Learning from Offline Foundation Features with Tensor Augmentations

Emir Konuk, Christos Matsoukas, Moein Sorkhei, Phitchapha Lertsiravarameth, Kevin Smith

Hypothesis

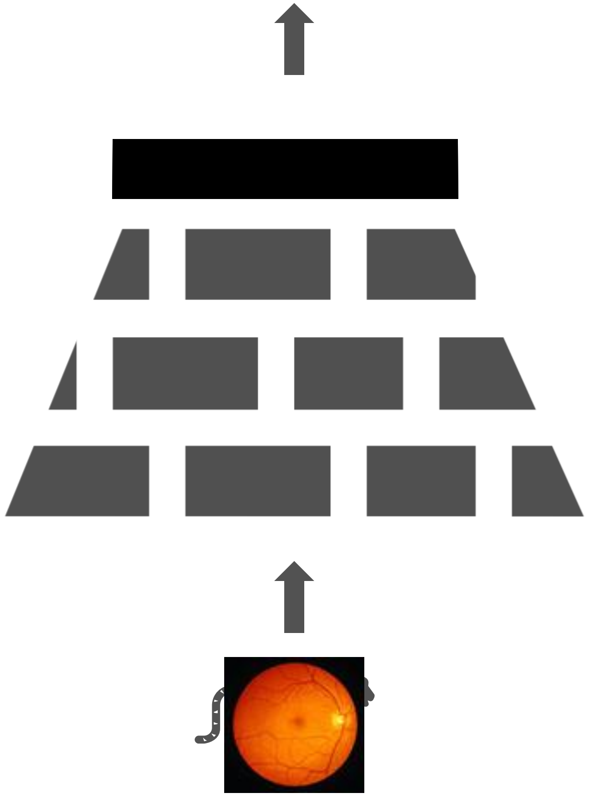
Foundation model representations are robust.

We should be able to train on them directly.

