Y-DATA 3rd Research Seminar

2025

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby (2021)

By Yair and Kai

Vision Transformers / ViT (2021) Results

1. **Finetune Accuracy** % After Pretraining on Different Datasets – Comparing to 'State-of-the-Art'.
    I. Top1 Finetune Accuracy % After Pretraining on Different Datasets
    II. VTAB Breakdown
2. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT and different dataset sizes.
3. Accuracy % Relative to Compute for Various Models
4. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser
5. **Attention Map**
6. **Self - Supervision**
7. **Position Embedding**, its Dimensions & Where to Add
8. Position Embedding Trained With Different Hyperparameters
9. **Attention Distance** at Various Network Depths
10. **Batch Size** for Models at Various Input Sizes

1

Finetune Accuracy % After Pretraining on Different Datasets

1. Finetune Accuracy % After Pretraining on Different Datasets

Model	Pretrained On	Remarks
BiT-L ResNet152x4	ImageNet21k	Baseline for all image datasets BiT = "Big Transfer" architecture
Noisy Student EfficientNet-L2	ImageNet21k	Baseline for ImageNet
ViT-L/16 (Large model)	ImageNet21k	Test performance ViT = Vision Transformer
ViT-L/16	JFT-300M (Google proprietary)	
ViT-H/14 (Huge model, bigger than Large model)	JFT-300M (Google proprietary)	

Notes
- No information on what the models were finetuned on
- Assuming finetuning was performed on ImageNet21k dataset

1. Finetune Accuracy % After Pretraining on Different Datasets

	Ours-JFT (ViT-L/16)		BiT-L (ResNet152x4)		Noisy Student (EfficientNet-L2)
ImageNet	87.76 ± 0.03	≈	87.54 ± 0.02	<	$88.4/88.5^*$
ImageNet ReaL	90.54 ± 0.03		90.54		90.55
CIFAR-10	99.42 ± 0.03		99.37 ± 0.06		–
CIFAR-100	93.90 ± 0.05		93.51 ± 0.08		–
Oxford-IIIT Pets	97.32 ± 0.11		96.62 ± 0.23		–
Oxford Flowers-102	$\mathbf{99.74} \pm 0.00$		99.63 ± 0.03		–
VTAB (19 tasks)	76.28 ± 0.46		76.29 ± 1.70		–

1. Finetune Accuracy % After Pretraining on Different Datasets

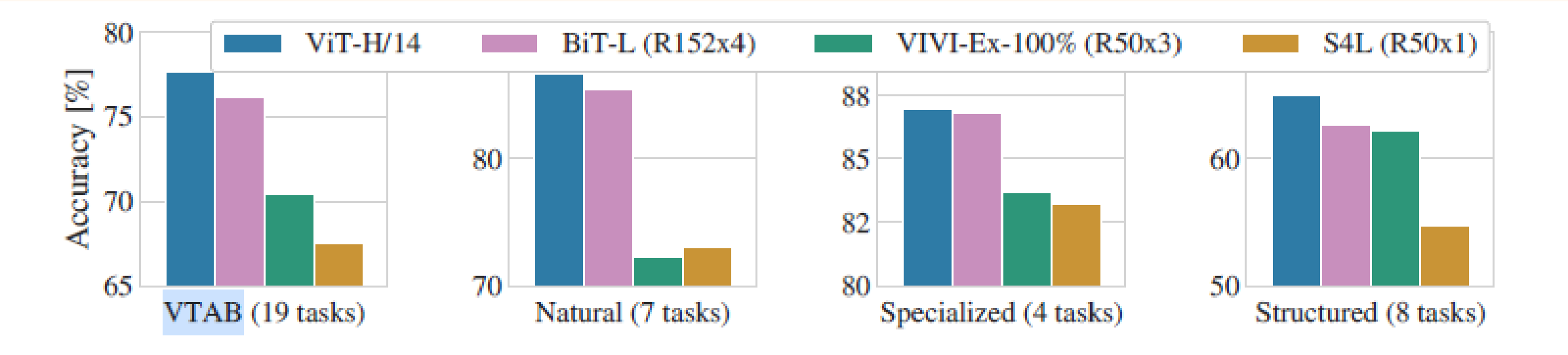
	Ours-JFT (ViT-H/14)		BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	>	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05		90.54	90.55
CIFAR-10	**99.50** ± 0.06		99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04		93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03		96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02		99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23		76.29 ± 1.70	–

1. Finetune Accuracy % After Pretraining on Different Datasets

Comparison – Higher Accuracy using Less Compute

	Ours-JFT (ViT-H/14)		BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04		87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05		90.54	90.55
CIFAR-10	**99.50** ± 0.06	>	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04		93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03		96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02		99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23		76.29 ± 1.70	–
TPUv3-core-days	2.5k	<	9.9k	12.3k

