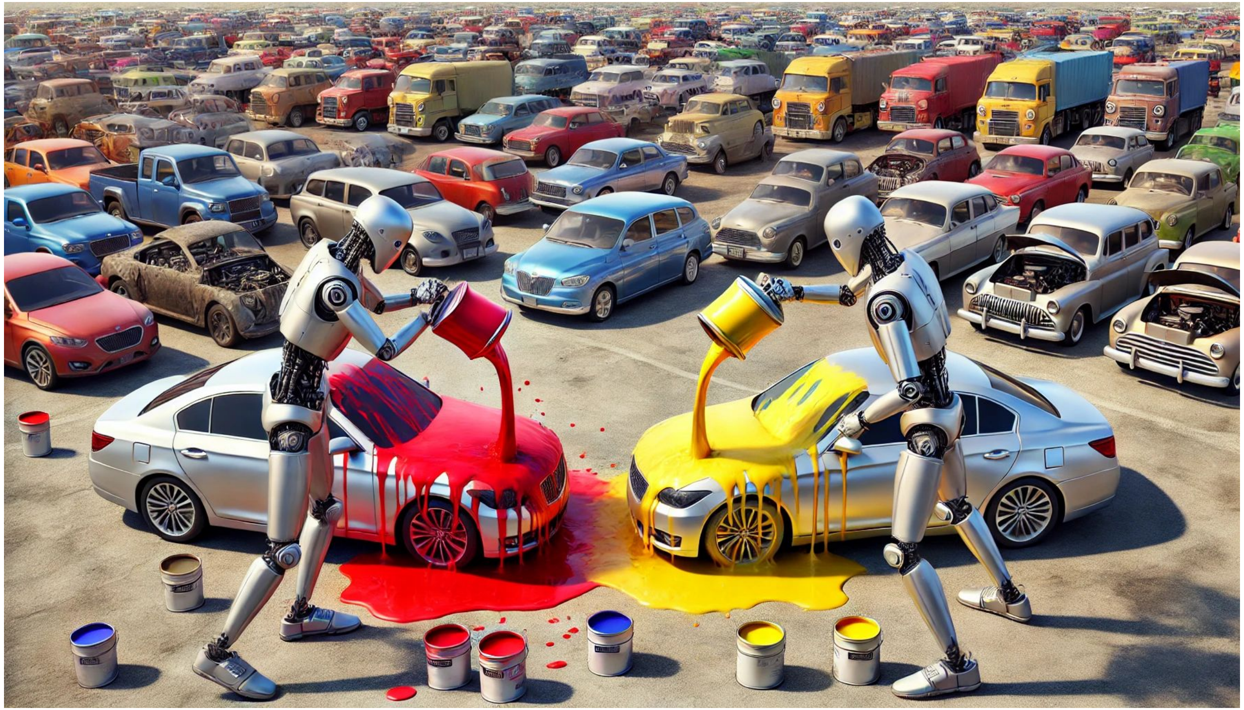
Rotation prediction

Image colorization

https://arxiv.org/pdf/1505.05192.pdf, https://richzhang.github.io/colorization/, https://arxiv.org/pdf/1803.07728.pdf,

Modern pretext tasks

~~Contrastive~~ ~~Siamese~~ Joint-Embedding models



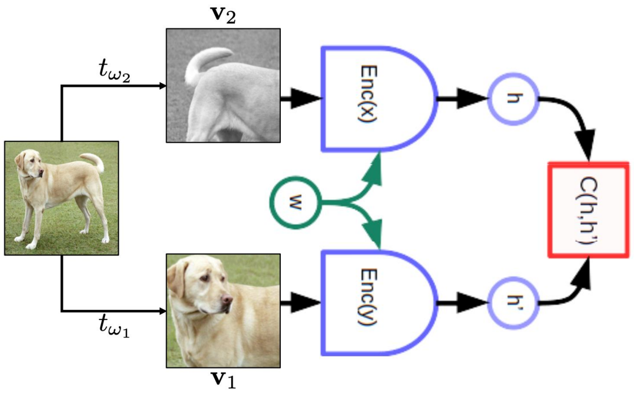
https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/ , https://arxiv.org/pdf/2111.09613.pdf