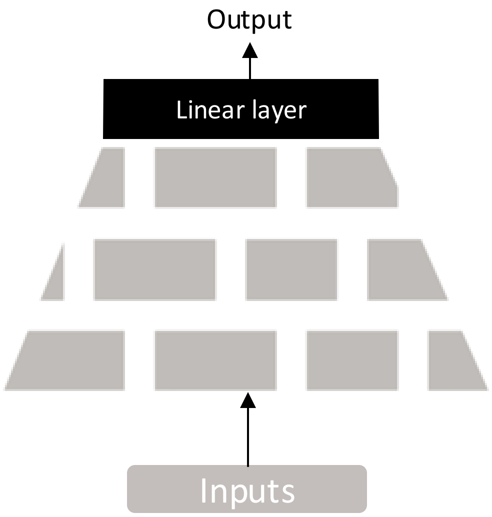
Transfer learning

Not a cat

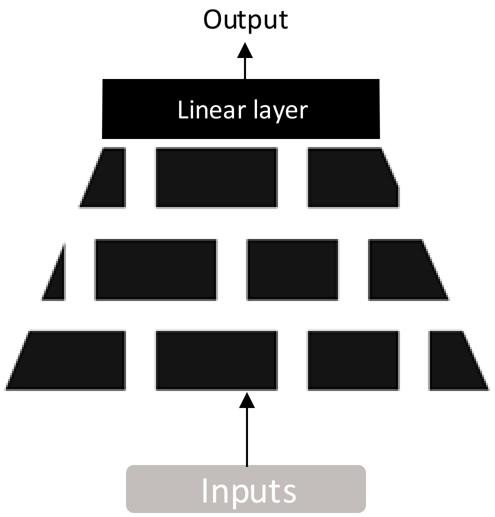
Not diabetes



Transfer learning

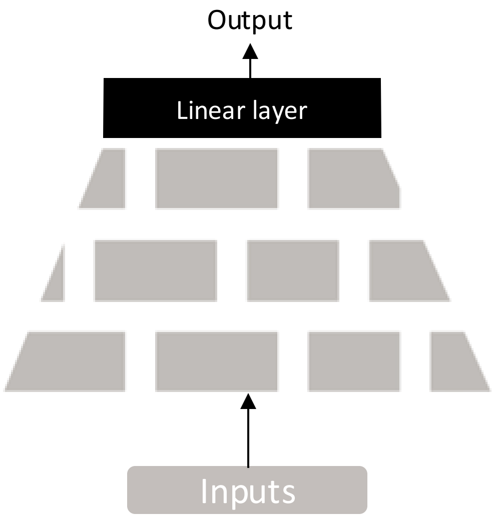


Linear probing



Fine tuning

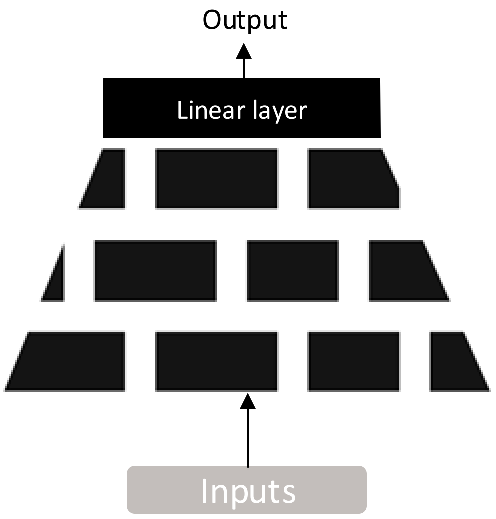
Efficient adaptation


Output

Linear layer

Inputs

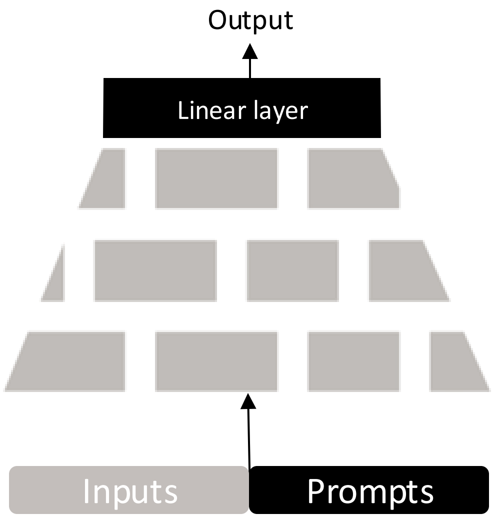
Linear probing


Output