1. VTAB break-down



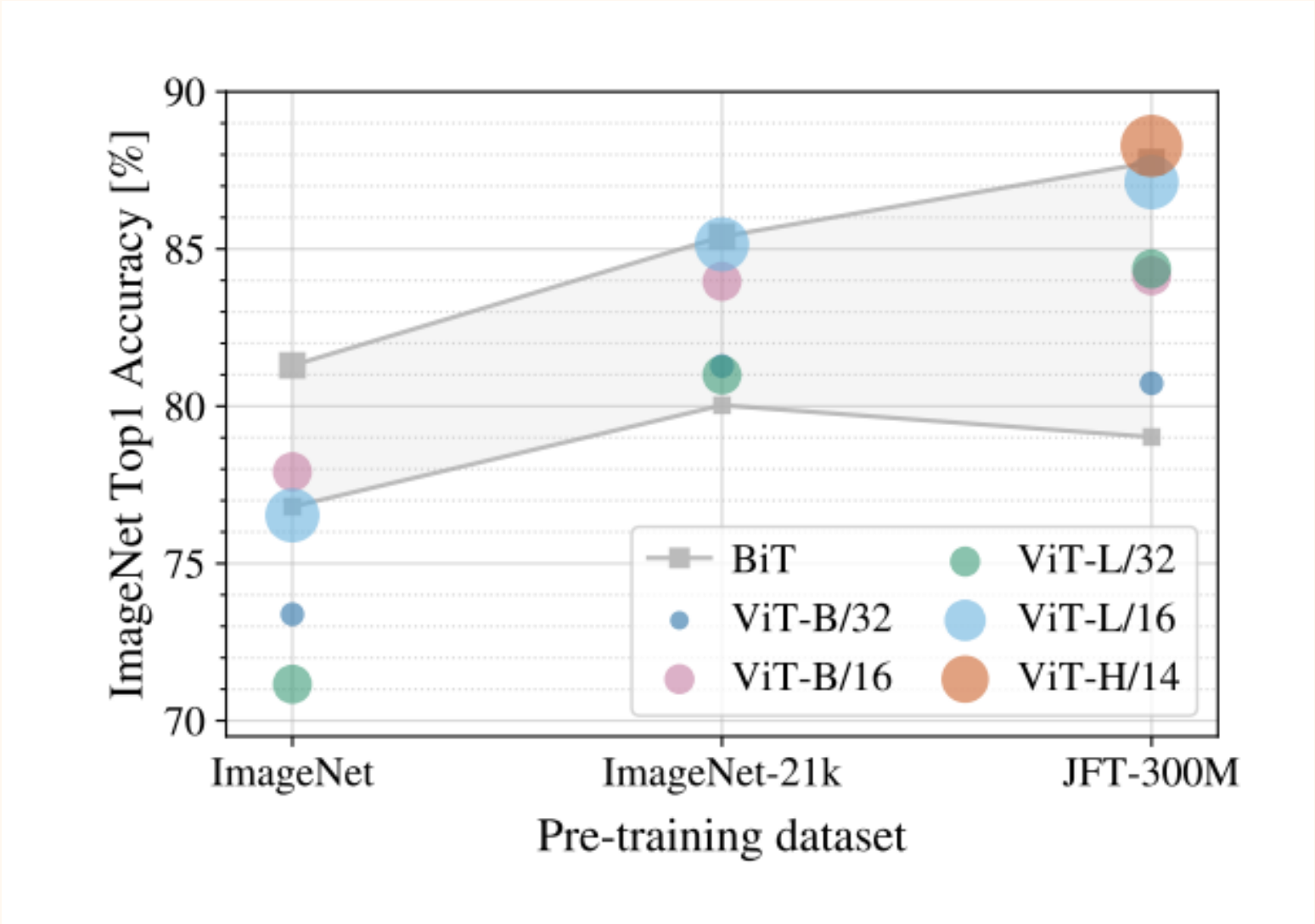
2

Top1 Finetune Accuracy % After Pretraining on Different Datasets

2. Top1 Finetune Accuracy % After Pretraining on Different Datasets

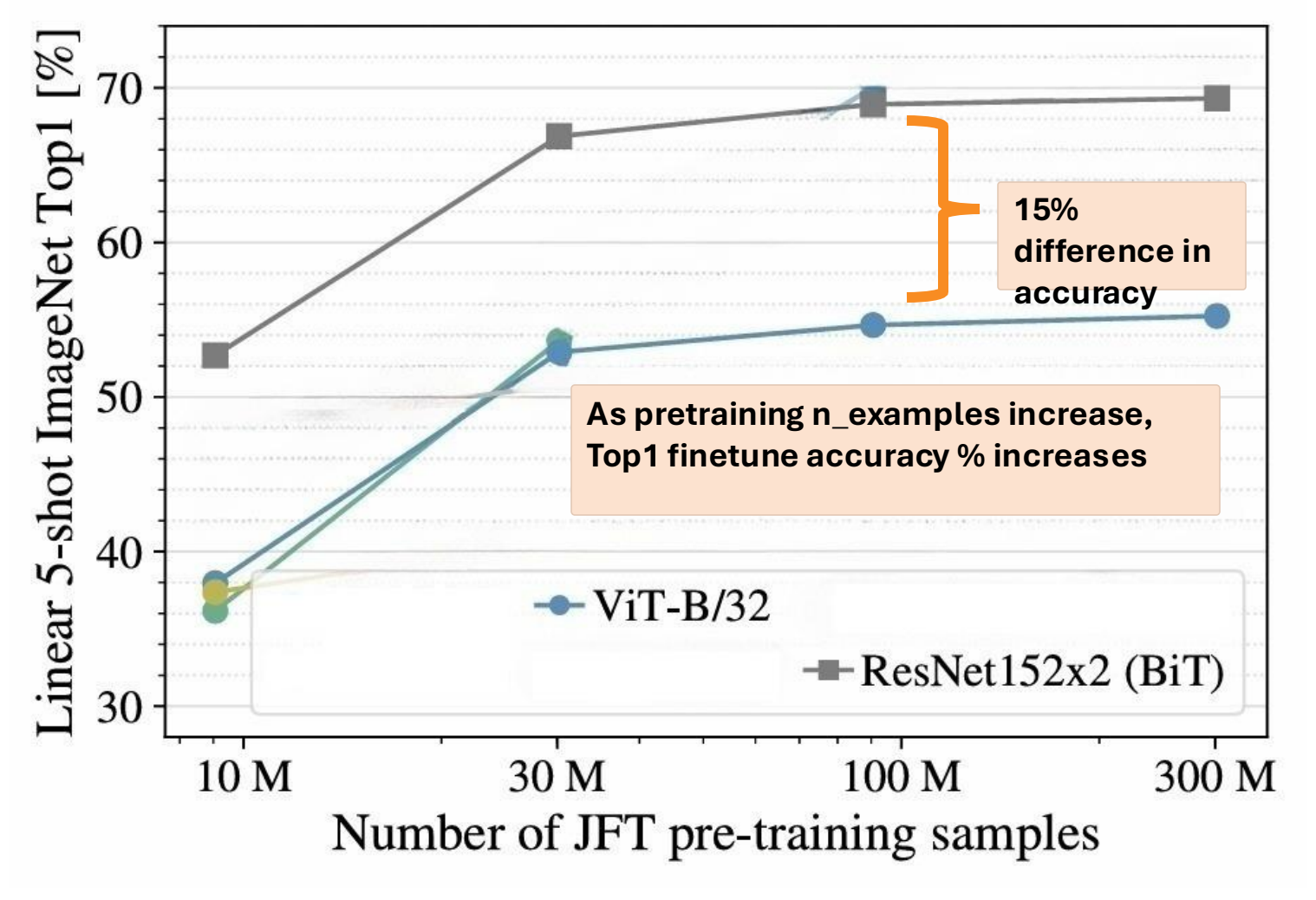
Model	Pretrained On	Remarks
ViT-B/16	ImageNet ImageNet21k JFT-300M (Google proprietary)	**Base model**
ViT-B/32		Base model pretrained on **lower resolution input images**
ViT-L/16		**Large model**
ViT-L/32		Large model pretrained on lower resolution input images
ViT-H/14		**Huge model, bigger than Large model**

2. Top1 Finetune Accuracy % After Pretraining on Different Datasets



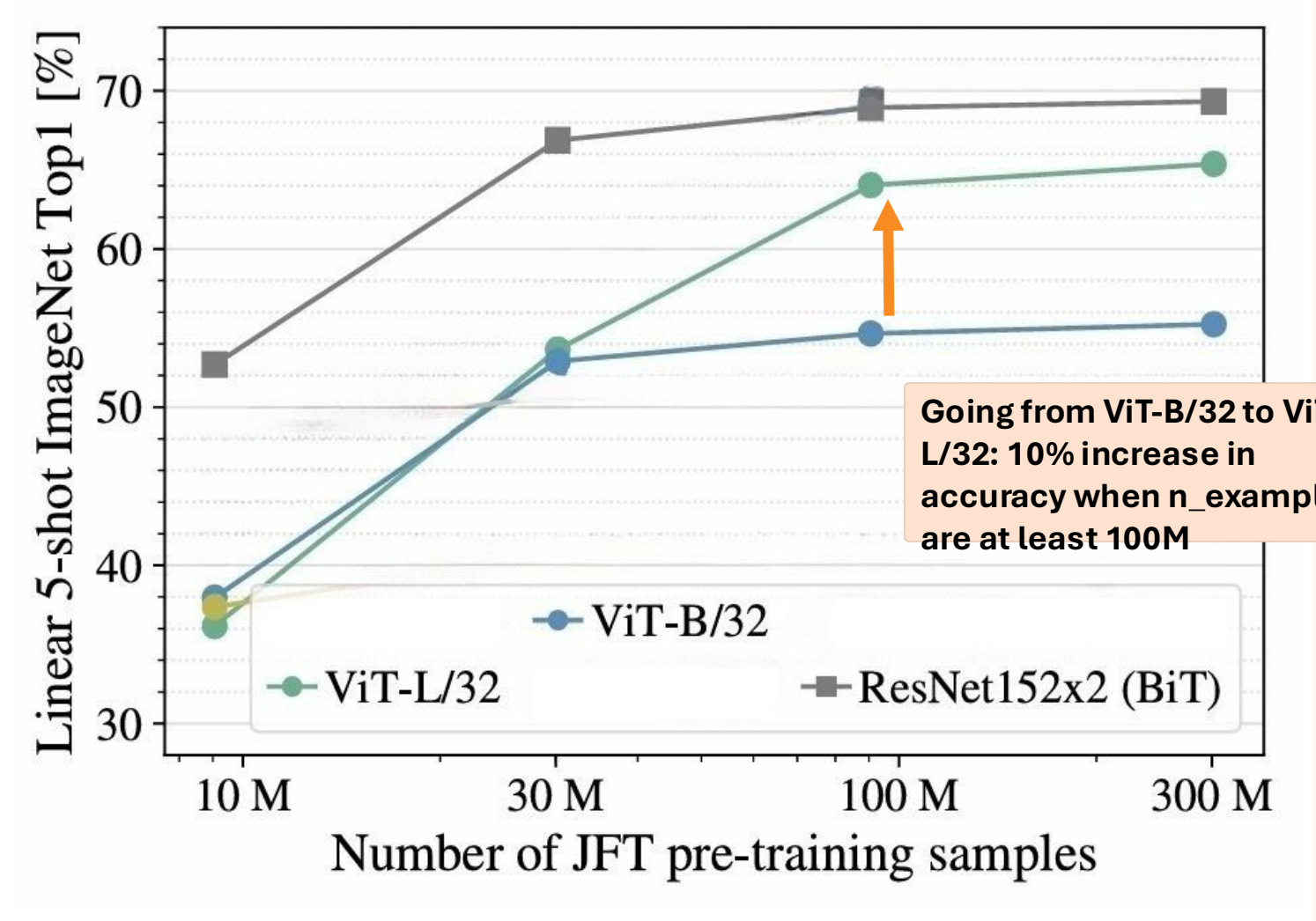
2. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT

Comparison 1



2. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT

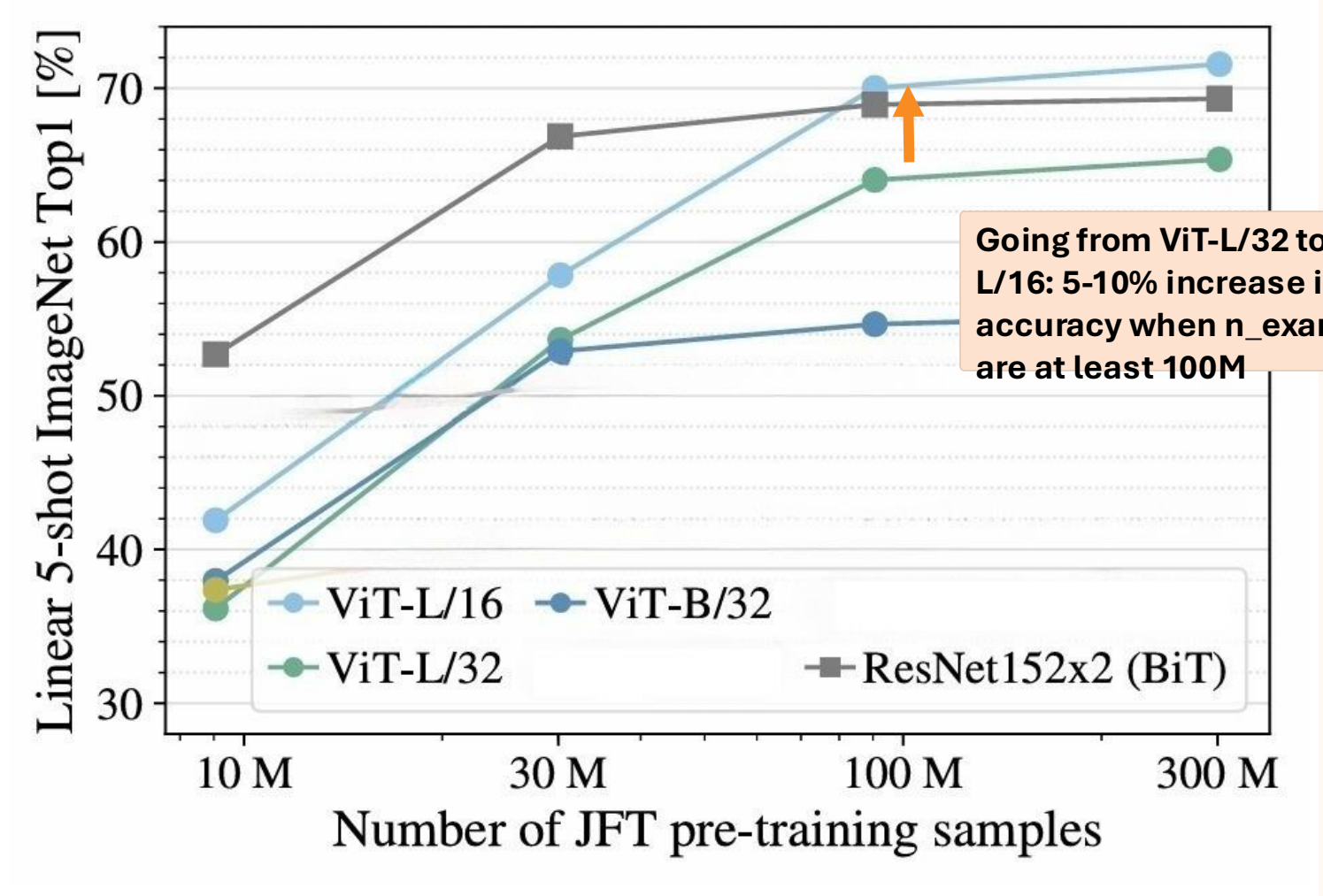
Comparison 2



Going from ViT-B/32 to ViT-L/32: 10% increase in accuracy when n_examples are at least 100M

2. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT

Comparison 3



Going from ViT-L/32 to ViT-L/16: 5-10% increase in accuracy when n_examples are at least 100M

