Fantastic Features and Where to Find Them:

A Probing Method to combine Features from Multiple Foundation Models

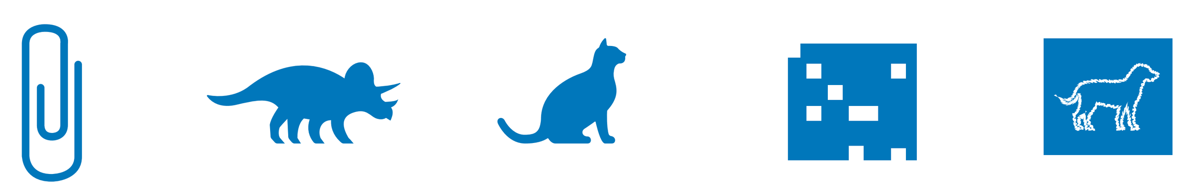
Benjamin Ramtoula, Pierre-Yves Lajoie, Paul Newman, Daniele De Martini

MRG MOBILE ROBOTICS GROUP OXFORD ROBOTICS INSTITUTE · UNIVERSITY OF OXFORD · POLYTECHNIQUE MONTREAL · NEURAL INFORMATION PROCESSING SYSTEMS

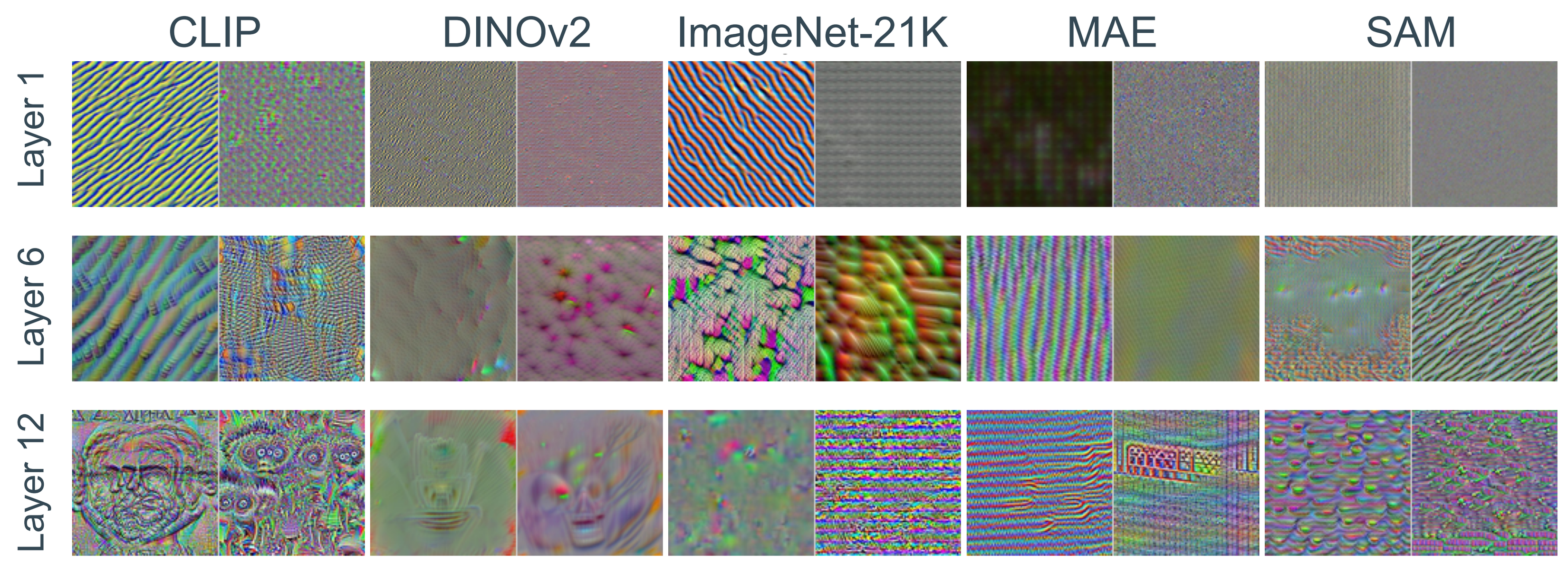
Different foundation models learn different representations

We now have access to different pre-trained models

Model	CLIP	DINOv2	ImageNet	MAE	SAM	…
Training objective	Align text and image embeddings	Produce consistent image embeddings	Classify the semantic content of images	Reconstruct missing pixels	Segment instances	…
Dataset	WIT	LVD142M	ImageNet	ImageNet	SA-1B	…