Linear layer

Inputs

Fine tuning

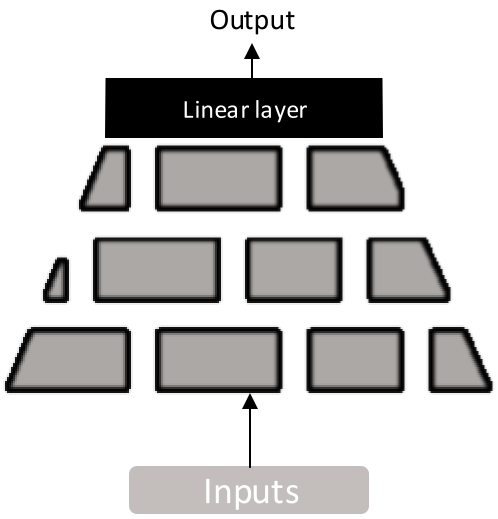

Output

Linear layer

Inputs

Prompts

Prompting

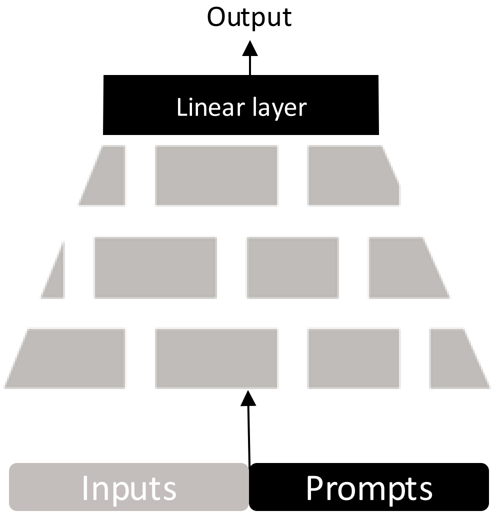

Output

Linear layer

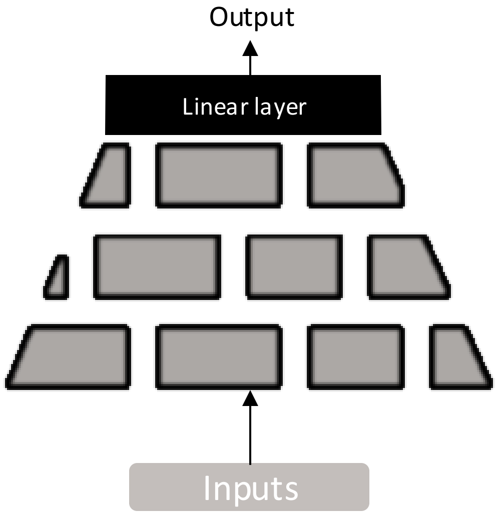
Inputs

Adapters

Motivation



Prompting

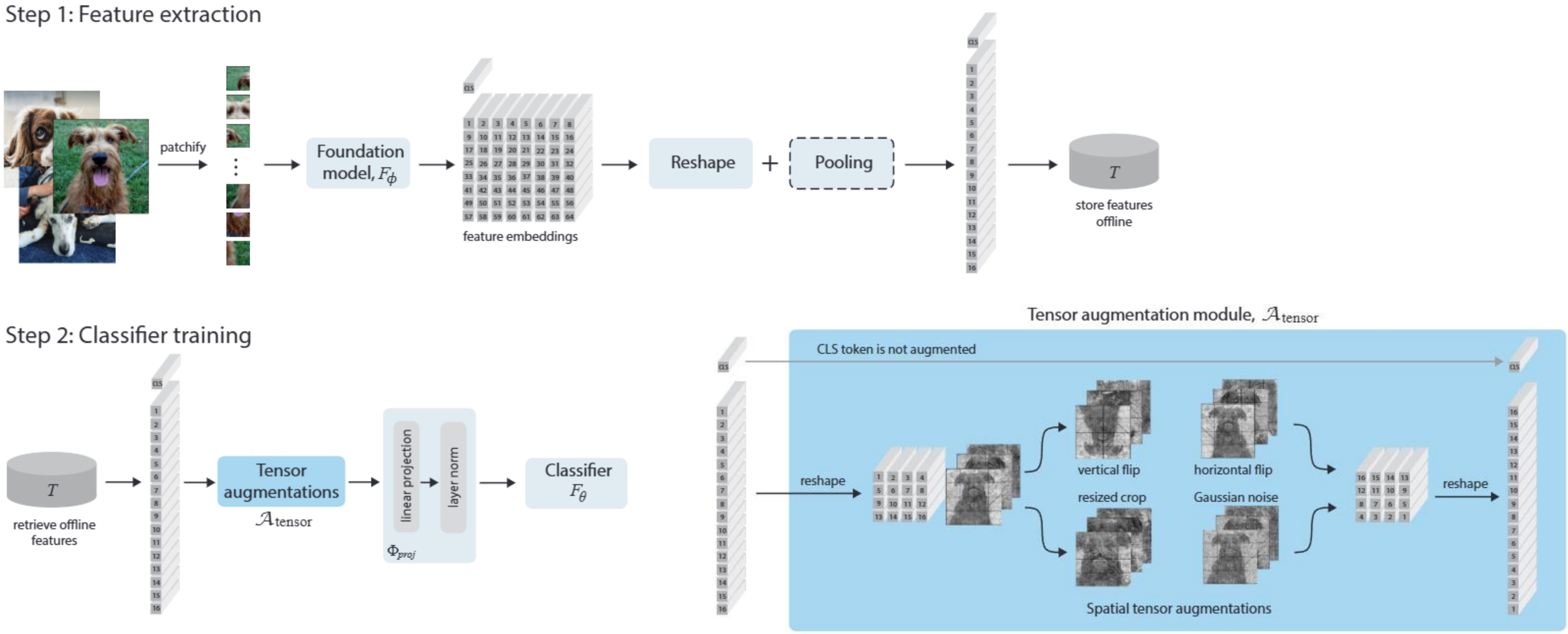


Adapters

Foundation models are still there.

Can we decouple them?