Modern pretext tasks

~~Contrastive~~ ~~Siamese~~ Joint-Embedding models

Intuition:

- augment an image in two different ways
- obtain network representations of two augmented images
- optimize the (pairwise) similarity of image representations **and** their diversity



$$\mathcal{L}_{\mathrm{SSL}} = \sum_{(\boldsymbol{x},\boldsymbol{x}')\in\mathrm{PositivePairs}} \mathrm{Distance}(f_\theta(\boldsymbol{x}), f_\theta(\boldsymbol{x}')) - \mathrm{Diversity}(\{f_\theta(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{X}\}),$$

https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/ , https://arxiv.org/pdf/2111.09613.pdf

What are the problems with Joint-Embedding SSL?

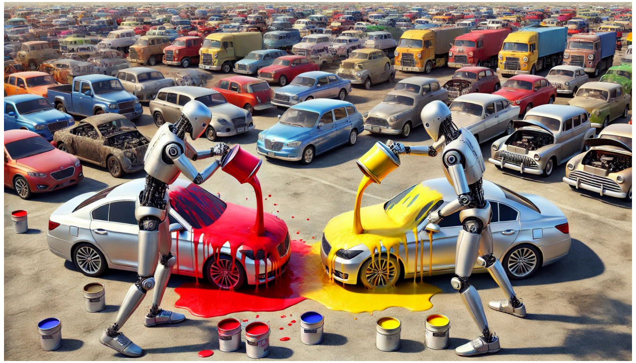
Joint-embedding SSL methods are inherently bound to augmentations

Augmentations need to be carefully selected for pretraining datasets

- solved for ImageNet, but what about other datasets?

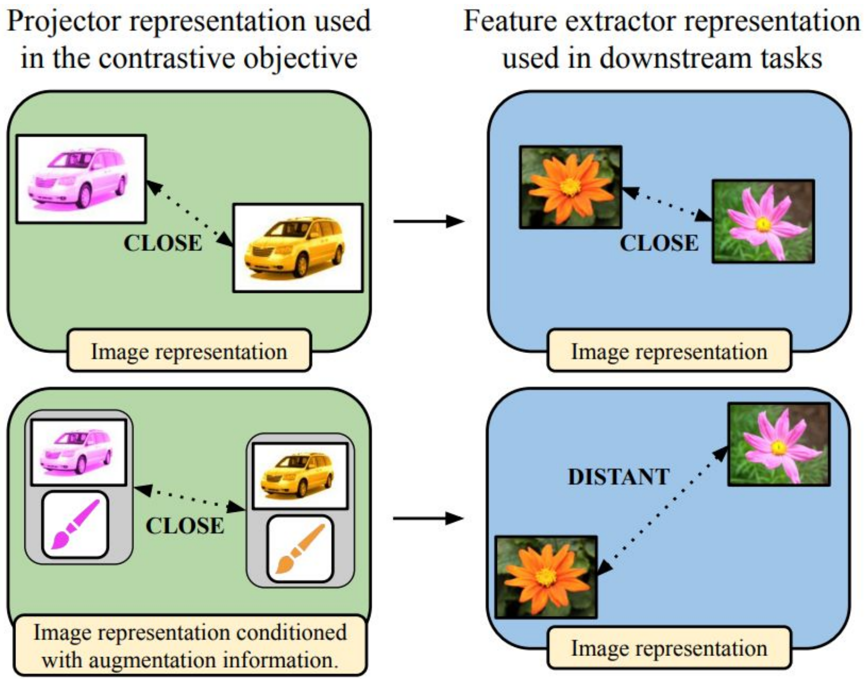
Invariance to augmentations can be detrimental for downstream tasks