3

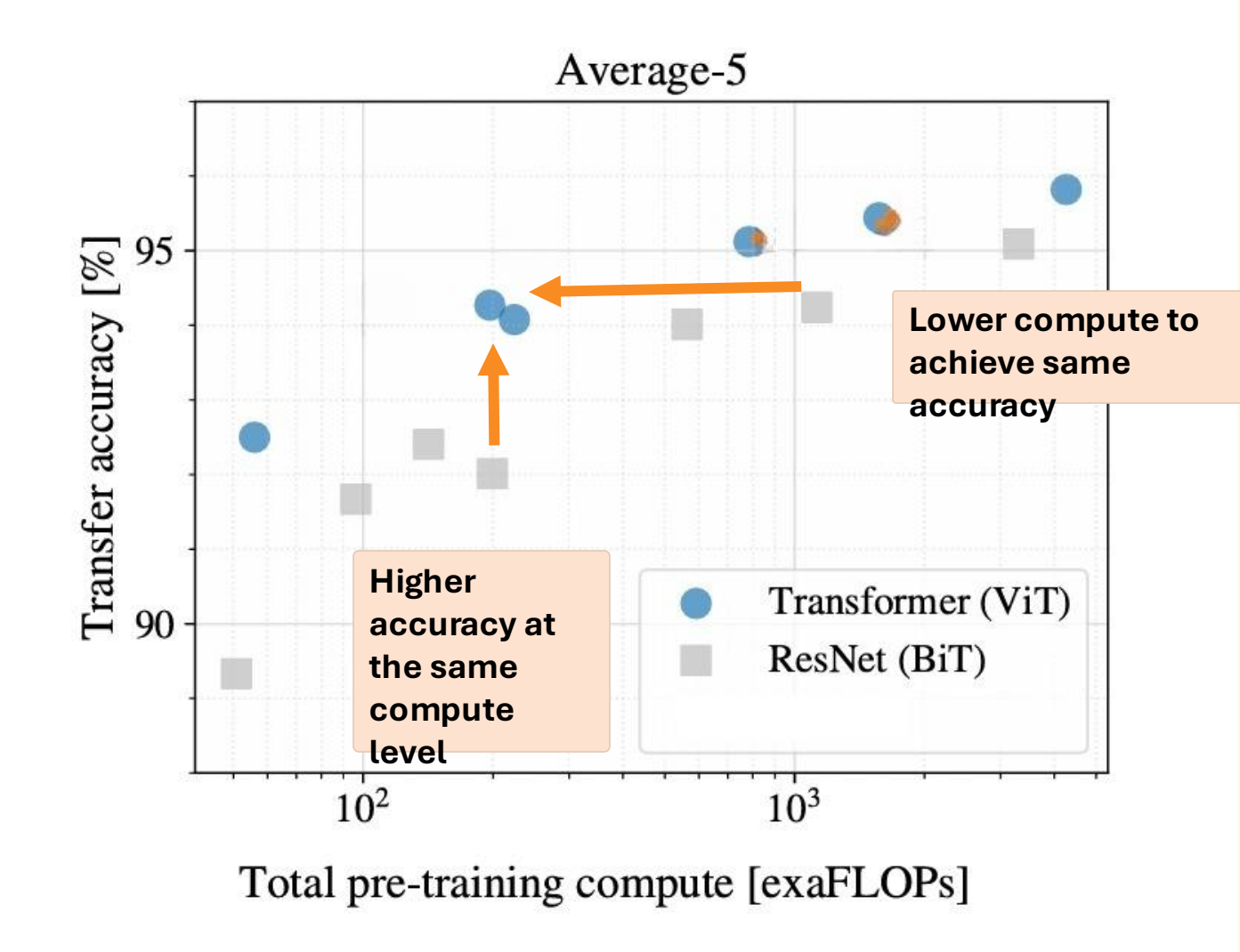
Accuracy % Relative to Compute for Various Models

3. Accuracy % Relative to Compute for Various Models

Model	Pretrained On	Remarks
ResNet (BiT)	Not Applicable	
Vision Transformer (ViT)		
Hybrid		Hybrid model with ResNet CNN output feature map to ViT

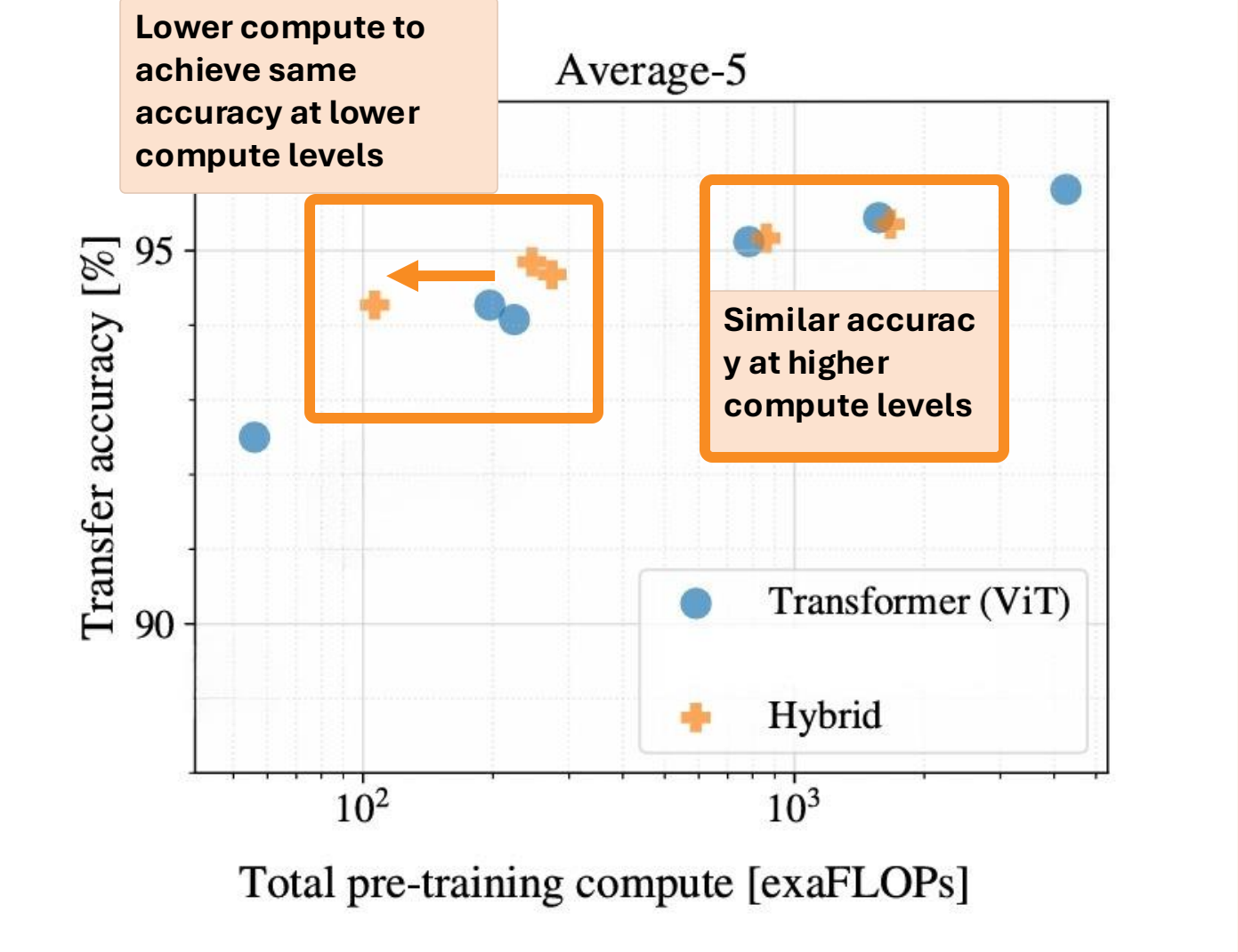
3. Accuracy % Relative to Compute for Various Models

Comparison 1



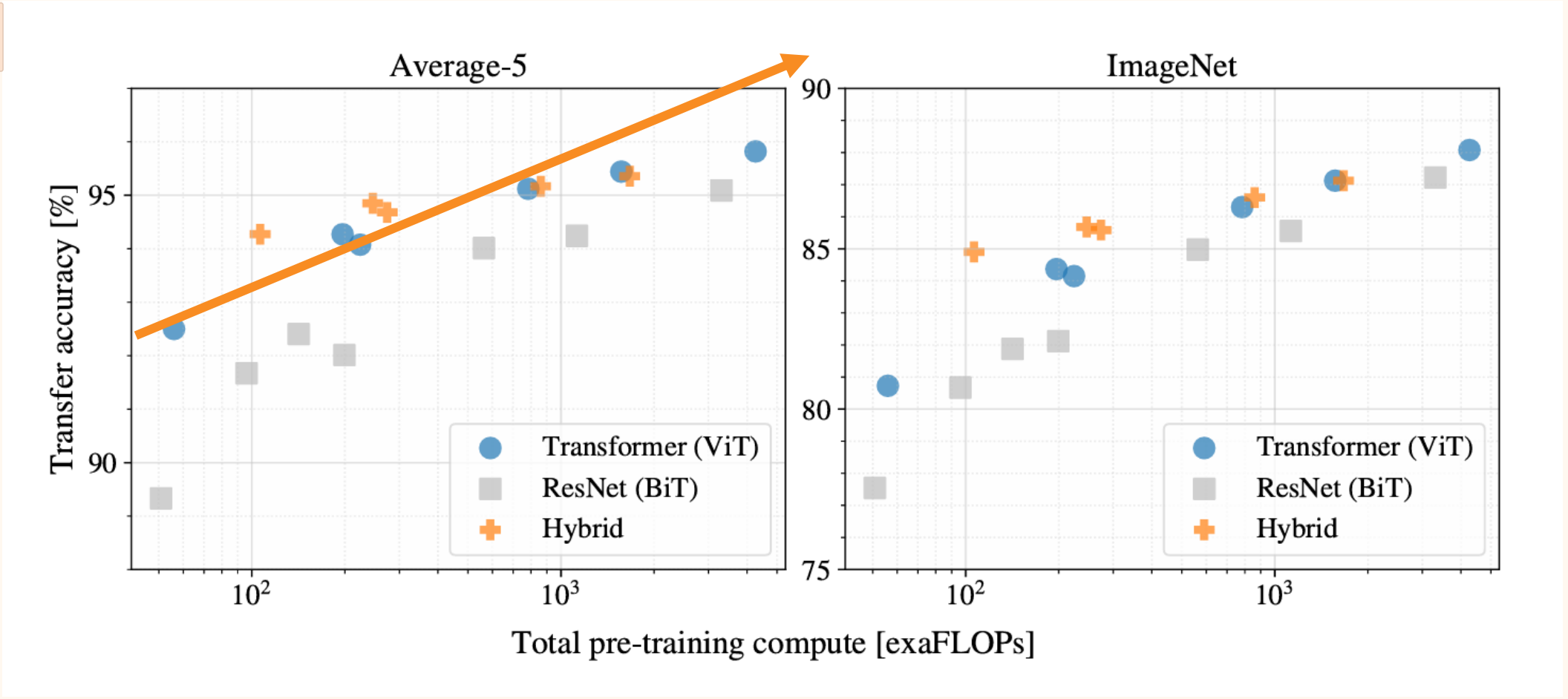
3. Accuracy % Relative to Compute for Various Models

Comparison 2



3. Accuracy % Relative to Compute for Various Models

Comparison 3



Similar increasing trend and pattern for Average-5 & ImageNet dataset
Increasing trend might continue even beyond 1e4

Notes

- Average-5 might be referring to 5 non-ImageNet datasets: CIFAR-10, CIFAR-100, Oxford-IIIT Pets, Oxford Flowers-102, VTAB (19 tasks)

4

Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

4. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

Model	Pretrained On	Finetuned on
ResNet50	Unknown Dataset	ImageNet CIFAR10 CIFAR100 Oxford-IIIT Pets Oxford Flowers-102
ResNet152x2		

4. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

Comparison 1

Dataset	ResNet50 Adam	ResNet50 SGD
ImageNet	77.54	78.24
CIFAR10	97.67	97.46
CIFAR100	86.07	85.17
Oxford-IIIT Pets	91.11	91.00
Oxford Flowers-102	94.26	92.06
Average	89.33	88.79

>

Table 7: Fine-tuning ResNet models pre-trained with Adam and SGD.

4. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

Comparison 2

Dataset	ResNet152x2 Adam	SGD
ImageNet	84.97	84.37
CIFAR10	99.06	99.07
CIFAR100	92.05	91.06
Oxford-IIIT Pets	95.37	94.79
Oxford Flowers-102	98.62	99.32
Average	94.01	93.72

>

Table 7: Fine-tuning ResNet models pre-trained with Adam and SGD.

4. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

Comparison 3

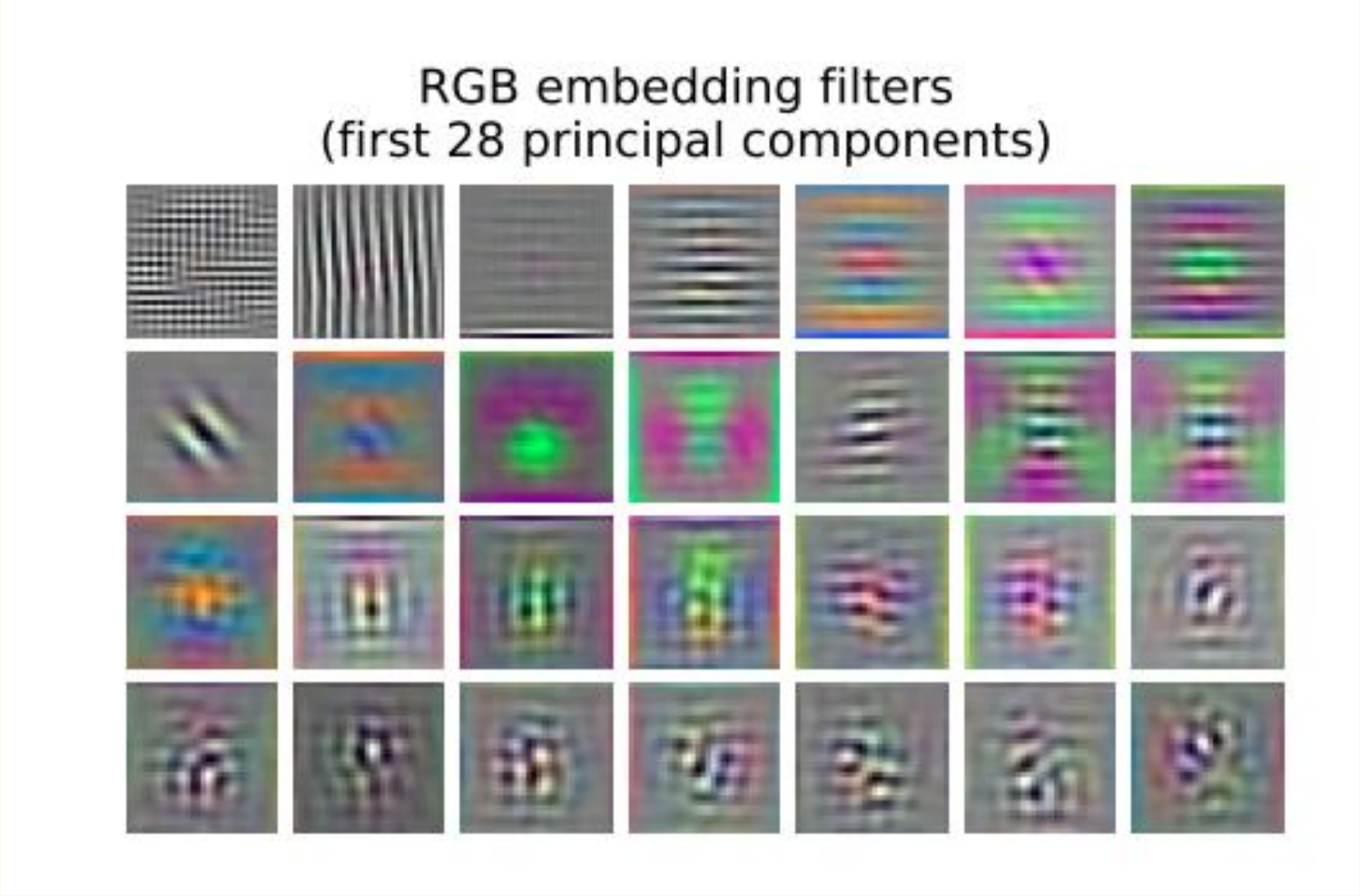
Dataset	ResNet50 Adam		ResNet152x2 Adam
ImageNet	77.54		84.97
CIFAR10	97.67		99.06
CIFAR100	86.07	<	92.05
Oxford-IIIT Pets	91.11		95.37
Oxford Flowers-102	94.26		98.62
Average	89.33		94.01

Table 7: Fine-tuning ResNet models pre-trained with Adam and SGD.