LOFFTA



Step 1: Feature extraction

patchify

Foundation model, F_ϕ

feature embeddings

Reshape + Pooling

T

store features offline

Step 2: Classifier training

T

retrieve offline features

Tensor augmentations $\mathcal{A}_{\text{tensor}}$

linear projection

layer norm

Φ_{proj}

Classifier F_θ

Tensor augmentation module, $\mathcal{A}_{\text{tensor}}$

CLS token is not augmented

reshape

vertical flip

horizontal flip

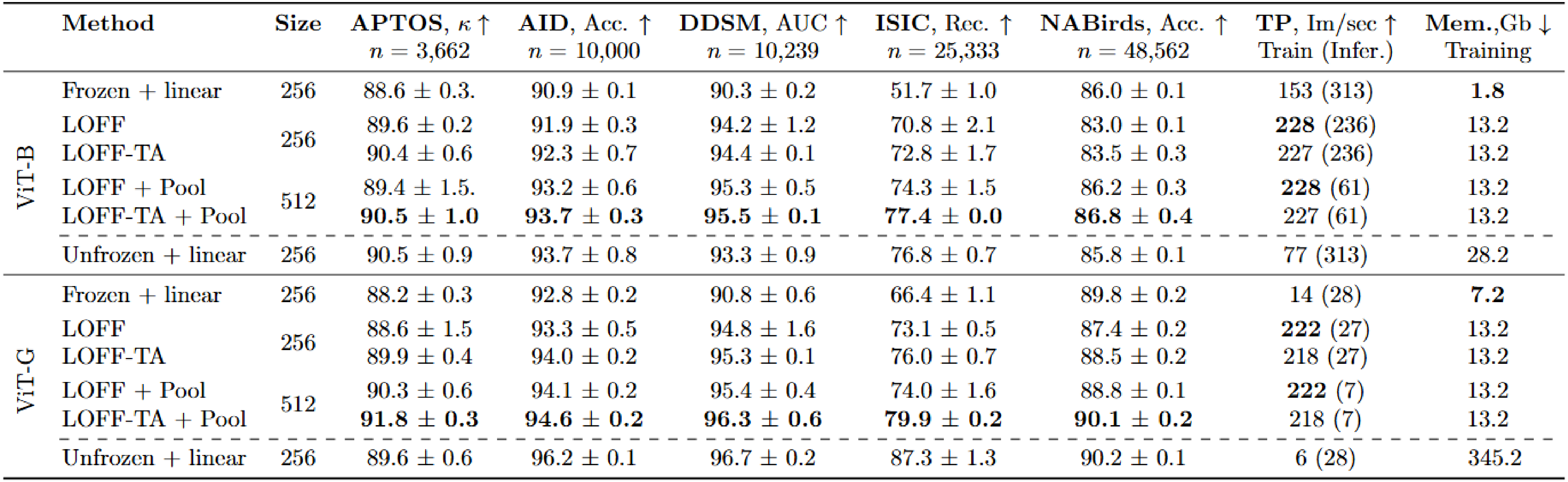
resized crop

Gaussian noise

reshape

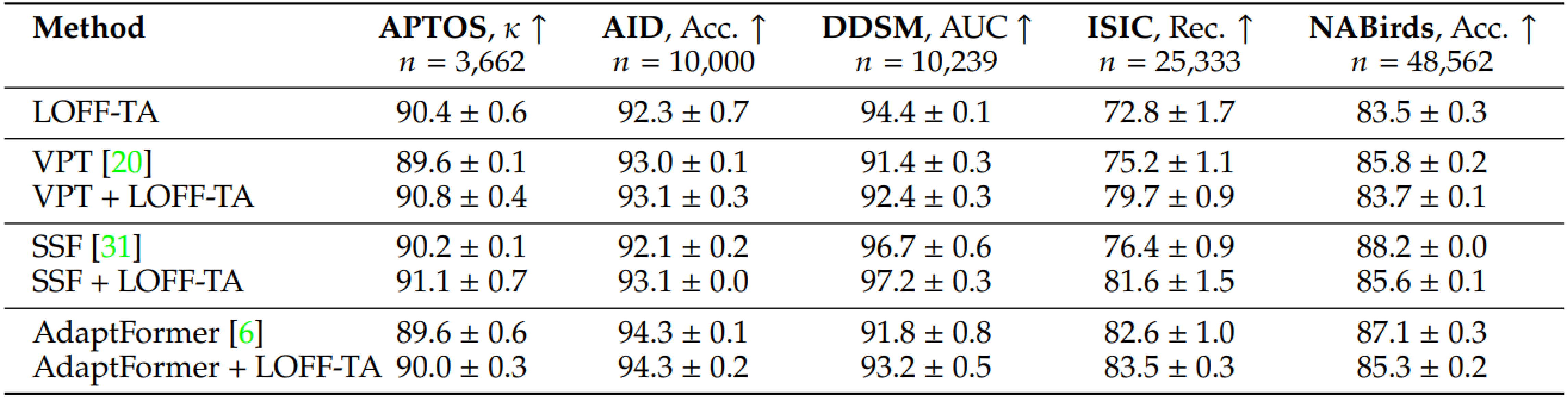
Spatial tensor augmentations

Results



	Method	Size	**APTOS**, $\kappa \uparrow$ n = 3,662	**AID**, Acc. ↑ n = 10,000	**DDSM**, AUC ↑ n = 10,239	**ISIC**, Rec. ↑ n = 25,333	**NABirds**, Acc. ↑ n = 48,562	**TP**, Im/sec ↑ Train (Infer.)	**Mem.**,Gb ↓ Training
ViT-B	Frozen + linear	256	88.6 ± 0.3.	90.9 ± 0.1	90.3 ± 0.2	51.7 ± 1.0	86.0 ± 0.1	153 (313)	**1.8**
	LOFF	256	89.6 ± 0.2	91.9 ± 0.3	94.2 ± 1.2	70.8 ± 2.1	83.0 ± 0.1	**228** (236)	13.2
	LOFF-TA		90.4 ± 0.6	92.3 ± 0.7	94.4 ± 0.1	72.8 ± 1.7	83.5 ± 0.3	227 (236)	13.2
	LOFF + Pool	512	89.4 ± 1.5.	93.2 ± 0.6	95.3 ± 0.5	74.3 ± 1.5	86.2 ± 0.3	**228** (61)	13.2
	LOFF-TA + Pool		**90.5 ± 1.0**	**93.7 ± 0.3**	**95.5 ± 0.1**	**77.4 ± 0.0**	**86.8 ± 0.4**	227 (61)	13.2
	Unfrozen + linear	256	90.5 ± 0.9	93.7 ± 0.8	93.3 ± 0.9	76.8 ± 0.7	85.8 ± 0.1	77 (313)	28.2
