Master's Thesis Opportunity

Adaptive Computation through Nested Model Architectures

Deep learning models have expanded dramatically in recent years, often incorporating billions of parameters. Yet, empirical evidence demonstrates diminishing accuracy gains when simply scaling models larger. This indicates that not every input requires the full computational capacity of these massive models; instead, different inputs demand varying levels of complexity.

To address this challenge, there is an increasing need for models that dynamically adjust their computational resources based on input complexity. This research explores nested model architectures, which allow distinct subsets of a model to operate independently. Such architectures facilitate adaptive computation tailored to specific input demands.

Recent innovations, such as MatFormer, have demonstrated the practicality of nested transformer architectures. MatFormer allows extraction of multiple high-performing submodels without additional training overhead, enabling elastic inference across various deployment constraints and has notably been integrated into Google's Gemma 3n model. Similarly, LayerSkip utilizes early-exit inference and self-speculative decoding to reduce inference latency effectively, without sacrificing accuracy.

Building on these foundational advances, this research aims to further optimize nested model architectures. The primary focus is on enhancing methods for dynamically allocating computational resources, thereby achieving superior efficiency and scalability in diverse deep learning applications.

Industrial Supervisor: **Dr. Mostafa Elhoushi**, Cerebras (formerly FAIR at Meta and Huawei).

If you are interested in this opportunity, please email **Prof. Dr. Amr Alanwar** at alanwar@tum.de with your transcript and CV for consideration.