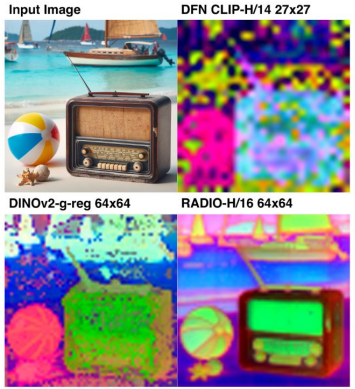
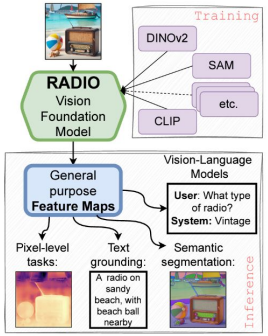
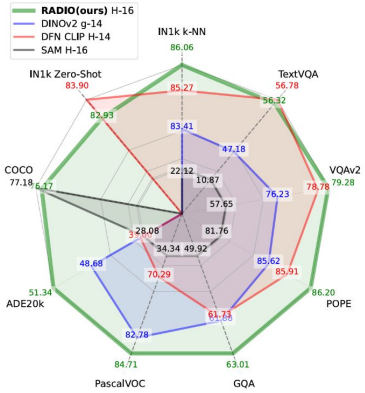
AM-RADIO:Agglomerative Vision Foundation Model Reduce All Domains Into One



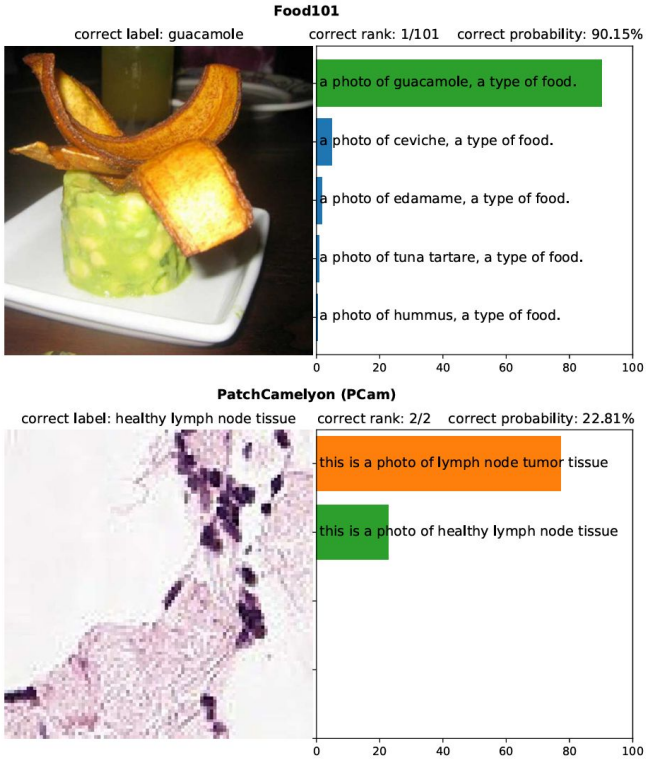




San Diego Machine Learning
Ryan Chesler

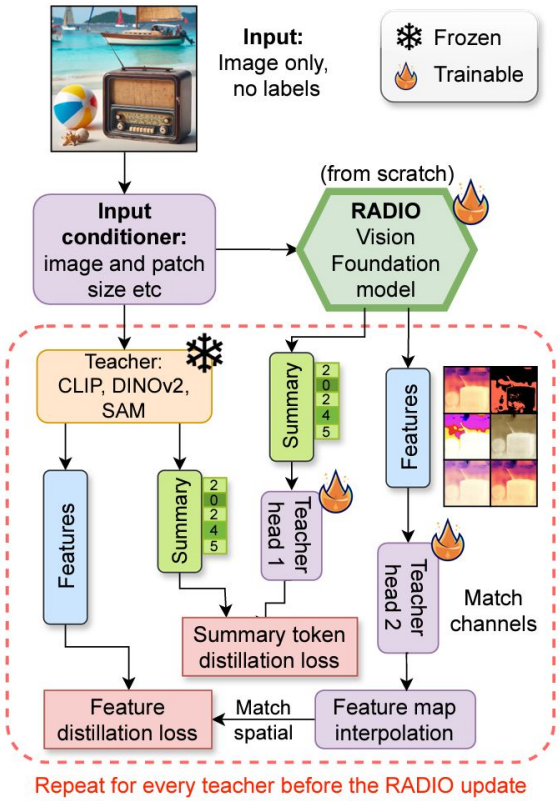
Purpose

- New foundation vision models keep popping up
  - Strong models that take a lot of effort to train
  - Data++ and compute++
  - Bring a variety of awesome capabilities
  - https://segment-anything.com/
  - Zero-shot classification with CLiP
  - Powerful representation from DINOv2
- Would be interesting if we could merge capabilities into a single model
- Maybe overlap in knowledge makes it better at all tasks



