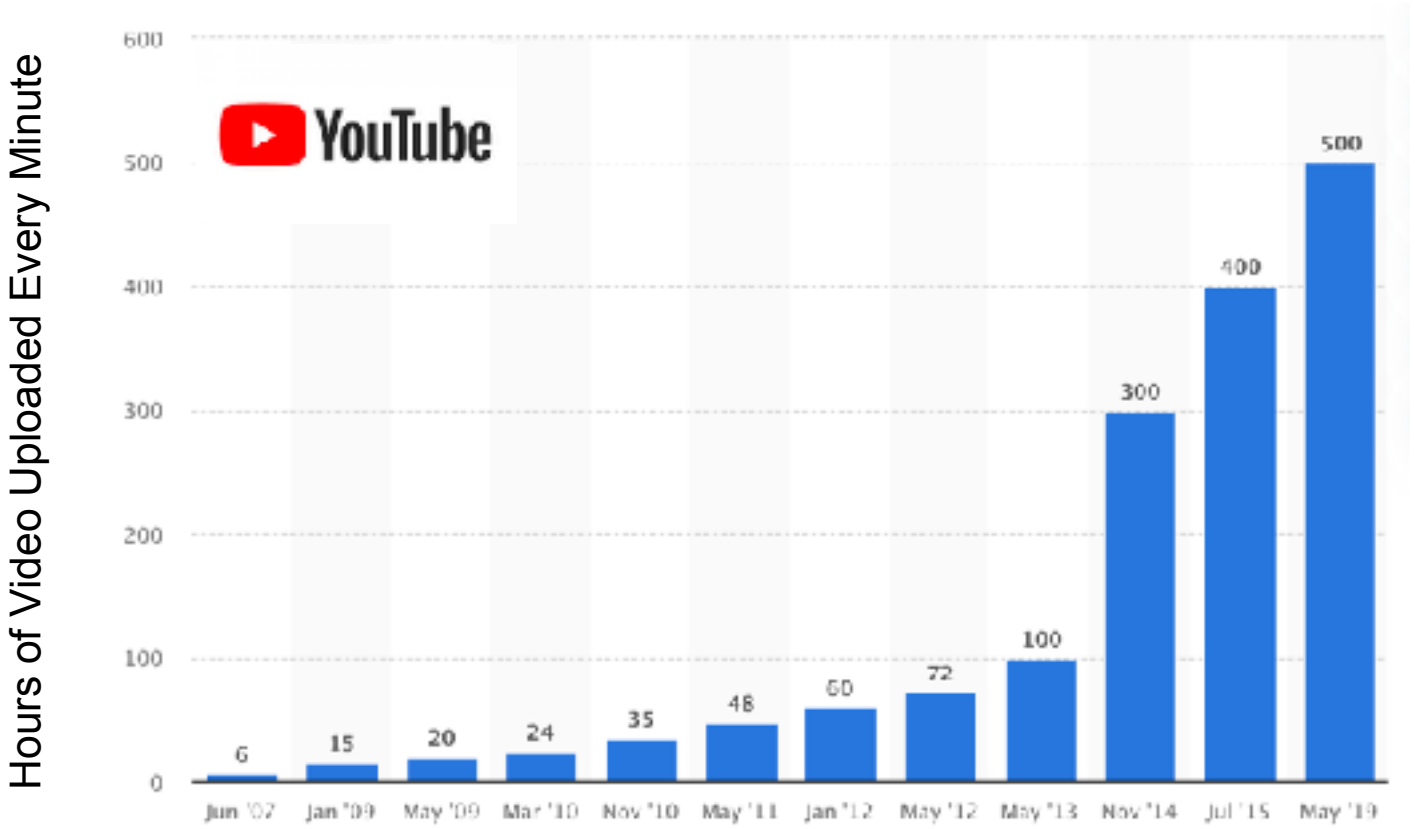
Scaling and Benchmarking Self-Supervised Visual Representation Learning

ICCV 2019

Priya Goyal, Dhruv Mahajan, Abhinav Gupta, Ishan Misra

Why Self-Supervised Learning?

- Availability of vast amount of unlabelled image/video data.



Self-Supervised Imagenet Pretraining

- However, most SSL methods thus far, used Imagenet (without labels) for pretraining.

Self-Supervised Imagenet Pretraining

- However, most SSL methods thus far, used Imagenet (without labels) for pretraining.

Imagenet (~1M images) represents a tiny fraction of our entire visual world

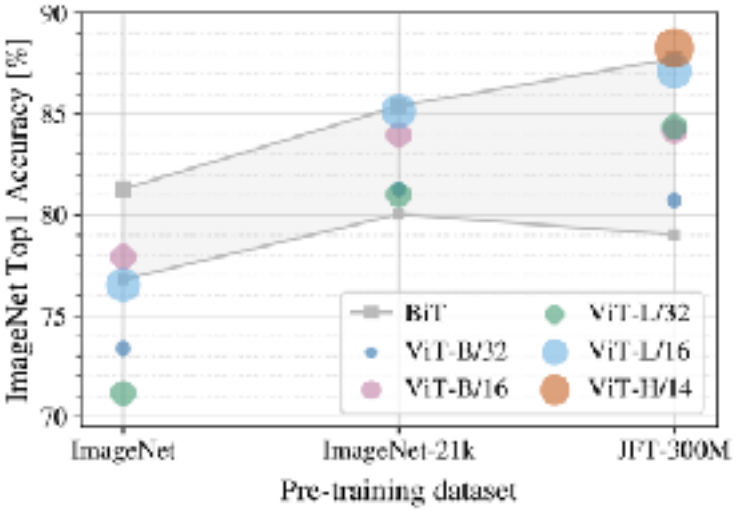
Self-Supervised Imagenet Pretraining

- However, most SSL methods thus far, used Imagenet (without labels) for pretraining.

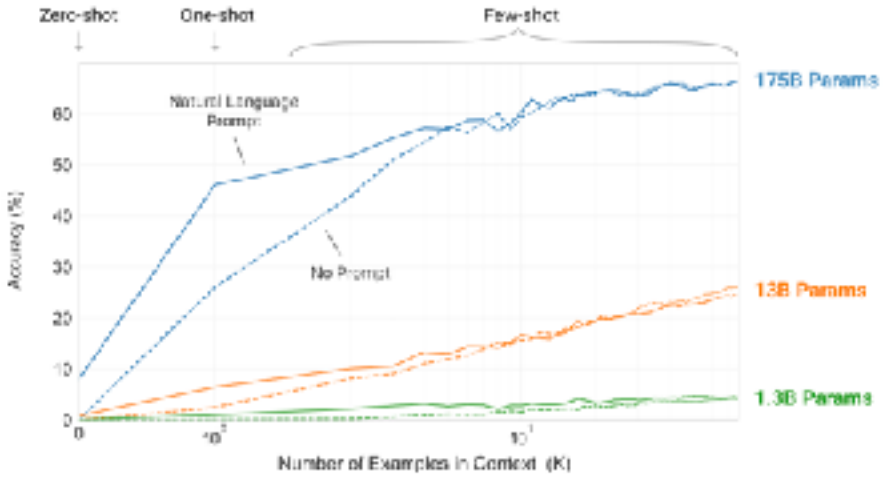
Why constrain ourselves with Imagenet when SSL enables us to use any dataset that we want?

The Importance of Scaling

- In the recent years, scaling has been shown to be extremely important for good performance.



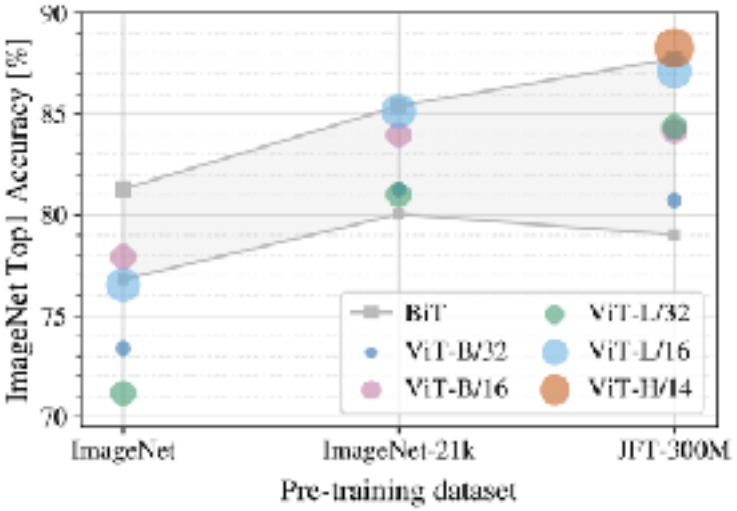
Dosovitskiy et al. “An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale”, ICLR 2020



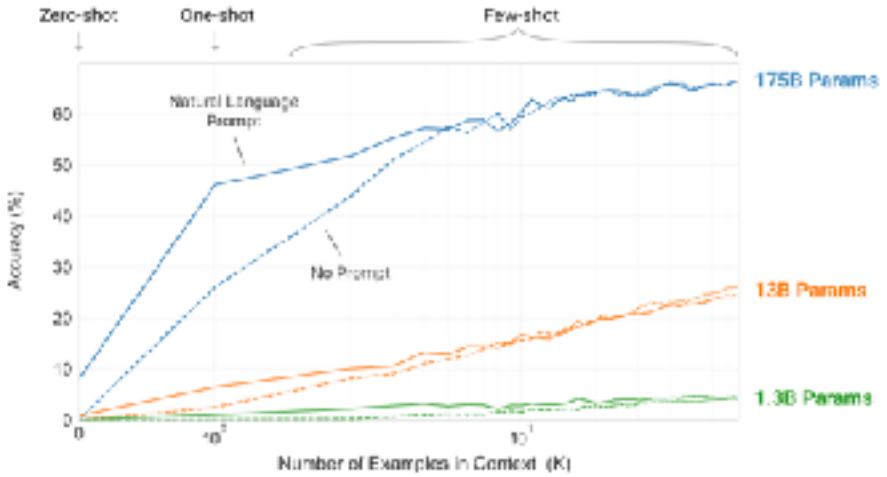
Brown et al. “Language Models are Few-Shot Learners”, NeurIPS 2020

The Importance of Scaling

- In the recent years, scaling has been shown to be extremely important for good performance.



