Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship project – April/May – August

Subject

Diffusion models in infinite dimensional spaces : application to structured data generation.

Possibility to continue as a PhD candidate

YES (Funding to be confirmed)

About the chair

The Chair Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS), lead by Florence d'Alché-Buc, a Professor in the department Image, Data, Signal of Telecom Paris, unites five industrial partners: Airbus Defence & Space, Engie, Idemia, Safran et Valeo. It's general objective is to develop, in collaboration with the partners, teaching and research of the international level.

Its four principal research directions are:

1. Building predictive analytics on time series and data streams.
2. Exploiting large scale, heterogeneous, partially labeled data.
3. Machine Learning for trusted and robust decision.
4. Learning through interactions with environment.

Context and goal :

Introduced by Sohl-Dickstein et al. in 2015, denoising diffusion models provide a powerful approach to generative modeling that has now replaced GAN for many usages. While their training is based on a forward and backward process that consists in adding and removing noise, at inference time they allow to generate data from noise. Recent works on diffusion models have highlighted their links with SDE (Meng et al. 2021) as well as with gradient flows (Khrulkov et al. 2022). In another branch of the Machine Learning community, kernel methods are known to tackle nonlinear and structured data by leveraging the kernel trick, i.e. by using Reproducing Kernel Hilbert Spaces, as hypothesis spaces. In order to build a « generic » generative model for structured data, we propose to study how to extend diffusion models to Reproducing Kernel Hilbert Spaces opening the door to tackle structured data generation with the use of a decoding function. The novel approach will studied in terms of statistical properties and applied on structured data generation. The internship will take place in the Signal, Statistics and Learning group with a collaboration where both kernel methods and diffusion models are explored. The intern will rely on the experience of the group on learning in RKHS for structured prediction (see for instance Brogat-Motte et al. 2022), signal processing and more recently optimal transport (Staerman et al. 2020, Brogat-Motte et al. 2022). The internship can later give rise to a PhD thesis on modeling dynamics of structured data.

References

Luc Brogat-Motte, Alessandro Rudi, Céline Brouard, Juho Rousu, Florence d'Alché-Buc:
Vector-Valued Least-Squares Regression under Output Regularity Assumptions. J. Mach. Learn. Res. 23:

344:1-344:50 (2022)
Céline Brouard, Marie Szafranski, Florence d'Alché-Buc:Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. J. Mach. Learn. Res. 17: 176:1-176:48 (2016)

Luc Brogat-Motte, Rémi Flamary, Céline Brouard, Juho Rousu, Florence d'Alché-Buc:
Learning to Predict Graphs with Fused Gromov-Wasserstein Barycenters. ICML 2022: 2321-2335.

Hagemann, P., Ruthotto, L., Steidl, G., & Yang, N. T. (2023). Multilevel diffusion: Infinite dimensional score-based diffusion models for image generation. *arXiv preprint arXiv:2303.04772*.
Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In
H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), Advances in Neural
Information Processing Systems (NeurIPS), volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.
Khrulkov, Valentin, et al. "Understanding DDPM Latent Codes Through Optimal Transport." *The Eleventh International Conference on Learning Representations (ICLR)*. 2022.
Koo, H. , Lim, T.E. (2023). A Survey on Generative Diffusion Models for Structured Data. *arXiv preprint arXiv:2306.04139*.
Alex Lambert, Dimitri Bouche, Zoltán Szabó, Florence d'Alché-Buc:
Functional Output Regression with Infimal Convolution: Exploring the Huber and ε-insensitive Losses. ICML 2022: 11844-11867.
Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit:
Image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073,
2021. 1, 3
Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny
Burnaev. Large-scale wasserstein gradient flows. Advances in Neural Information Processing
Systems (NeurIPS), 34:15243–15256, 2021. 9
Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. ICML, 202
Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." *International conference on machine learning (ICML)*. PMLR, 2015.
Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, Florence d'Alché-Buc:
When OT meets MoM: Robust estimation of Wasserstein Distance. AISTATS 2021: 136-144.
Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems (NeurIPS), 32, 2019.

Supervisors:
Florence d'Alché-Buc, https://perso.telecom-paristech.fr/fdalche/ : florence.dalche@telecom-paris.fr
Mathieu Fontaine, https://matfontaine.github.io/, mathieu.fontaine@telecom-paris.fr
Location:
LTCI, S2A team, Télécom Paris, 19 place Marguerite Perey, 91120 Palaiseau

Candidate profile
Very motivated student wishing to spend time in the lab and having master 2 in applied maths such as MVA, Data Science …

Solid background in Statistical learning, probabilistic models, generative modeling, kernel methods

- Very Good level of programming (Python)

- Very Good command of English

Application

To send on florence.dalche@telecom-paris.fr and mathieu.fontaine@telecom-paris.fr

- Curriculum Vitae

- Personalized motivation letter that explains interest of the candidate in the subject (can be directly in the body of the email)

- Grade reports for recent years

- Contact of a person willing to give recommendation

Incomplete applications will not be considered.