

Lecture 22: Self-Supervised Learning

Today: Self-Supervised Learning

Recall: Supervised vs Unsupervised Learning

Supervised Learning

Data: (x, y)

x is data, y is label

Goal: Learn a *function* to map $x \rightarrow y$

Examples: Classification, regression,
object detection, semantic
segmentation, image captioning, etc.

Unsupervised Learning

Data: x

Just data, no labels!

Goal: Learn some underlying
hidden *structure* of the data

Examples: Clustering,
dimensionality reduction, feature
learning, density estimation, etc.

Problem: Supervised Learning is Expensive!

Assume you want to label 1M images. How much will it cost?

Problem: Supervised Learning is Expensive!

Assume you want to label 1M images. How much will it cost?

(1,000,000 images)

(Small to medium sized dataset)

× (10 seconds/image)

(Fast annotation)

× (1/3600 hours/second)

× (\$15 / hour)

(Low wage paid to annotator)

Problem: Supervised Learning is Expensive!

Assume you want to label 1M images. How much will it cost?

(1,000,000 images)

(Small to medium sized dataset)

× (10 seconds/image)

(Fast annotation)

× (1/3600 hours/second)

× (\$15 / hour)

(Low wage paid to annotator)

= **\$41,667**

(Other assumptions: one annotator per image, no benefits / payroll tax / crowdsourcing fee for annotators; not accounting for time to set up tasks for annotators, etc. Real costs could easily be 3x this or more)

Problem: Supervised Learning is Expensive!

Assume you want to label **1B** images. How much will it cost?

(1,000,000,000 images)

(Large dataset)

× (10 seconds/image)

(Fast annotation)

× (1/3600 hours/second)

× (\$15 / hour)

(Low wage paid to annotator)

= **\$41,666,667**

(Other assumptions: one annotator per image, no benefits / payroll tax / crowdsourcing fee for annotators; not accounting for time to set up tasks for annotators, etc. Real costs could easily be 3x this or more)

Problem: Supervised Learning is Not How We Learn

Babies don't get supervision
for everything they see!



[Baby image](#) is [CC0 public domain](#)

Solution: Self-Supervised Learning

Lets build methods that learn from "raw" data – no annotations required

Unsupervised Learning: Model isn't told what to predict. Older terminology, not used as much today.

Self-Supervised Learning: Model is trained to predict some naturally-occurring signal in the raw data rather than human annotations.

Solution: Self-Supervised Learning

Lets build methods that learn from "raw" data – no annotations required

Unsupervised Learning: Model isn't told what to predict. Older terminology, not used as much today.

Self-Supervised Learning: Model is trained to predict some naturally-occurring signal in the raw data rather than human annotations.

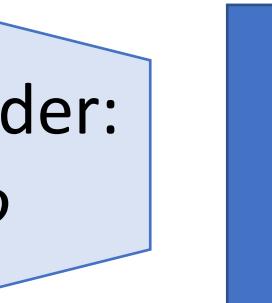
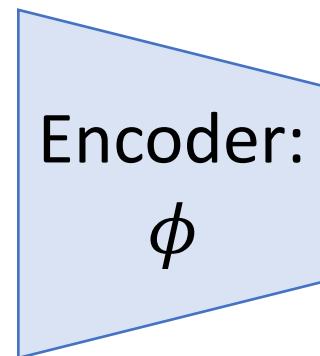
Semi-Supervised Learning: Train jointly with some labeled data and (a lot) of unlabeled data.

Self-Supervised Learning: Pretext then Transfer

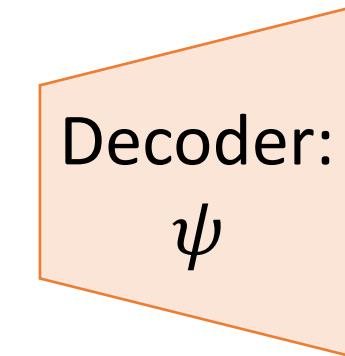
Step 1: Pretrain a network on a pretext task that doesn't require supervision



Input Image: x



Features: $\phi(x)$



Prediction: \hat{y}

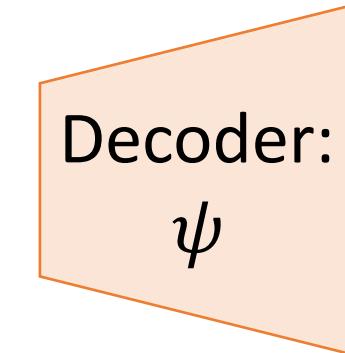
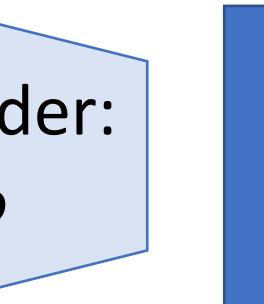
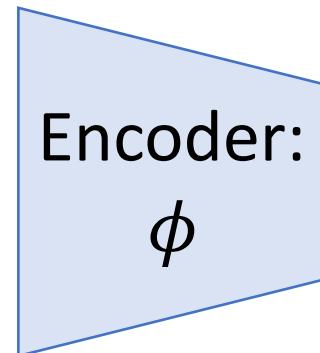
Loss:
 $L(\hat{y}, y)$

Self-Supervised Learning: Pretext then Transfer

Step 1: Pretrain a network on a pretext task that doesn't require supervision



Input Image: x



Loss:
 $L(\hat{y}, y)$

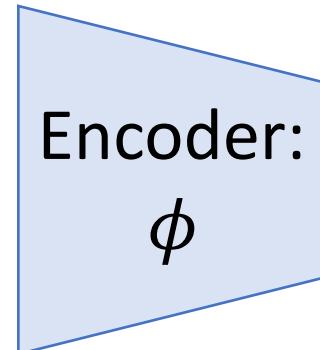
Features: $\phi(x)$

Prediction: \hat{y}

Step 2: Transfer encoder to downstream tasks via linear classifiers, KNN, finetuning



Input Image: x



Features: $\phi(x)$

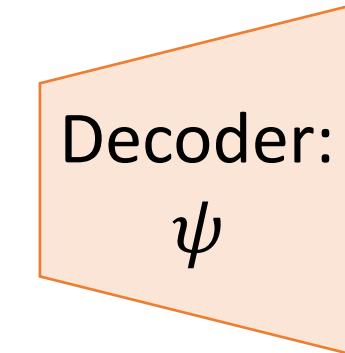
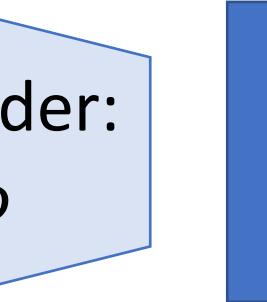
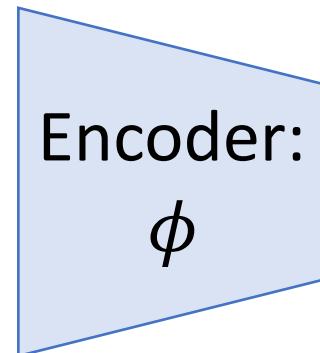
Downstream tasks:
Image classification,
object detection,
semantic segmentation

Self-Supervised Learning: Pretext then Transfer

Step 1: Pretrain a network on a pretext task that doesn't require supervision



Input Image: x



Loss:
 $L(\hat{y}, y)$

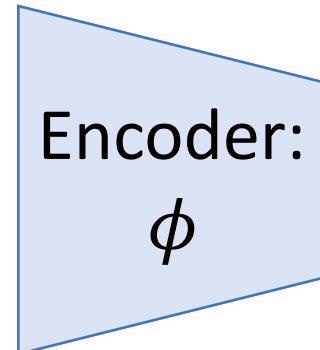
Features: $\phi(x)$

Prediction: \hat{y}

Step 2: Transfer encoder to downstream tasks via linear classifiers, KNN, finetuning



Input Image: x



Features: $\phi(x)$

Goal: Pretrain + Transfer does better than supervised pretraining, and better than directly training on downstream

Self-Supervised Learning: Pretext Tasks

Generative: Predict part of the input signal

- Autoencoders (sparse, denoising, masked)
- Autoregressive
- GANs
- Colorization
- Inpainting

Discriminative: Predict something about the input signal

- Context prediction
- Rotation
- Clustering
- Contrastive

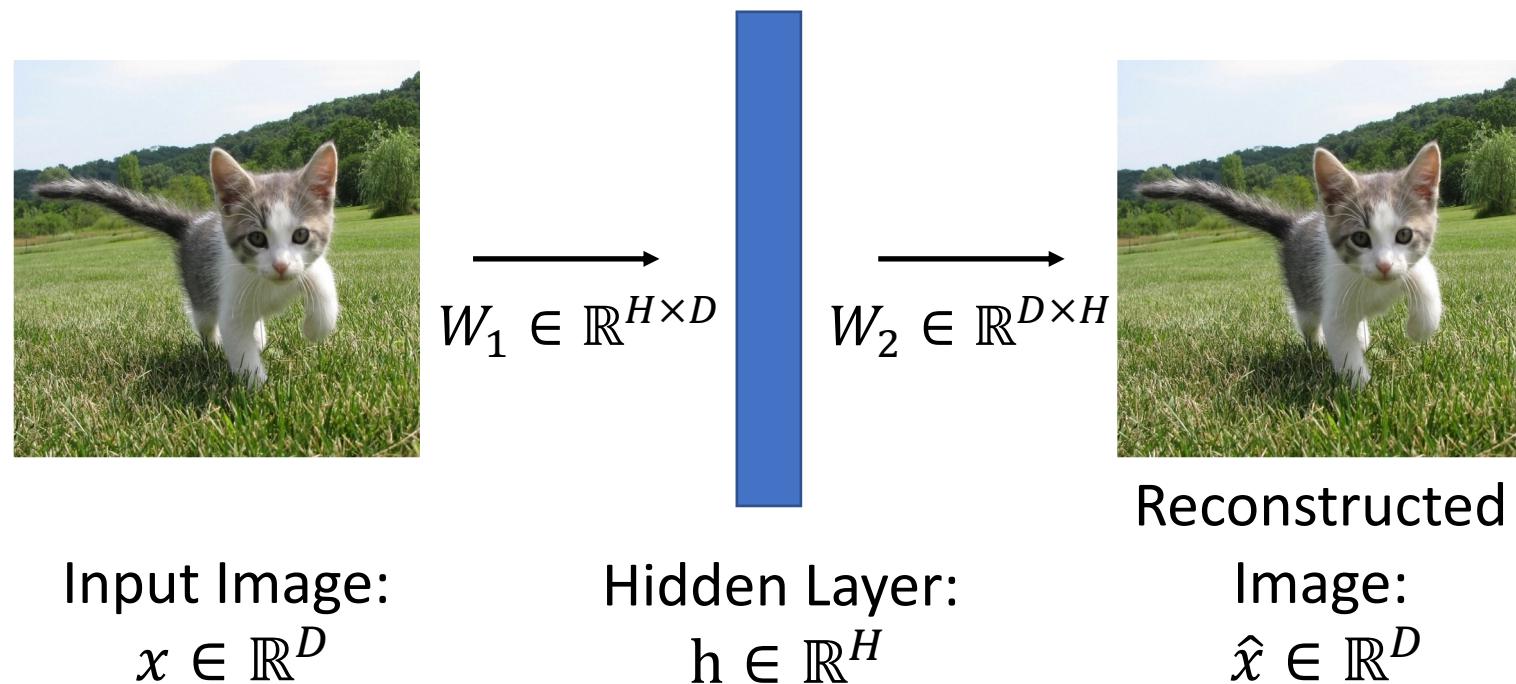
Multimodal: Use some additional signal in addition to RGB images