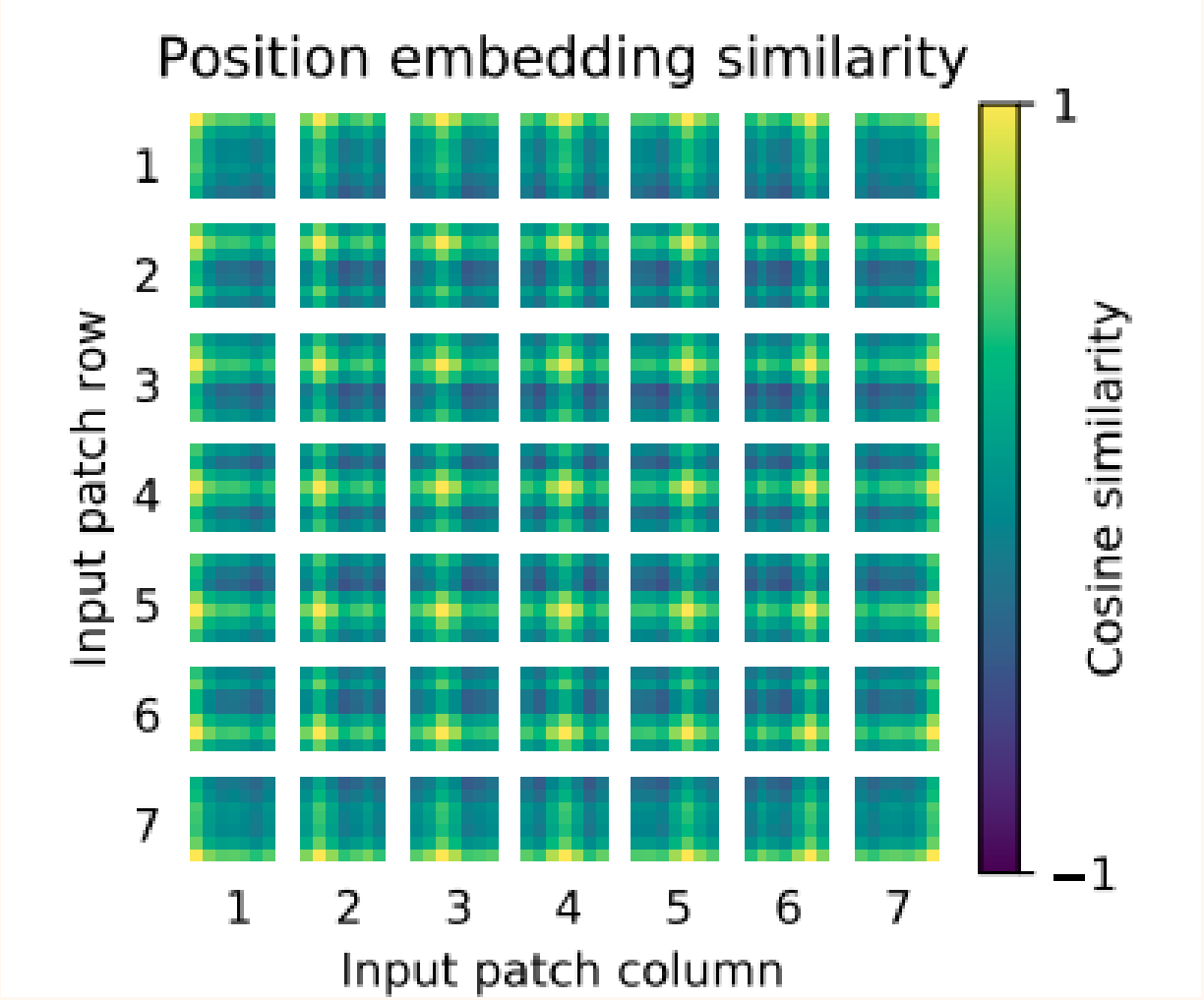
5

Attention Map

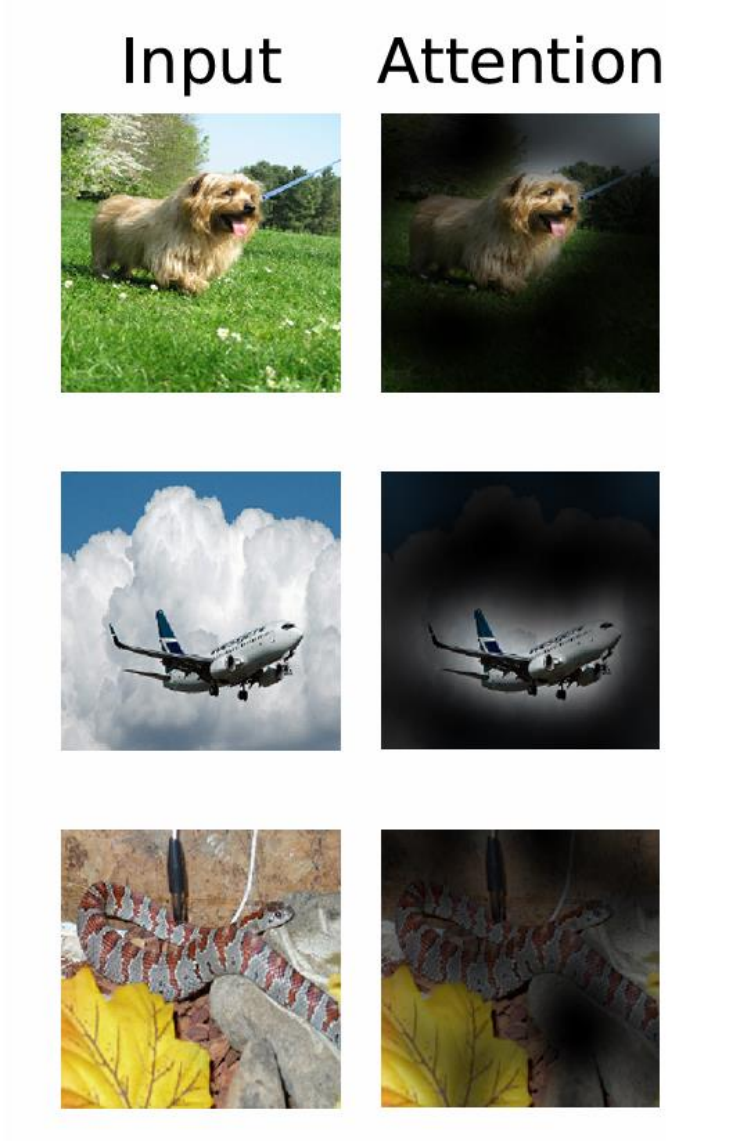
5. Attention Map



5. Attention Map



5. Attention Map



6

Self - Supervision

6. Self - Supervision

Model	Accuracy %	
ViT-B/16	79.90	Self-Supervised Pre-training
	77.9	Training from Scratch
	83.9	Supervised Pre-training

7

Position Embedding, its Dimensions & Where to Add

7. Position Embedding, its Dimensions & Where to Add

Comparison 1

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292

Position embedding increases accuracy

Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

7. Position Embedding, its Dimensions & Where to Add

Comparison 2

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022

1D vs 2D - Not much difference in accuracy

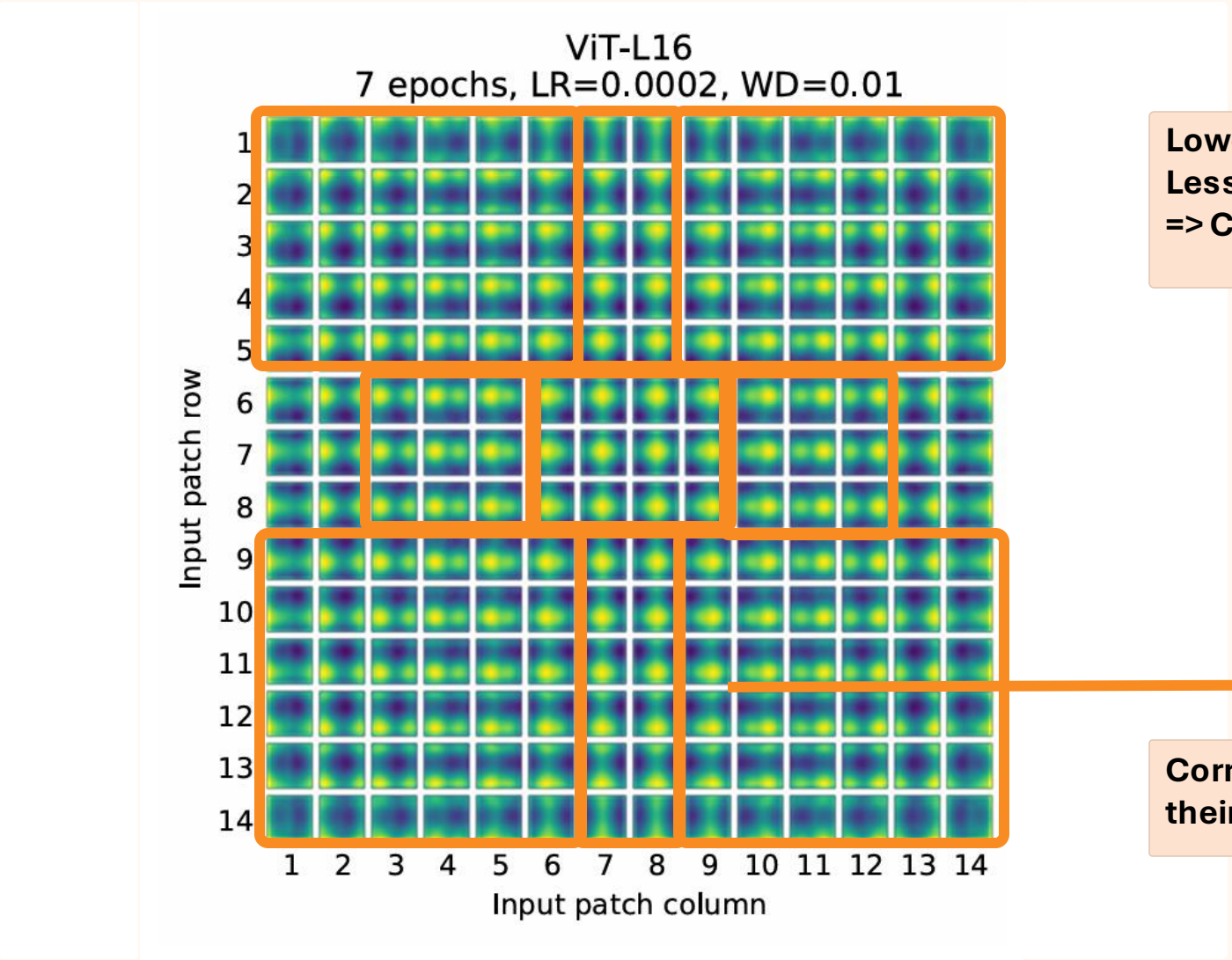
Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

8

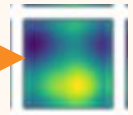
Position Embedding Trained With Different Hyperparameters

8. Position Embedding Trained With Different Hyperparameters

Comparison 1



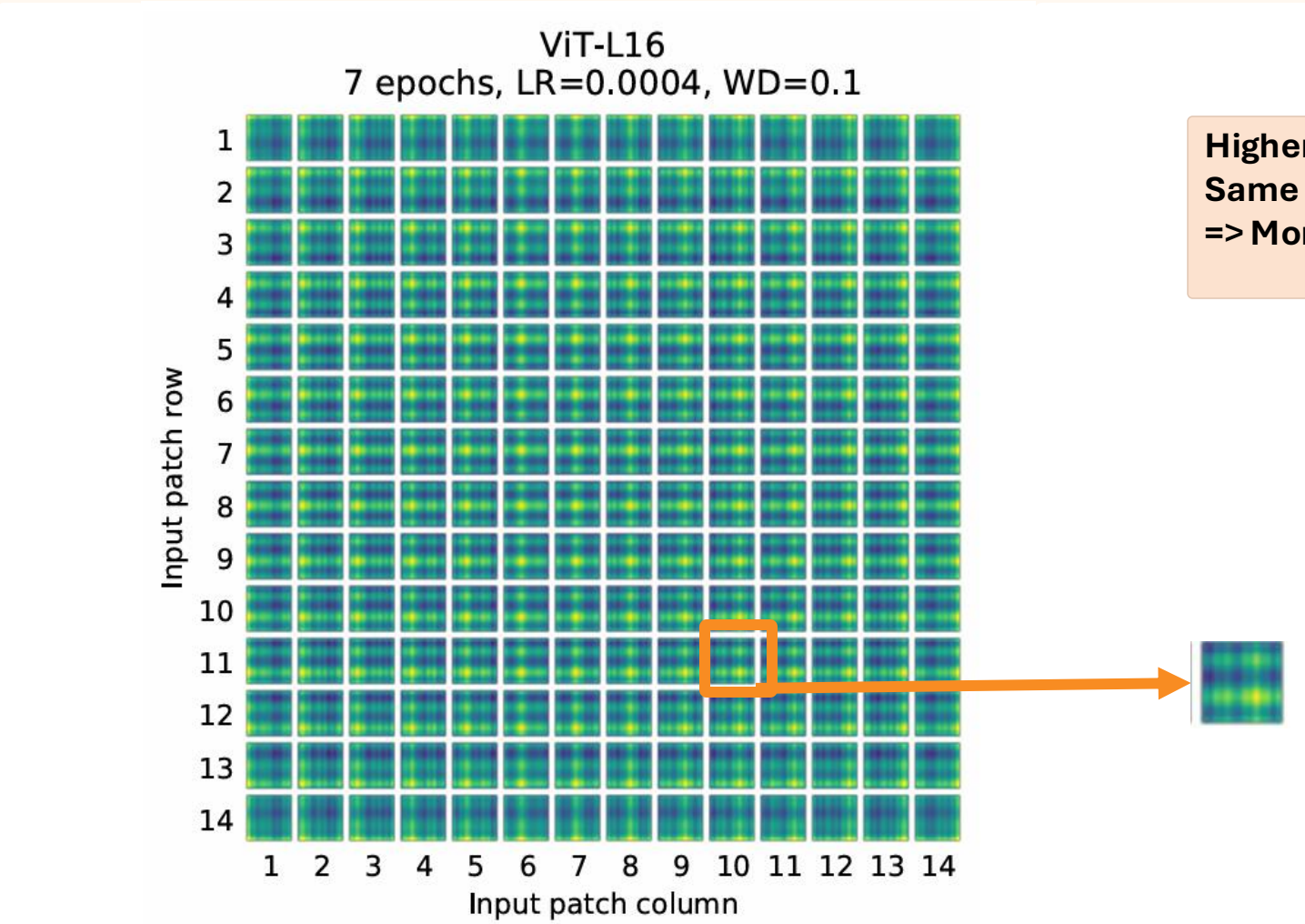
Lower Learning Rate
Less Epochs
=> Clearer Patterns