Dosovitskiy et al. “An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale”, ICLR 2020

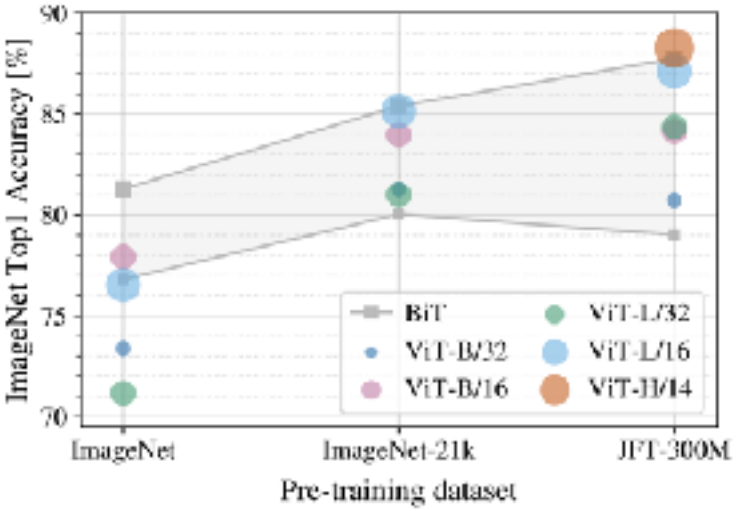


Brown et al. “Language Models are Few-Shot Learners”, NeurIPS 2020

Scaling the amount of training data

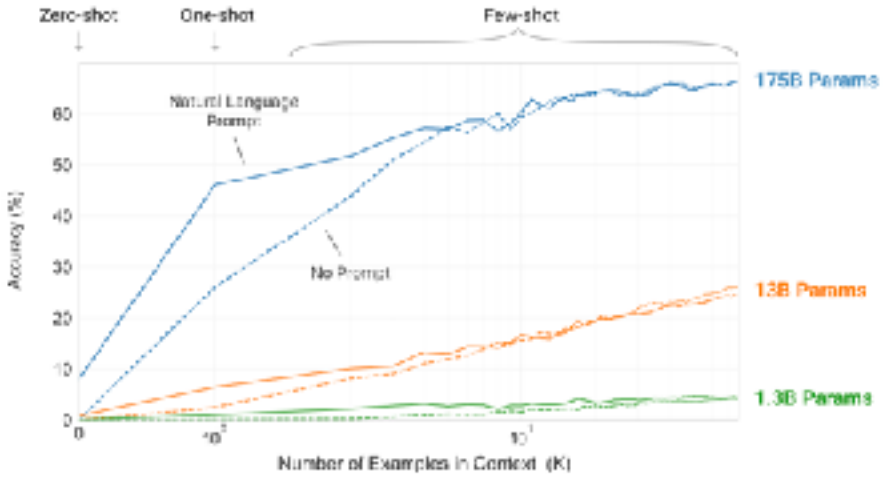
The Importance of Scaling

- In the recent years, scaling has been shown to be extremely important for good performance.



Dosovitskiy et al. “An Image Is Worth 16X16 Words: Transformers for Image Recognition at Scale”, ICLR 2020

Scaling the amount of training data

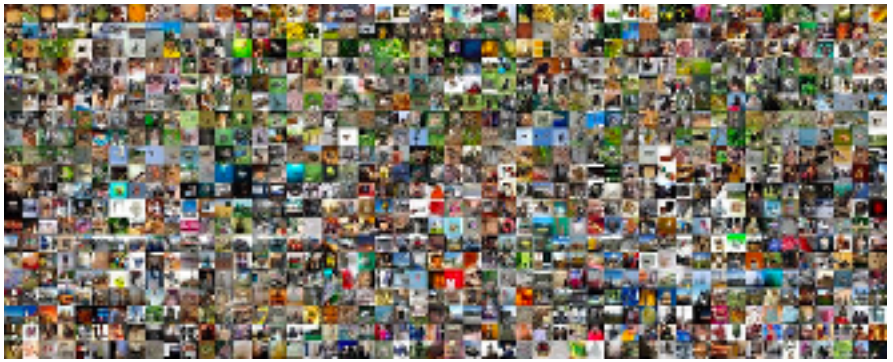


Brown et al. “Language Models are Few-Shot Learners”, NeurIPS 2020

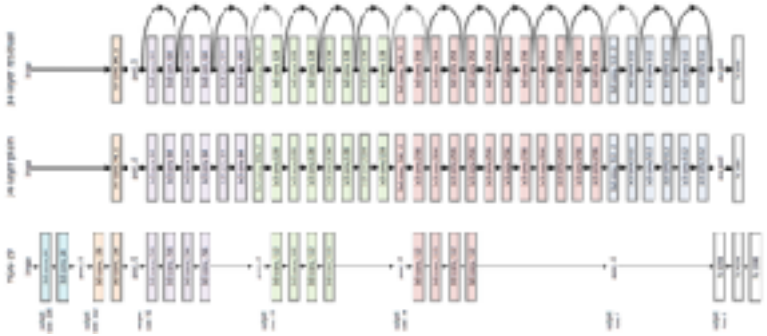
Scaling the model capacity (i.e., # of parameters)

Outline

1. Scaling pre-training data



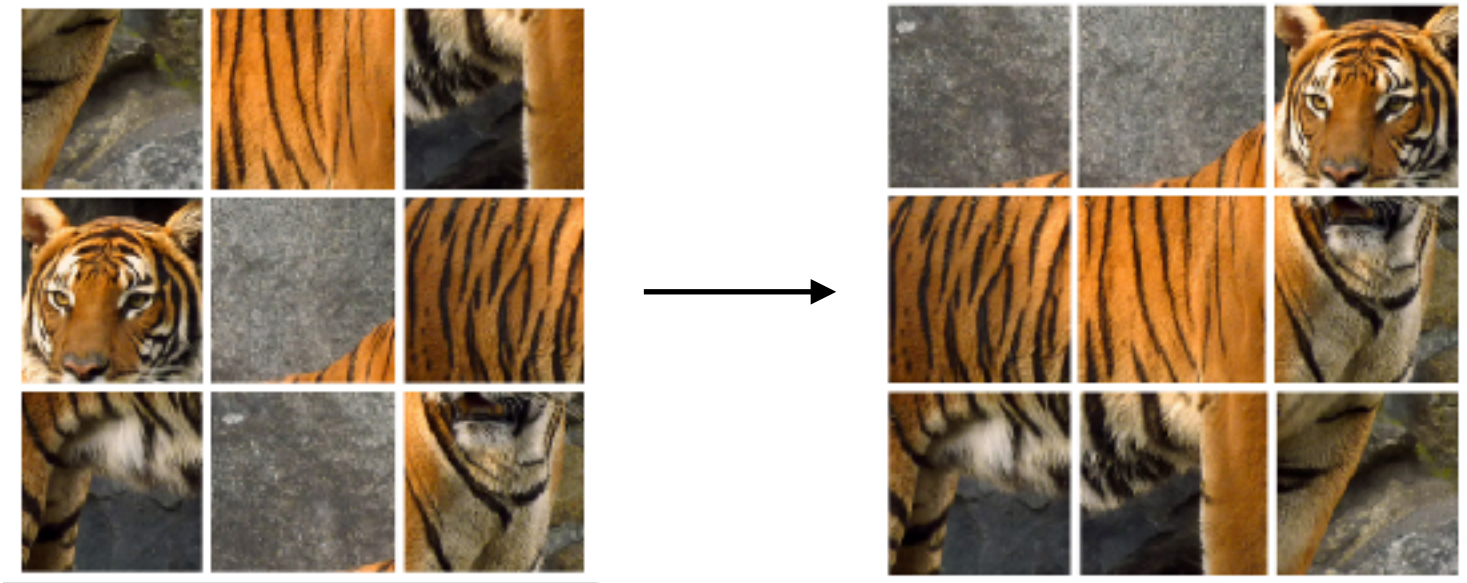
2. Scaling model capacity



3. Scaling problem complexity

Pretext Tasks (Jigsaw Puzzle)

- Given scrambled patches of an image the goal is to correctly unscramble the image.