- Video
- 3D
- Sound
- Language

Recall: Autoencoder

Autoencoder tries to reconstruct inputs. Hidden layer (hopefully) learns good representations.
Generative pretraining task!

$$\begin{aligned} L(x) &= R(x, \hat{x}) \\ &= \|x - \hat{x}\|_2^2 \end{aligned}$$



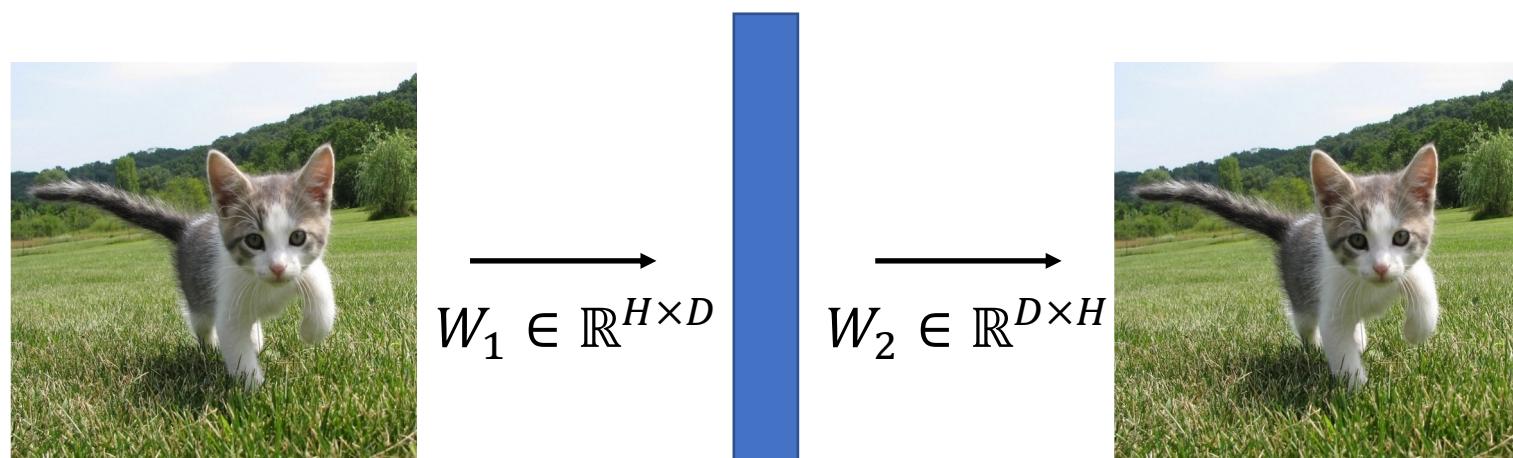
Lee et al, "Efficient Sparse Coding Algorithms", NeurIPS 2006; Ranzato et al, "Efficient Learning of Sparse Representations with an Energy-Based Model", NeurIPS 2006;
Lee et al, "Sparse deep belief net models for visual area V2", NeurIPS 2007; Ng, "Sparse Autoencoder", CS294A Lecture Notes

Recall: Autoencoder

Autoencoder tries to reconstruct inputs. Hidden layer (hopefully) learns good representations

$$\begin{aligned} L(x) &= R(x, \hat{x}) \\ &= \|x - \hat{x}\|_2^2 \end{aligned}$$

H < D is the only thing forcing non-trivial hidden representations...



Input Image:
 $x \in \mathbb{R}^D$

Hidden Layer:
 $h \in \mathbb{R}^H$

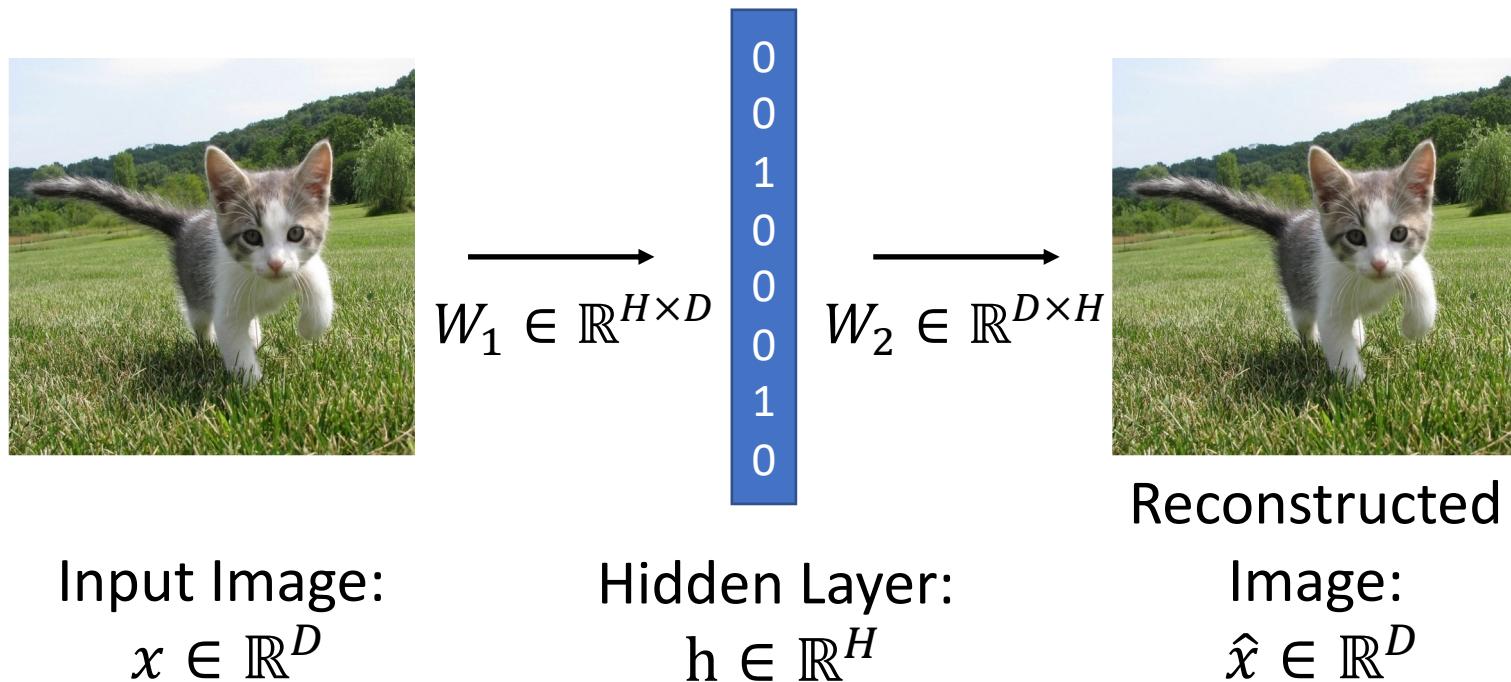
Reconstructed
Image:
 $\hat{x} \in \mathbb{R}^D$

Lee et al, "Efficient Sparse Coding Algorithms", NeurIPS 2006; Ranzato et al, "Efficient Learning of Sparse Representations with an Energy-Based Model", NeurIPS 2006;
Lee et al, "Sparse deep belief net models for visual area V2", NeurIPS 2007; Ng, "Sparse Autoencoder", CS294A Lecture Notes

Sparse Autoencoder

Train an autoencoder to **reconstruct inputs** with **sparse activations** (mostly 0). Many ways to implement sparsity penalties!

$$\begin{aligned} L(x) &= R(x, \hat{x}) + \lambda S(h) \\ &= \|x - \hat{x}\|_2^2 + \lambda \|h\|_1 \end{aligned}$$



Lee et al, "Efficient Sparse Coding Algorithms", NeurIPS 2006; Ranzato et al, "Efficient Learning of Sparse Representations with an Energy-Based Model", NeurIPS 2006; Lee et al, "Sparse deep belief net models for visual area V2", NeurIPS 2007; Ng, "Sparse Autoencoder", CS294A Lecture Notes; Le et al, "Building high-level features using large-scale unsupervised learning", ICML 2012

Denoising Autoencoder

Train an autoencoder to
reconstruct noisy inputs
(pixels randomly set to zero)

$$\begin{aligned} L(x) &= R(x, \hat{x}) \\ &= \|x - \hat{x}\|_2^2 \end{aligned}$$



