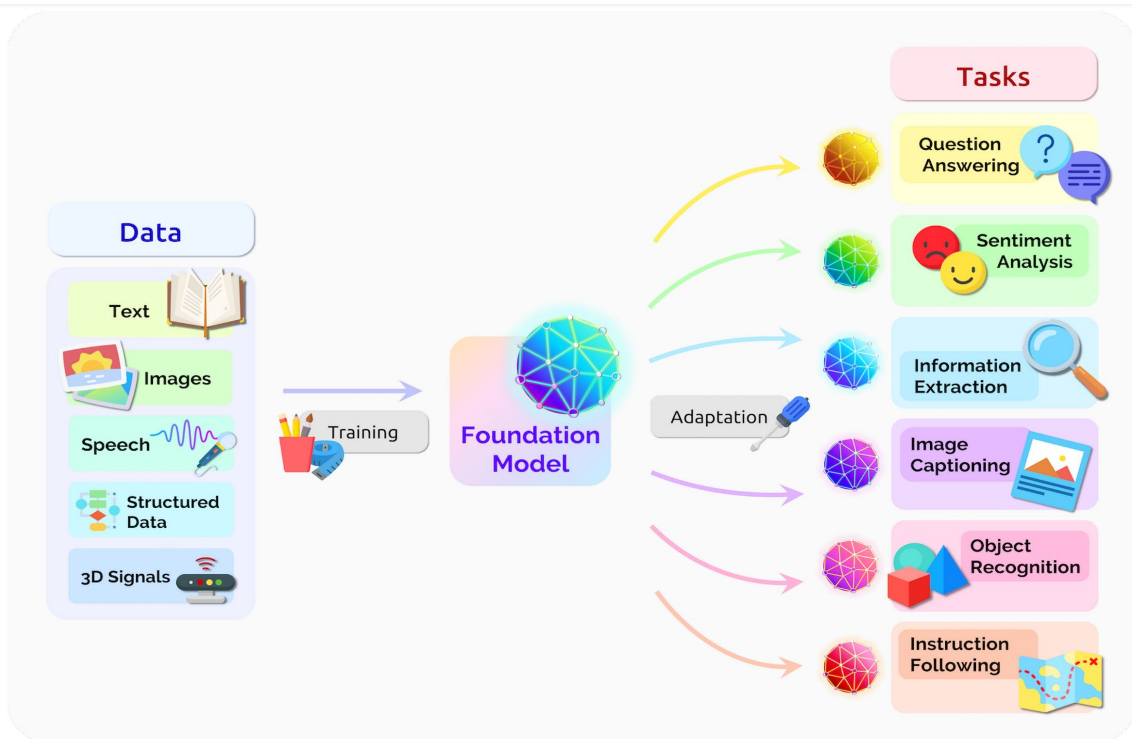
AdaInf: Adaptive Inference for Resource-Constrained Foundation Models