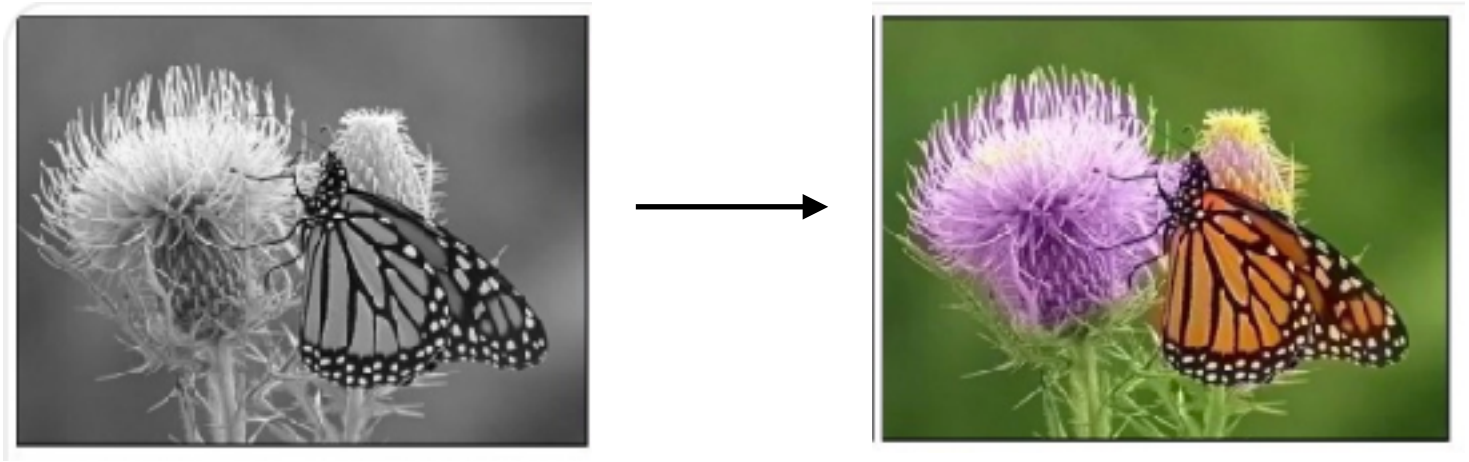
Scrambled Patches of an Image

Unscrambled Image

Noroozi et al., ”Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles,” ECCV 2016

Pretext Tasks (Colorization)

- Given a grayscale image the goal is to predict the correct color for every single pixel



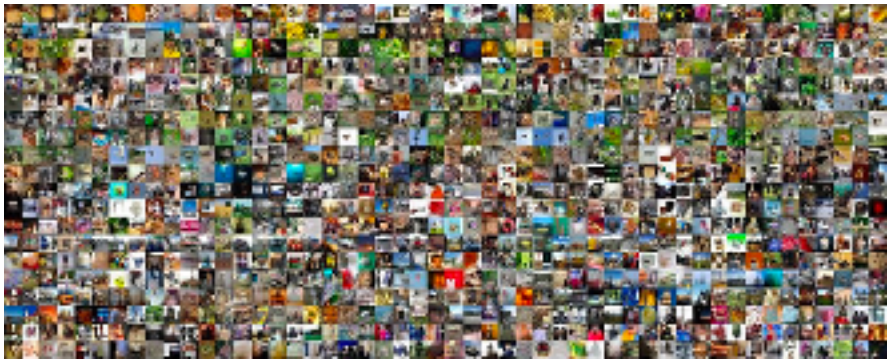
A Grayscale Image

Prediction

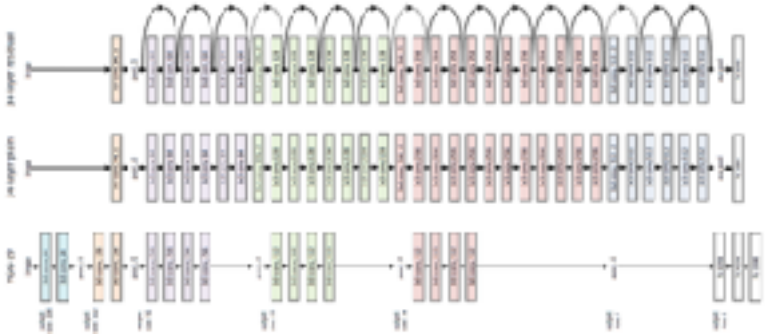
Zhang et al., ”Colorful Image Colorization,” ECCV 2016

Outline

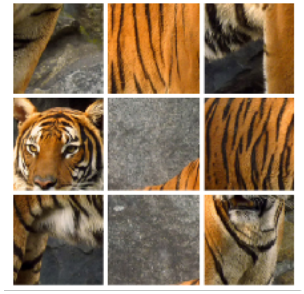
1. Scaling pre-training data



2. Scaling model capacity

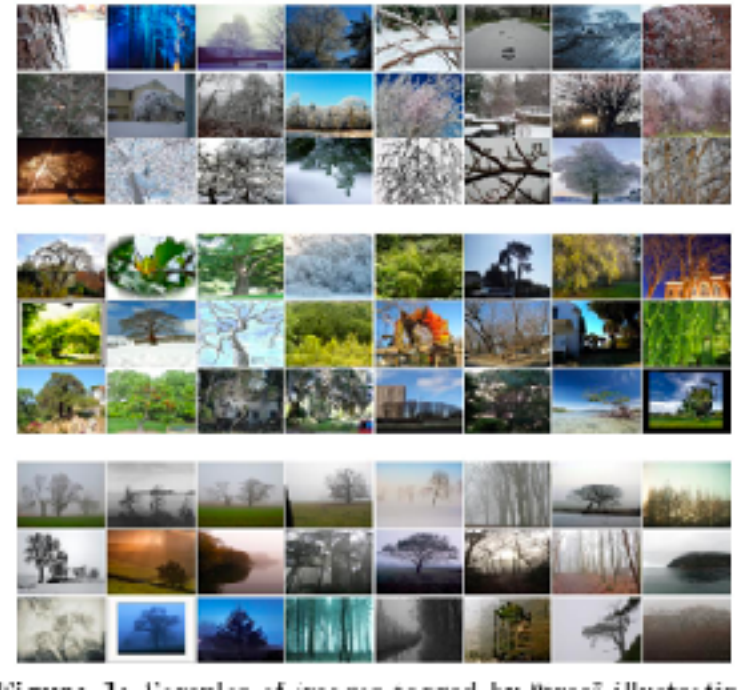


3. Scaling problem complexity



YFCC100M

- A dataset consisting of ~100 million photos uploaded to Flickr between 2004 and 2014.



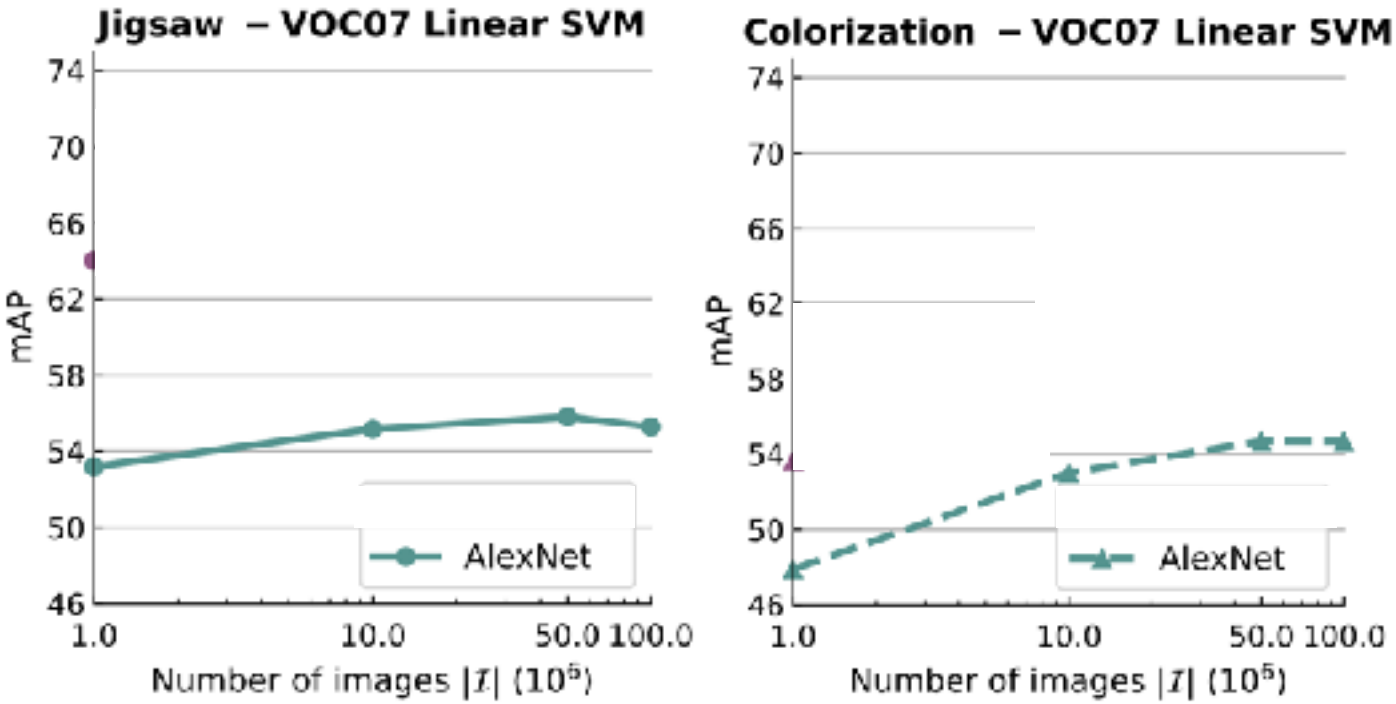
Thomee et al., ”YFCC100M: The New Data in Multimedia Research,” ACM MM 2016

Experimental Setup

- The authors use various subsets of the YFCC-100M dataset (e.g., 1, 10, 50, 100 million images) for pretraining.
- Evaluation is done on VOC'07 classification task.
- AlexNet is used as a base model.
- Jigsaw and Colorization tasks are used as pretext tasks.