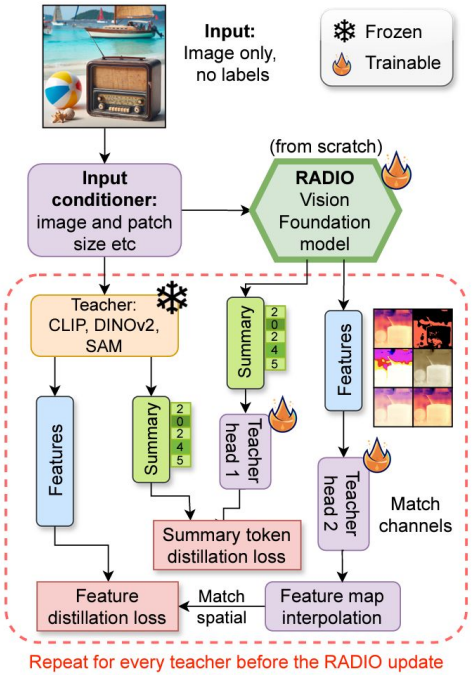
Knowledge Distillation

- Trying to teach one model from a previous model
  - Have the teacher model take an input and make a prediction
  - Give a student model the same input and learn to make the same prediction
- Can be done at the final task layer(classifier, segmenter, etc.)
- Can also be done at the feature level(trying to make some intermediate embedding)
- In this case it is complicated because models have different outputs, internal embedding sizes and even different input resolutions



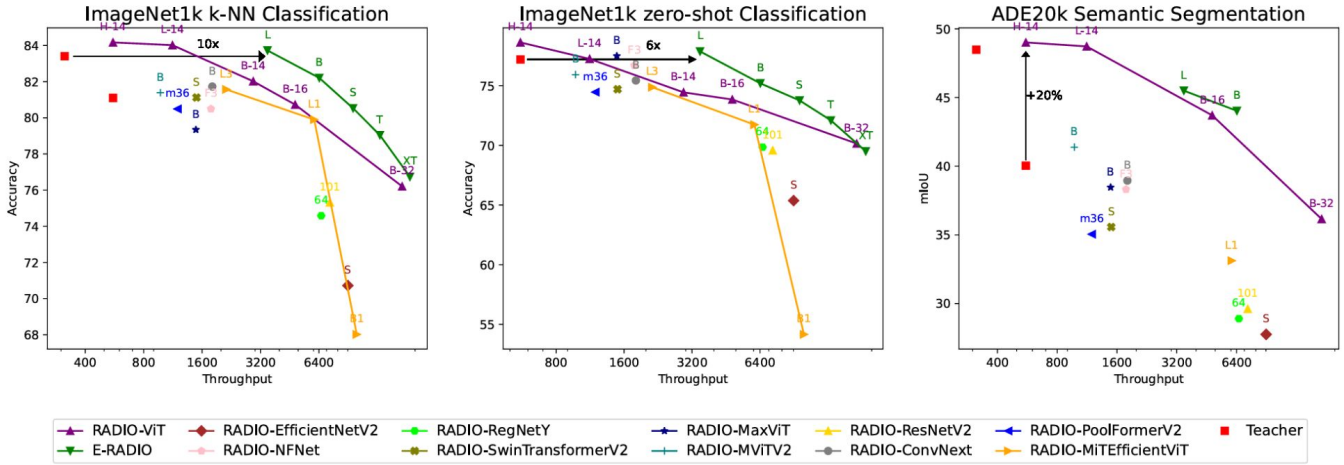
Merging different representations

- RADIO outputs some embedding with a fixed size
- Three two-layer adaptor heads map embedding to the same size as each target model separately
- This is done for both summary embeddings(CLS token) and feature embeddings
- The goal is to essentially make a single model that is a drop-in replacement for any of the models



Training Details

- 600k training steps, 614M images seen, 64 GPUs
- Half of GPUs ran CLiP + DINO and other half ran SAM
- Tried many different model architectures trying to find the best student



E-Radio

- Created a new architecture based on their experiments that had really good performance frontier for throughput vs accuracy
- Mixes some of the benefits of CNNs with the strengths of vision transformers
- First layer is strided convolutions that reduce resolution by 4x
  - Better than directly downsampling
  - Greatly reduces the attention cost of transformers later on
- Does both global and local attention, merging various different representations

Results

Backbone	Param. Count	Through-put	Zero Shot	k-NN	ADE20k	FD loss
Teachers						
DINOv2 G/14	1.14B	313	N/A	83.41	47.53	
OpenCLIP H/14	632M	556	77.19	81.10	40.04	
Existing Efficient Models						
EfficientNetV2-S	21M	9017	65.37	70.72	27.75	0.415
ResNetv2-101	44M	7283	69.58	75.32	29.61	0.405
RegNetY-064	30M	6573	69.84	74.59	28.9	0.394
EfficientViT-L1	38M	6048	71.73	79.90	33.12	0.376
ConvNext-B	88M	1805	75.43	81.73	38.95	0.358
NFNet-F3	254M	1777	76.93	80.50	38.31	0.340
SwinV2-S	49M	1497	74.70	81.12	35.57	0.364
MaxViT-B	119M	1486	77.49	79.34	38.46	0.340
PoolformerV2-M36	56M	1194	74.46	80.49	35.05	0.377
MViTV2-B	51M	975	75.92	81.39	41.39	0.345
Proposed architecture						
E-RADIO-B	118M	6422	75.19	82.21	44.03	0.319
↳ w/o upsample	113M	7040	75.45	82.05	41.26	0.353
E-RADIO-L	265M	3472	77.87	83.73	45.5	0.265

COCO 2017 drop-in SAM replacement at 1024x1024

Family	**Arch**	**mIOU**	**Throughput**
SAM	Base	75.78	50.94
	Large	77.02	20.62
	Huge	77.18	11.83
E-RADIO (ours)	Large	76.31	121.74
RADIO (ours)	ViTDet-H/16-W8†	76.09	29.09
	ViTDet-H/16-W16†	76.23	27.91