- invariance to color shifts may not transfer well to flower classification



https://openreview.net/pdf?id=CZ8Y3NzuVzO

Conditional Augmentation-aware Self-Supervised Learning (CASSLE)



Previously:

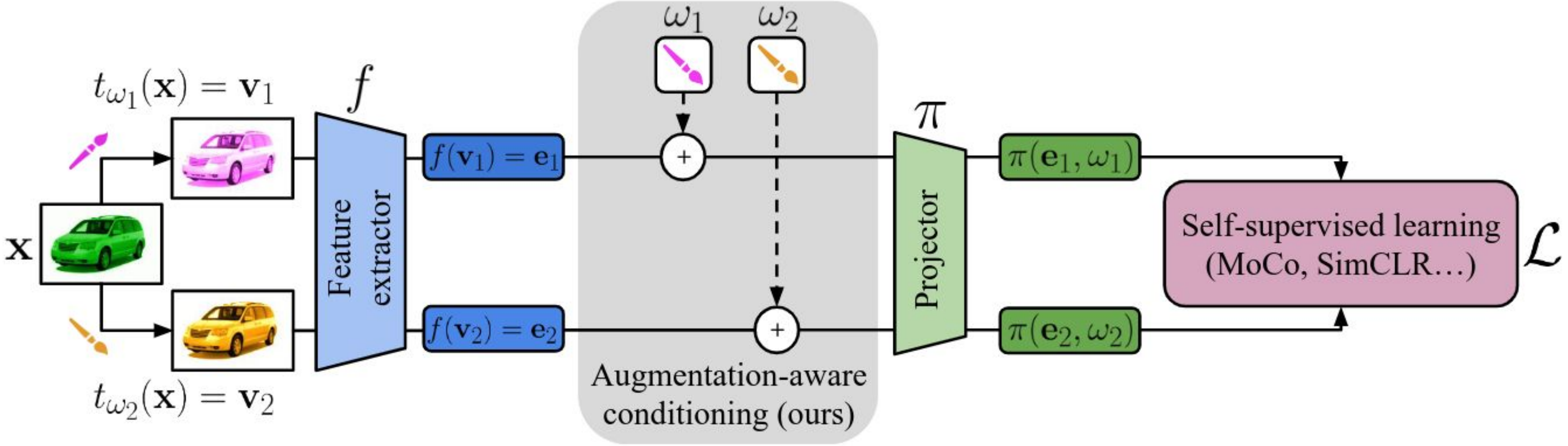
Join two image embeddings together

Now:

Join two image embeddings together ***on condition of knowing how they were augmented***

https://arxiv.org/pdf/2306.06082.pdf, joint work with M. Pyla, B. Zieliński, B. Twardowski, J. Tabor and M. Śmieja

Architecture of CASSLE



- Image and augmentation information are joined together before the projector
- Feature extractor (e.g. a ResNet) remains unaware of augmentation information
- In order for projector to act upon the knowledge of augmentations, feature extractor must learn to preserve information about features modified by them

CASSLE is applicable to all J-E architectures, regardless of their loss function.

What defines the augmentations?