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

ICML
International Conference On Machine Learning

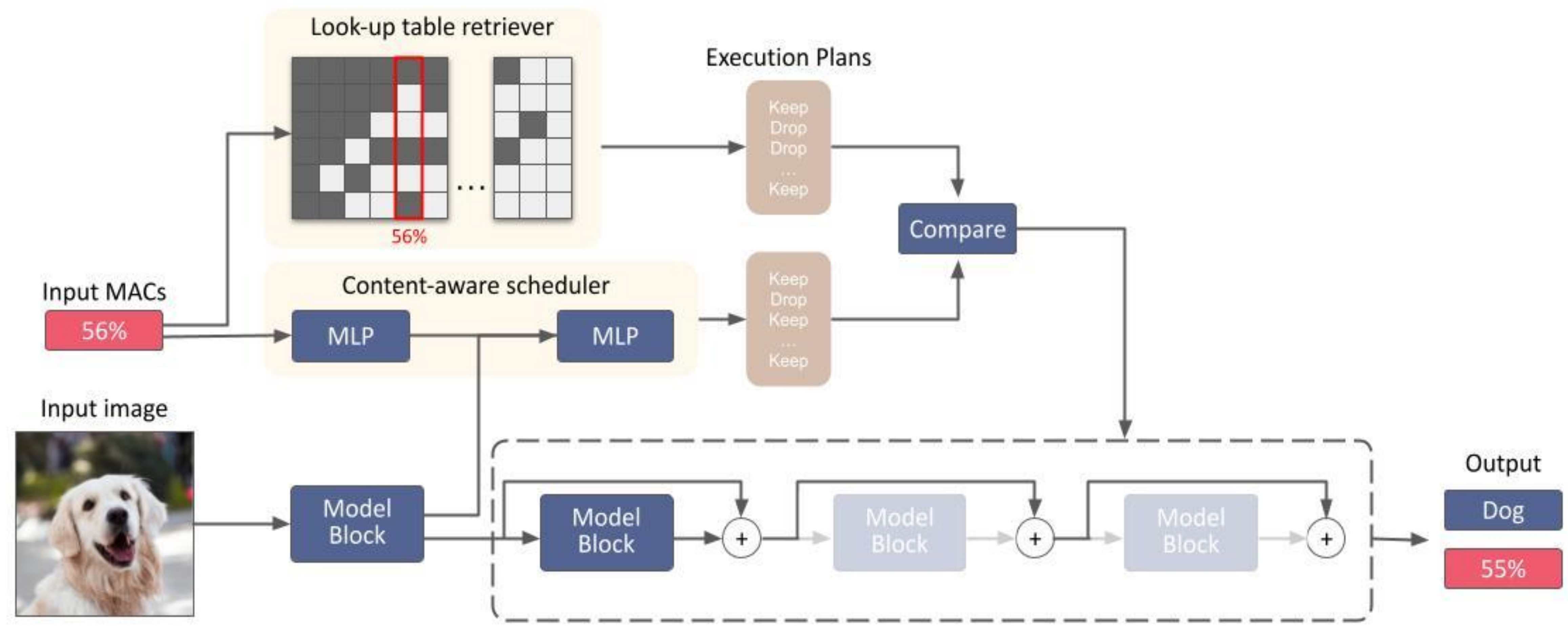
Zhuoyan Xu, Khoi Duc Nguyen, Preeti Mukherjee, Somali Chaterji, Yingyu Liang, Yin Li

Motivation



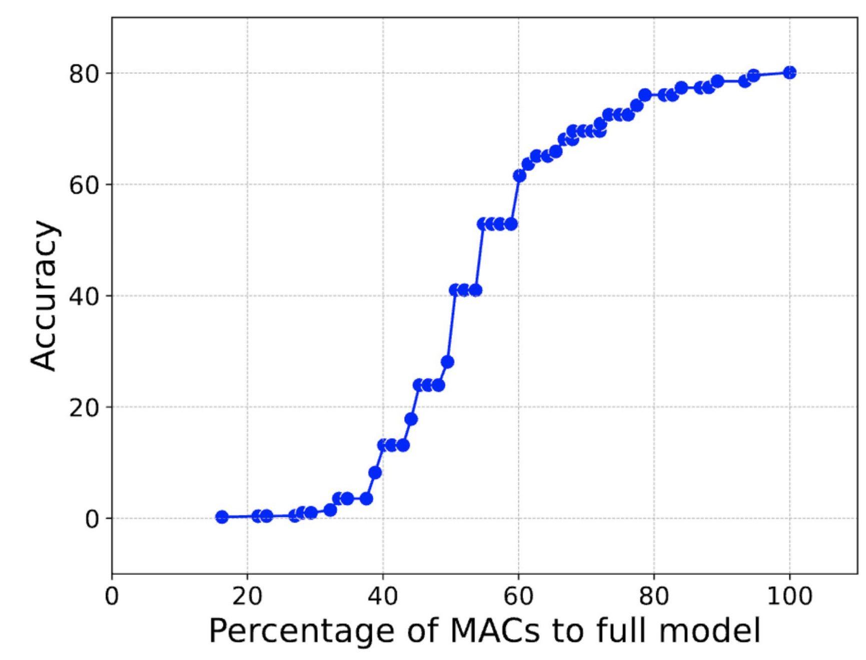
Foundation Model

Source "On the opportunities and risks of foundation models." (2021)



Build-in redundancy of pretrained RedNet50

Multiple execution branches of *ResNet50* pretrained on *ImageNet*. Each point refers a branch.



