Internship – Dynamic Inference in CNNs via Mixture of Experts and Early Exits

Context

The deployment of deep neural networks on edge devices such as smartphones or embedded systems poses significant challenges in terms of computational cost, energy consumption, and latency. Traditional models process all inputs with the same fixed architecture, regardless of their complexity, leading to inefficient use of resources. For instance, a clear image of a cat is processed with the same depth and width as a noisy, ambiguous scene, despite requiring far less computation.

To address this, recent research has explored dynamic neural networks, which adapt their computation based on input content. Two prominent strategies include:

- Mixture of Experts (MoE): activating only a subset of network parameters (e.g., specific convolutional branches) per input, enabling width adaptation.
- Early Exits: allowing simpler inputs to exit the network at intermediate layers, reducing inference depth.

While these approaches have been studied independently—MoE primarily in Transformers and early exits in CNNs—their combination remains underexplored, especially in convolutional architectures. Jointly leveraging both mechanisms could enable dual adaptation in width and depth, significantly improving efficiency without sacrificing accuracy.

This internship aims to design, implement, and evaluate a dynamic CNN architecture that integrates Mixture of Experts blocks with confidence-based early exits, enabling input-adaptive inference for vision tasks such as image classification. The work will contribute to the growing field of efficient and sustainable AI, with potential applications in mobile vision and real-time systems.

Objectives

The main goal is to design and validate a hybrid dynamic CNN that couples conditional activation (Mixture-of-Experts) with adaptive depth (early-exit). To reach this goal, the intern will first carry out a bibliographic survey on dynamic inference, covering MoE in CNNs, early-exit networks such as BranchyNet, and recent attempts at joint width-and-depth adaptation; key training difficulties—load balancing, confidence estimation, stability—will be identified. Next, a full architecture will be proposed: convolutional MoE blocks whose top-k gating network selects the most relevant experts for each input, and auxiliary classifiers inserted at several depths that can terminate inference as soon as a confidence threshold is exceeded; a single decision rule will be learnt that decides, at every stage, whether to route or to exit. The model will then be implemented in PyTorch on standard backbones (ResNet or VGG variants) and trained on CIFAR-10/100 or Tiny-ImageNet; knowledge distillation and load-balancing losses will be used to stabilise MoE training, while a cost-aware term will encourage both sparse expert selection and early termination. Finally, the system will be evaluated in terms of accuracy, average inference depth, FLOPs and latency and compared against strong baselines (standard CNN, BranchyNet, MoE-CNN without exits); a detailed analysis will correlate input difficulty with the chosen experts and the actual exit layer. If time permits, the intern will explore ultra-lightweight gating for on-device deployment and validate the approach on a mobile-oriented use-case.

Required Technical Skills

- Python, PyTorch
- Deep learning (CNNs, optimization)
- Familiarity with vision datasets and evaluation metrics
- (Preferred) Experience with model compression or dynamic networks

Supervision

The internship will be conducted at the LIRIS lab (www.liris.cnrs.fr) at INSA Lyon, France, and supervised by Stefan Duffner, IMAGINE team.

Contact

stefan.duffner@insa-lyon.fr