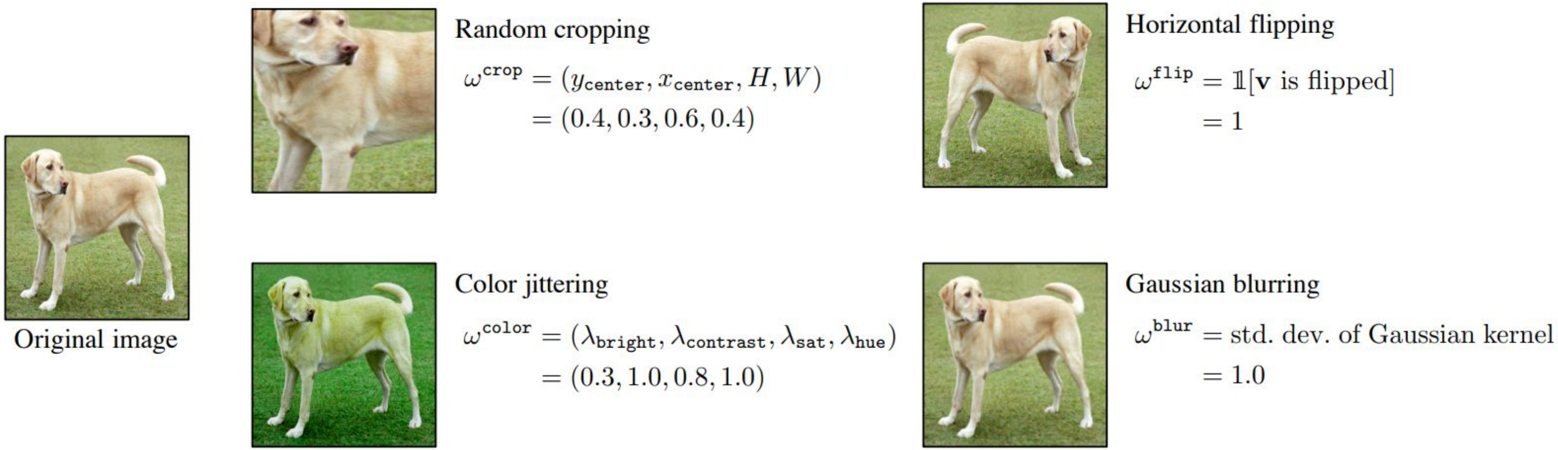
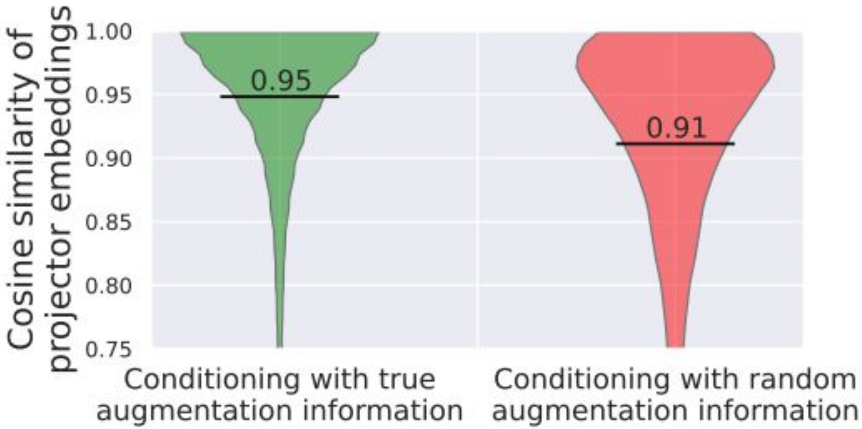


Figure 3: Examples of the commonly-used augmentations and their parameters $\omega^{\texttt{aug}}$.