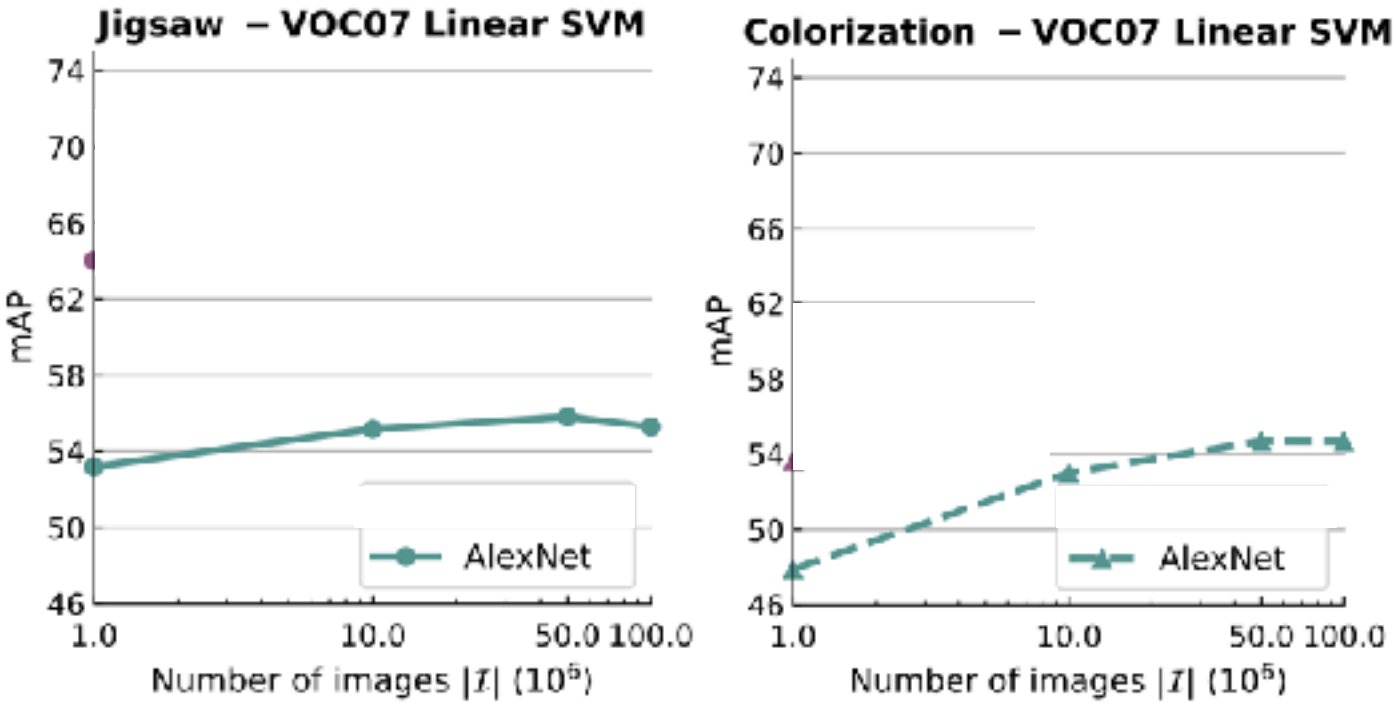
Results

- Increasing the size slightly improves the transfer learning performance for both the Jigsaw and Colorization methods.



Results

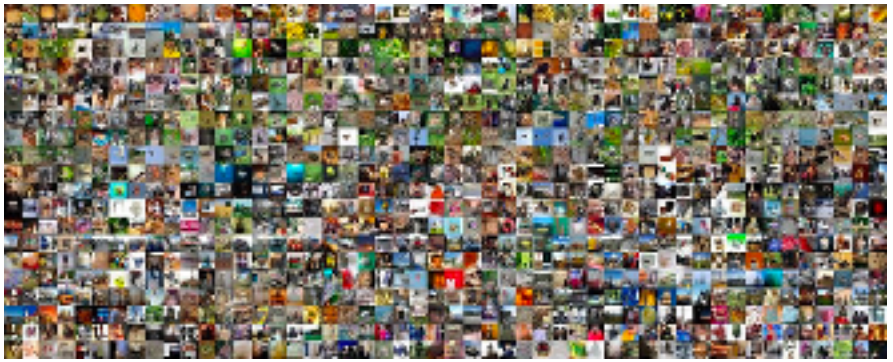
- Increasing the size slightly improves the transfer learning performance for both the Jigsaw and Colorization methods.



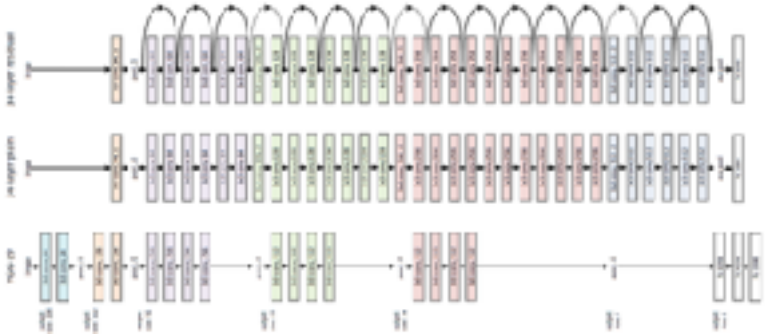
Why are the gains so small?

Outline

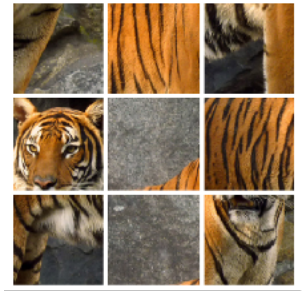
1. Scaling pre-training data



2. Scaling model capacity

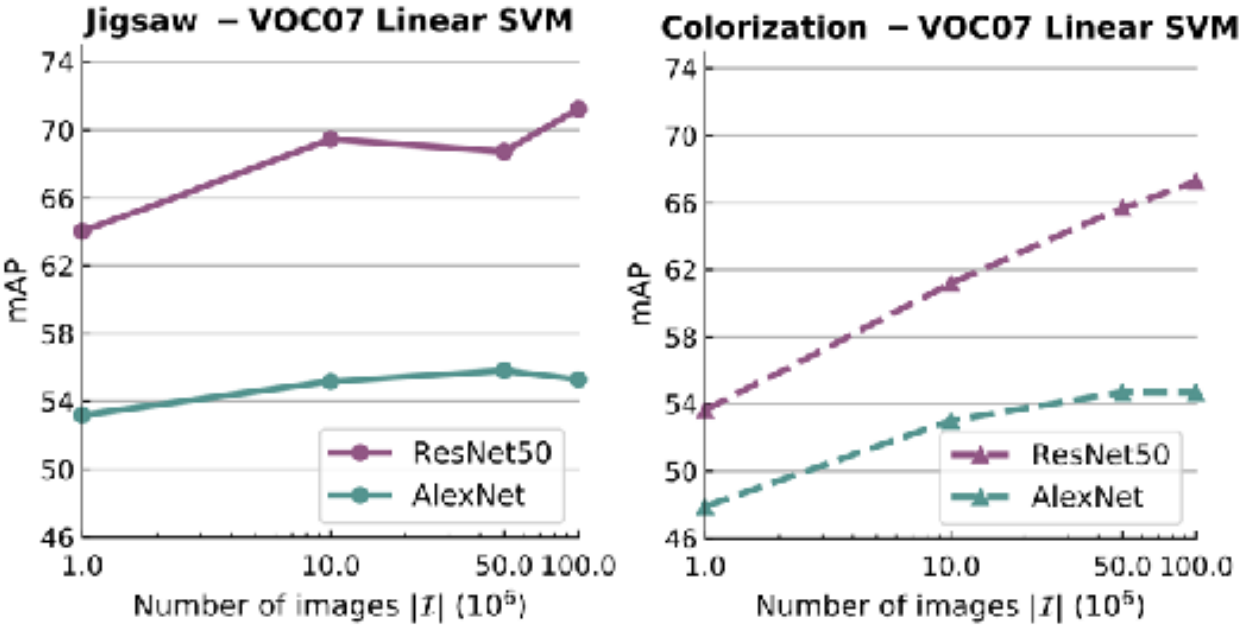


3. Scaling problem complexity



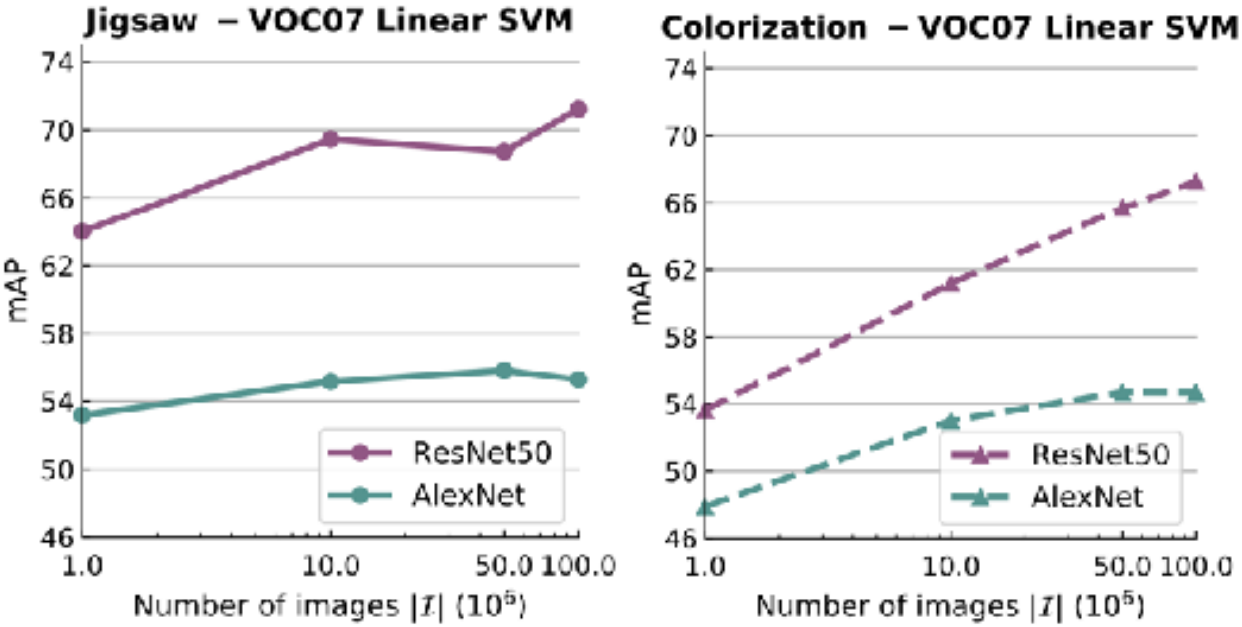
Results

- ResNet-50 shows much larger improvements as the data size increases.