Take-Home Message

We propose **AdaInf**—an adaptive inference framework that dynamically allocates and executes different parts of foundation models to reduce computation costs.

Key Intuition

- Existing large pretrained models has built-in redundancy, since modern training techniques adopt aggressive regularization (e.g., stochastic depth). Such redundancy allows us to treat a model as a collection of execution branches.
- Different execution branches can be tailored for runtime conditions, thereby achieving adaptive inference.

Contribution

- Our framework **AdaInf** learns a scheduler to decide on the branch to execute, based on a compute budget as well as the input data.
- We conduct preliminary experiments on CIFAR and ImageNet using pre-trained ResNet and CLIP models. We show AdaInf can achieve varying accuracy and latency trade-offs in response to the input data and the latency budget, outperforming baselines.

Problem Formulation

- Foundation model: f_θ
- light-weighted scheduler: $g_\beta(\cdot,\cdot)$
- Given latency requirement M, have execution plan $g_\beta(x, M)$
- Prediction $\widehat{f}(x,p)$, actual latency $\widehat{M}(x,p)$
- Loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(y, \widehat{f}(x,p)) + \lambda \mathcal{L}_{\text{macs}}\left(\widehat{M}, M\right)$$

- $\mathcal{L}_{\text{macs}}\left(\widehat{M}, M\right) = \max\{0, \widehat{M}(x,p) - M\}$

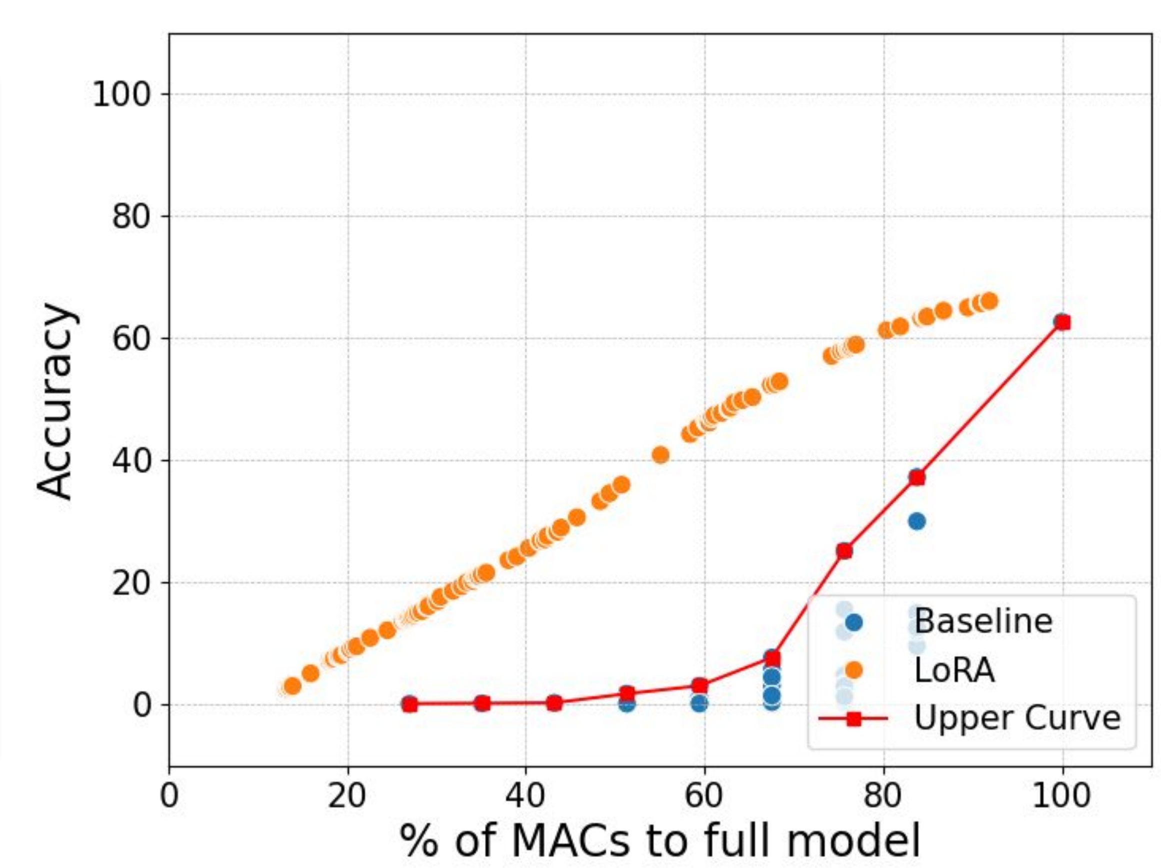
Experiments

Experimental Setup:

- Model:
  - ResNet18, ResNet32 pretrained on CIFAR100
  - CLIP (ViT-B) pretrained on LAION-400m
- Dataset:
  - CIFAR100
  - ImageNet

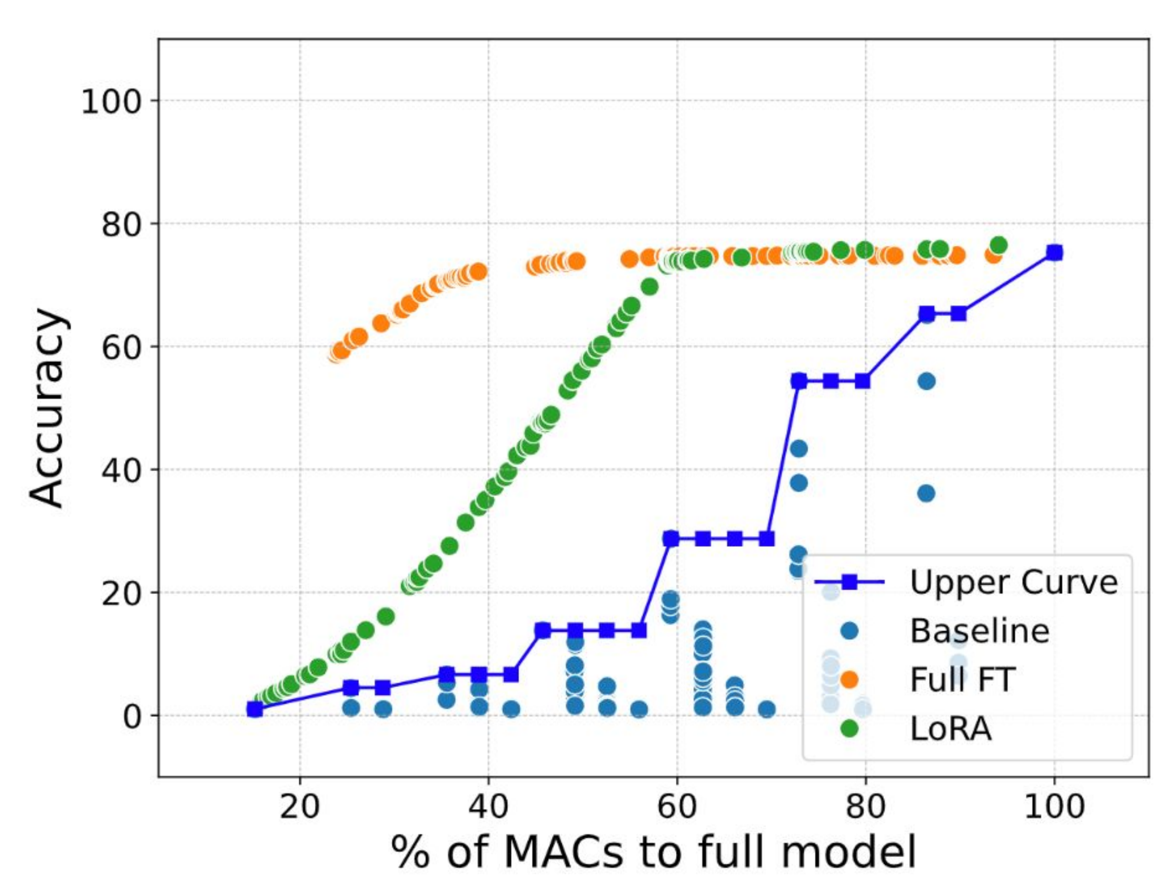
Results on ViT encoder of CLIP pretrained on LAION-400m.

