Fraunhofer IOSB

Interaktive Echtzeitsysteme

Bachelor/Master Thesis:
Gaussian Mixture Compression

Topic
Model compression for Gaussian mixture is compelling for several reasons. First, expectation maximization is nonconvex, often requiring multiple random restarts; compressing a well converged model preserves its hard won optimum and avoids repeated runs. Second, compression without retraining is a major advantage, delivering smaller footprints and faster inference while keeping the learned distribution intact. Third, maintaining multiple storage and compute tiers of the same model—full, medium, and ultra-light—mirrors the ChatGPT-4 and 4-mini pattern: a unified capability surface scaled for latency and cost. This tiering enables adaptive deployment, edge compatibility, and efficient A/B testing without duplicating training pipelines and simplifies fleet management.

Path
Our main goal is to develop an algorithm that can compress a model into a smaller model with minimal loss of information give a number of components.

Prerequisite
There are no hard constraints but the more programming and math you know the more you can have fun while doing the project.

What I offer

- A teammate/supervisor who is actually present.
- Possibility to be a co-author in a research level publication.
- A BSc/MSc thesis project that will be used in production level software for an enterprise level project.
- I can probably provide you with an office.
- Nice private IT infrastructure to implement whatever wild ideas you have in mind.

Contact:
Ali Darijani

ali.darijani@iosb.fraunhofer.de
ali.darijani@kit.edu