Results

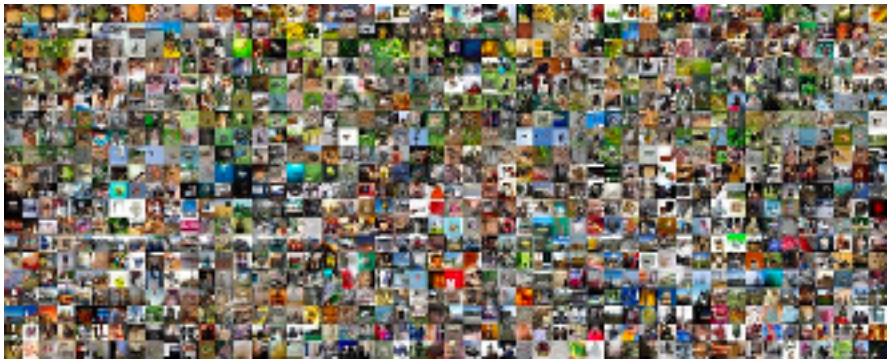
- ResNet-50 shows much larger improvements as the data size increases.



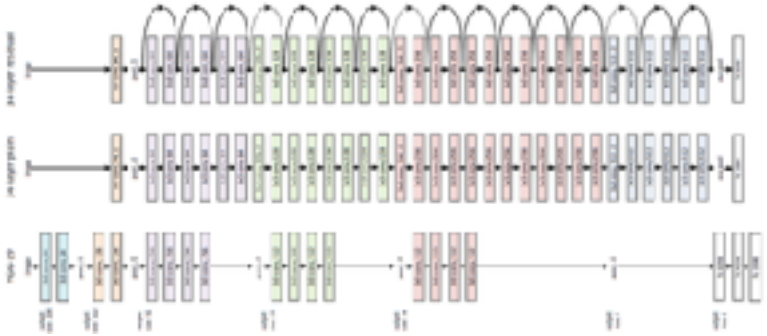
Low capacity models like AlexNet do not show much improvement with more data

Outline

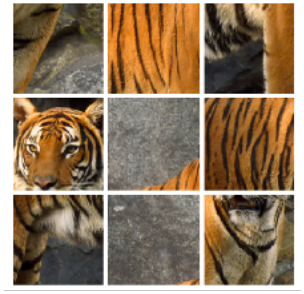
1. Scaling pre-training data



2. Scaling model capacity

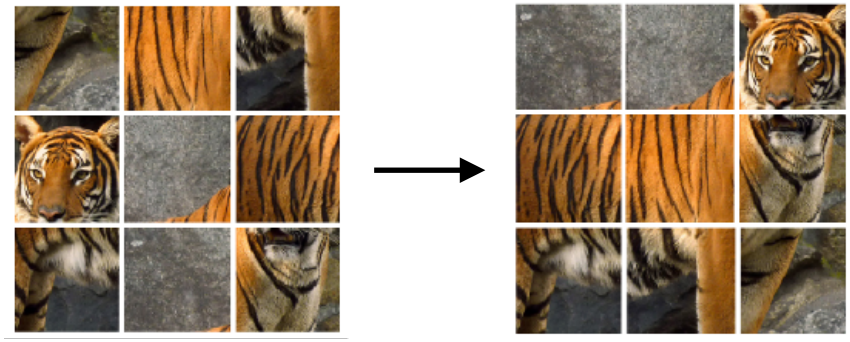


3. Scaling problem complexity



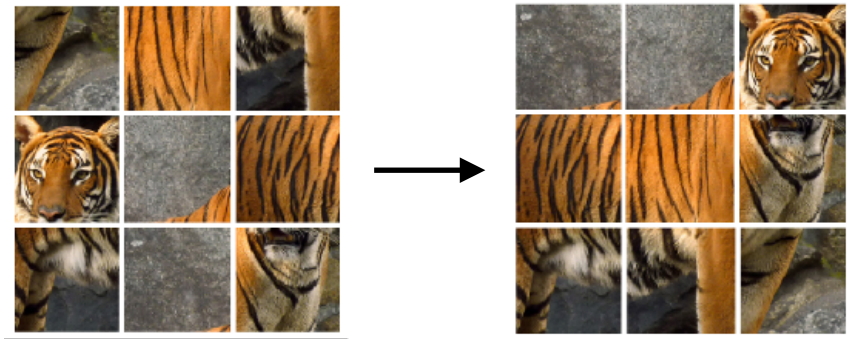
Increasing Task Difficulty (Jigsaw Puzzle)

- An input image is divided into N = 9 non-overlapping patches.
- A ‘puzzle’ is created by shuffling these patches randomly.
- A CNN is trained to predict the permutation used to create the puzzle.
- As the total number of permutations N! can be large, a fixed subset P of the total N! permutations is used.
- The prediction problem is reduced to classification into one of |P| classes.



Increasing Task Difficulty (Jigsaw Puzzle)

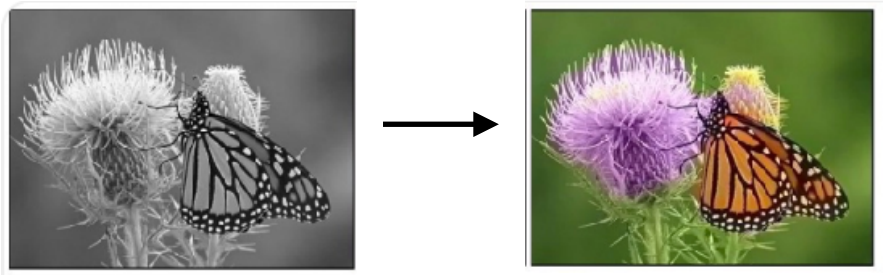
- An input image is divided into N = 9 non-overlapping patches.
- A ‘puzzle’ is created by shuffling these patches randomly.
- A CNN is trained to predict the permutation used to create the puzzle.
- As the total number of permutations N! can be large, a fixed subset P of the total N! permutations is used.
- The prediction problem is reduced to classification into one of |P| classes.



The prediction problem can be made more difficult by increasing P

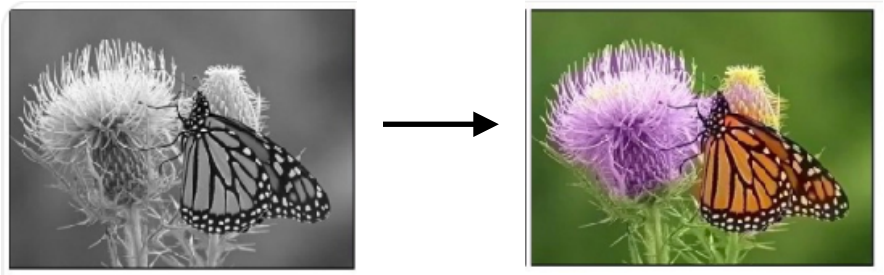
Increasing Task Difficulty (Colorization)

- Given a grayscale image the goal is to predict the correct color for every single pixel.
- The output color space is quantized into a set of discrete bins Q = 313.
- This reduces the problem to a |Q|-way classification problem.
- The target image is soft-encoded into |Q| bins by looking at the K-nearest neighbor bins.



Increasing Task Difficulty (Colorization)

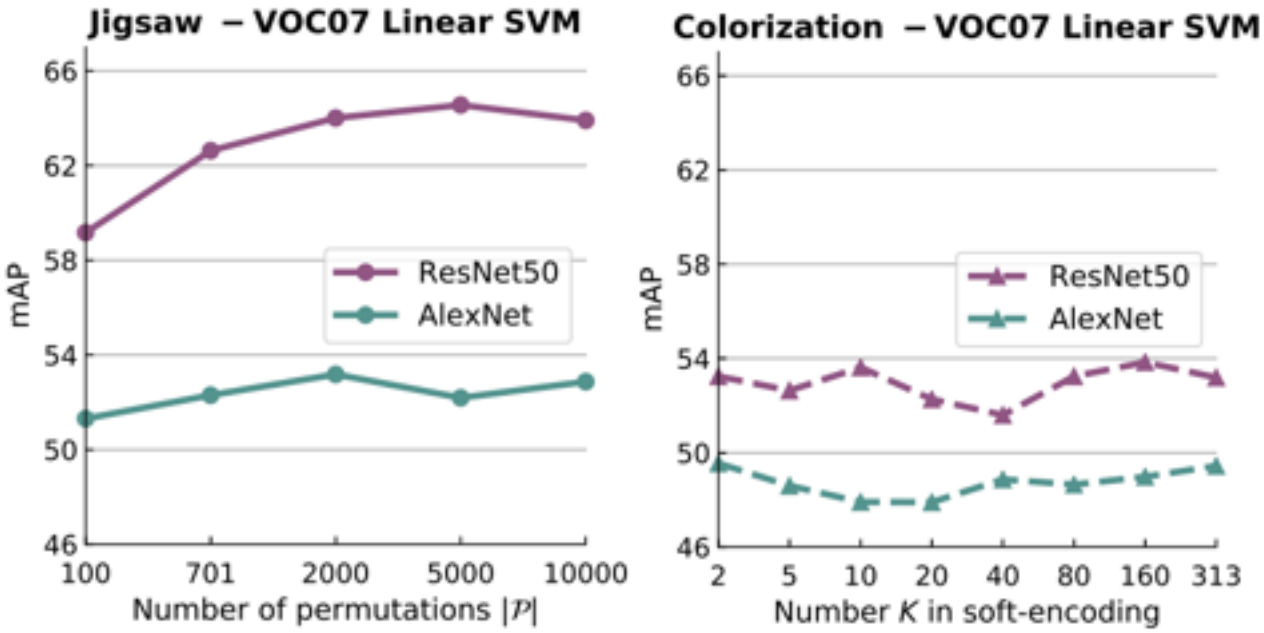
- Given a grayscale image the goal is to predict the correct color for every single pixel.
- The output color space is quantized into a set of discrete bins Q = 313.
- This reduces the problem to a |Q|-way classification problem.
- The target image is soft-encoded into |Q| bins by looking at the K-nearest neighbor bins.