$W_1 \in \mathbb{R}^{H \times D}$



$W_2 \in \mathbb{R}^{D \times H}$



Reconstructed

Input Image:
 $x \in \mathbb{R}^D$

Corrupted Image:
 $x \in \mathbb{R}^D$

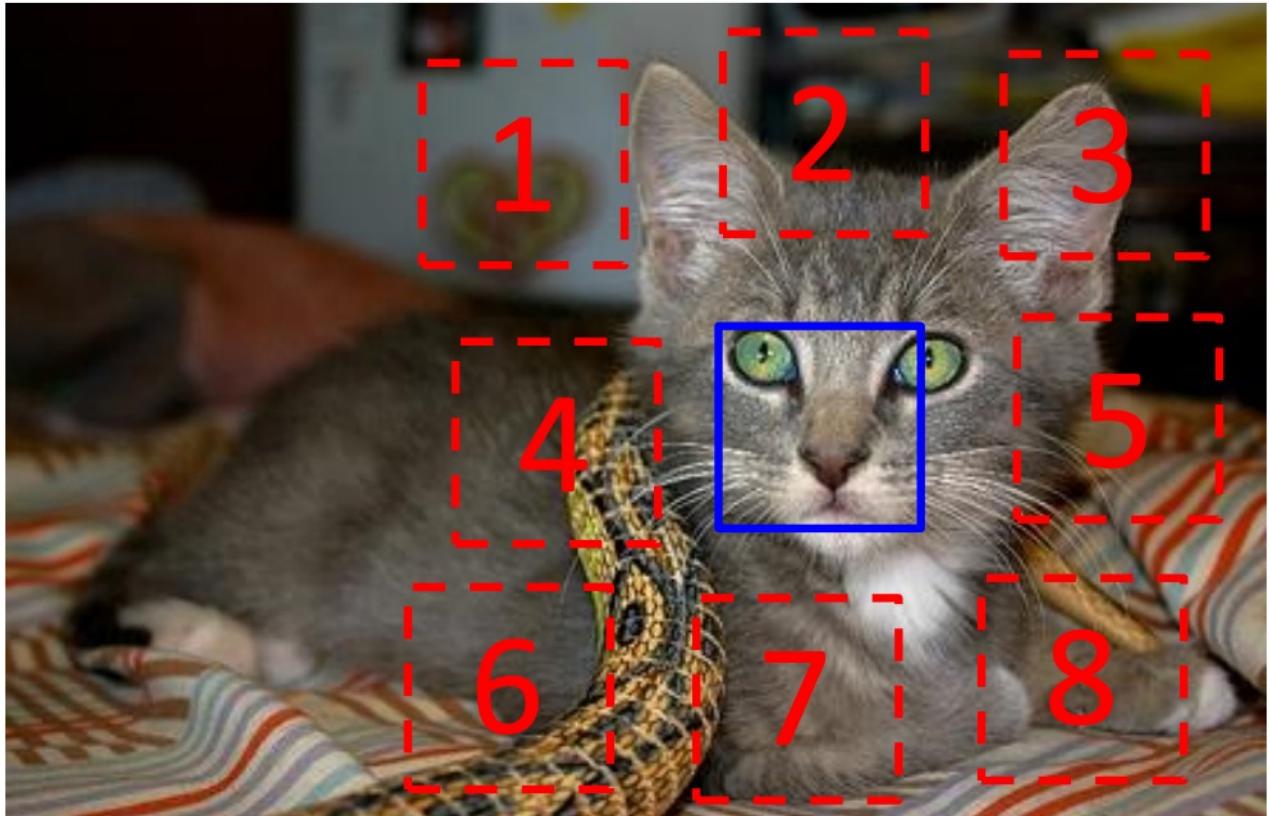
Hidden Layer:
 $h \in \mathbb{R}^H$

Image:
 $\hat{x} \in \mathbb{R}^D$

Context Prediction

Model predicts relative location of two patches from the same image.
Discriminative pretraining task

Intuition: Requires understanding objects and their parts



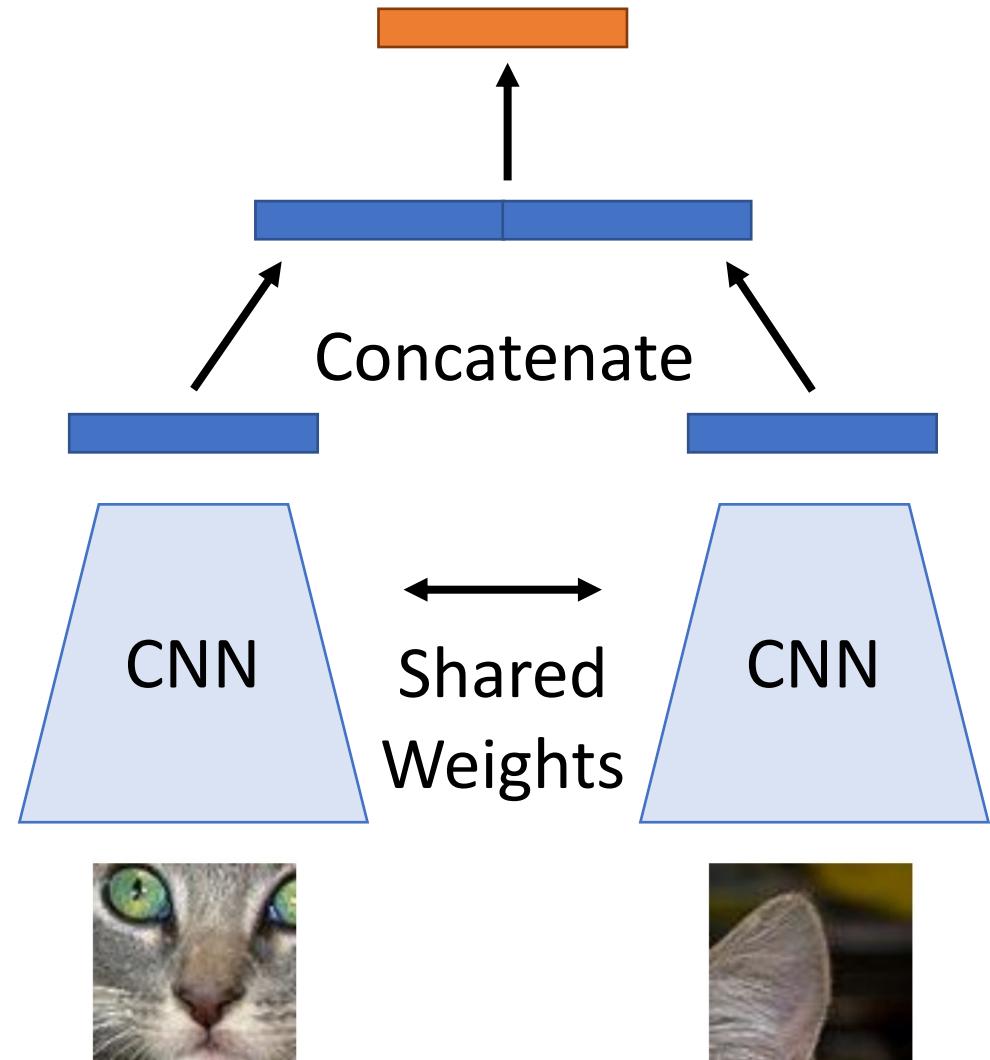
$$X = \left(\begin{array}{c} \text{Patch 5 (cat's face)} \\ , \\ \text{Patch 3 (cat's ear)} \end{array} \right); Y = 3$$

Context Prediction

Model predicts relative location of two patches from the same image.
Discriminative pretraining task

Intuition: Requires understanding objects and their parts

Classification over 8 positions



Doersch et al, "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

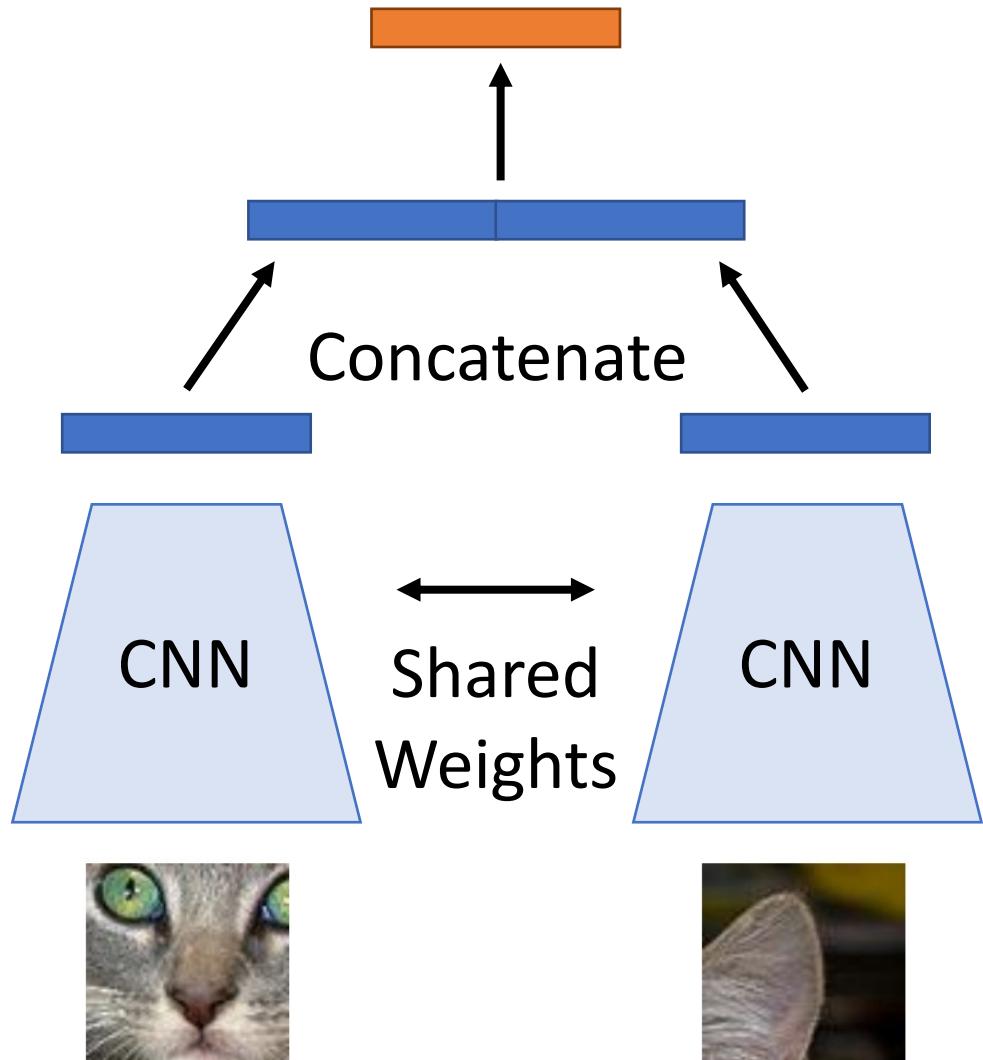
Context Prediction

Model predicts relative location of two patches from the same image.
Discriminative pretraining task

Intuition: Requires understanding objects and their parts

Two networks with shared weights sometimes called a “Siamese network” – but I don’t really like this term

