FYS-3810 Special curriculum in Generative AI

Course description and objective:

Generative AI, short for Generative Artificial Intelligence, is a cutting-edge technology that uses machine learning algorithms to create or generate data, content, or information autonomously. It can produce text, images, music, and even other types of media by learning from existing examples and then generating new, often highly realistic, and contextually relevant content. Generative AI has a wide range of applications, from natural language processing and image synthesis to creative content generation and data augmentation, revolutionizing various industries and offering innovative solutions in fields such as art, entertainment, healthcare, and more.
The course will cover Large language models, Generative AI for images as well as generative AI for audio.

The goal of the course is to give hands-on experience about these new technologies. The students will practice with the new open-source tools for generative AI. At the end of the course the students will understand the concepts behind generative AI and be able to program and interact with their own generative models.

The course is for students passionate about AI and generative AI. This knowledge will be extremely valuable in the industry sector as well as in academia and research. Some emphasis will be put on creating AI tools, which could lead to possible startup ideas.

A good knowledge of programming and programming in Python is required.

Knowledge of deep learning is not mandatory but recommended.

Teaching methods:

The course will consist of lectures presenting the concepts and frameworks, together with practice sessions and hands-on programming.

Evaluation:

The evaluation will be based on a 3-weeks project at the end of the course. they will run a practical experiment in which they perform and analyze one or more methods discussed in the course. The students will deliver a report and the code about it.

Plan:

Total 14 Weeks: 11 Weeks (lectures) + 3 Weeks (project)

First attendance: Wednesday, February 7th 2024

Last attendance: Wednesday, May 15th 2024

Report deadline: Wednesday, June 12th 2024

Date	Topic	Lecturer
W6	Introduction to Generative models+ tools Git, Python, Huggingface, google colabs. Introduction https://www.youtube.com/watch?v=zjkBMFhNj_g + Innovation and startups	Benjamin Ricaud + Anna Dranovska (Innovation leader)
W7	Large language models, Practice: NanoGPT https://github.com/karpathy/nanoGPT [1]	Erland Grimstad + watch nanoGPT youtube video and comment
W8	Monday: Large language model, principles, evaluation, ethics [2,7,8]. Practice: paper presentations on evaluations and ethics. Thursday: Interacting with LLMs. Langchain and other tools	Monday: Benjamin Ricaud + Thursday: Erland Grimstad
W9	Large language models and Langchain +Retrieval Augmented Generation [3]	Erland Grimstad
W10	Large language models and vector databases, Practice: movie recommender system from text, https://github.com/weaviate-tutorials/awesome-moviate	Erland Grimstad
W11	Monday: Fine-tuning models with LoRa [4] + Thursday: Generative AI for audio part 1 [6]	Monday: Iver Martinsen Thursday: Jørgen Lund
W12	Monday: Generative AI for audio part 2 [6] + Thursday: project brainstorming and making teams	Jørgen Lund + Benjamin Ricaud
W13 & W14	Easter break	
W15	Generative AI for images Stable diffusion [5] (introduction, VAE, CLIP) CLIP:	Theory: Kristoffer Wickstrøm

	https://dataflowr.github.io/website/modules/19-clip/	
W16	Monday: Theory behind diffusion models + programming the vanilla diffusion model from scratch on Pytorch Tutorial 2: introduction to the Diffusers library from Huggingface to implement advances SOTA diffusion models	Filippo Bianchi
W17	Monday: Generative AI for videos, Thursday: Reasoning and AGI	Monday: Kristoffer Wickstrøm + Thursday : Benjamin Ricaud
W18	Monday: Project work Thursday visit to Tvibit and generative image practice with Comfi.ui	Benjamin + Tvibit

References:

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

[4] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).

[5] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

[6] Copet, Jade, et al. "Simple and Controllable Music Generation." *arXiv preprint arXiv:2306.05284* (2023).

[7] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., ... & Xie, X. (2023). A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

[8] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).