Deep Learning Engineer for Quantization Aware Training (m/f/d)

Do you get excited by working on cutting-edge computer vision and NLP models to bring Artificial Intelligence applications to the next level? We are expanding our Benelux team with a Deep Learning Engineer to optimize the latest deep neural networks for our state-of-the-art neural-network accelerator chip.

Most Deep Learning Engineers work with pretrained models and few still train their own. However, Deep Learning builds on *Learning*, hence training. Yet good training demands a deep insight in your model and an intuition about what methods facilitate learning. Quantization-aware-training allows the neural network to compensate for the information loss caused by quantization and can be vital for optimally deploying networks. Are you the expert in this Deep Learning field? Then this is THE job for you!

Your role

Your role will mainly include the following:

- Understanding and implementing the latest deep learning networks (Pytorch, Tensorflow, etc.)
- Optimizing these networks so they run efficiently on our chip using state-of-the-art quantization-aware-training algorithms
- Keeping track of the deep learning literature on training algorithms, novel models and quantization
- Working with clusters of GPUs
- Communicating your results to team members

Your profile

- At least 3 years of experience with Python (Numpy, Scipy)
- At least 3 years of experience with a deep learning framework like Pytorch or Tensorflow
- Strong understanding of the inner-workings and applications of recent CNNs, MLPs, RNNs and transformers
- Preferably experience with quantizing neural networks and network optimizations for ASICs. Having previous experience with quantization-aware-training is clearly a plus
- Preferably knowledgeable about distributed multi-GPU training
- Problem-solving mindset, capable of debugging and patching software issues
- Good oral and written communication skills
- Fluent in English, both in speaking and writing
- You are a team player and you are also able to autonomously plan and perform research tasks
- You have a strong sense of responsibility and want to realize high ambitions

Who we are

Axelera AI is a truly European deep-tech Start Up company which is developing a game-changing hardware and software platform for AI at the edge that will make the industry more integrated, efficient and accessible. Our mission is to spread artificial intelligence for a green, fair, trusted and safe world enabling new application of AI in diverse sectors like smart cities, retail and other markets. Our company is a spin-off from a multinational deep-tech group and is backed by a strong syndicate of institutional investors. We have an extraordinary and international team of top talented researchers and developers working in the headquarter in Eindhoven (NL) and in the branch offices in Leuven (BE), Zurich (CH) and Pisa (IT).

What we offer

Take the chance to become part of a dynamic, fast-growing, international organization. We offer an attractive compensation package, including a pension plan, extensive employee insurances and the option to get company shares.

An open culture that not only supports creativity and continual innovation is awaiting you. Collaborative ownership and freedom with responsibility is characteristic for the way we act and work as a team.

Interested?

Great! We are looking forward to receiving your application! Feel free to contact us whenever you have any questions or for further information @Heike Wilfling, HR Business Partner +31621929159. For further information on Axelera AI please also have a look on our website: www.axelera.ai

Do you want to stay informed about our journey and our current vacancies? Follow us on Linkedin