The prediction problem can be made more difficult by increasing K

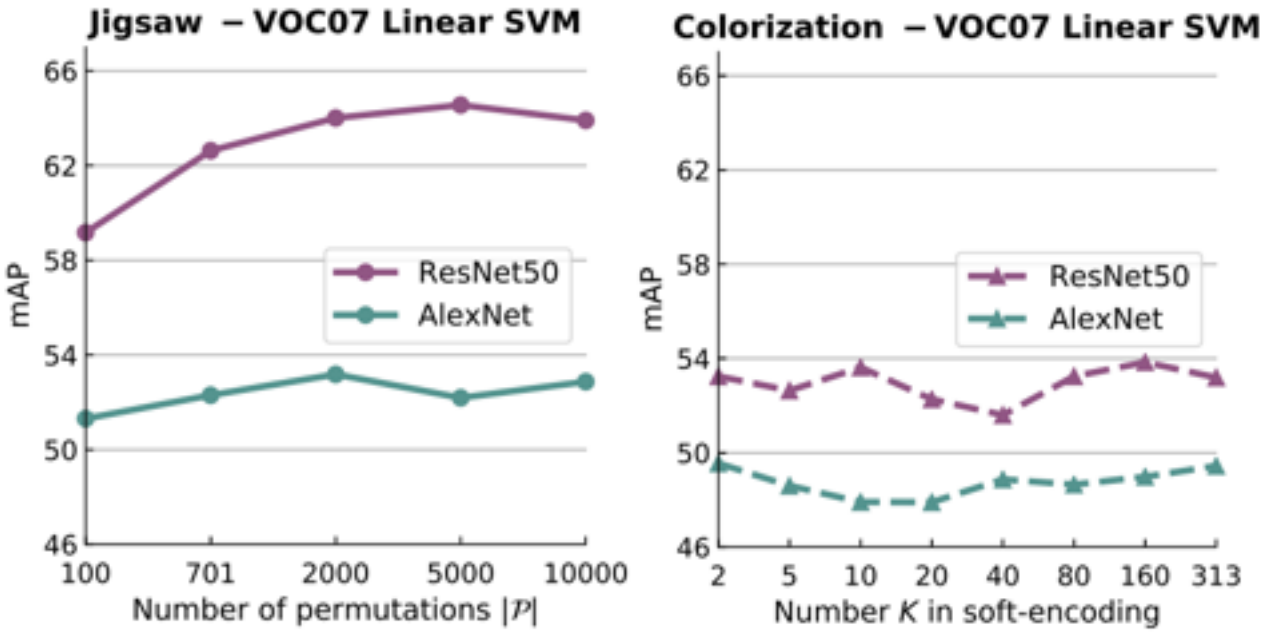
Results

- The results on the VOC07 classification task when scaling the problem complexity.



Results

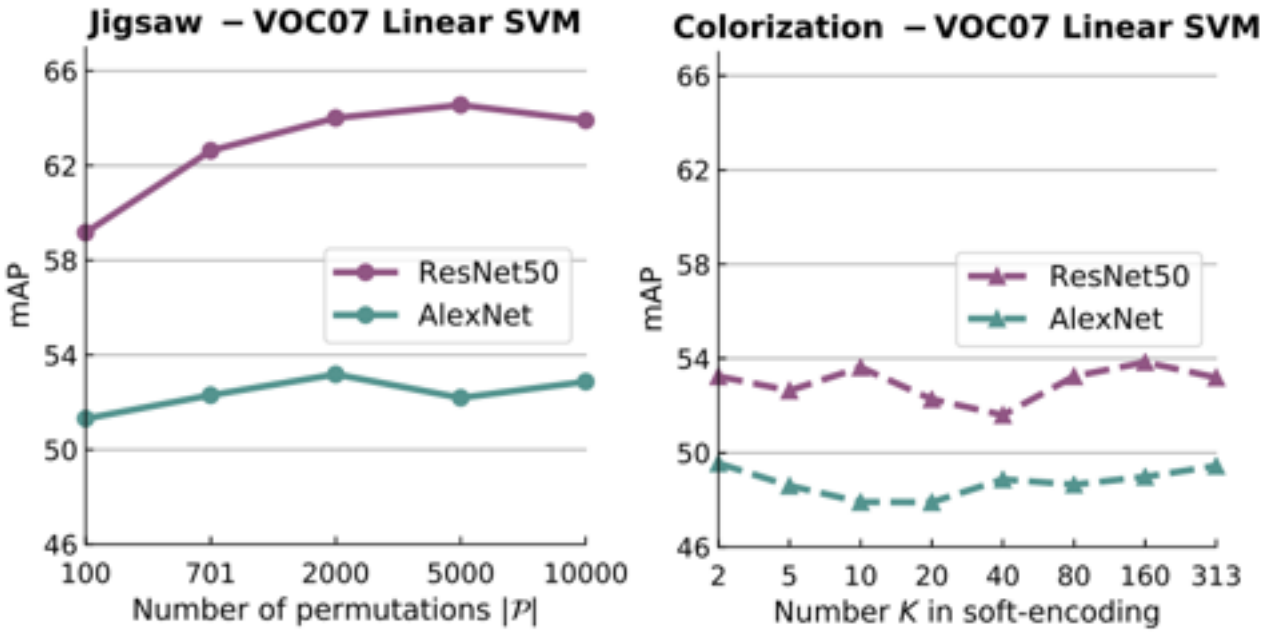
- The results on the VOC07 classification task when scaling the problem complexity.



For the Jigsaw Puzzle, there is an improvement in transfer learning performance as the size of the permutation set increases.

Results

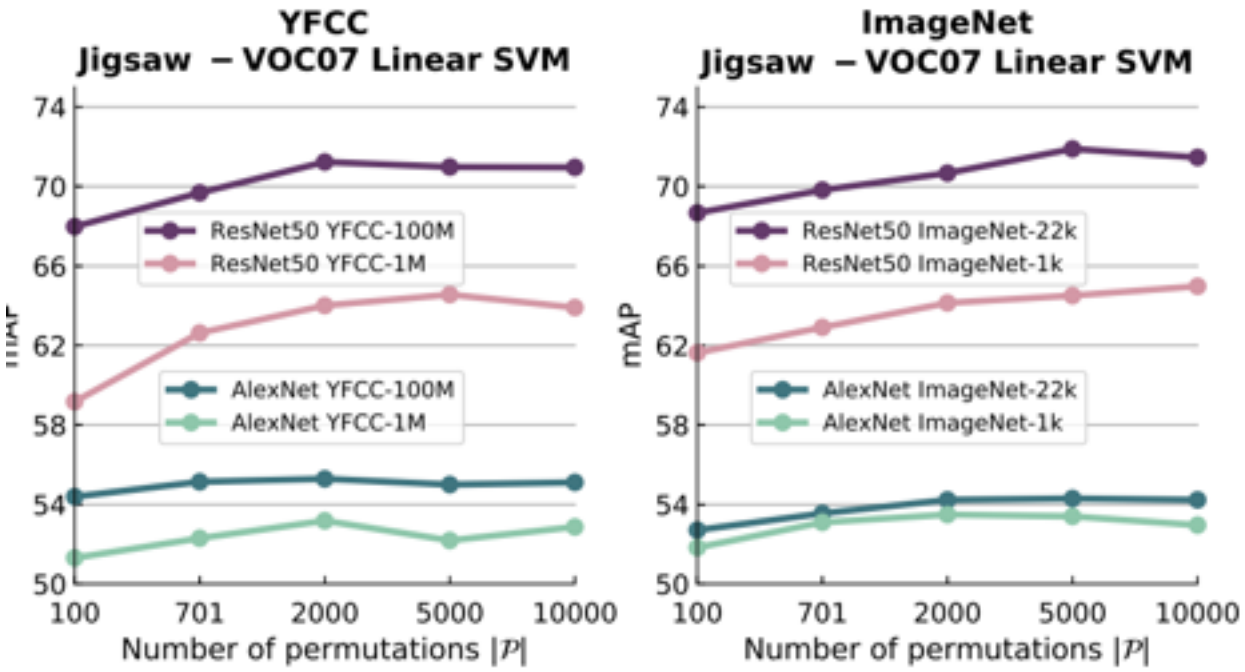
- The results on the VOC07 classification task when scaling the problem complexity.



The Colorization approach appears to be less sensitive to changes in problem complexity.

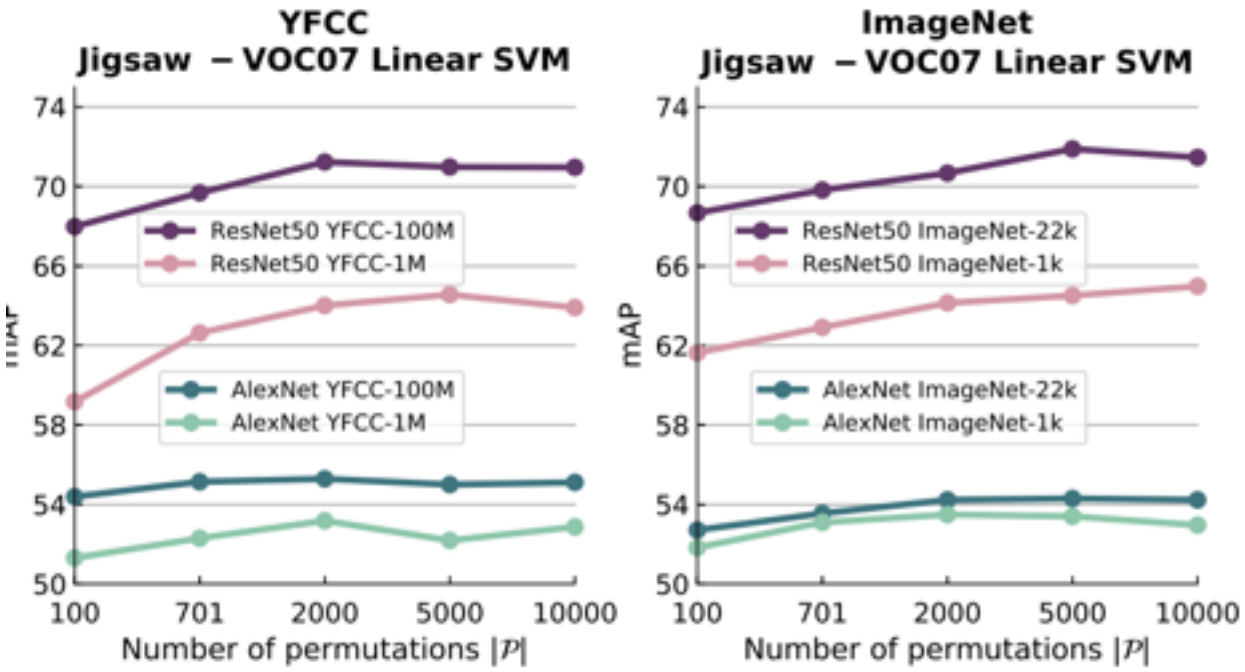
Scaling All Three Axes Together

- The authors explore the relationship between all the three axes of scaling.