Classification over 8 positions



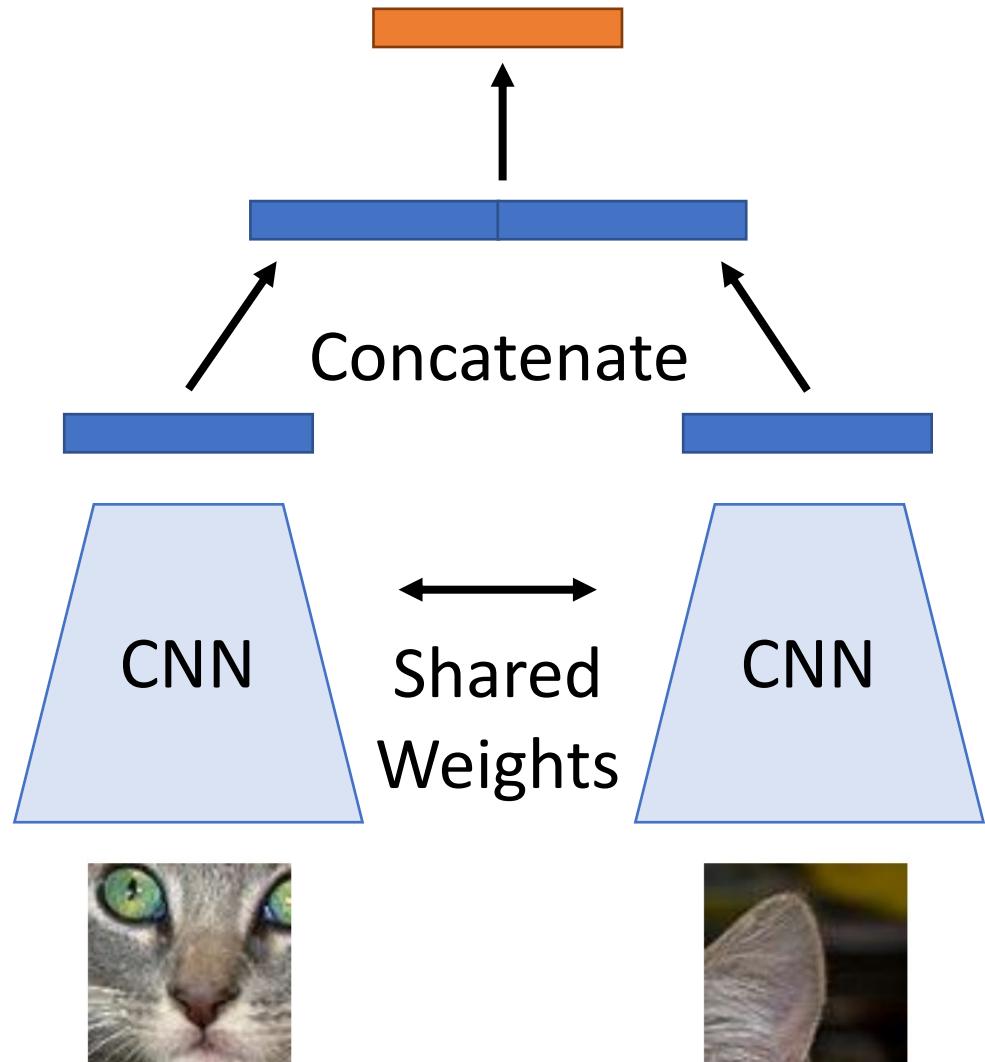
Context Prediction

Model predicts relative location of two patches from the same image.
Discriminative pretraining task

Intuition: Requires understanding objects and their parts

“For experiments, we use a ConvNet trained on a K40 GPU for approximately four weeks.”

Classification over 8 positions



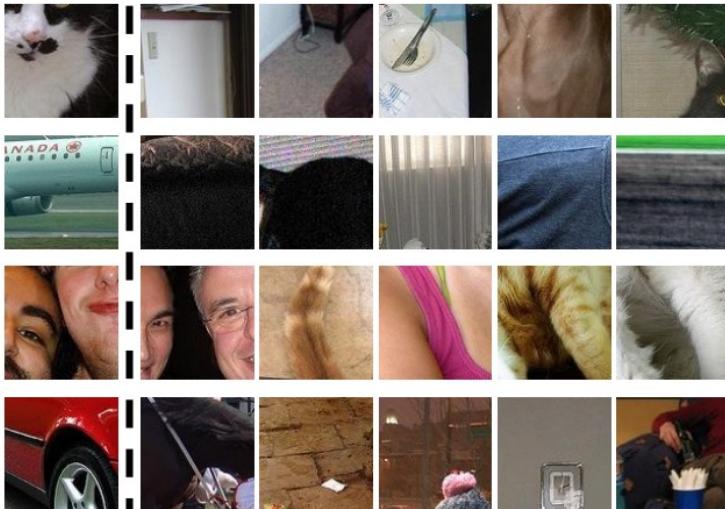
Context Prediction: Nearest Neighbors in Feature Space

Input Patch



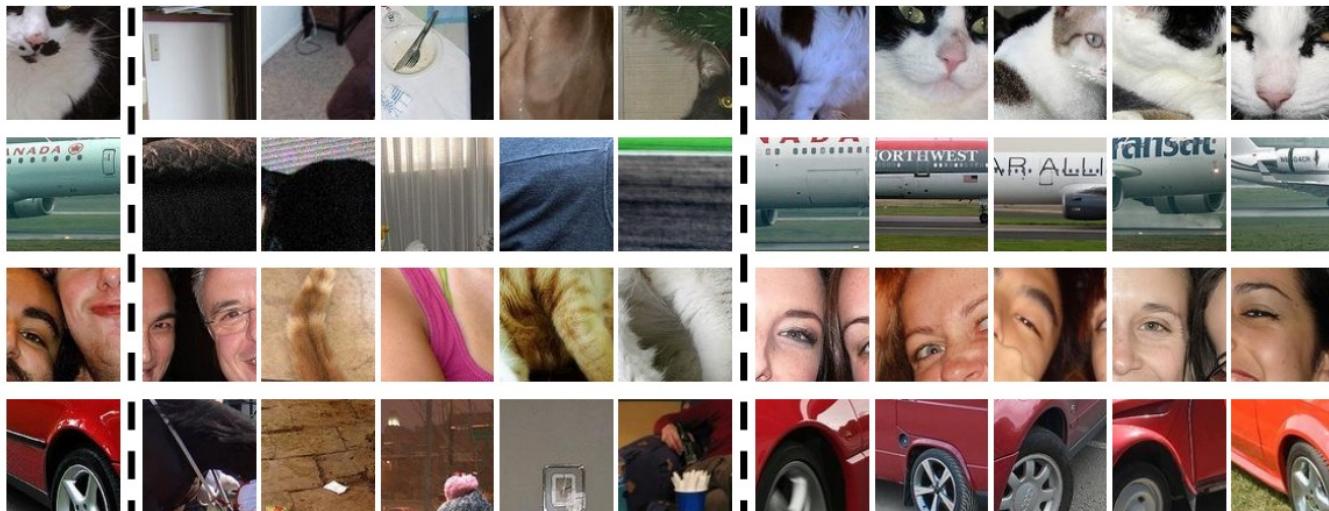
Context Prediction: Nearest Neighbors in Feature Space

Input Patch Random Init



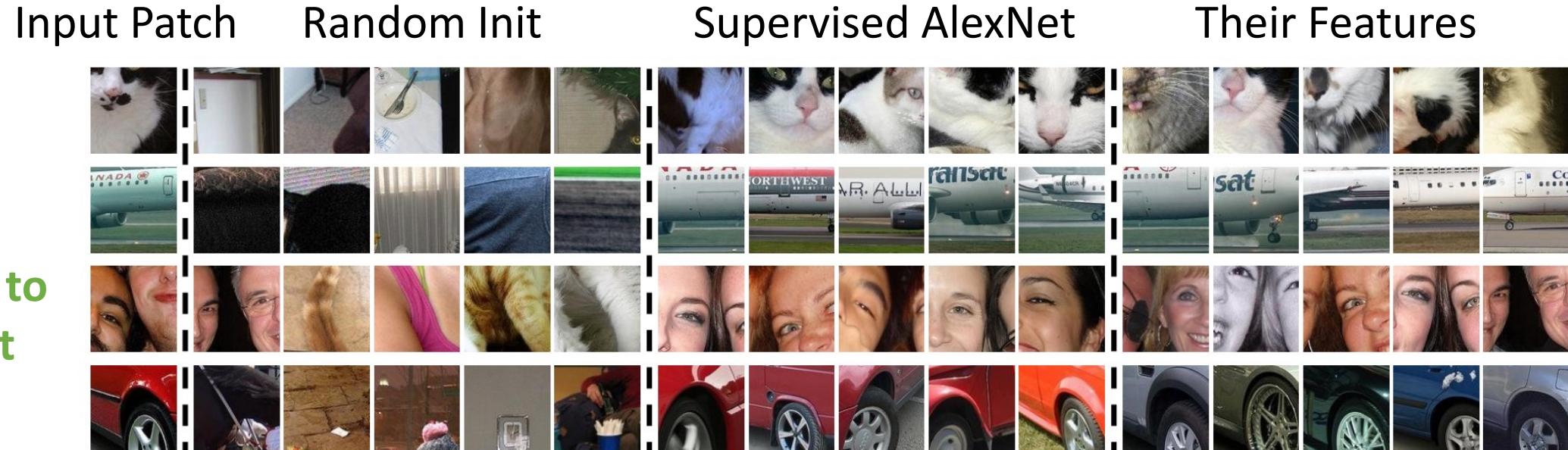
Context Prediction: Nearest Neighbors in Feature Space