ViT-G	Frozen + linear	256	88.2 ± 0.3	92.8 ± 0.2	90.8 ± 0.6	66.4 ± 1.1	89.8 ± 0.2	14 (28)	**7.2**
	LOFF	256	88.6 ± 1.5	93.3 ± 0.5	94.8 ± 1.6	73.1 ± 0.5	87.4 ± 0.2	**222** (27)	13.2
	LOFF-TA		89.9 ± 0.4	94.0 ± 0.2	95.3 ± 0.1	76.0 ± 0.7	88.5 ± 0.2	218 (27)	13.2
	LOFF + Pool	512	90.3 ± 0.6	94.1 ± 0.2	95.4 ± 0.4	74.0 ± 1.6	88.8 ± 0.1	**222** (7)	13.2
	LOFF-TA + Pool		**91.8 ± 0.3**	**94.6 ± 0.2**	**96.3 ± 0.6**	**79.9 ± 0.2**	**90.1 ± 0.2**	218 (7)	13.2
	Unfrozen + linear	256	89.6 ± 0.6	96.2 ± 0.1	96.7 ± 0.2	87.3 ± 1.3	90.2 ± 0.1	6 (28)	345.2

Results



Method	**APTOS**, κ ↑ n = 3,662	**AID**, Acc. ↑ n = 10,000	**DDSM**, AUC ↑ n = 10,239	**ISIC**, Rec. ↑ n = 25,333	**NABirds**, Acc. ↑ n = 48,562
LOFF-TA	90.4 ± 0.6	92.3 ± 0.7	94.4 ± 0.1	72.8 ± 1.7	83.5 ± 0.3
VPT [20]	89.6 ± 0.1	93.0 ± 0.1	91.4 ± 0.3	75.2 ± 1.1	85.8 ± 0.2
VPT + LOFF-TA	90.8 ± 0.4	93.1 ± 0.3	92.4 ± 0.3	79.7 ± 0.9	83.7 ± 0.1
SSF [31]	90.2 ± 0.1	92.1 ± 0.2	96.7 ± 0.6	76.4 ± 0.9	88.2 ± 0.0
SSF + LOFF-TA	91.1 ± 0.7	93.1 ± 0.0	97.2 ± 0.3	81.6 ± 1.5	85.6 ± 0.1
AdaptFormer [6]	89.6 ± 0.6	94.3 ± 0.1	91.8 ± 0.8	82.6 ± 1.0	87.1 ± 0.3
AdaptFormer + LOFF-TA	90.0 ± 0.3	94.3 ± 0.2	93.2 ± 0.5	83.5 ± 0.3	85.3 ± 0.2

Limitations

- Slower during inference
- LOFFTA is competitive but not consistently better in performance

Conclusions

- Foundation models as fixed feature extractors
- Spatial tensor augmentations

Thank you!

ekonuk@kth.se