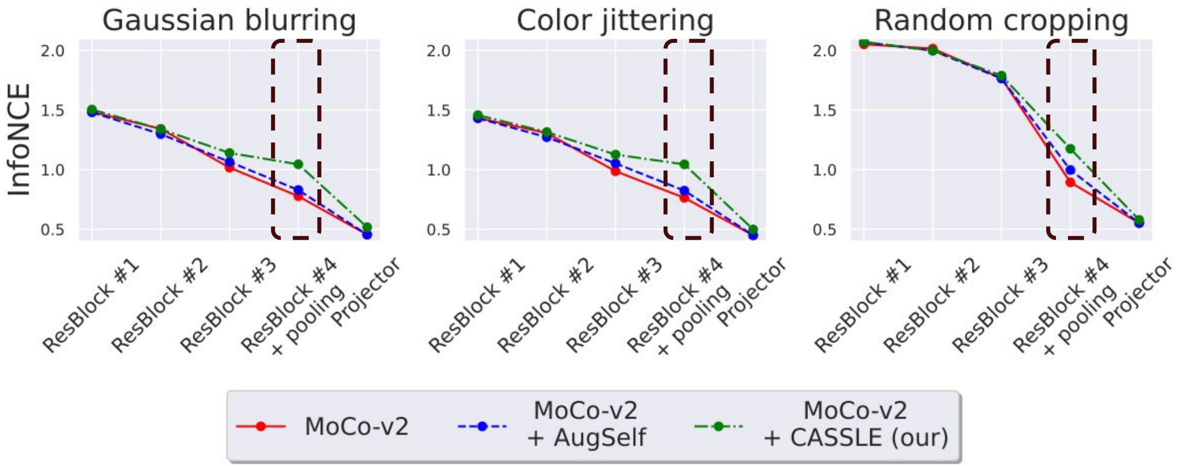
https://arxiv.org/pdf/2111.09613.pdf

Is the knowledge of augmentations useful for SSL tasks?

- We ***do not*** explicitly force the model to utilize the augmentation information.
- Conditioning the CASSLE projector with wrong augmentation information decreases its ability to draw image pairs together.
- CASSLE projector indeed relies on augmentation information to perform its task.



CASSLE learns an augmentation-aware data representation



- we measure the error of matching embeddings of augmented image pairs
- embeddings generated by CASSLE are **hardest** to match together (the highest InfoNCE value)
- CASSLE preserves the largest amount of augmentation-induced noise

CASSLE improves the transferability of SSL models

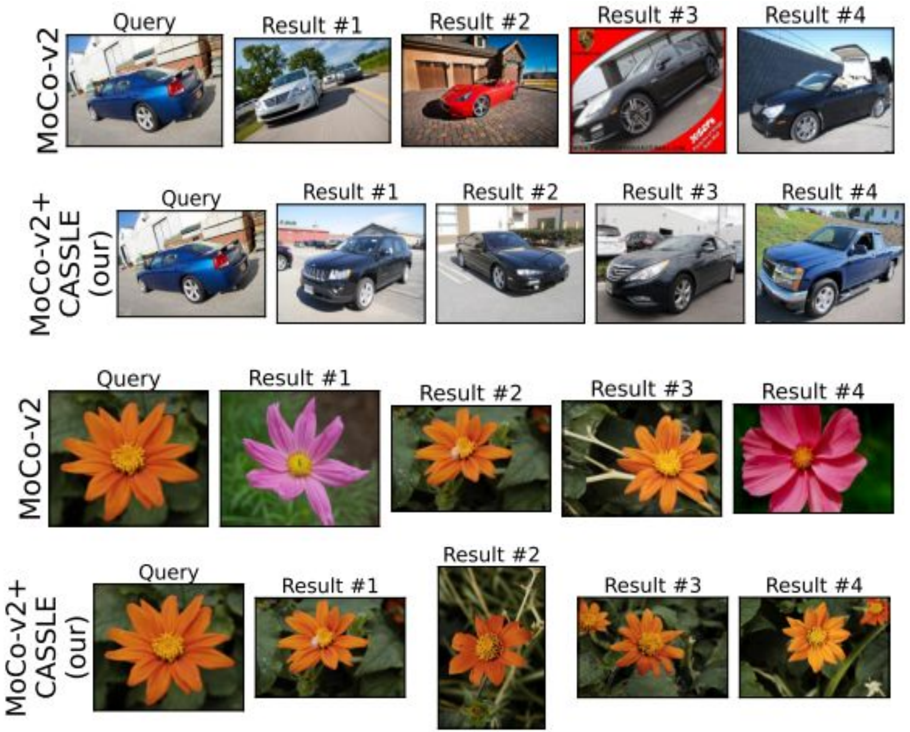
Table 1: Linear evaluation on downstream classification and regression tasks. CASSLE consistently improves representations formed by vanilla SSL approaches and performs better or comparably to other techniques of increasing sensitivity to augmentations [69, 40, 14].