Input Patch Random Init Supervised AlexNet



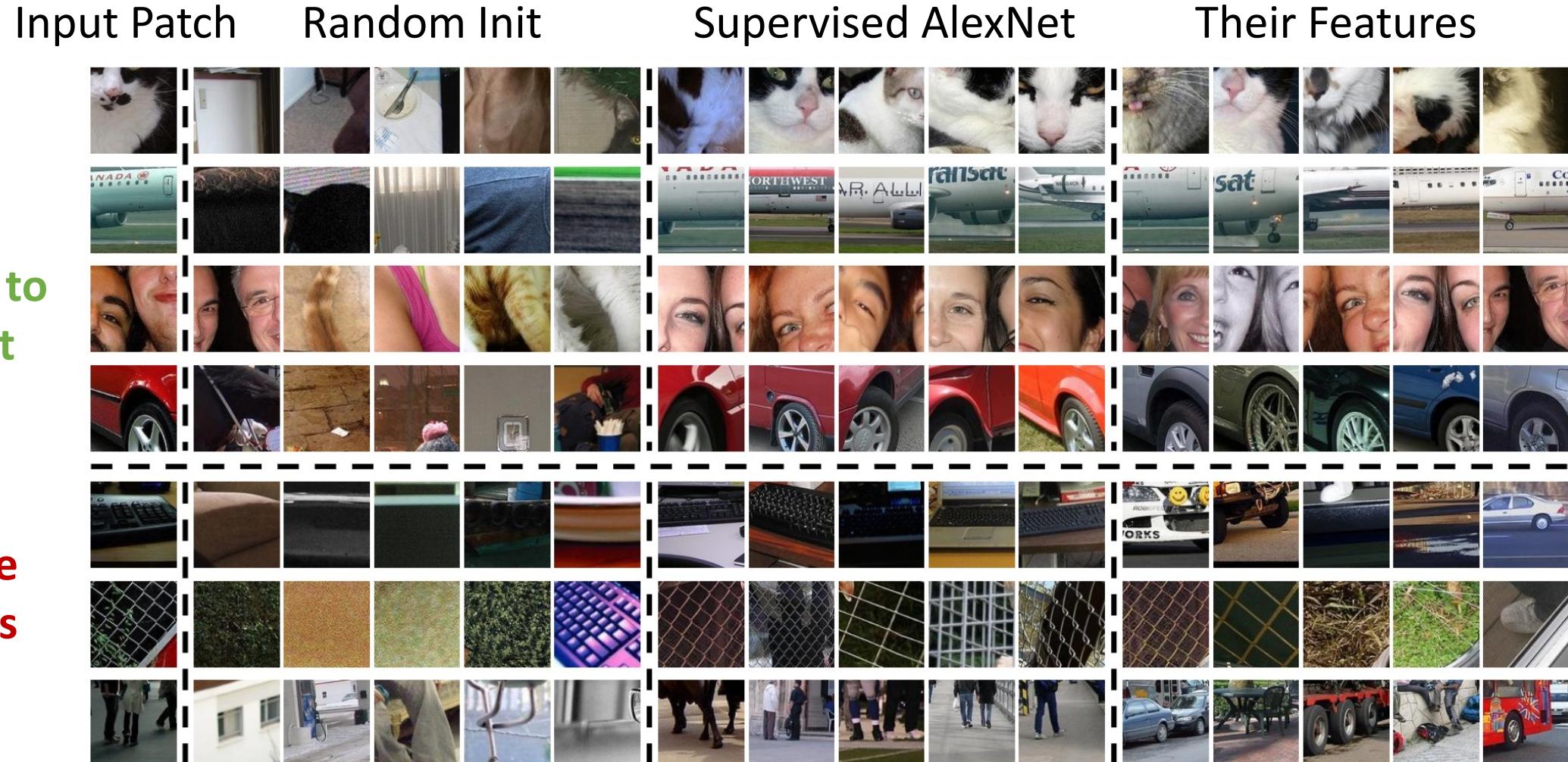
Doersch et al, "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Context Prediction: Nearest Neighbors in Feature Space



Doersch et al, "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

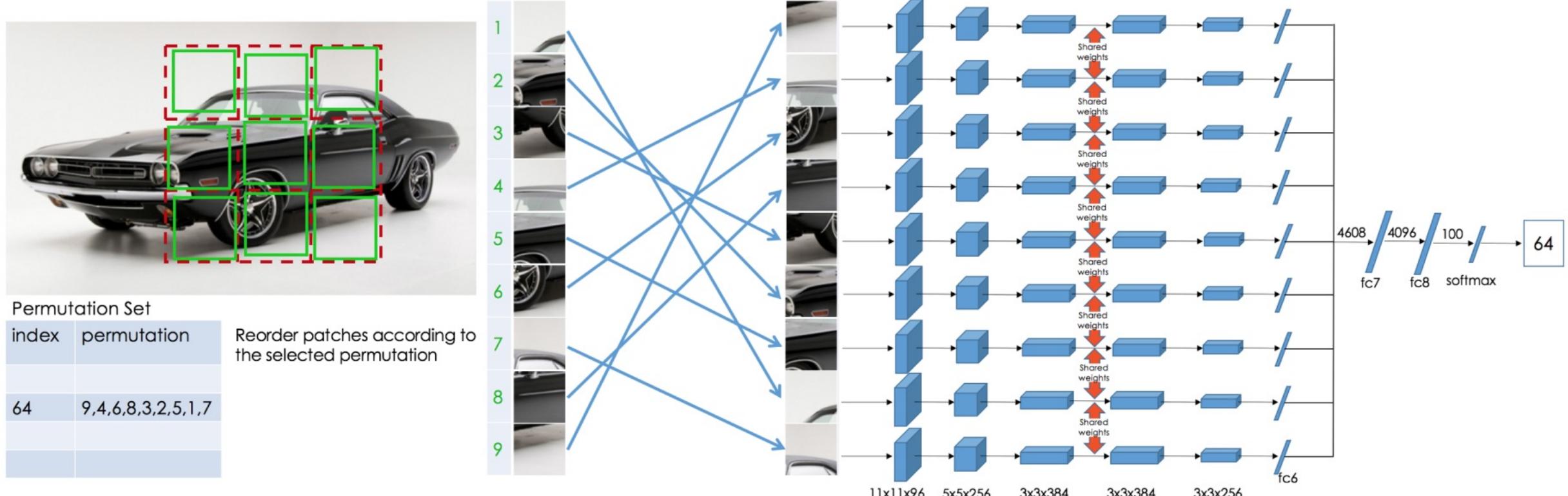
Context Prediction: Nearest Neighbors in Feature Space



Doersch et al, "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Extension: Solving Jigsaw Puzzles

Rather than predict relative position of two patches, instead predict permutation to “unscramble” 9 shuffled patches

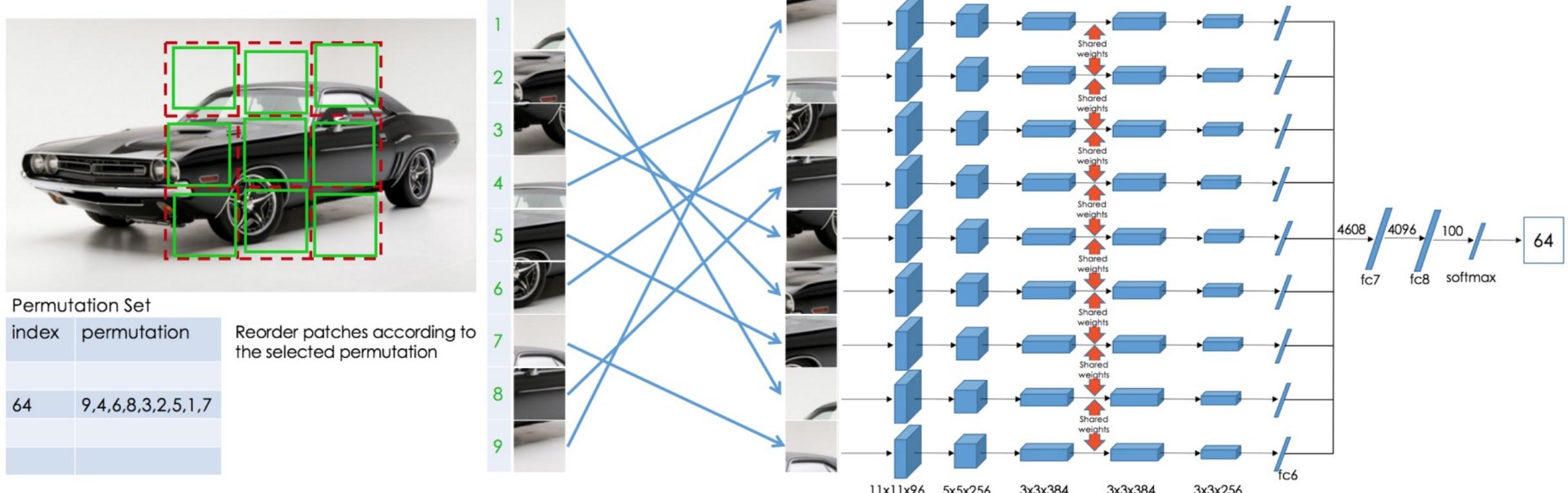


Noroozi and Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles”, ECCV 2016

Extension: Solving Jigsaw Puzzles

Problem: These methods
only work on patches,
not whole images!

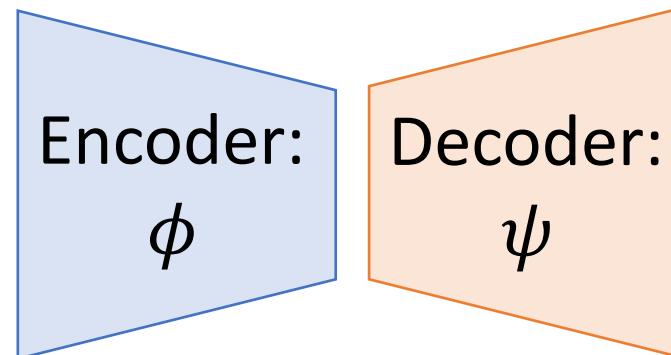
Rather than predict relative position of two patches, instead
predict permutation to “unscramble” 9 shuffled patches



Noroozi and Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles”, ECCV 2016

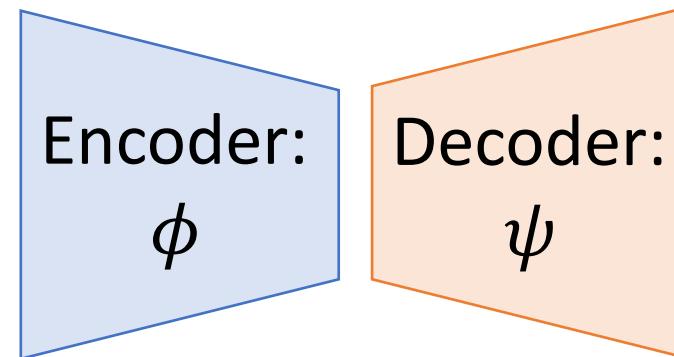
Context Encoders: Learning by Inpainting