- Baseline: Look-up-table baseline constructed in look-up-table
- Upper Curve: The upper curve of the baseline.

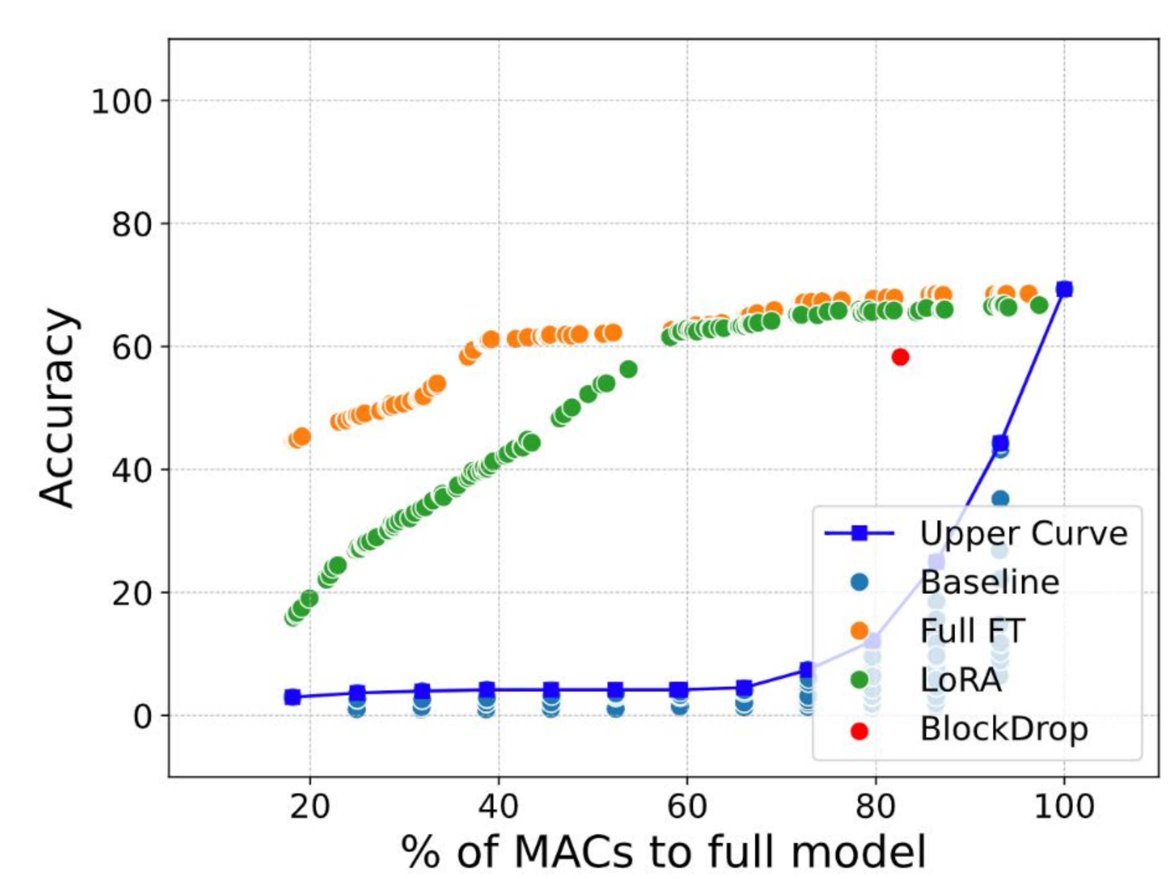


Results on ResNet pretrained on CIFAR100.

- Baseline: Look-up-table baseline.
- Upper Curve: The upper curve of the baseline.
- Full FT: Results on fully finetune the ResNet.
- LoRA: LoRA finetune on ResNet.
- BlockDrop: results in [ZTA+18]



(a) Results of ResNet18 on CIFAR100.



(b) Results of ResNet32 on CIFAR100.

[ZTA+18]Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L. S., Grauman, K., and Feris, R. Blockdrop: Dynamic inference paths in residual networks. CVPR 2018.