Supervision and data differences affect the representations learned throughout their layers

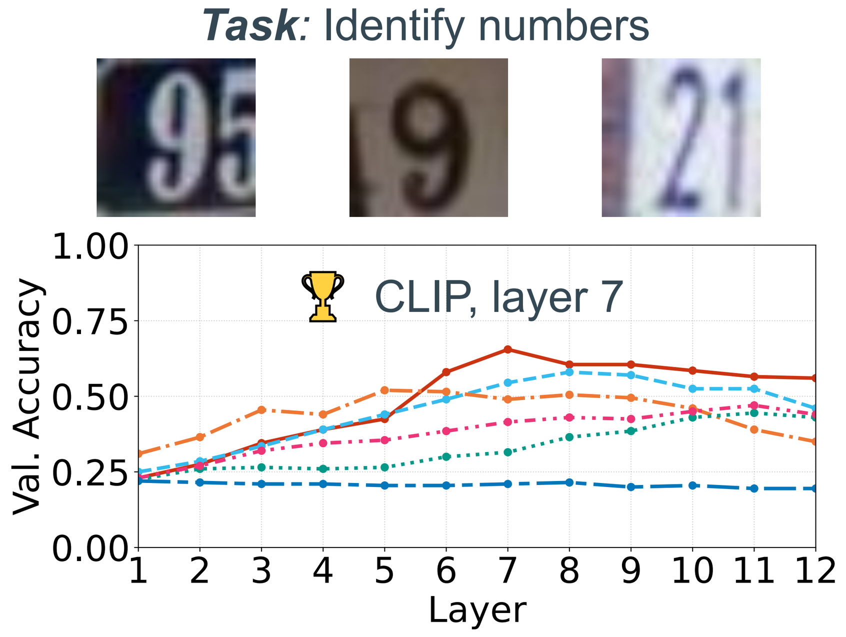


Images that maximise activations of different neurons

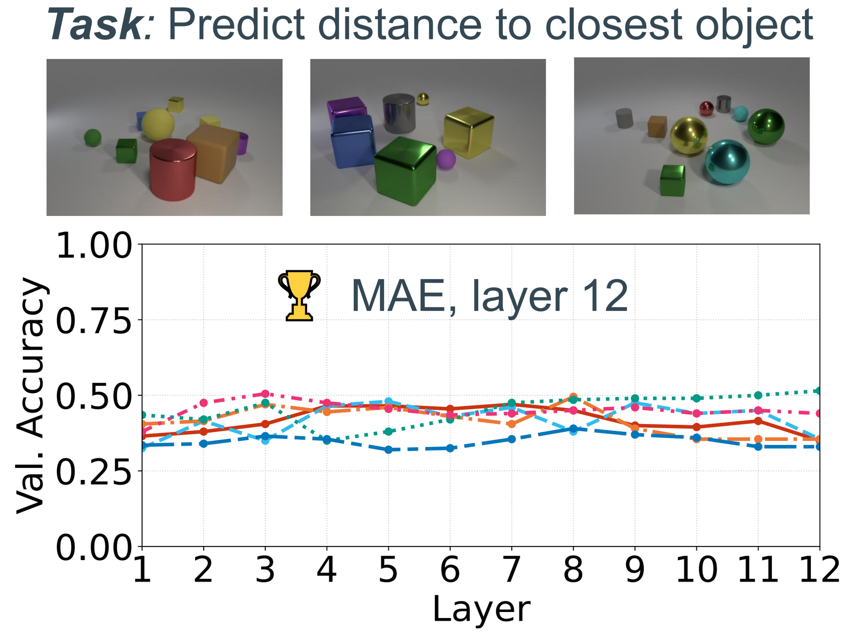
The model and layer producing the most relevant features depend on the task

Linear probing on top of features from different layers of different models

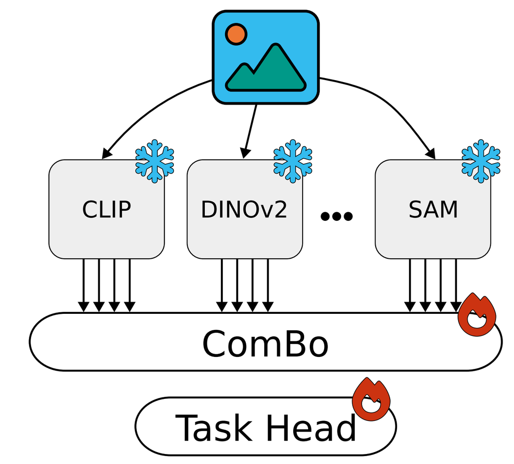
VTAB-1k SVHN
Task: Identify numbers



VTAB-1k Clevr/distance
Task: Predict distance to closest object



We propose an architecture for efficient multi-layer, multi-model feature probing