Input Image



Context Encoders: Learning by Inpainting

Input Image



Predict Missing Pixels

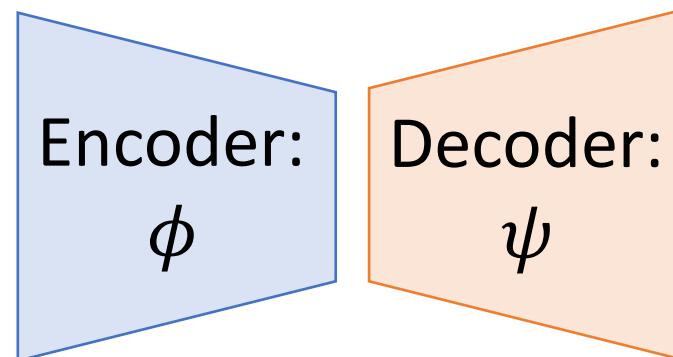


Human Artist

Pathak et al, "Context Encoders: Feature Learning by Inpainting", CVPR 2016

Context Encoders: Learning by Inpainting

Input Image



Predict Missing Pixels

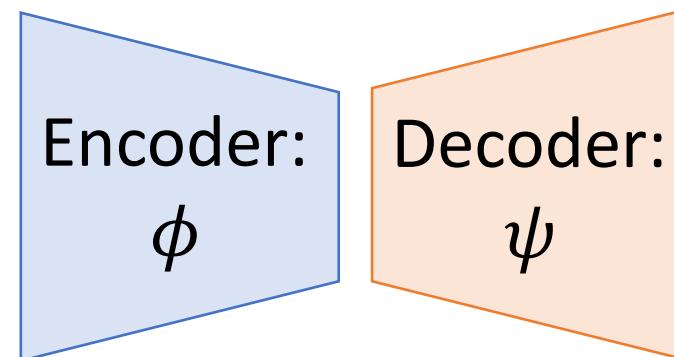


L2 Loss
(Best for feature learning)

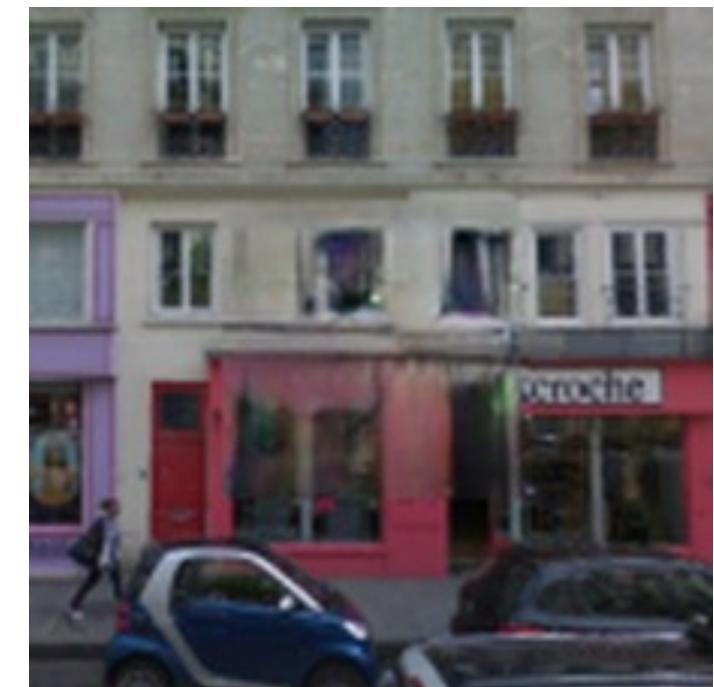
Pathak et al, "Context Encoders: Feature Learning by Inpainting", CVPR 2016

Context Encoders: Learning by Inpainting

Input Image



Predict Missing Pixels



L2 + Adversarial Loss
(Best for nice images)

Pathak et al, "Context Encoders: Feature Learning by Inpainting", CVPR 2016

Colorization

Intuition: A model must be able to identify objects to be able to colorize them



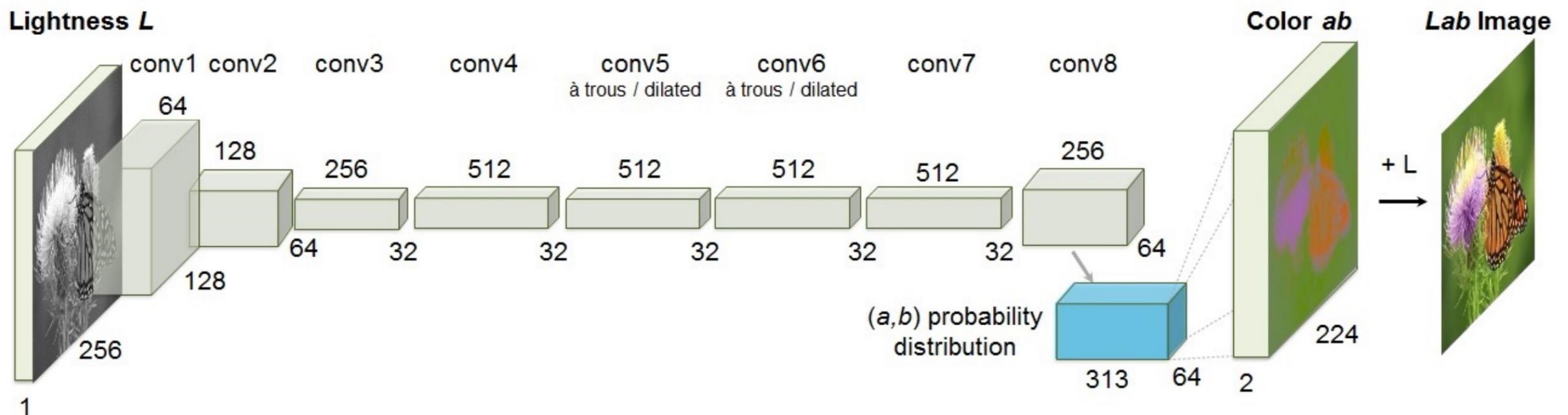
Input: Grayscale Image



Output: Color Image

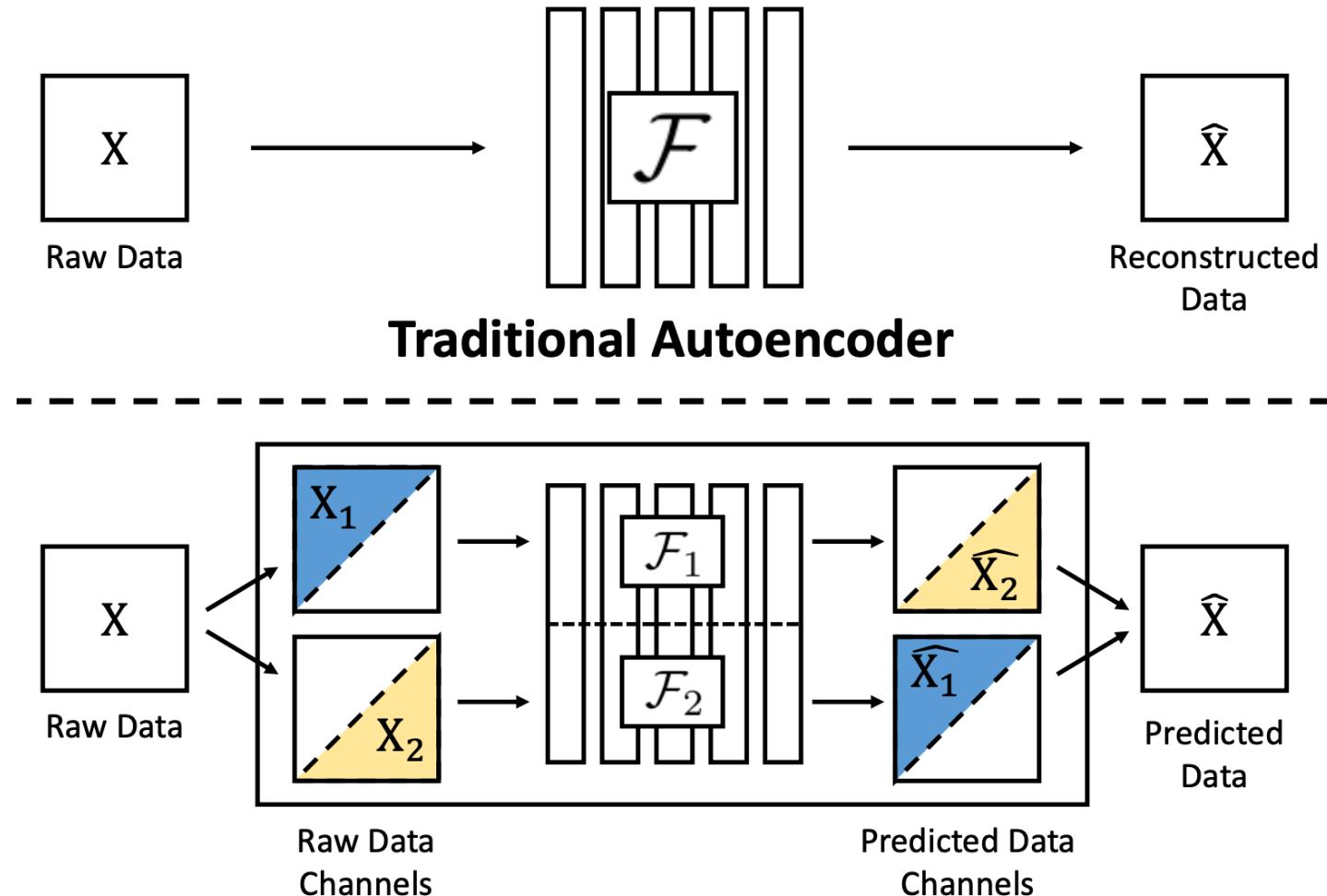
Zhang et al, "Colorful Image Colorization", ECCV 2016