Correlation is highest for pixels in their respective '9x9 subgrids

8. Position Embedding Trained With Different Hyperparameters

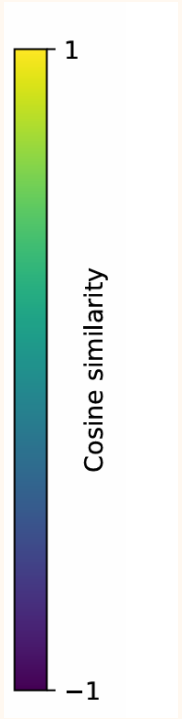
Comparison 2

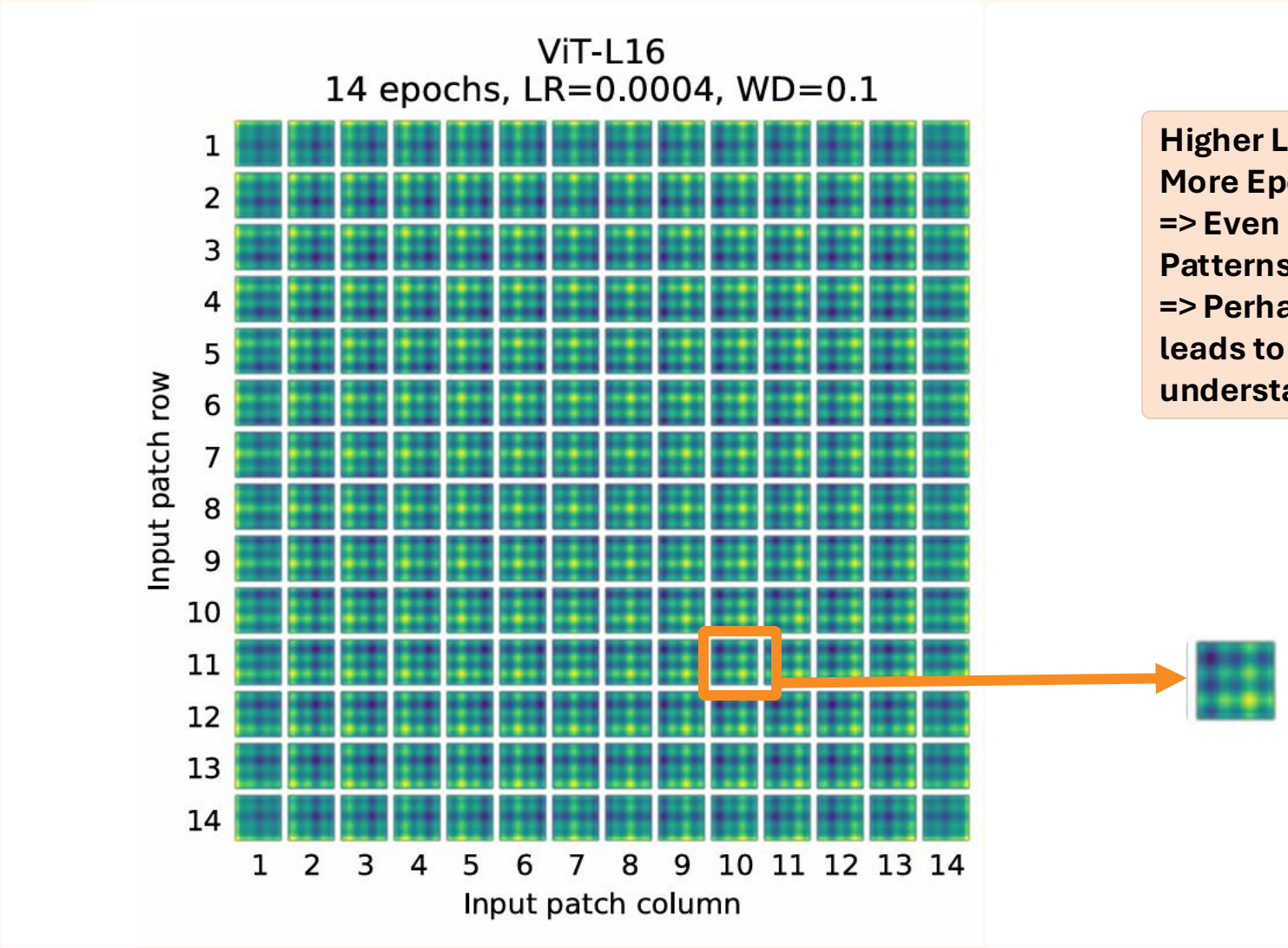


Higher Learning Rate
Same Epochs
=> More Fine-Grained Patterns

8. Position Embedding Trained With Different Hyperparameters

Comparison 3





Higher Learning Rate
More Epochs
=> Even More Fine-Grained Patterns
=> Perhaps more training leads to better understanding?