Existing solutions have limitations

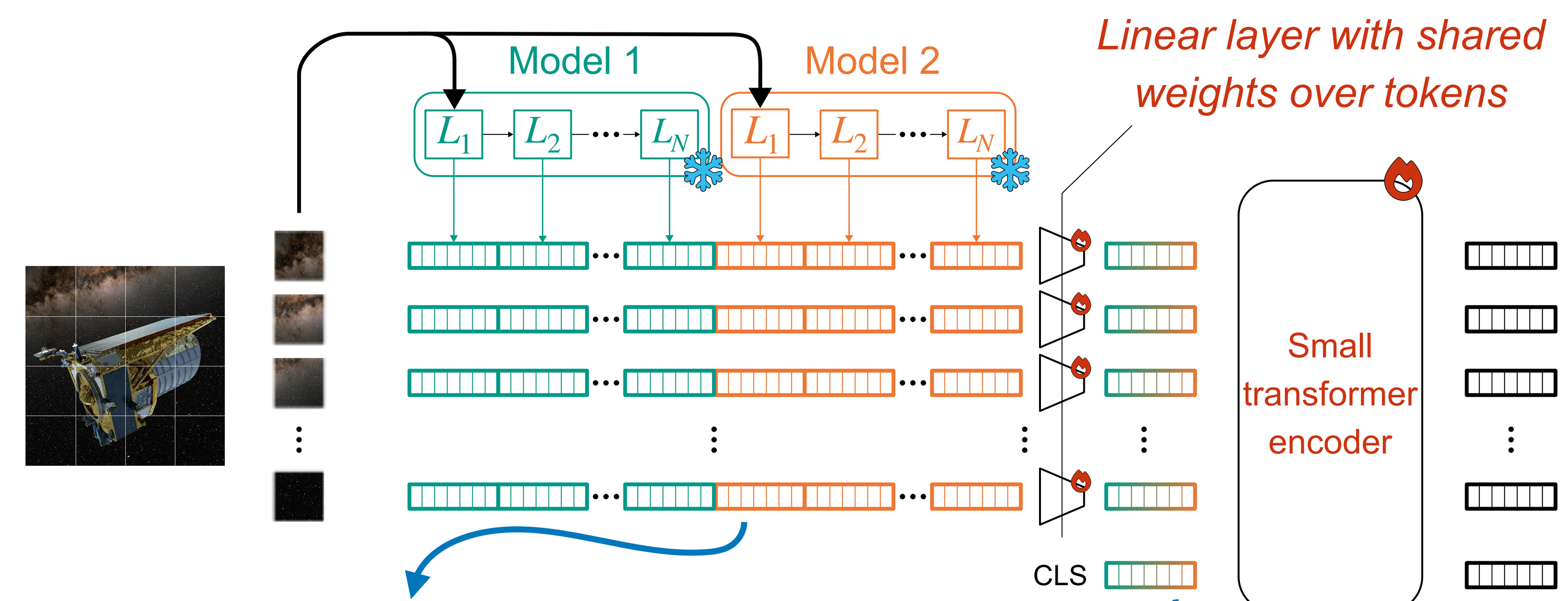
Existing ways to adapt pre-trained models	Examples	Scales to multiple backbones?	Can easily use new backbones?	Can easily be adapted to a new task?
Fine-tuning-based approaches	LoRA, Adapter+	✗	✓	✓
Multi-layer probing of frozen features	Head2Toe, SMP	✓	✓	✗
Distillation + adaptation	RADIOv2.5 + Adapter+	✓	✗	✓

We address them with **ComBo**, our probing approach to **Com**bine back**Bo**nes

✓No backpropagation through backbones

✓Can easily include new models

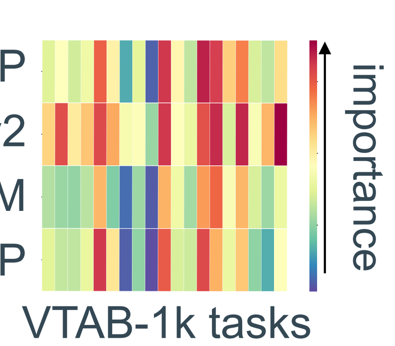
✓No task-specific hyperparameter tuning



We can also use ComBo to identify and keep only the most task-relevant models

Using the norm of learned linear layer weights associated with each model to measure their importance:

1. Train ComBo while minimising each model's importance
2. Inspect weights to measure task-relevance
3. Retrain using only the most relevant backbones

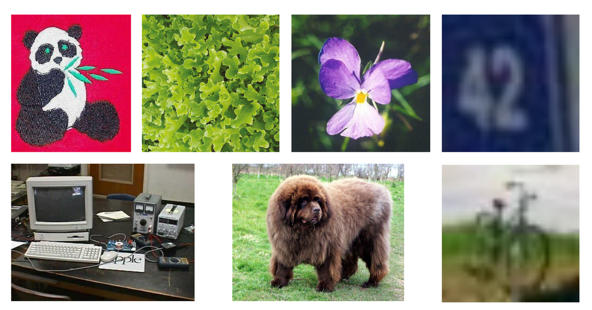


Why is this useful?

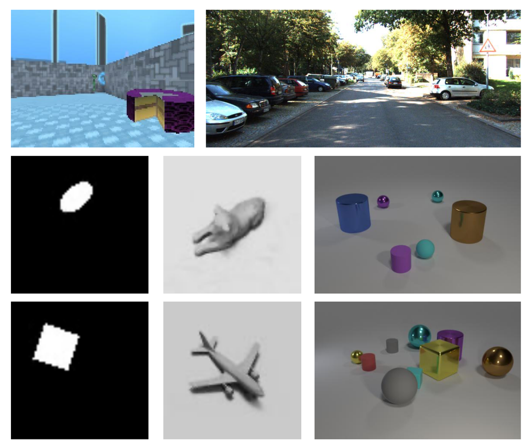
Experimental setting

- **VTAB-1k benchmark:**
  - 19 tasks framed as classification
  - Only 1000 training images per task
  - Fixed hyperparameters