Colorization



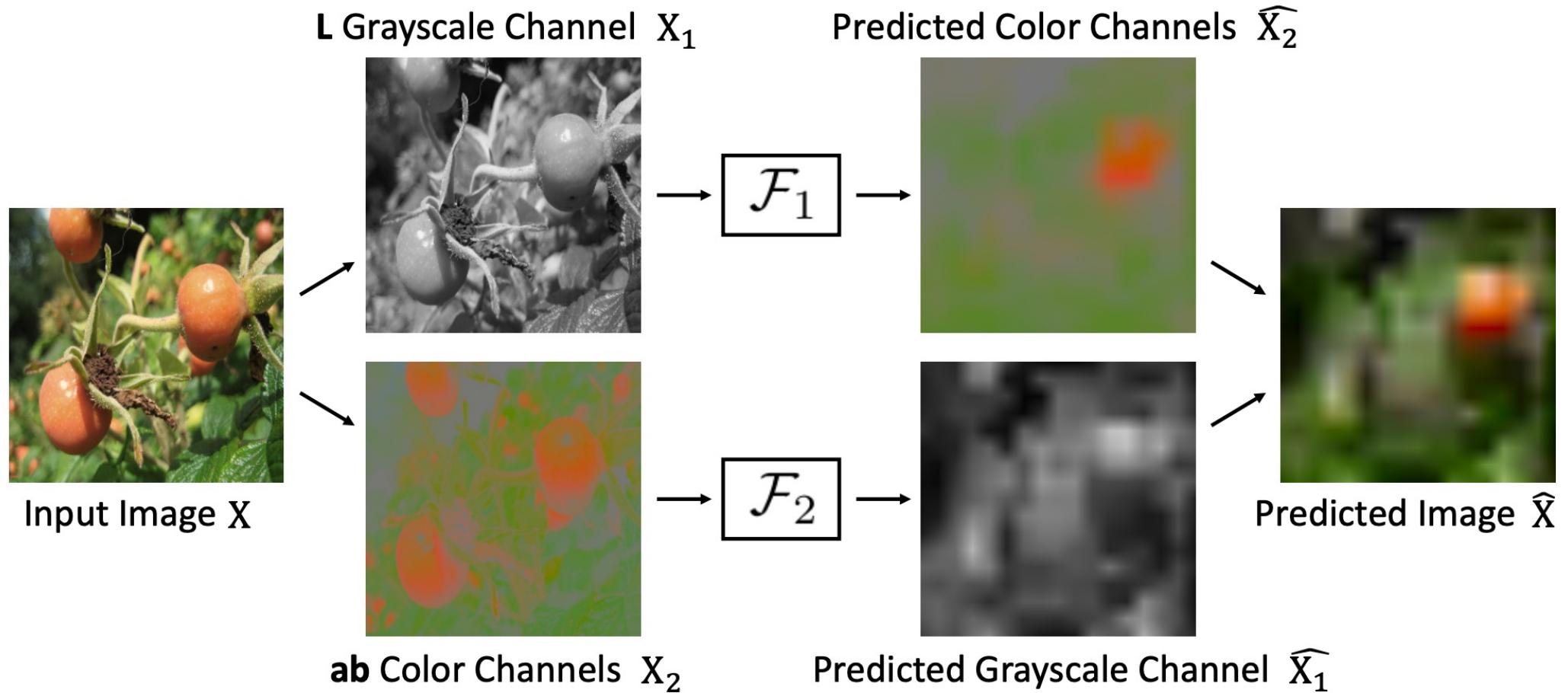
Zhang et al, "Colorful Image Colorization", ECCV 2016

Colorization Extension: Split-Brain Autoencoder



Zhang et al, "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction", CVPR 2017

Colorization Extension: Split-Brain Autoencoder



Colorization Extension: Split-Brain Autoencoder

Concern: Generative pretexts encourage spending model capacity on details unimportant for downstream tasks (e.g. regressing exact right shade of orange)



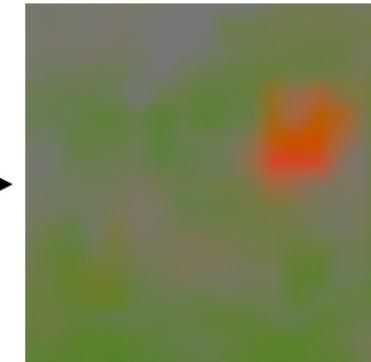
Input Image X

L Grayscale Channel X_1



$$\rightarrow \mathcal{F}_1 \rightarrow$$

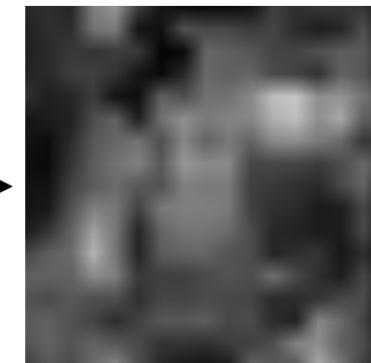
Predicted Color Channels \widehat{X}_2



ab Color Channels X_2

Predicted Grayscale Channel \widehat{X}_1

$$\rightarrow \mathcal{F}_2 \rightarrow$$



Predicted Image \widehat{X}

Colorization Extension: Split-Brain Autoencoder

Concern: Generative pretexts encourage spending model capacity on details unimportant for downstream tasks (e.g. regressing exact right shade of orange)



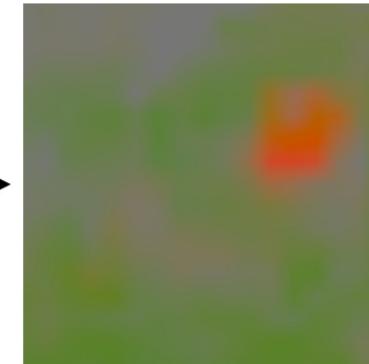
Input Image X

L Grayscale Channel X_1



$$\rightarrow \mathcal{F}_1 \rightarrow$$

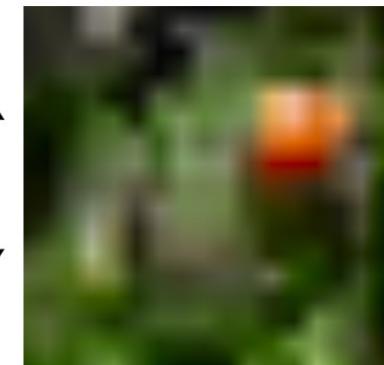
Predicted Color Channels \widehat{X}_2



ab Color Channels X_2

$$\rightarrow \mathcal{F}_2 \rightarrow$$

Predicted Grayscale Channel \widehat{X}_1

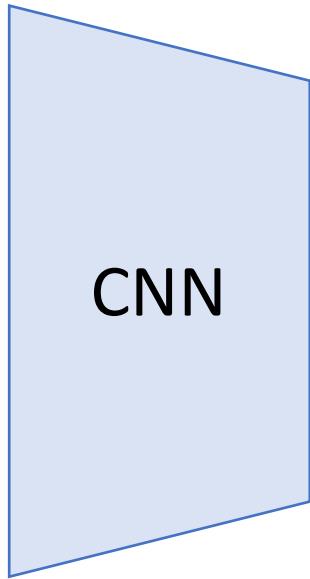


Predicted Image \widehat{X}

Solution: Discriminative pretext tasks that require classification, not generation

Deep Clustering

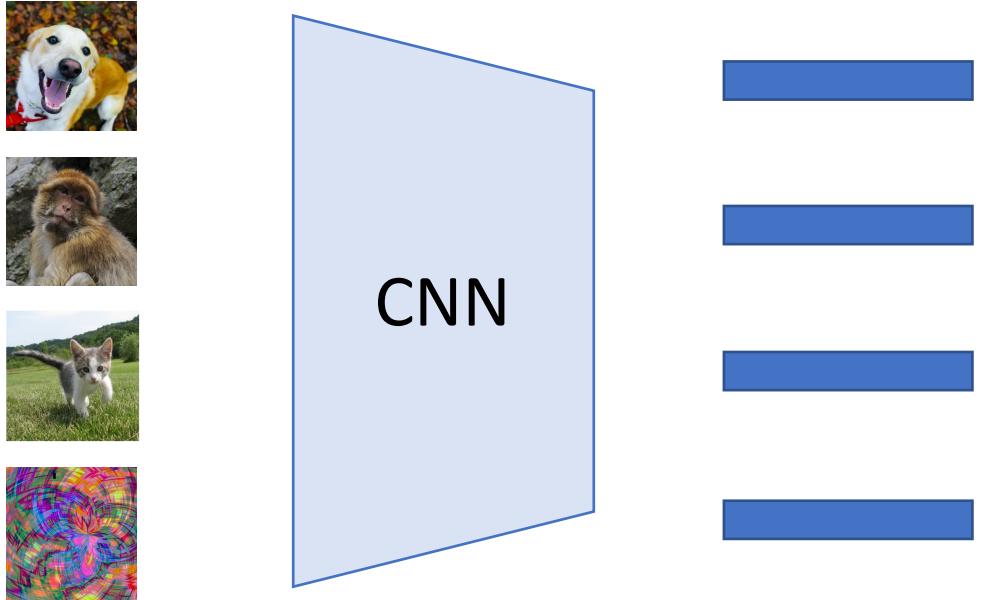
(1) Randomly initialize a CNN



Caron et al, "Deep Clustering for Unsupervised Learning of Visual Features", ECCV 2018
Caron et al, "Unsupervised Pre-Training of Image Features on Non-Curated Data", ICCV 2019
Yan et al, "ClusterFit: Improving Generalization of Visual Representations", CVPR 2020
Caron et al, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", NeurIPS 2020

Deep Clustering

(1) Randomly initialize a CNN

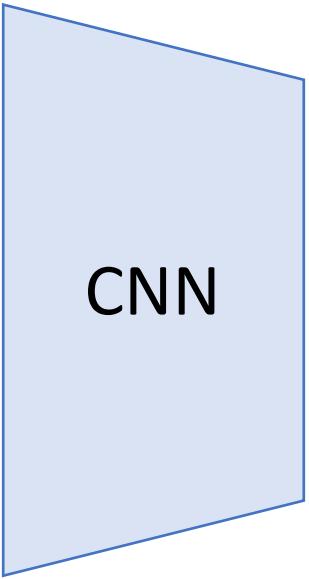


(2) Run many images through
CNN, get their final-layer features

Caron et al, "Deep Clustering for Unsupervised Learning of Visual Features", ECCV 2018
Caron et al, "Unsupervised Pre-Training of Image Features on Non-Curated Data", ICCV 2019
Yan et al, "ClusterFit: Improving Generalization of Visual Representations", CVPR 2020
Caron et al, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", NeurIPS 2020