Method	C10	C100	Food	MIT	Pets	Flowers	Caltech	Cars	FGVCA	DTD	SUN	CUB	300W
						SimCLR [15]							
Vanilla	84.41†	61.40	57.48†	63.10†	71.60†	83.37†	**79.67**†	35.14†	40.03†	64.90	46.92†	30.98†	88.59†
AugSelf [40]†	84.45	62.67	59.96	63.21	70.61	**85.77**	77.78	37.38	42.86	65.53	**49.18**	34.24	88.27
AI [14]	83.90	63.10	–	–	69.50	68.30	74.20	–	–	53.70	–	**38.60**	88.00
CASSLE	**86.31**	**64.36**	**60.67**	**63.96**	**72.33**	85.22	79.62	**39.86**	**43.10**	**65.96**	48.91	33.21	**88.88**
						MoCo-v2 [32, 17]							
Vanilla	84.60	61.60	59.67	61.64	70.08	82.43	77.25	33.86	41.21	64.47	46.50	32.20	88.77†
AugSelf [40]	85.26	63.90	60.78	63.36	73.46	85.70	78.93	37.35	39.47	66.22	48.52	37.00	89.49†
AI [14]	81.30	64.60	–	–	**74.00**	81.30	78.90	–	–	**68.80**	–	**41.40**	**90.00**
LooC [69]	–	–	–	–	–	–	–	–	–	–	39.60	–	–
IFM [57]†	83.36	60.22	59.86	60.60	72.99	85.73	78.77	36.54	41.05	62.34	47.48	35.90	88.92
CASSLE	**86.32**	**65.29**	**61.93**	**63.86**	72.86	**86.51**	**79.63**	**38.82**	**42.03**	66.54	**49.25**	36.22	88.93
						MoCo-v3 [19] with ViT-Small [23] pretrained on the full ImageNet dataset.							
Vanilla†	83.17	62.40	56.15	53.28	62.29	81.48	69.63	28.63	32.84	57.18	42.16	35.00	87.42
AugSelf [40]†	84.25	64.12	**58.28**	**56.12**	**63.93**	**83.13**	72.45	29.64	32.54	**60.27**	43.22	**37.16**	87.85
CASSLE	**85.13**	**64.67**	57.30	55.90	63.88	82.42	**73.53**	**30.92**	**35.91**	58.24	**43.37**	36.09	**88.53**

Table 2: Average Precision of object detection on VOC dataset [25, 42]. CASSLE extension of MoCo-v2 and SimCLR outperforms the vanilla approaches and AugSelf extension by a small margin.

Method	*MoCo-v2*	*SimCLR*
Vanilla	45.12	44.74
AugSelf [40]	45.20	44.50
CASSLE	**45.90**	**45.60**

What data is similar for SSL?



Summary

- augmentation invariance is a key component of modern Self-Supervised Learning
- it can lead to learning representations that are suboptimal for downstream tasks which rely on features of data modified by augmentations
- we propose to increase augmentation-awareness of SSL methods by conditioning them with information about used augmentations

paper

Thank you for your attention!

group of machine learning research

JAGIELLONIAN UNIVERSITY IN KRAKÓW

IDEAS NCBR

gmum.net

paper

ideas-ncbr.pl

Some pics by DALLE-3

Check out our talks during ML in PL 2024!

Friday:

Session 2 / Lecture Hall B / 10:35

Deep learning for effective analysis of high content screening
Adriana Borowa

Session 4 / Lecture Hall A / 14:30

Efficient fine-tuning of LLMs: exploring PEFT methods and LORA-XS insights
Klaudia Bałazy

Session 5 / Lecture Hall B / 14:30

Current trends in intrinsically interpretable Deep Learning
Dawid Rymarczyk

Neural rendering: the future of 3D modeling
Przemysław Spurek

JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine learning research gmum

Saturday:

Session 7 / Lecture Hall A / 12:00

AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale
Adam Pardyl

Session 8 / Lecture Hall B / 12:00

Augmentation-aware Self-supervised Learning with Conditioned Projector
Marcin Przewięźlikowski

gmum.net