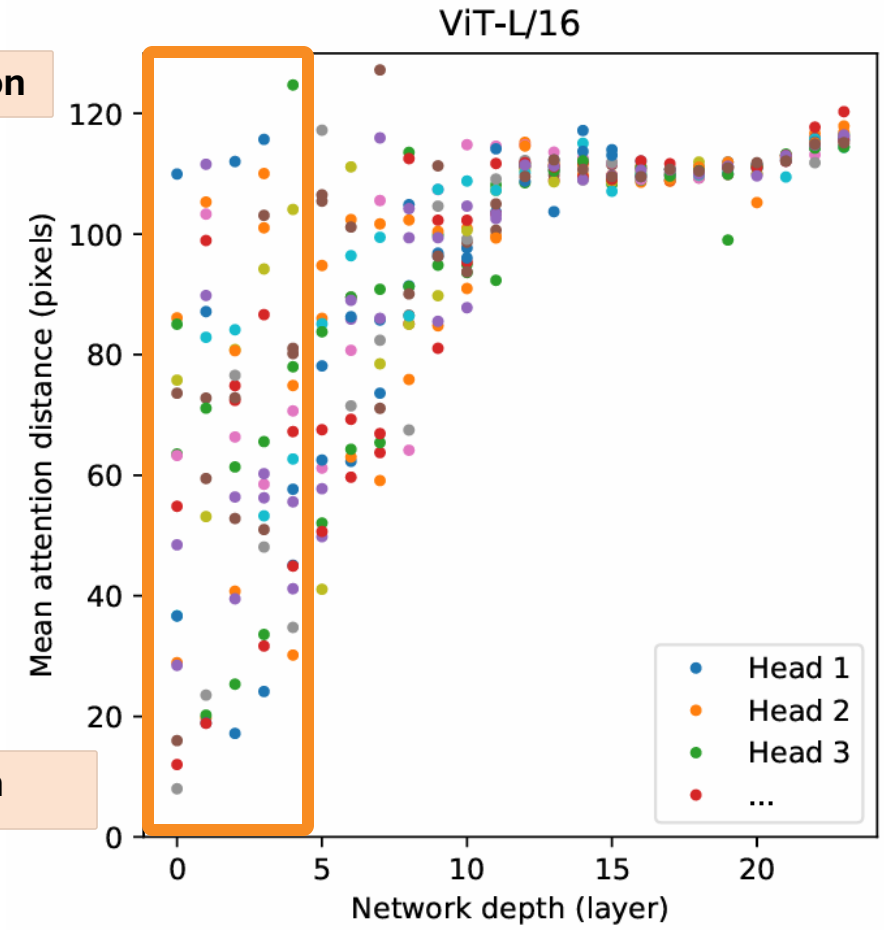
9

Attention Distance at Various Network Depths

9. Attention Distance at Various Network Depths

Comparison 1

Global attention



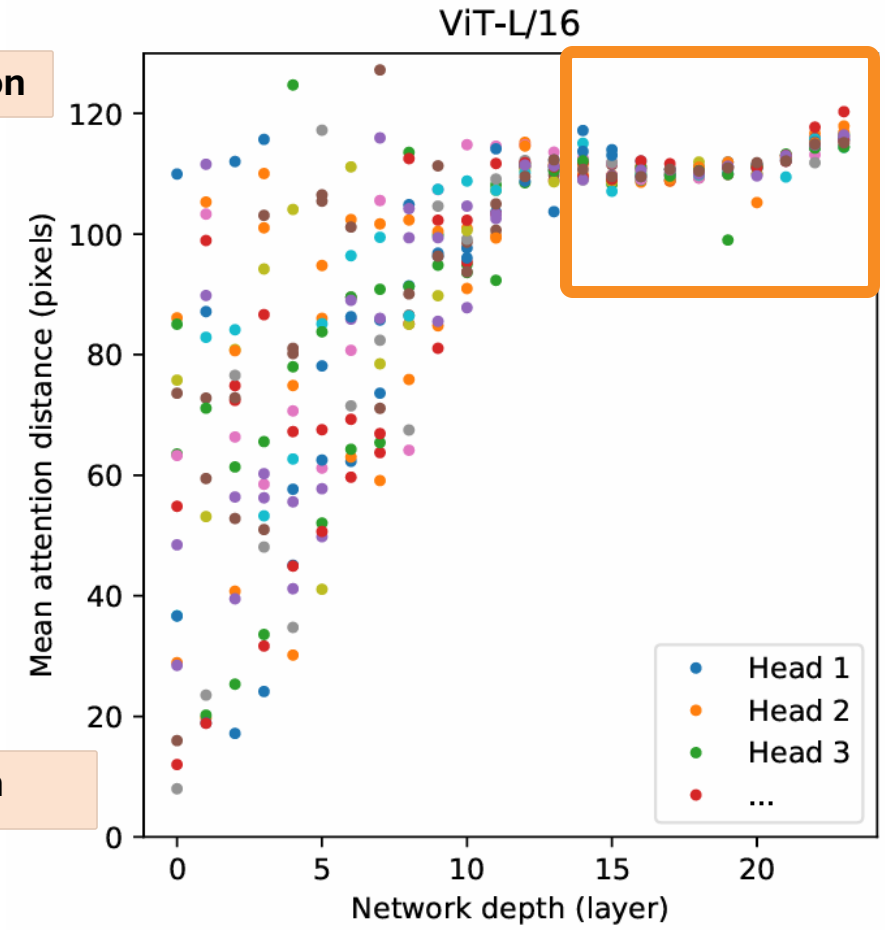
Local attention

Earlier network depths – Attention distance can range from low to high

9. Attention Distance at Various Network Depths

Comparison 2

Global attention



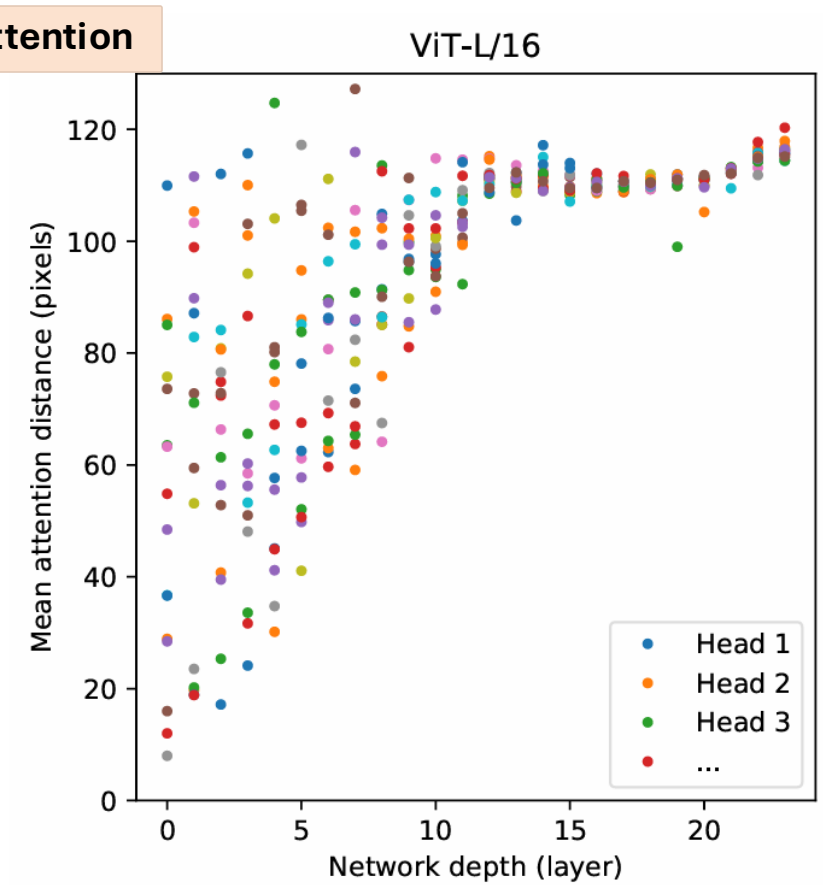
Local attention

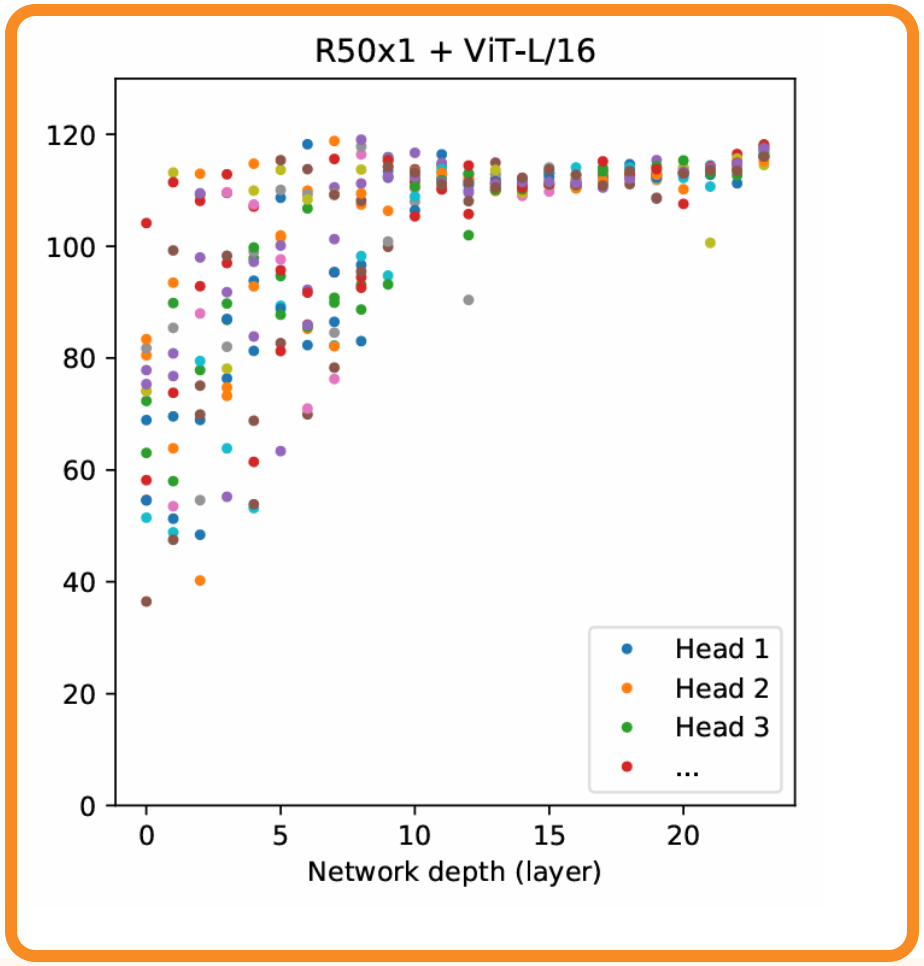
Deeper layers – Attention heads focus on global attention

9. Attention Distance at Various Network Depths

Comparison 3

Global attention





Local attention

Similar phenomenon

10

Batch Size for Models at Various Input Sizes

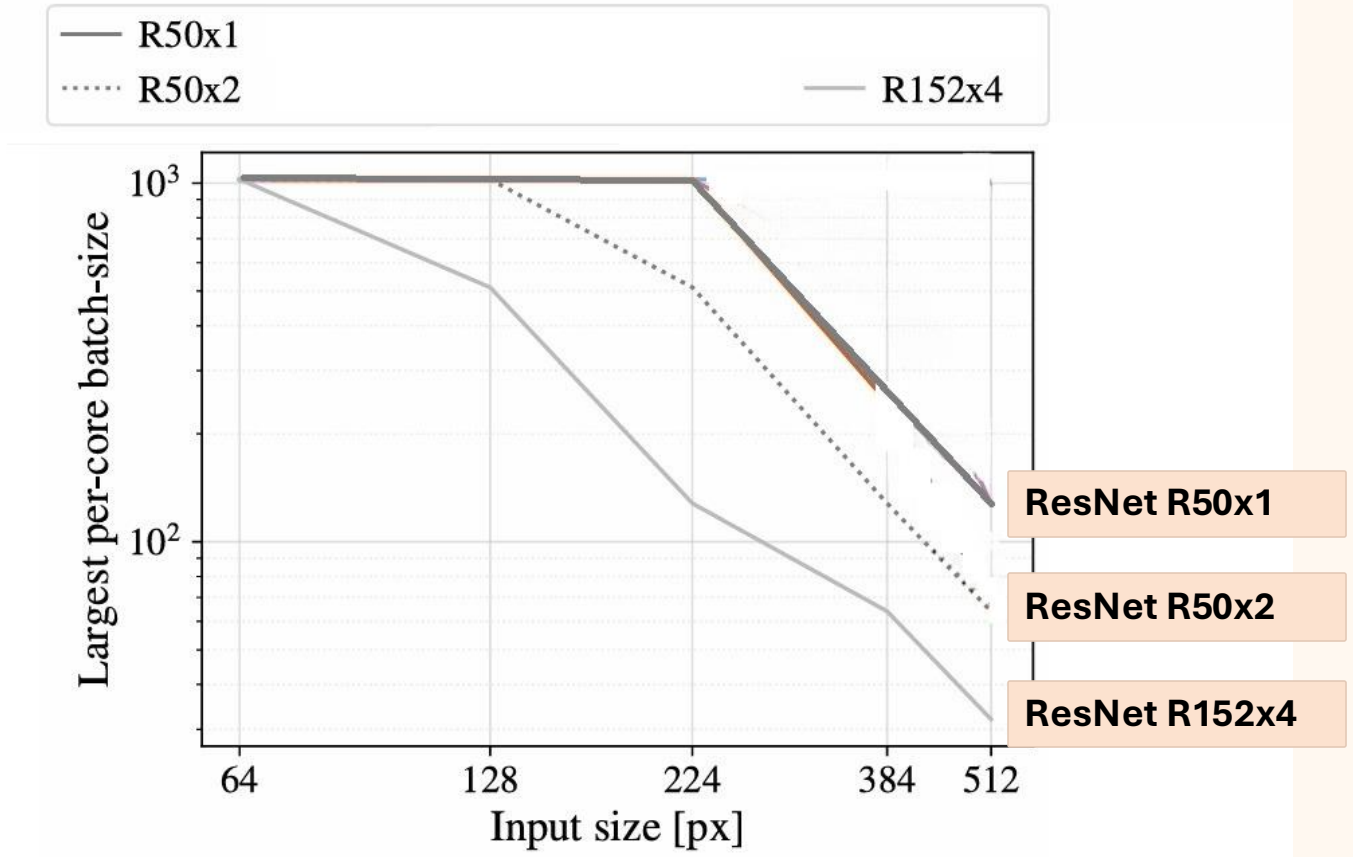
10. Batch Size for Models at Various Input Sizes

Model	Pretrained On	Remarks
ResNet R50x1	Unknown	
ResNet R50x2		
ResNet R152x4		
ViT-B/16		Base model
ViT-B/32		Base model with lower resolution inputs
ViT-H/14		Huge model, bigger than Large model