Natural tasks

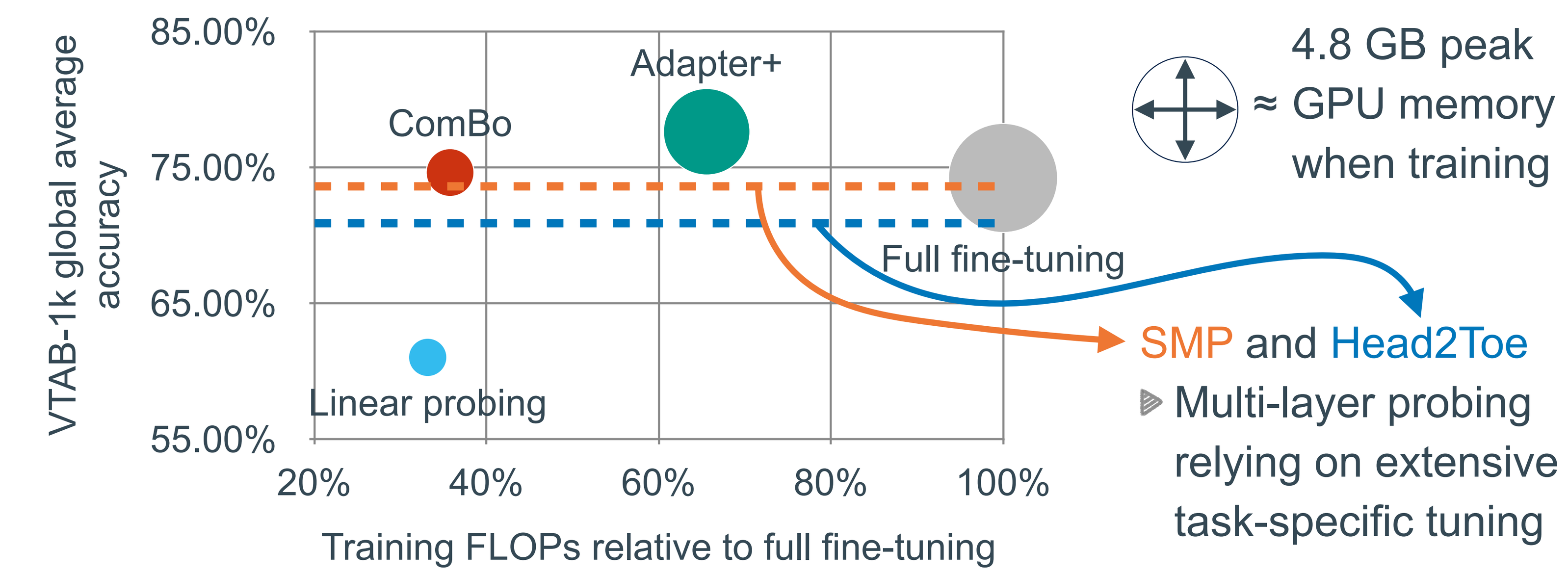


Structured tasks



Specialised tasks

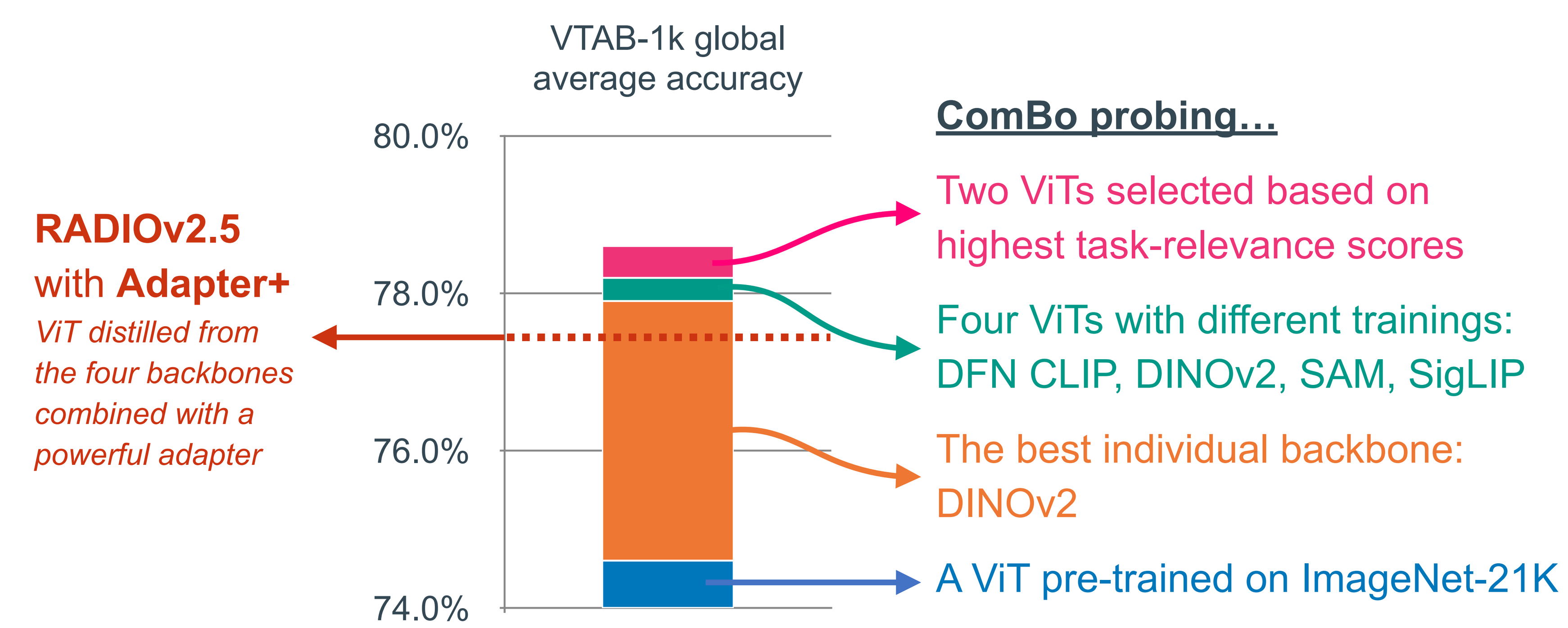
Adapting a ViT-B/16 pre-trained on ImageNet-21K



✓Good performance and minimal compute (5min to train on an RTX 3090 Ti GPU)

- Enables scaling to multiple backbones

Probing multiple foundation models at once



✓Combining multiple models improves performance

✓ComBo outperforms costly distillation

✓Further improvements when focusing on the most relevant models

This work was supported by EPSRC Programme Grant "From Sensing to Collaboration" (EP/V000748/1), the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (EP/S024050/1), Oxa, and the Digital Research Alliance of Canada.

UKRI Engineering and Physical Sciences Research Council · AIMS EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems