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
2. Top1 Finetune Accuracy % After Pretraining on Different Datasets
3. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT
4. Accuracy % Relative to Compute for Various Models
5. Finetune Accuracy % for ResNet using Adam vs SGD Optimiser
6. **Attention Map**
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

Vision Transformers / ViT (2021) Results Summary

#	Topic	Key Results
1	**Finetune Accuracy** % After Pretraining on Different Datasets	**ViT-L/16** pretrained on JFT-300M **≈ ResNet Bit-L** **ViT-H/14** pretrained on JFT-300M **> ResNet Bit-L**
2	Top1 Finetune Accuracy % After Pretraining on Different Datasets	ViT-B or L/**16 (higher resolution inputs) is better than 32** **ViT-H/14 > ViT-L/16 > ViT-B/16**
3	ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT	As pretraining n_examples increase -> accuracy increases **ViT-L/16 > ResNet**152x2 BiT **when n_examples are >= 100M**
4	Accuracy % Relative to Compute for Various Models	ViT vs ResNet – ViT **uses lesser compute** to achieve same accuracy
5	Finetune Accuracy % for ResNet using Adam vs SGD Optimiser	Adam optimiser > SGD optimiser
6	**Attention Map**	Visualise how ViT model focuses on object position
7	**Position Embedding**, its Dimensions & Where to Add	**Position embedding increases accuracy** 1D vs 2D embedding – No difference Where we add position embedding – No difference
8	Position Embedding Trained With Different Hyperparameters	**Higher learning rate or more training epochs** **=> More fine-grained patterns** => Perhaps better understanding?
9	**Attention Distance** at Various Network Depths	**Earlier** network depths – Attention distance can **range from low to high** **Deeper** layers – Attention heads **focus on global attention**
10	**Batch Size** for Models at Various Input Sizes	**ViT-B/32 is so memory efficient** that it can maintain high batch_size

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

Comparison 1

	Ours-I21k (ViT-L/16)		BiT-L (ResNet152x4)		Noisy Student (EfficientNet-L2)
ImageNet	85.30 ± 0.02	<	87.54 ± 0.02	<	$88.4/88.5^*$
ImageNet ReaL	88.62 ± 0.05		90.54		90.55
CIFAR-10	99.15 ± 0.03		99.37 ± 0.06		–
CIFAR-100	93.25 ± 0.05		93.51 ± 0.08		–
Oxford-IIIT Pets	94.67 ± 0.15		96.62 ± 0.23		–
Oxford Flowers-102	99.61 ± 0.02		99.63 ± 0.03		–
VTAB (19 tasks)	72.72 ± 0.21		76.29 ± 1.70		–

1. Finetune Accuracy % After Pretraining on Different Datasets

Comparison 2

	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

Comparison 3

	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 4 – Higher Accuracy using Lesser Compute

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

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

Comparison 1

		ViT-B/16		ViT-B/32
ImageNet	CIFAR-10	98.13		97.77
	CIFAR-100	87.13		86.31
	ImageNet	77.91		73.38
	ImageNet ReaL	83.57	>	79.56
	Oxford Flowers-102	89.49		85.43
	Oxford-IIIT-Pets	93.81		92.04
ImageNet-21k	CIFAR-10	98.95		98.79
	CIFAR-100	91.67		91.97
	ImageNet	83.97		81.28
	ImageNet ReaL	88.35	>	86.63
	Oxford Flowers-102	99.38		99.11
	Oxford-IIIT-Pets	94.43		93.02
JFT-300M	CIFAR-10	99.00		98.61
	CIFAR-100	91.87		90.49
	ImageNet	84.15	>	80.73
	ImageNet ReaL	88.85		86.27
	Oxford Flowers-102	99.56		99.27
	Oxford-IIIT-Pets	95.80		93.40

2. Top1 Finetune Accuracy % After Pretraining on Different Datasets

Comparison 2

		ViT-L/16	ViT-L/32
ImageNet	CIFAR-10	97.86	97.94
	CIFAR-100	86.35	87.07
	ImageNet	76.53	71.16
	ImageNet ReaL	82.19	77.83
	Oxford Flowers-102	89.66	86.36
	Oxford-IIIT-Pets	93.64	91.35
ImageNet-21k	CIFAR-10	99.16	99.13
	CIFAR-100	93.44	93.04
	ImageNet	85.15	80.99
	ImageNet ReaL	88.40	85.65
	Oxford Flowers-102	99.61	99.19
	Oxford-IIIT-Pets	94.73	93.09
JFT-300M	CIFAR-10	99.38	99.19
	CIFAR-100	94.04	92.52
	ImageNet	87.12	84.37
	ImageNet ReaL	89.99	88.28
	Oxford Flowers-102	99.56	99.45
	Oxford-IIIT-Pets	97.11	95.83

2. Top1 Finetune Accuracy % After Pretraining on Different Datasets

Comparison 3

		ViT-B/16		ViT-L/16
ImageNet	CIFAR-10	98.13		97.86
	CIFAR-100	87.13		86.35
	ImageNet	77.91	<	76.53
	ImageNet ReaL	83.57		82.19
	Oxford Flowers-102	89.49		89.66
	Oxford-IIIT-Pets	93.81		93.64
ImageNet-21k	CIFAR-10	98.95		99.16
	CIFAR-100	91.67		93.44
	ImageNet	83.97	<	85.15
	ImageNet ReaL	88.35		88.40
	Oxford Flowers-102	99.38		99.61
	Oxford-IIIT-Pets	94.43		94.73
JFT-300M	CIFAR-10	99.00		99.38
	CIFAR-100	91.87		94.04
	ImageNet	84.15	<	87.12
	ImageNet ReaL	88.85		89.99
	Oxford Flowers-102	99.56		99.56
	Oxford-IIIT-Pets	95.80		97.11

2. Top1 Finetune Accuracy % After Pretraining on Different Datasets

Comparison 4

		ViT-L/16		ViT-H/14
ImageNet	CIFAR-10	97.86		-
	CIFAR-100	86.35		-
	ImageNet	76.53		-
	ImageNet ReaL	82.19		-
	Oxford Flowers-102	89.66		-
	Oxford-IIIT-Pets	93.64		-
ImageNet-21k	CIFAR-10	99.16		99.27
	CIFAR-100	93.44		93.82
	ImageNet	85.15		85.13
	ImageNet ReaL	88.40	<	88.70
	Oxford Flowers-102	99.61		99.51
	Oxford-IIIT-Pets	94.73		94.82
JFT-300M	CIFAR-10	99.38		99.50
	CIFAR-100	94.04		94.55
	ImageNet	87.12	<	88.04
	ImageNet ReaL	89.99		90.33
	Oxford Flowers-102	99.56		99.68
	Oxford-IIIT-Pets	97.11		97.56

3

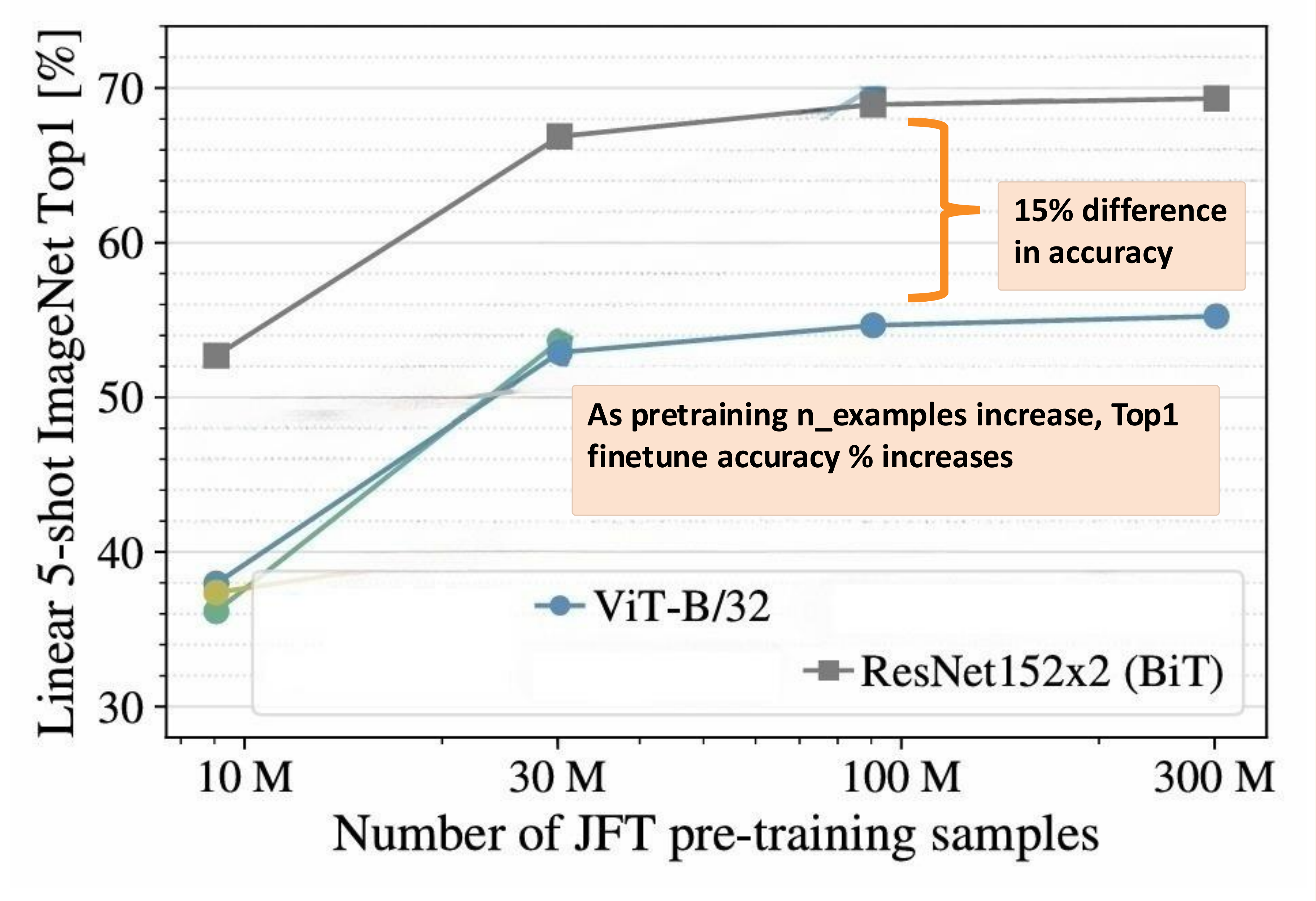
ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT

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

Model	Pretrained On	Remarks
ResNet152x2 (BiT)	JFT-300M (Google proprietary)	
ViT-B/32		Base model pretrained on **lower resolution input images**
ViT-L/32		Large model pretrained on **lower resolution input images**
ViT-L/16		**Large model**

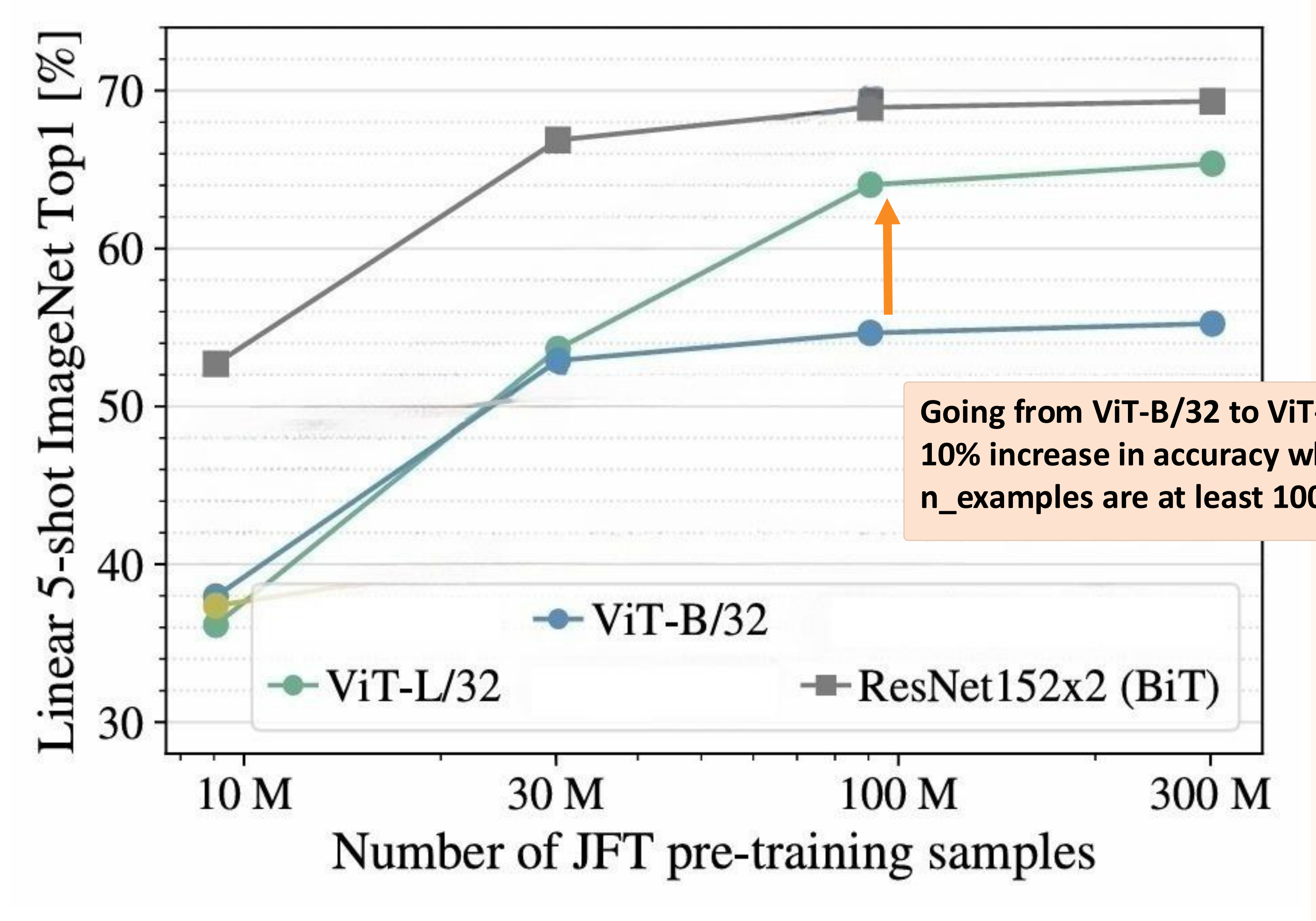
3. ImageNet Top1 Finetune Accuracy % After Pretraining on Various n_examples on JFT

Comparison 1



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

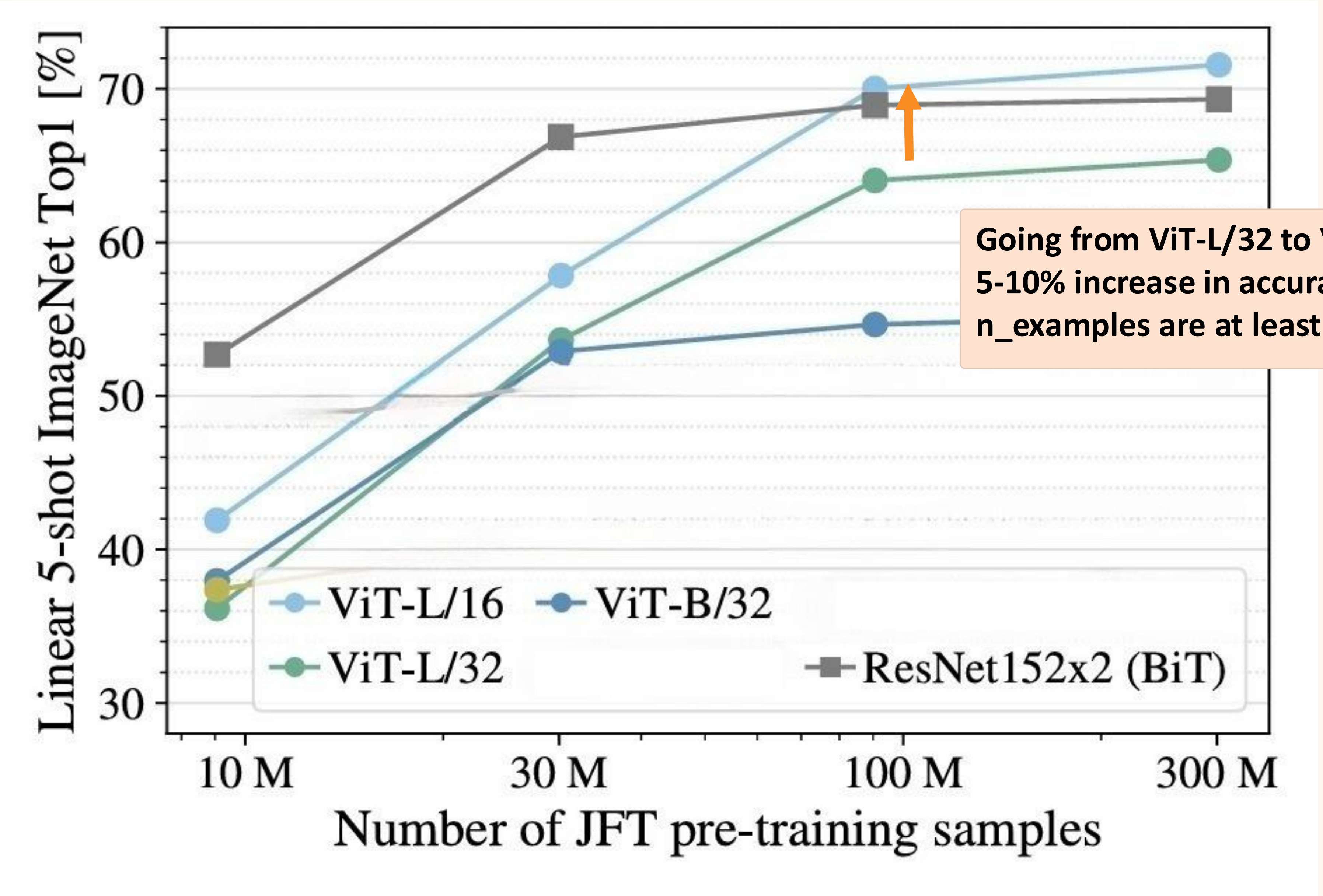
Comparison 2



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

3. 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

4

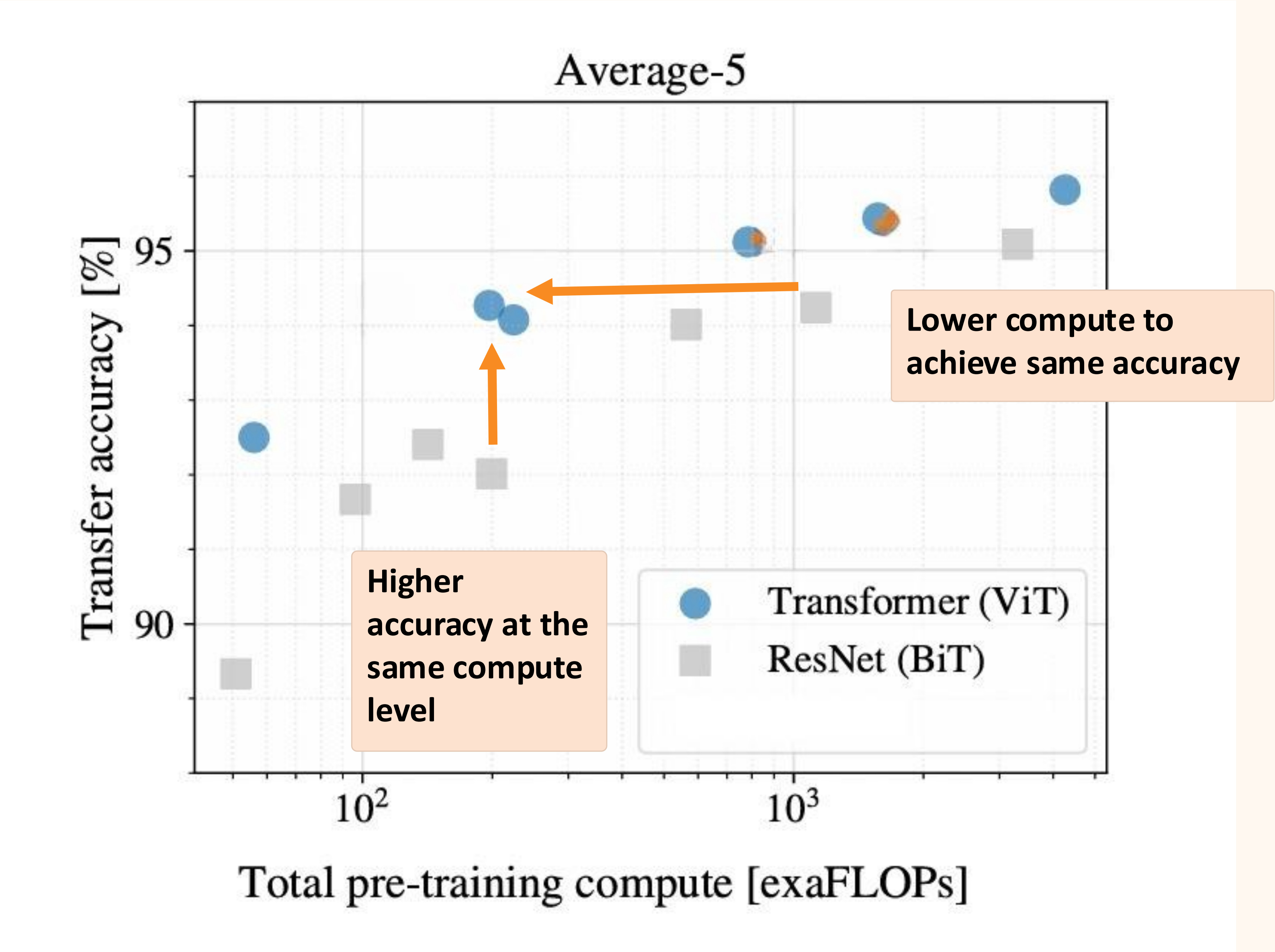
Accuracy % Relative to Compute for Various Models

4. 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

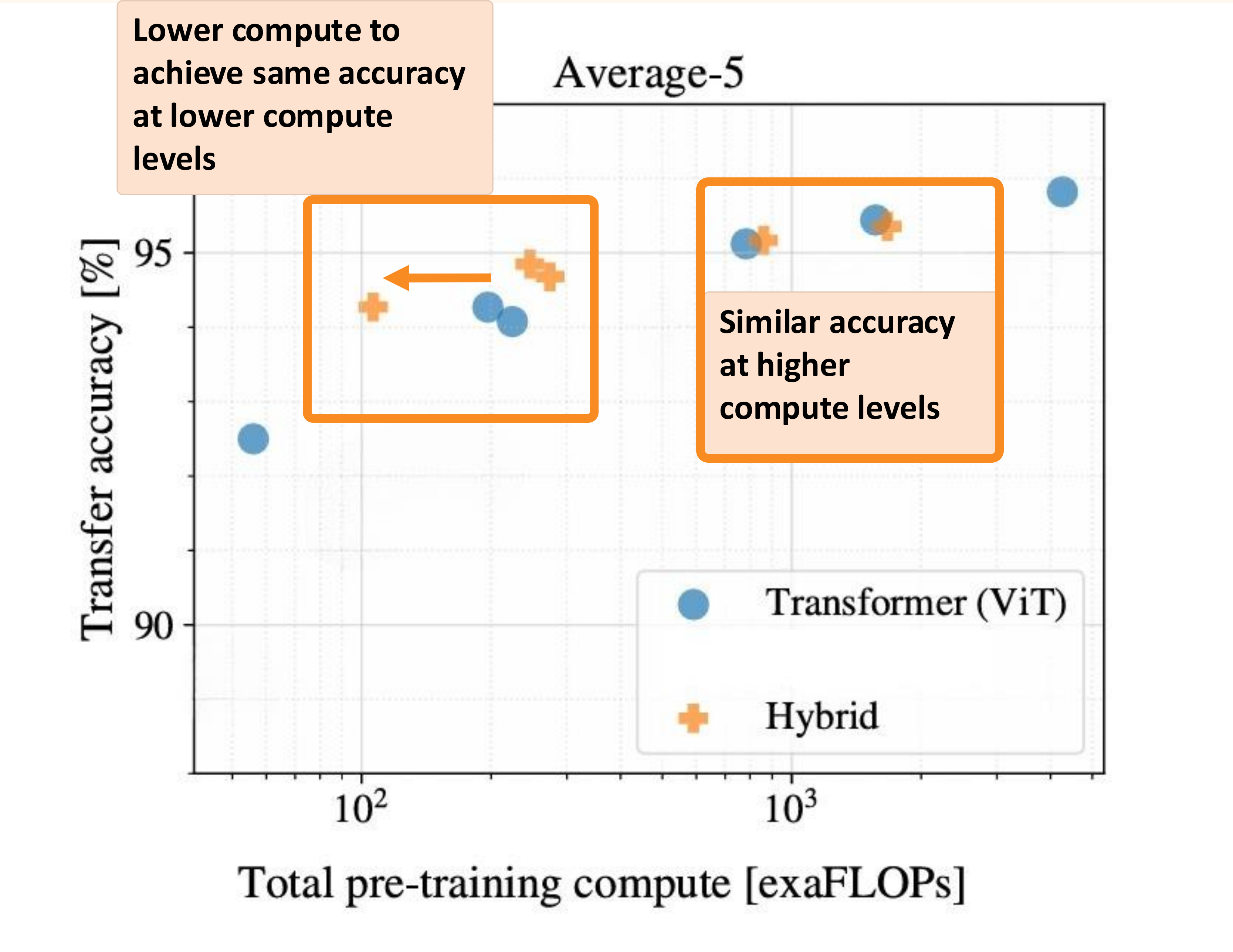
4. Accuracy % Relative to Compute for Various Models

Comparison 1



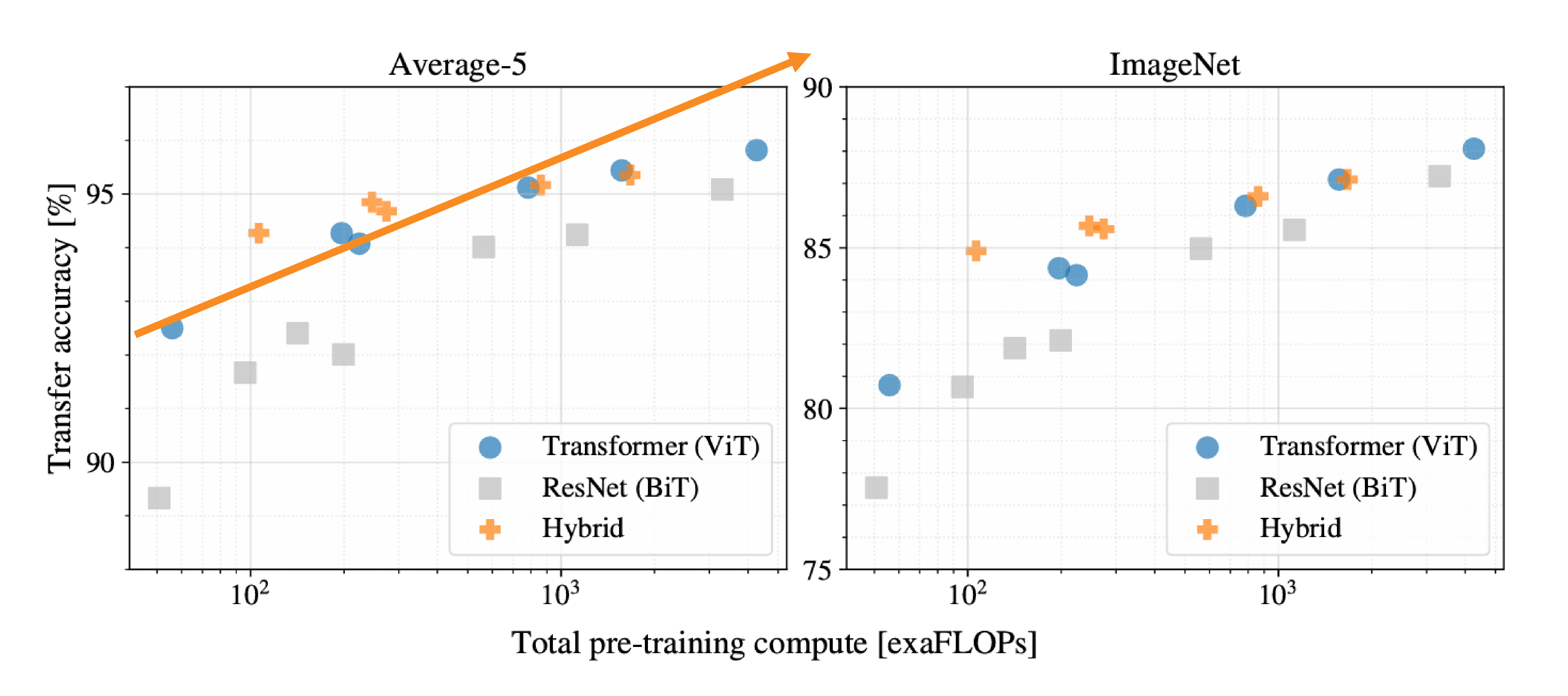
4. Accuracy % Relative to Compute for Various Models

Comparison 2



4. 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)

5

Finetune Accuracy % for ResNet using Adam vs SGD Optimiser

5. 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		

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

Comparison 1

	ResNet50	
Dataset	Adam	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.

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

Comparison 2

Dataset	ResNet152x2 Adam	ResNet152x2 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.

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

Comparison 3

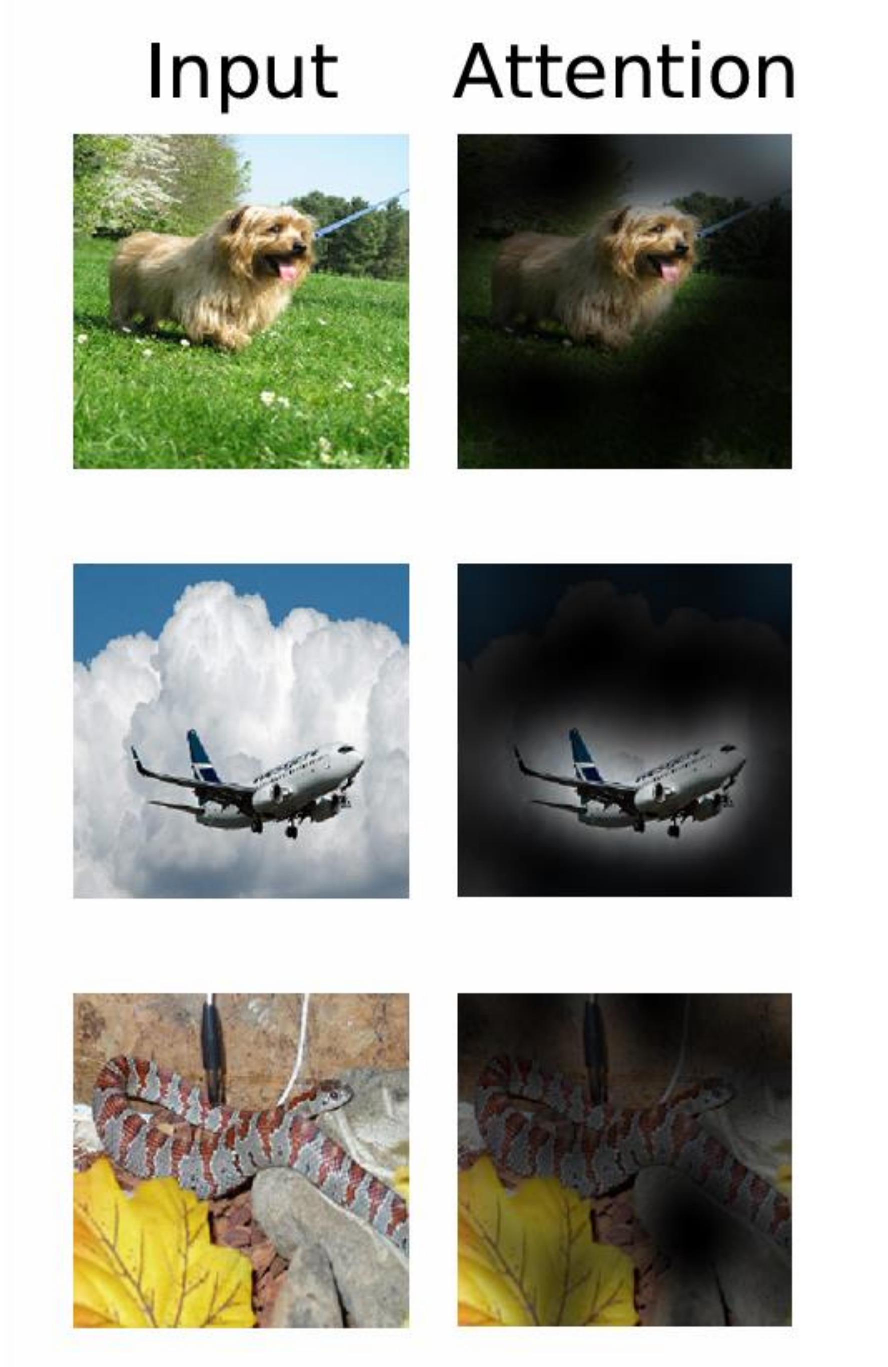
	ResNet50		ResNet152x2
Dataset	Adam		Adam
ImageNet	77.54		84.97
CIFAR10	97.67		99.06
CIFAR100	86.07	<	92.05
Oxford-IIIT Pets	91.11		95.37
Oxford Flowers-102	94.26		98.62
Average	89.33		94.01

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

6

Attention Map

6. Attention Map



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

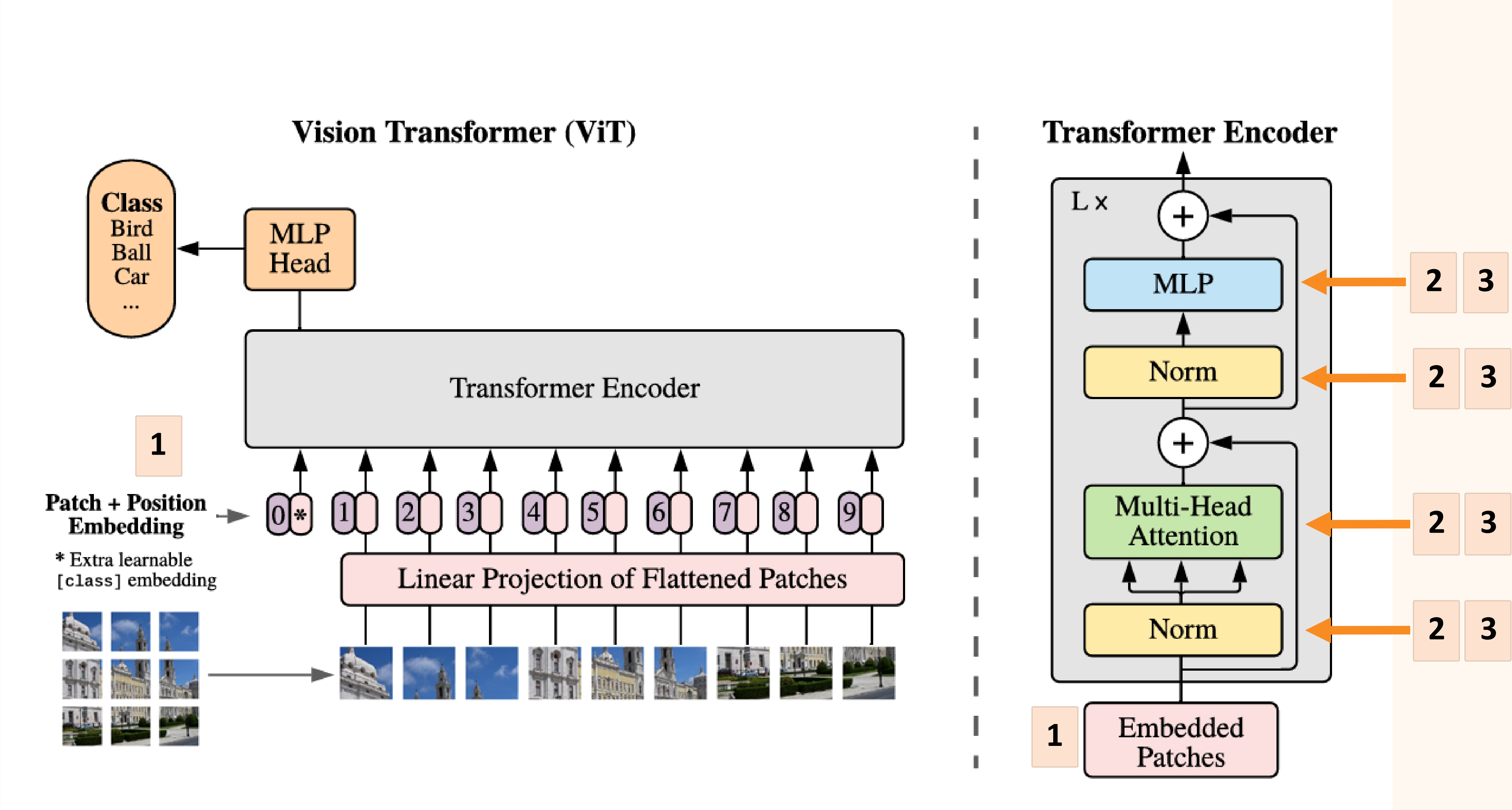
1D vs 2D - Not much difference in accuracy

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

7. Position Embedding, its Dimensions & Where to Add

Comparison 3

1	2	3
Input to encoder	Input separately to each layer	Input 1 common position embedding to each layer



7. Position Embedding, its Dimensions & Where to Add

Comparison 3

	1	**2**	**3**
	Input to encoder	**Input separately to each layer**	**Input 1 common position embedding to each layer**

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

Where we add position embedding - Not much difference in accuracy

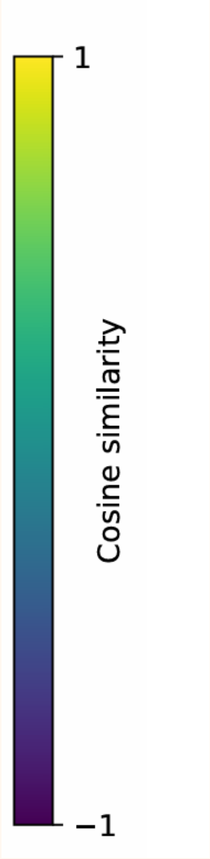
Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

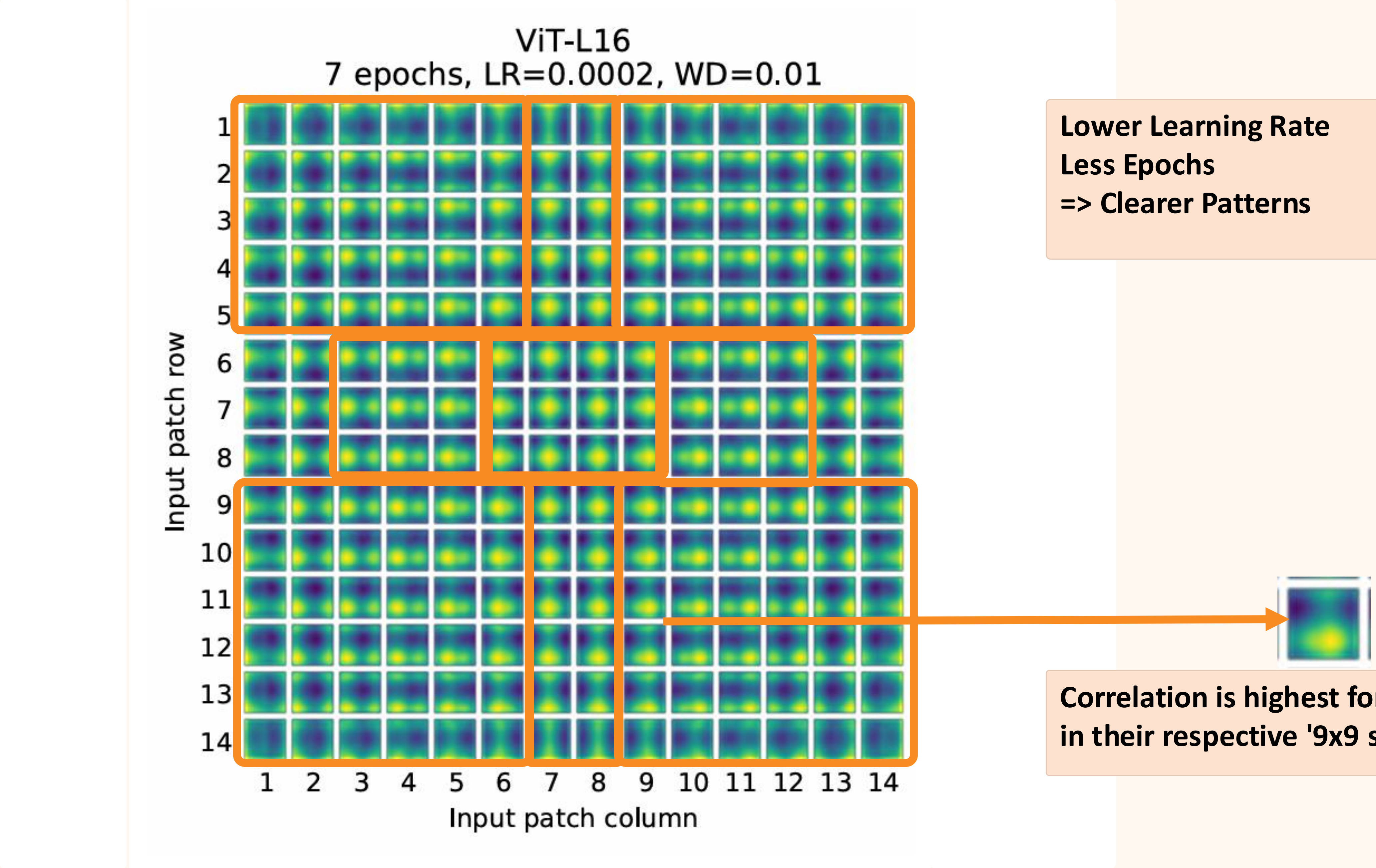
8

Position Embedding Trained With Different Hyperparameters

8. Position Embedding Trained With Different Hyperparameters

Comparison 1



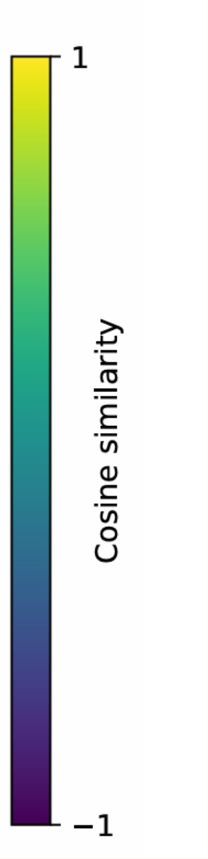


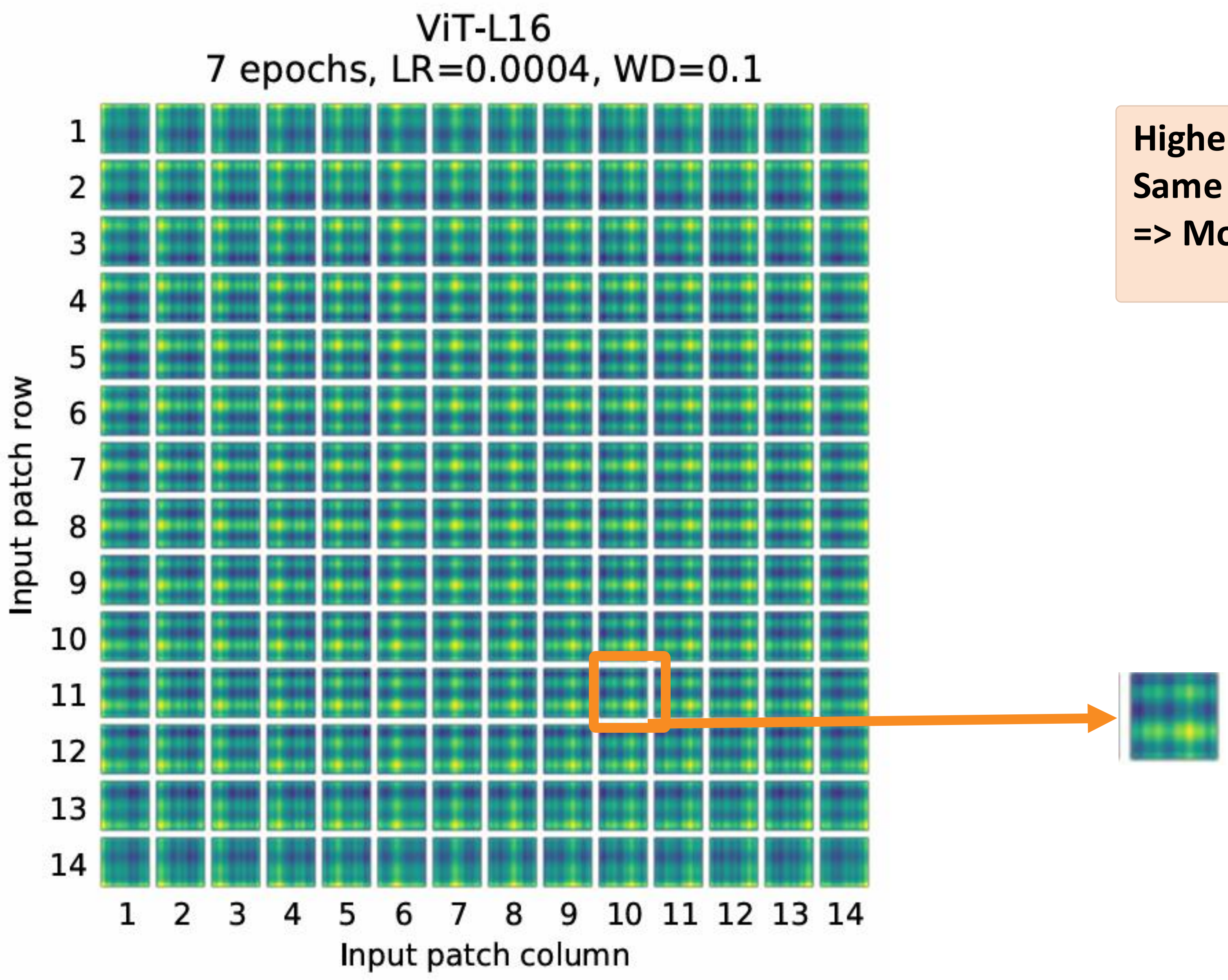
Lower Learning Rate
Less Epochs
=> Clearer Patterns

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

8. Position Embedding Trained With Different Hyperparameters

Comparison 2

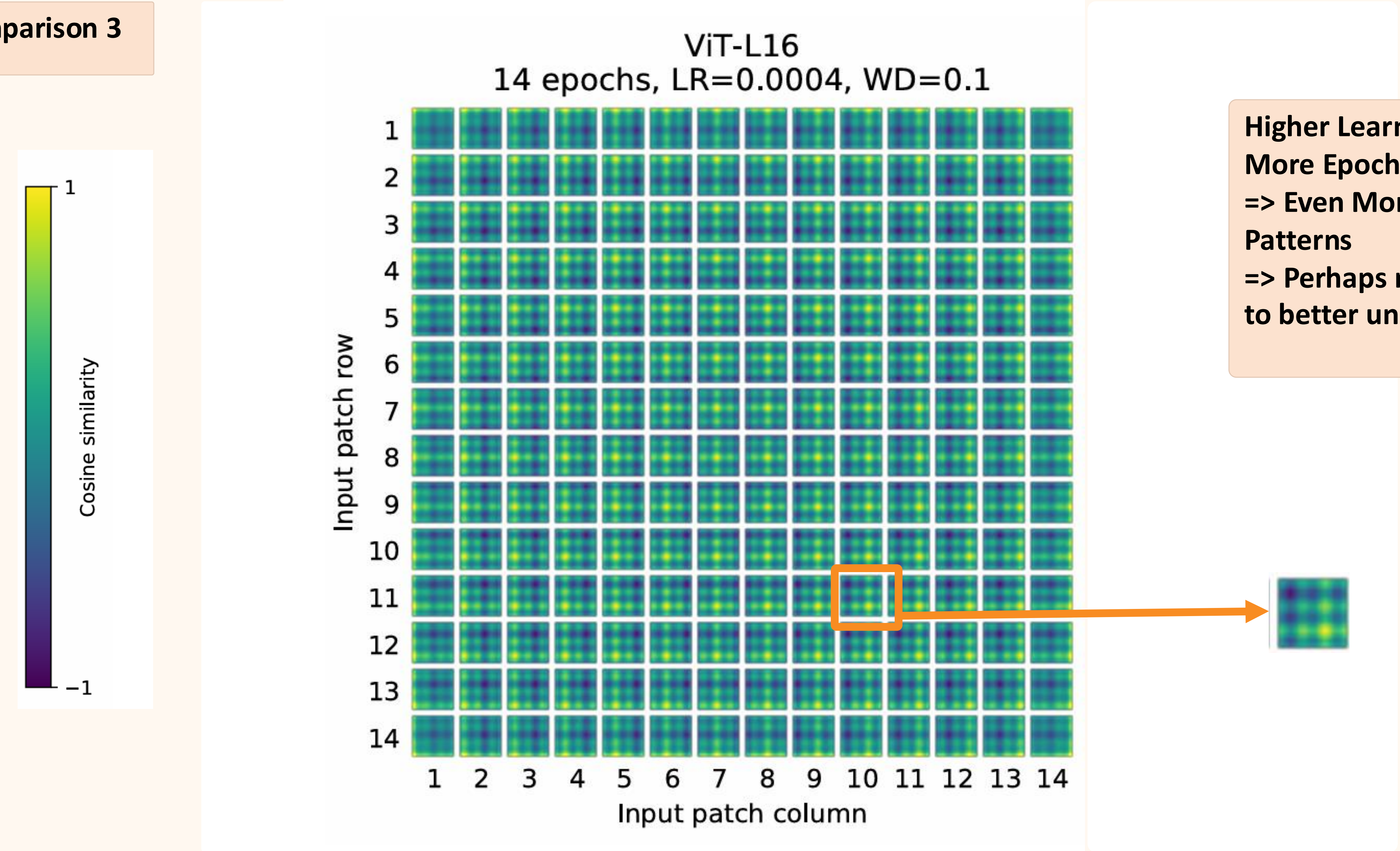




Higher Learning Rate
Same Epochs
=> More Fine-Grained Patterns

8. Position Embedding Trained With Different Hyperparameters

Comparison 3



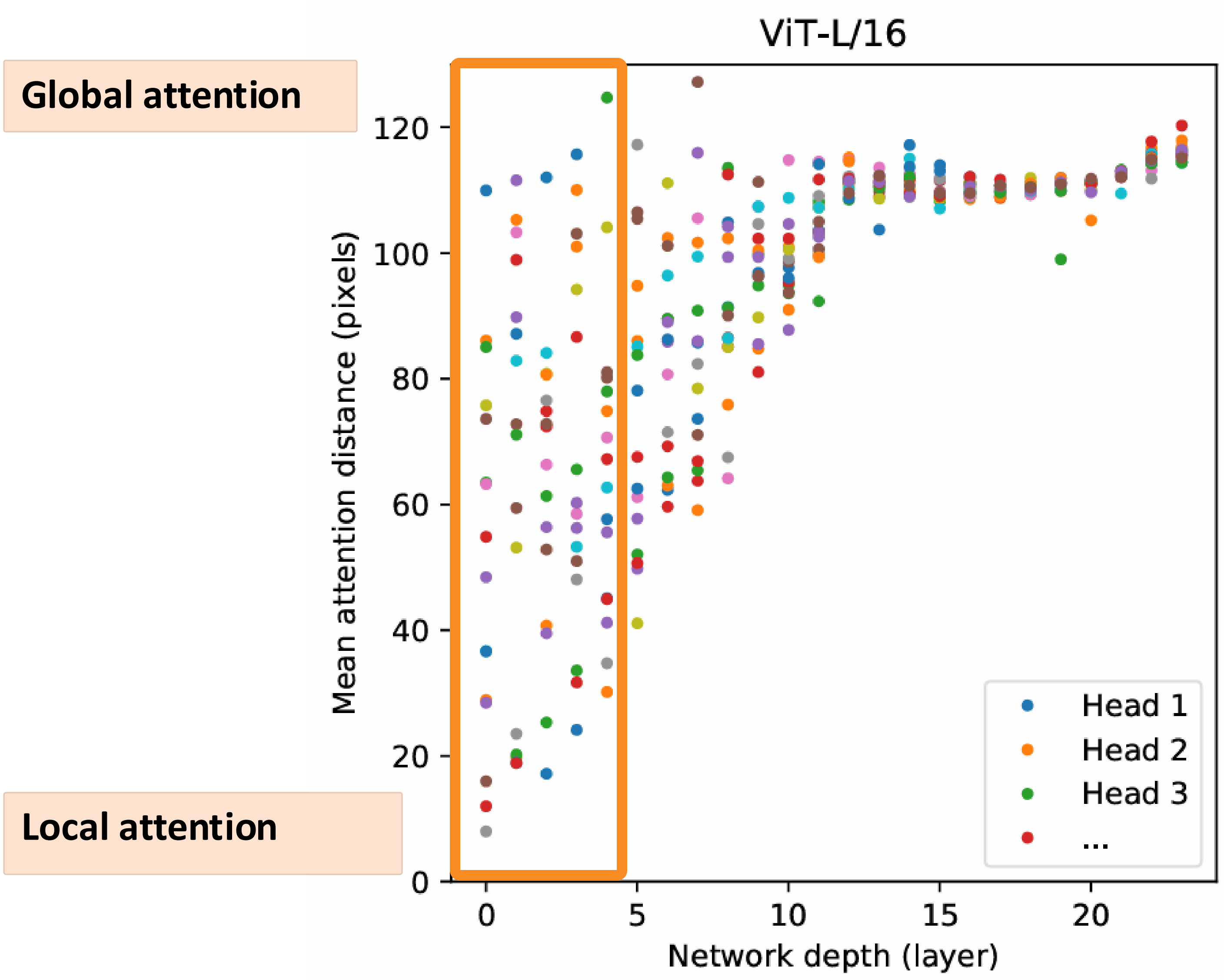
Higher Learning Rate
More Epochs
=> Even More Fine-Grained Patterns
=> Perhaps more training leads to better understanding?

9

Attention Distance at Various Network Depths

9. Attention Distance at Various Network Depths

Comparison 1



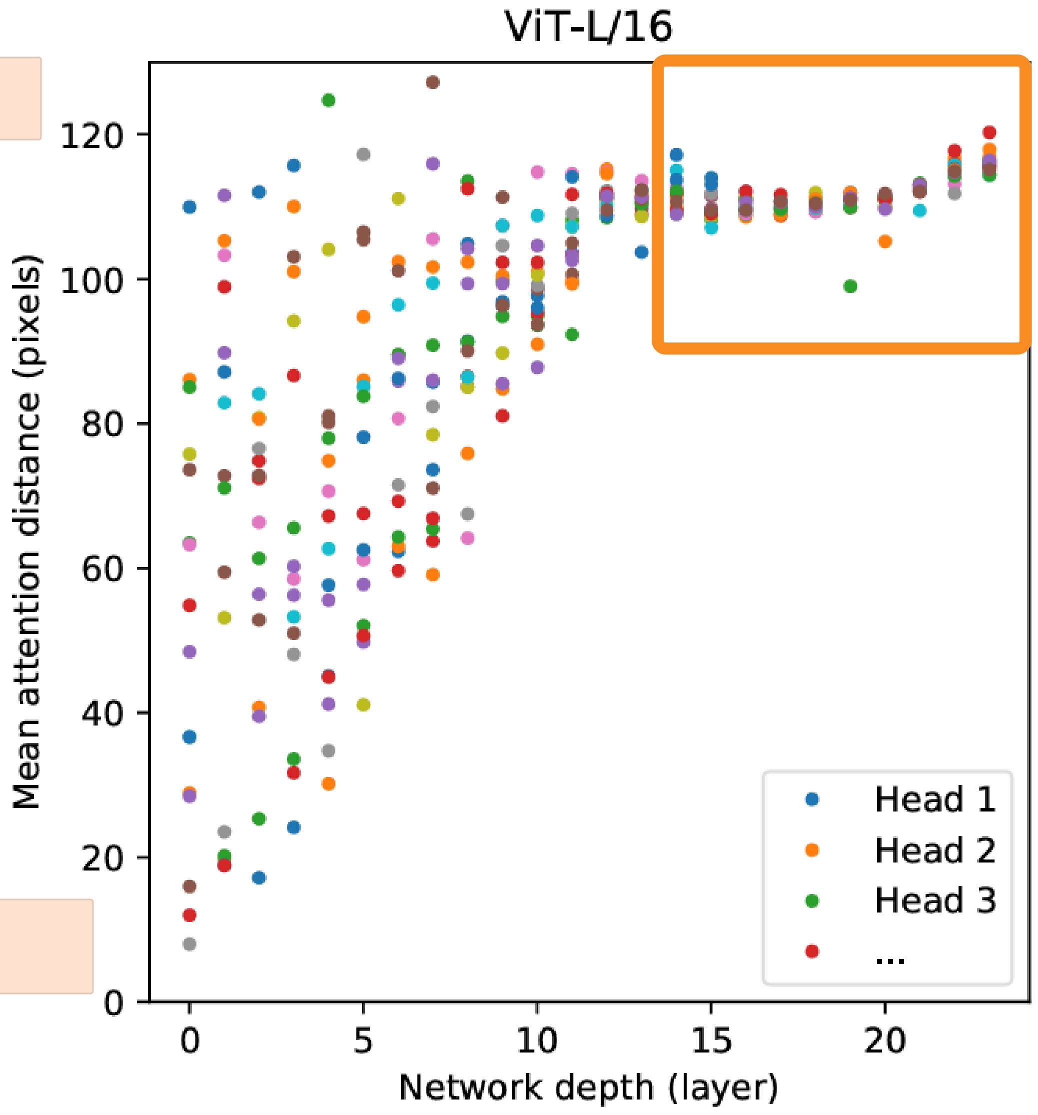
Earlier network depths – Attention distance can range from low to high

9. Attention Distance at Various Network Depths

Comparison 2

Global attention

Local attention

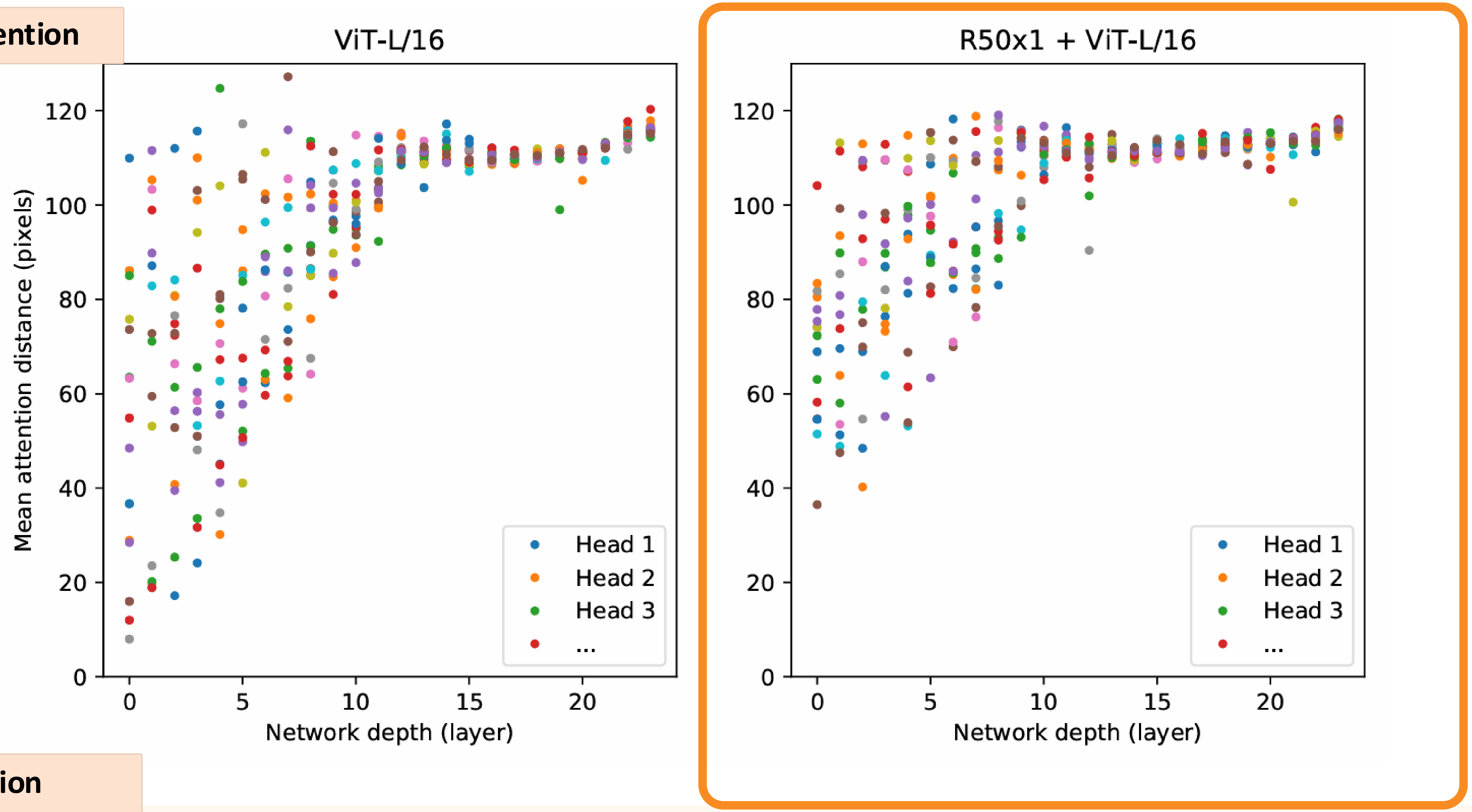


Deeper layers – Attention heads focus on global attention

9. Attention Distance at Various Network Depths

Comparison 3

Global attention



Local attention

Similar phenomenon

10

Batch Size for Models at Various Input Sizes

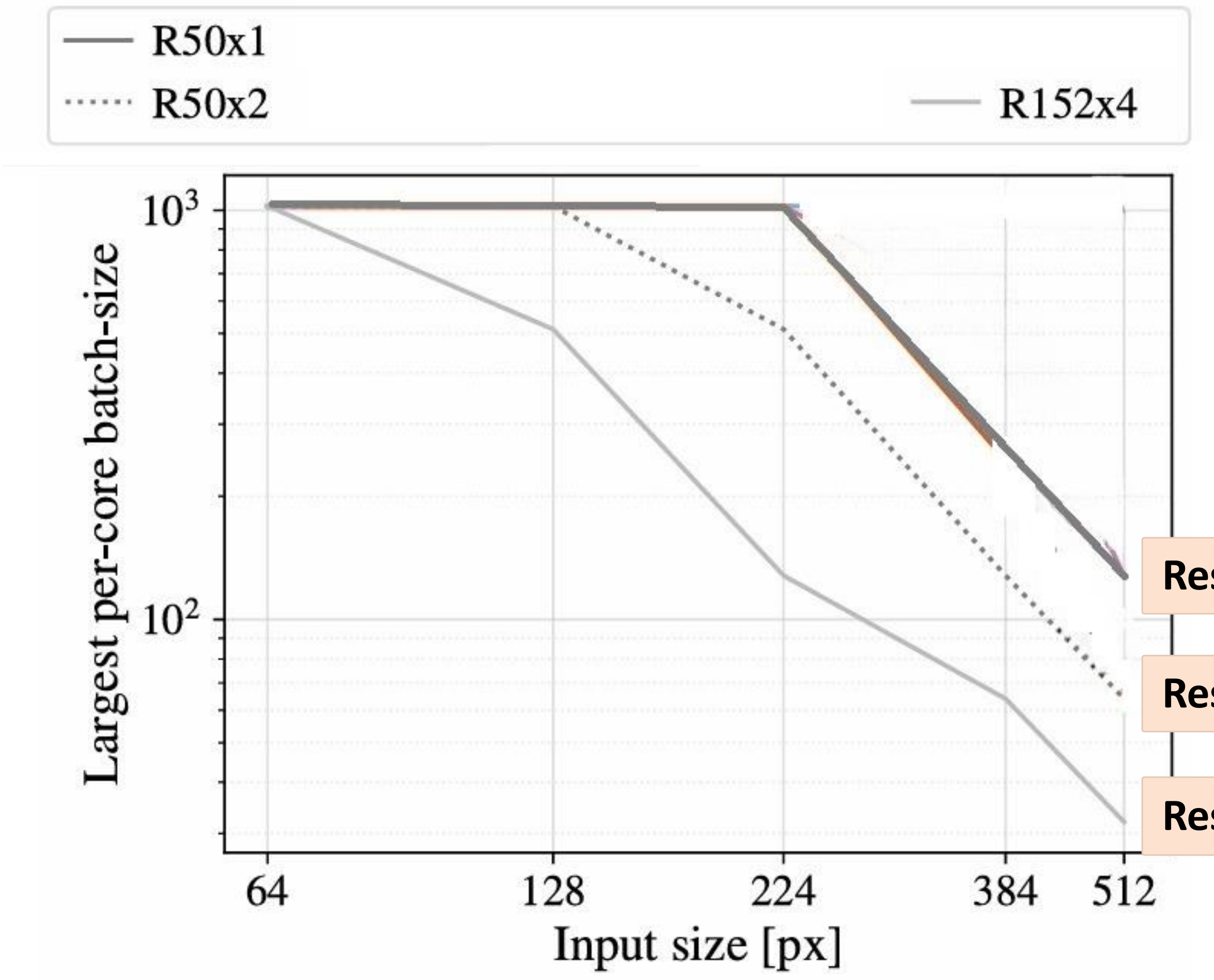
10. Batch Size for Models at Various Input Sizes

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

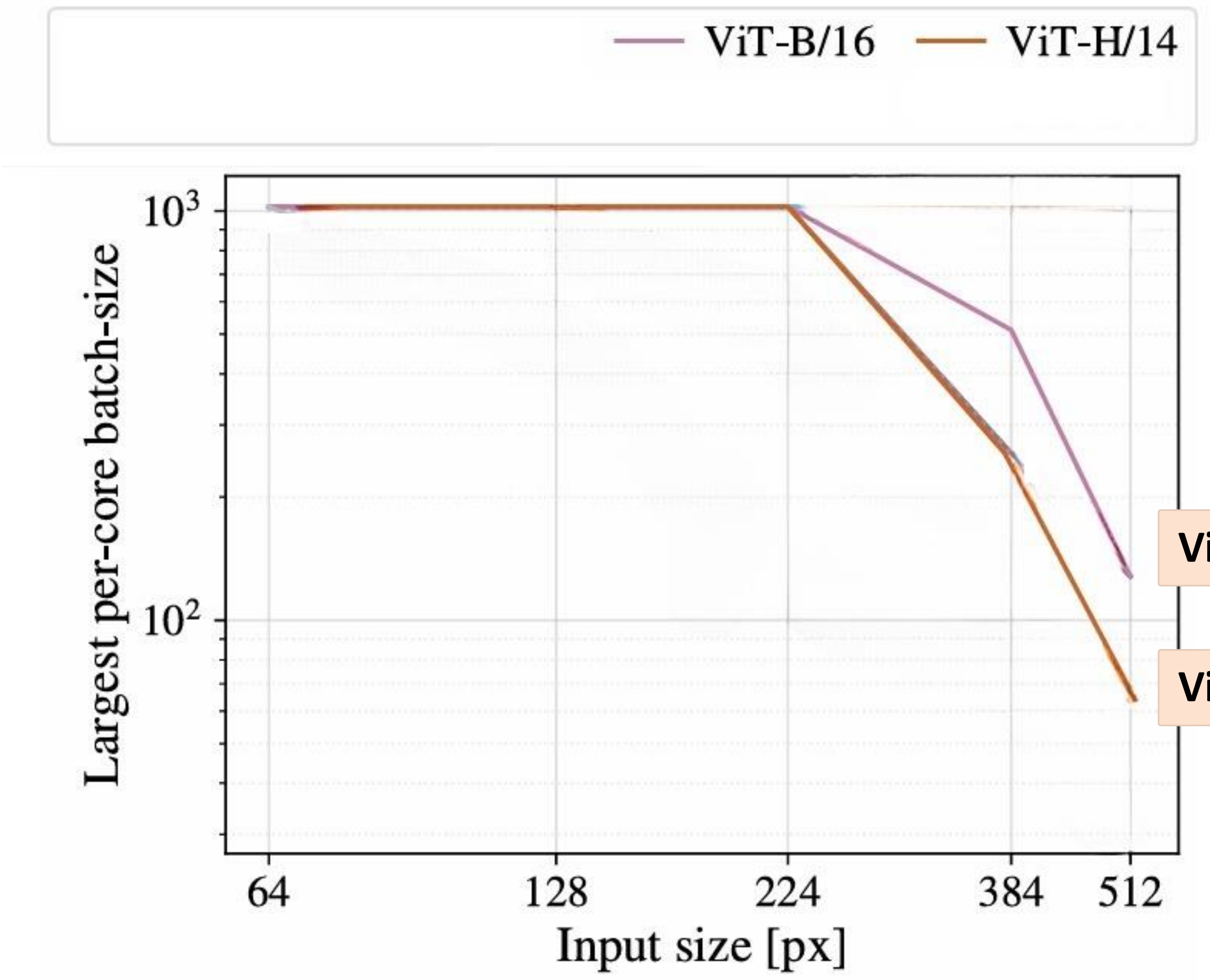
ResNet R50x2

ResNet R152x4

Larger Model, Lower Batch Size

10. Batch Size for Models at Various Input Sizes

Comparison 2



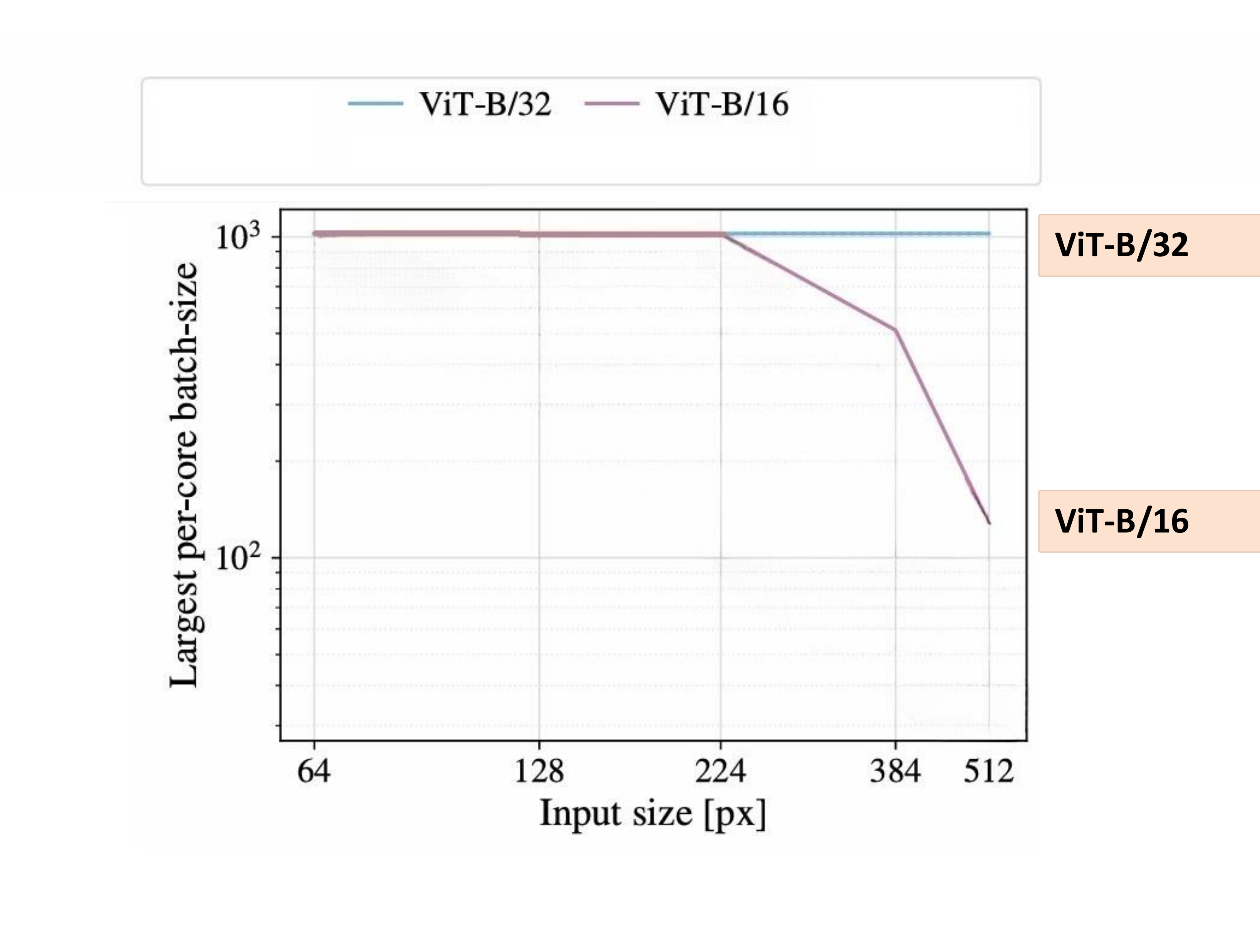
ViT-B/16

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