A U S H A N G

FREIE UNIVERSITÄT BERLIN

Fachbereich Mathematik und Informatik

Promotionsbüro, Arnimallee 14, 14195 Berlin

D I S P U T A T I O N

Freitag, 4. April 2025, 10:00 Uhr

Ort: Seminarraum 2006

(Zuse-Institut Berlin, Takustr. 7, 14195 Berlin)

Disputation über die Doktorarbeit von

Kuba Weimann

Thema der Dissertation:

Deep Learning Under Constraints: Techniques for Data-Limited Medical and Recommender System Applications

Thema der Disputation:

Is ImageNet Worth One Video?

Die Arbeit wurde unter der Betreuung von **Prof. Dr. T. Conrad** durchgeführt.

Abstract: Large-scale vision models are typically trained on millions of static images, often sourced from datasets like ImageNet [1], which provide a diverse range of visual examples. However, collecting and curating such large datasets is expensive, as it relies on careful selection of samples to ensure diversity and often extensive manual annotation. Self-supervised learning has helped mitigate some of these costs by training models to extract meaningful representations from visual data without human annotations [2,3], making it possible to scale training far beyond labeled datasets. But how efficiently do models actually use these massive collections of images during training? Unlike traditional vision models, which learn from disconnected static images, humans absorb visual information from their environment in a continuous stream. In this talk, I explore whether deep learning models can adopt a similar approach. I will examine “Discover and Track Objects over Time” (DoRA) [4], a method that trains a strong visual model using just a single long video instead of millions of static images. I will begin with a brief introduction to deep learning and then explain how DoRA discovers and tracks objects across frames, learning rich visual representations without human annotations. Finally, I will show that training on a single long video can achieve competitive results on various downstream visual tasks, challenging the conventional reliance on massive image datasets.

[1] Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, & Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248-255).
[2] Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9650-9660).
[3] He, K., Chen, X., Xie, S., Li, Y., Dollar, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16000-16009).
[4] Venkataramanan, S., Rizve, M. N., Carreira, J., Asano, Y. M., & Avrithis, Y. (2024). Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video. In The Twelfth International Conference on Learning Representations.

Die Disputation besteht aus dem o. g. Vortrag, danach der Vorstellung der Dissertation einschließlich jeweils anschließenden Aussprachen.

Interessierte werden hiermit herzlich eingeladen

Der Vorsitzende der Promotionskommission
Prof. Dr. T. Conrad