Scaling All Three Axes Together

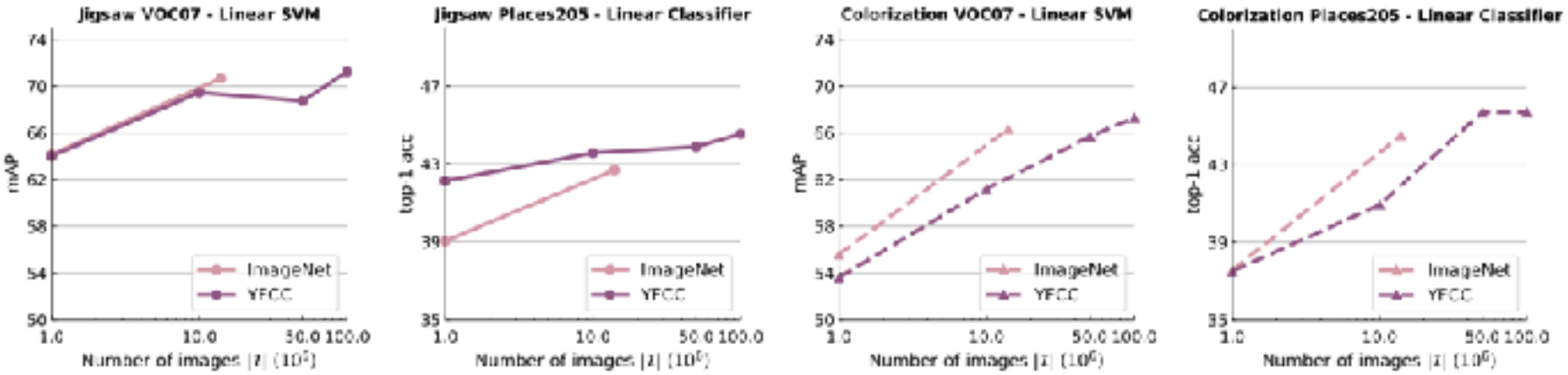
- The authors explore the relationship between all the three axes of scaling.



The three axes of scaling are complementary.

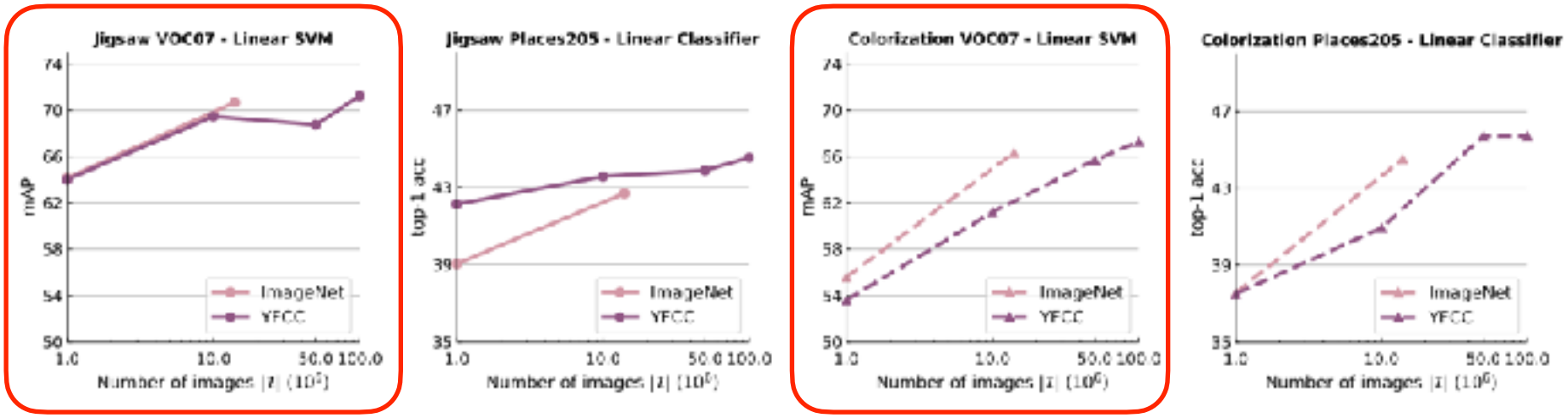
Pre-training and Transfer Domain Relation

- The authors vary the pre-training and transfer domains.



Pre-training and Transfer Domain Relation

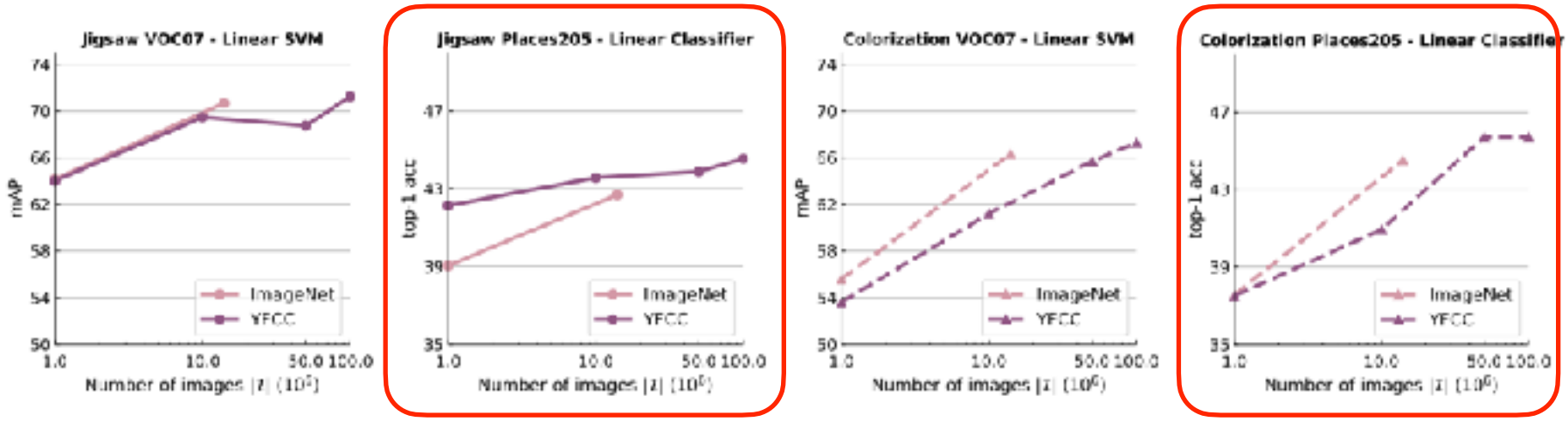
- The authors vary the pre-training and transfer domains.



Pre-training on ImageNet, rather than YFCC, provides a greater benefit when transferring to VOC07 classification.

Pre-training and Transfer Domain Relation

- The authors vary the pre-training and transfer domains.



YFCC is better for pretraining when transferring to the Places dataset.

Benchmarking Suite for SSL

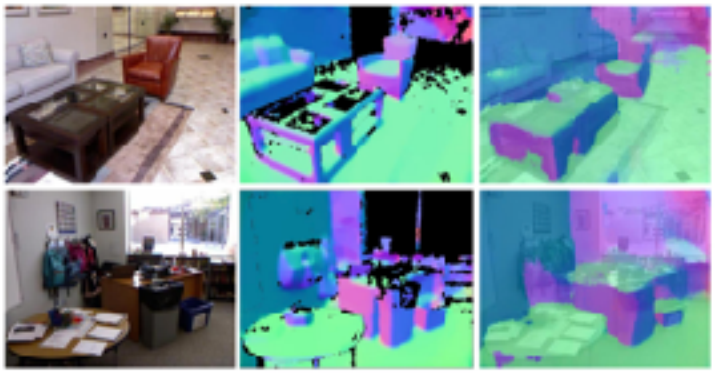
- The authors choose 9 different tasks to evaluate their learned representation.
- The tasks range from semantic classification/detection to 3D and actions (specifically, navigation).

Task	Datasets	Description
Image classification		
§ 6.1	Places205	Scene classification. 205 classes.
(Linear Classifier)	VOC07	Object classification. 20 classes.
	COCO2014	Object classification. 80 classes.
Low-shot image classification		
§ 6.2	VOC07	$\leq$ 96 samples per class
(Linear Classifier)	Places205	$\leq$ 128 samples per class
Visual navigation		
§ 6.3 (Fixed ConvNet)	Gibson	Reinforcement Learning for navigation.
Object detection		
§ 6.4	VOC07	20 classes.
(Frozen conv body)	VOC07+12	20 classes.
Scene geometry (3D)		
§ 6.5 (Frozen conv body)	NYUv2	Surface Normal Estimation.

Surface Normal Estimation

- Given a single image, the task is to predict the surface normal at each pixel.
- Normal legend: blue → X; green → Y; red → Z.