10. Batch Size for Models at Various Input Sizes

Comparison 1

Between ResNets



ResNet R50x1

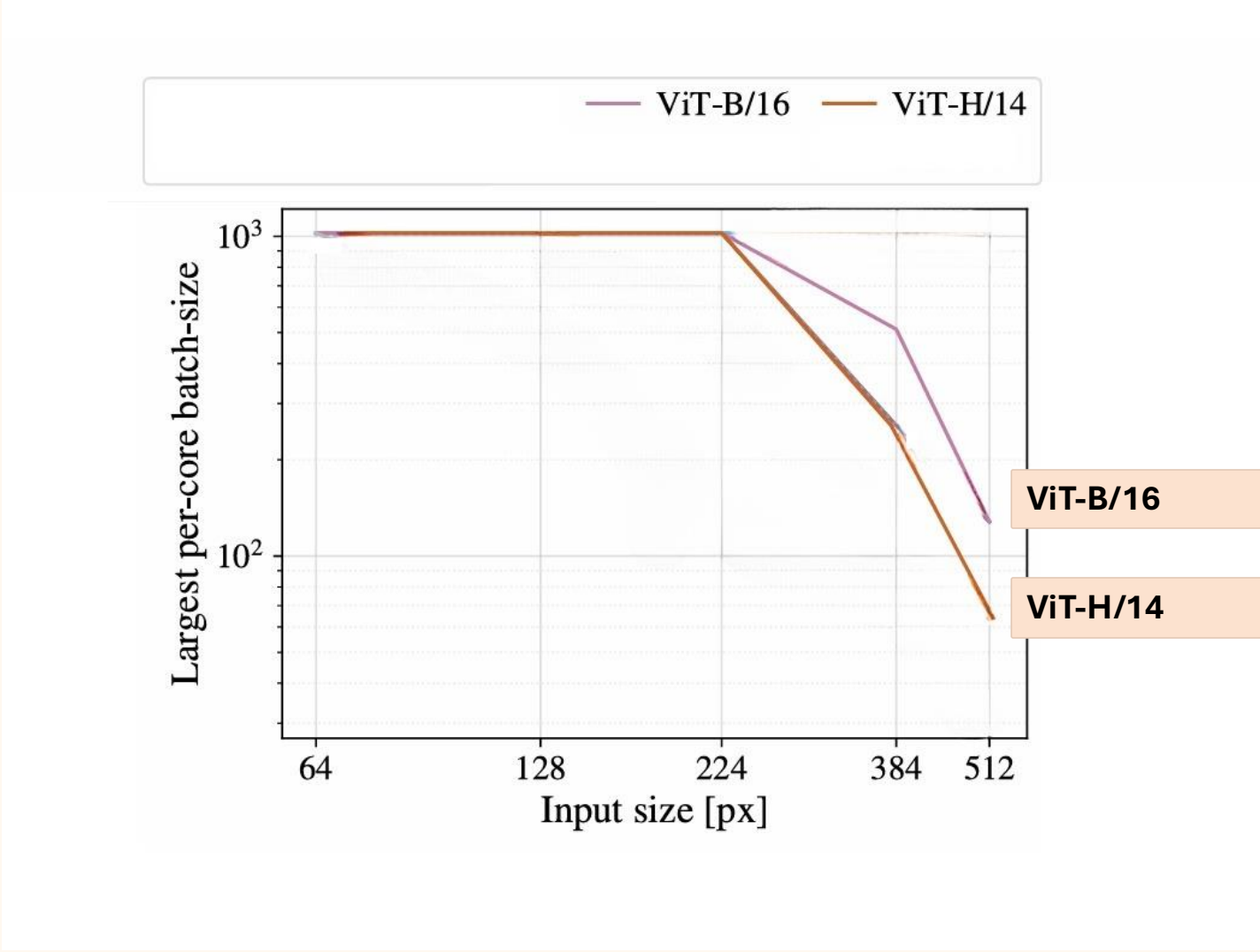
ResNet R50x2

ResNet R152x4

Larger Model, Lower Batch Size

10. Batch Size for Models at Various Input Sizes

Comparison 2



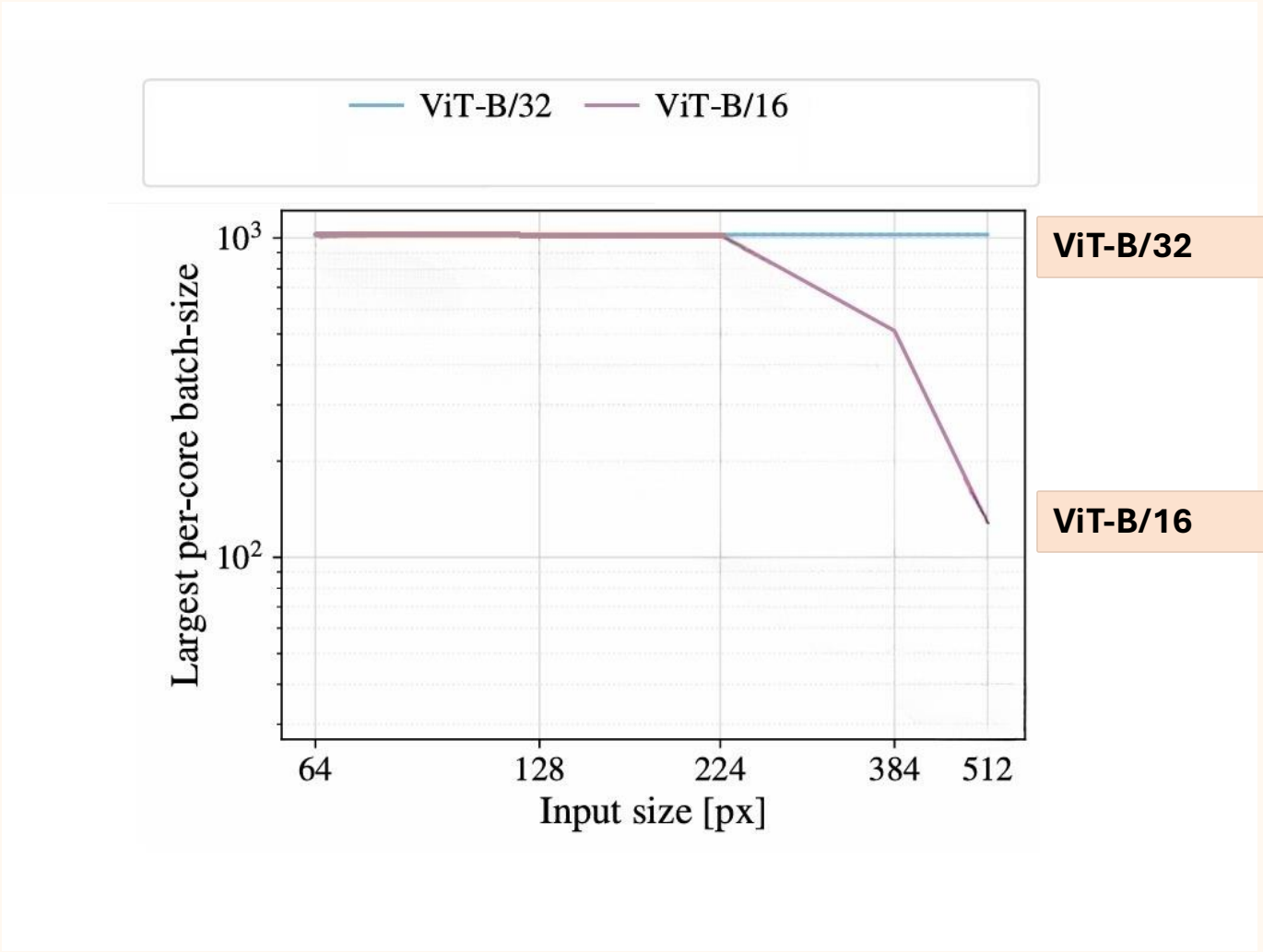
ViT-B/16

ViT-H/14

Larger Model, Lower Batch Size

10. Batch Size for Models at Various Input Sizes

Comparison 3



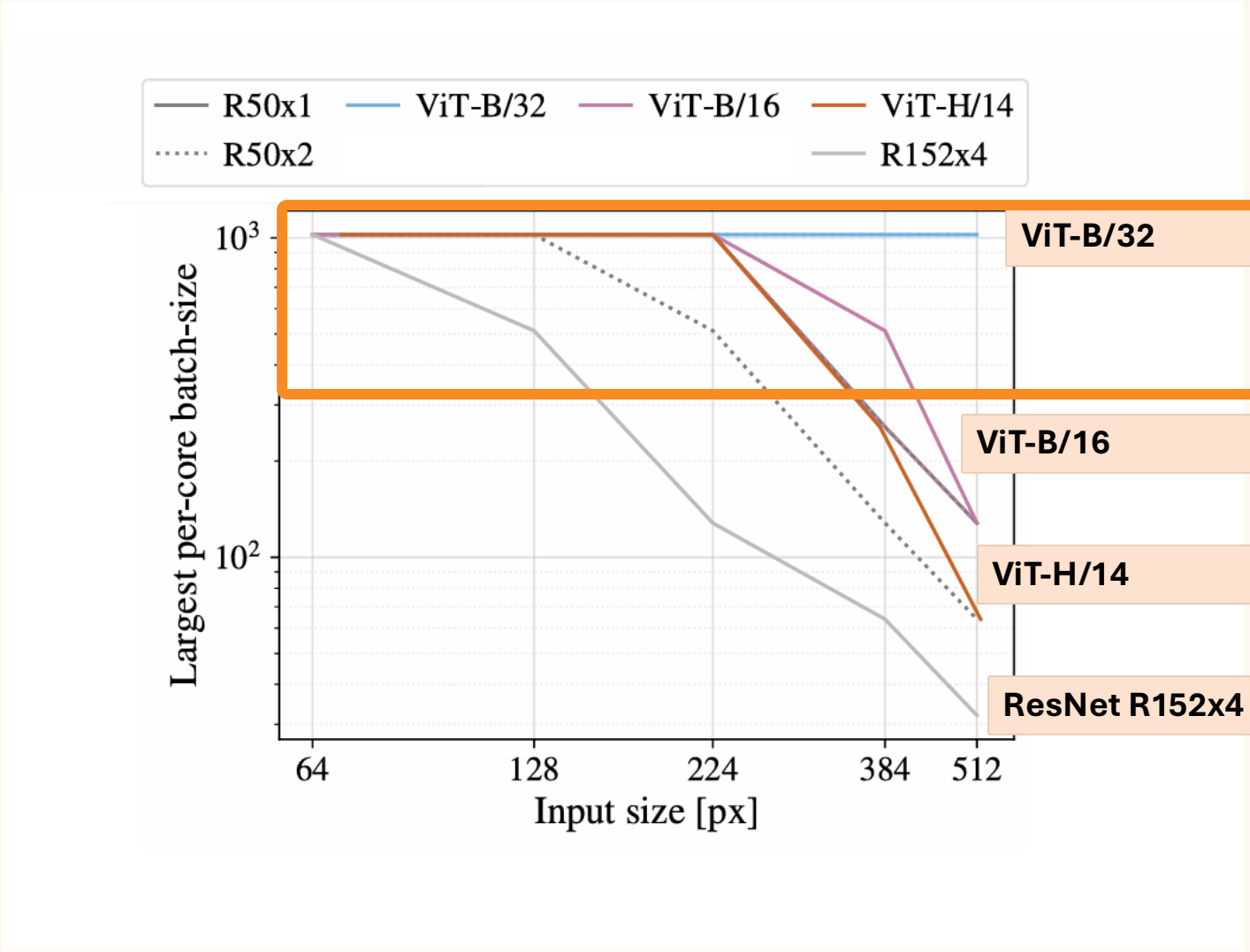
ViT-B/32

ViT-B/16

More Input Patches & at a Higher Resolution => Lower Batch Size

10. Batch Size for Models at Various Input Sizes

Comparison 4



Other Results: Feel Free to Check Out From the Paper

Transformer shape

Head Type and class token

Axial Attention

ObjectNet Results