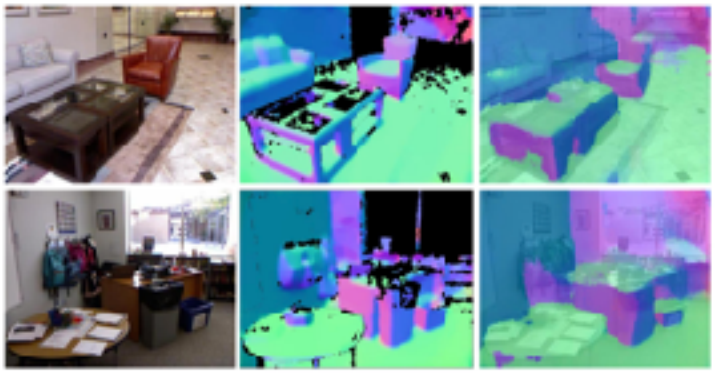
Initialization	Angle Distance		Within t°		
	Mean	Median	11.25	22.5	30
	(Lower is better)		(Higher is better)		
ResNet-50 ImageNet-1k supervised	26.4	17.1	36.1	59.2	68.5
ResNet-50 Places205 supervised	23.3	14.2	41.8	65.2	73.6
ResNet-50 Scratch	26.3	16.1	37.9	60.6	69.0
ResNet-50 Jigsaw ImageNet-1k	24.2	14.5	41.2	64.2	72.5
ResNet-50 Jigsaw ImageNet-22k	22.6	13.4	43.7	66.8	74.7
ResNet-50 Jigsaw YFCC-100M	**22.4**	**13.1**	**44.6**	**67.4**	**75.1**



Surface Normal Estimation

- Given a single image, the task is to predict the surface normal at each pixel.
- Normal legend: blue → X; green → Y; red → Z.

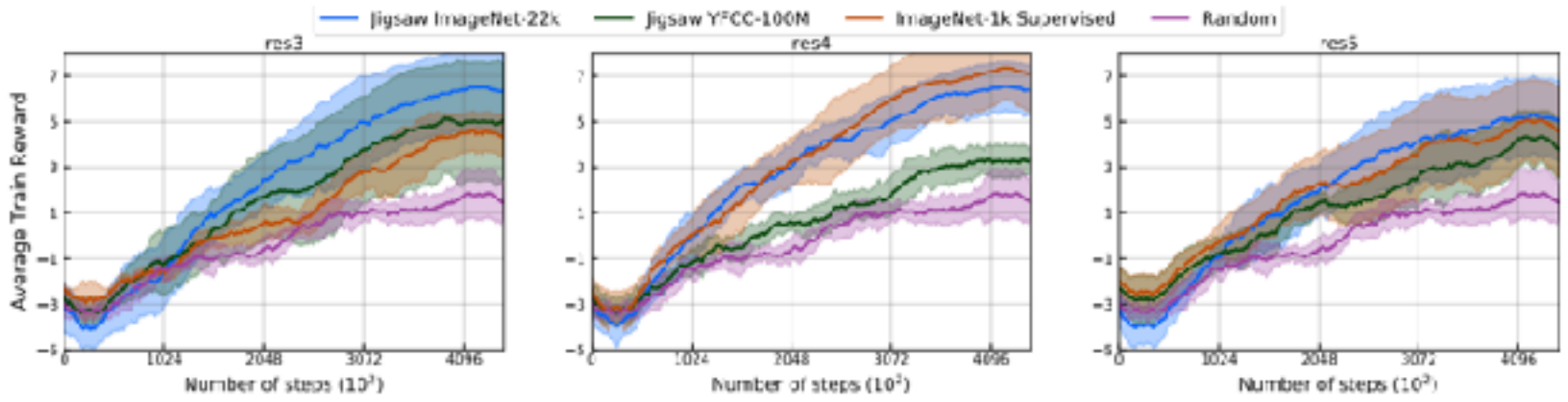
Initialization	Angle Distance		Within t°		
	Mean	Median	11.25	22.5	30
	(Lower is better)		(Higher is better)		
ResNet-50 ImageNet-1k supervised	26.4	17.1	36.1	59.2	68.5
ResNet-50 Places205 supervised	23.3	14.2	41.8	65.2	73.6
ResNet-50 Scratch	26.3	16.1	37.9	60.6	69.0
ResNet-50 Jigsaw ImageNet-1k	24.2	14.5	41.2	64.2	72.5
ResNet-50 Jigsaw ImageNet-22k	22.6	13.4	43.7	66.8	74.7
ResNet-50 Jigsaw YFCC-100M	**22.4**	**13.1**	**44.6**	**67.4**	**75.1**



Jigsaw YFCC-100M self-supervised model outperforms both supervised models across all the metrics by a significant margin.

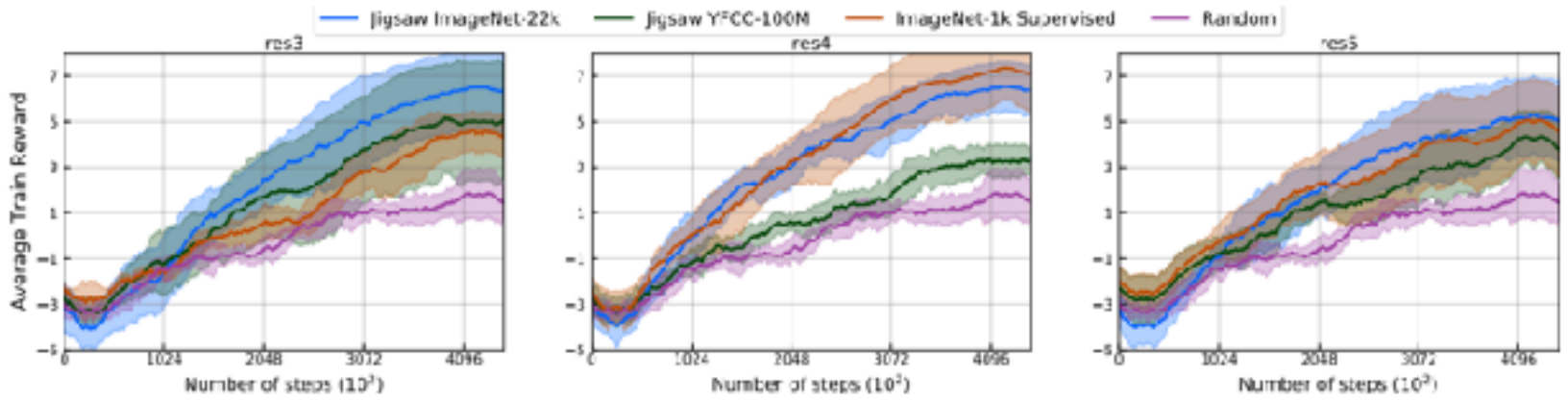
Visual Navigation

- An agent receives a stream of images as input and learns to navigate to a pre-defined location to get a reward.
- The agent is spawned at random locations and must build a contextual map in order to be successful at the task.



Visual Navigation

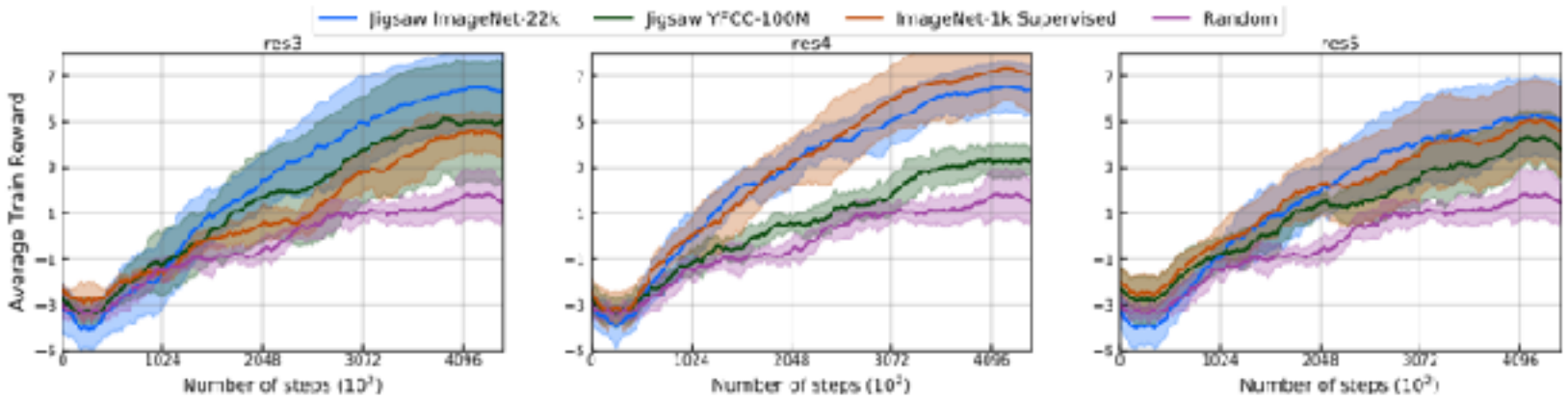
- An agent receives a stream of images as input and learns to navigate to a pre-defined location to get a reward.
- The agent is spawned at random locations and must build a contextual map in order to be successful at the task.



Using res3 layer features, Jigsaw ImageNet model gives a much higher training reward than its supervised counterpart.

Visual Navigation

- An agent receives a stream of images as input and learns to navigate to a pre-defined location to get a reward.
- The agent is spawned at random locations and must build a contextual map in order to be successful at the task.



The self-supervised approach is also a lot more sample-efficient (i.e., higher reward with fewer steps).

Summary

- The paper shows that transfer performance increases with the data size.
- The quality of the representations also improves with higher capacity models and problem complexity.
- Performance improvements on the the three axes are complementary.
- Scaling self-supervision is crucial for surpassing supervised pre-training-based approaches.