Deep Clustering

(1) Randomly initialize a CNN



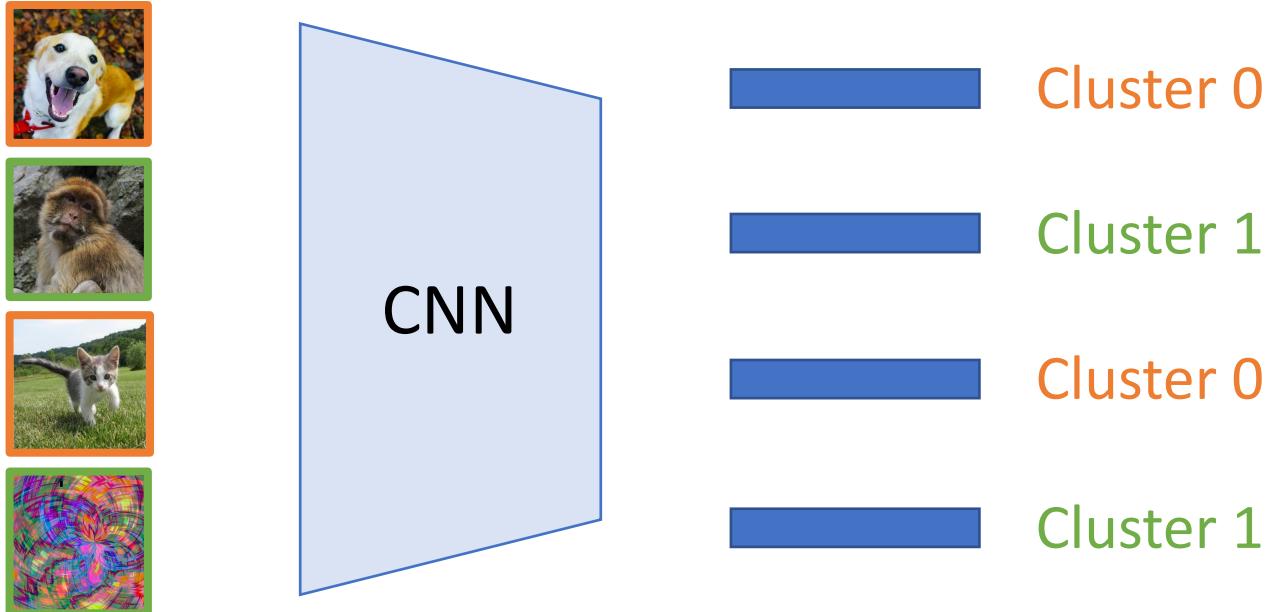
(3) Cluster the features with K-Means;
record cluster for each feature

(2) Run many images through
CNN, get their final-layer features

Caron et al, "Deep Clustering for Unsupervised Learning of Visual Features", ECCV 2018
Caron et al, "Unsupervised Pre-Training of Image Features on Non-Curated Data", ICCV 2019
Yan et al, "ClusterFit: Improving Generalization of Visual Representations", CVPR 2020
Caron et al, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", NeurIPS 2020

Deep Clustering

(1) Randomly initialize a CNN



(2) Run many images through
CNN, get their final-layer features

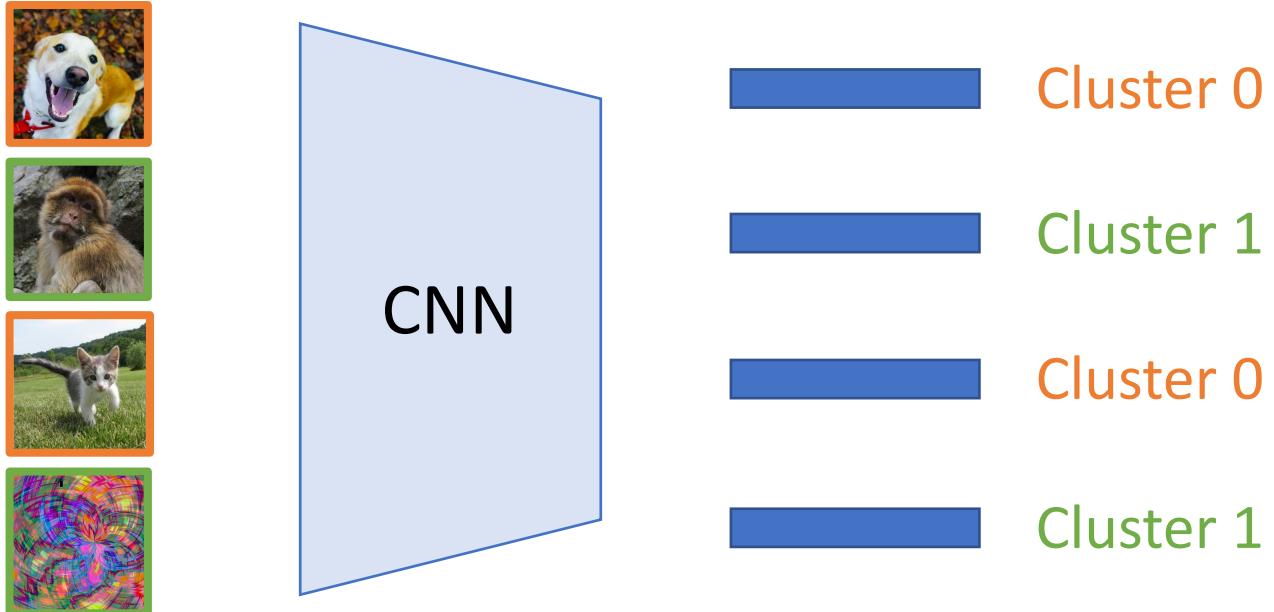
(3) Cluster the features with K-Means;
record cluster for each feature

(4) Use cluster assignments as pseudo-
labels for each image; train the CNN to
predict cluster assignments

Caron et al, "Deep Clustering for Unsupervised Learning of Visual Features", ECCV 2018
Caron et al, "Unsupervised Pre-Training of Image Features on Non-Curated Data", ICCV 2019
Yan et al, "ClusterFit: Improving Generalization of Visual Representations", CVPR 2020
Caron et al, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", NeurIPS 2020

Deep Clustering

(1) Randomly initialize a CNN



(2) Run many images through
CNN, get their final-layer features

(3) Cluster the features with K-Means;
record cluster for each feature

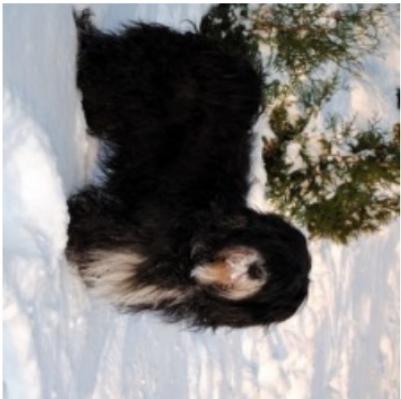
(4) Use cluster assignments as pseudo-
labels for each image; train the CNN to
predict cluster assignments

(5) Repeat: GOTO (2)

Caron et al, "Deep Clustering for Unsupervised Learning of Visual Features", ECCV 2018
Caron et al, "Unsupervised Pre-Training of Image Features on Non-Curated Data", ICCV 2019
Yan et al, "ClusterFit: Improving Generalization of Visual Representations", CVPR 2020
Caron et al, "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments", NeurIPS 2020

RotNet: Predict Rotation

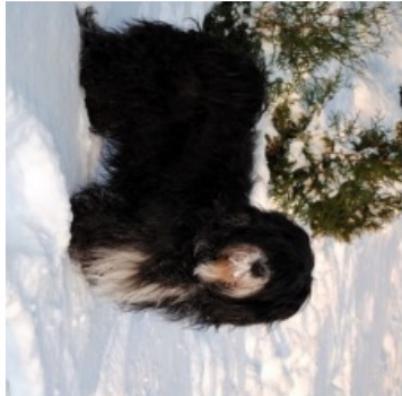
4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)



Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

RotNet: Predict Rotation

4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)

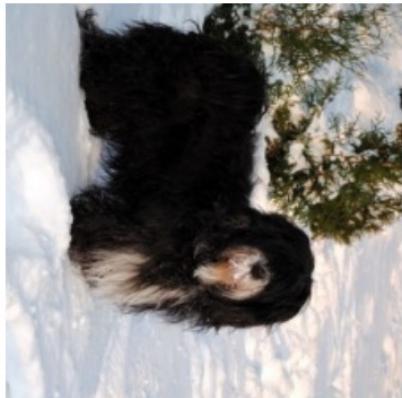


90

Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

RotNet: Predict Rotation

4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)



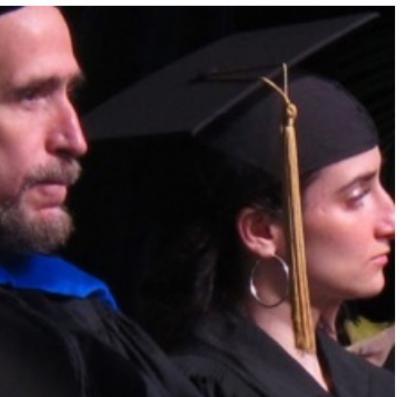
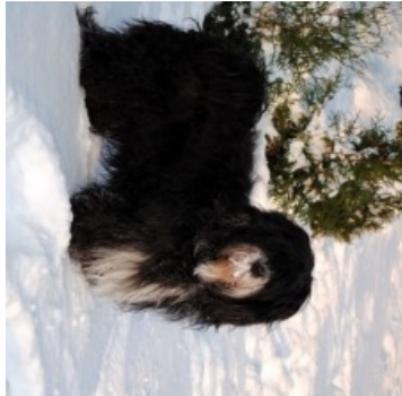
90

270

Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

RotNet: Predict Rotation

4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)



90

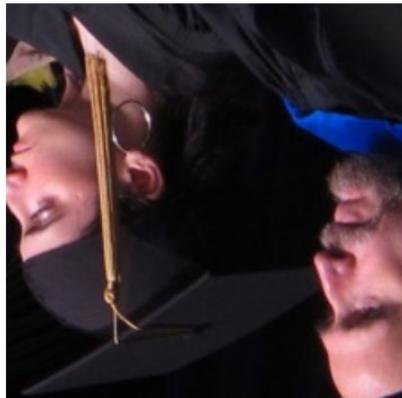
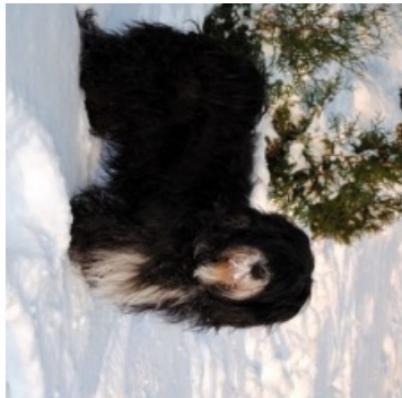
270

180

Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

RotNet: Predict Rotation

4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)



90

270

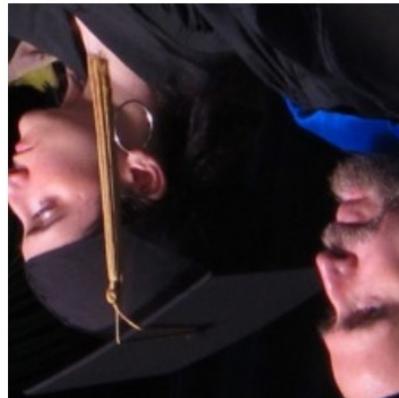
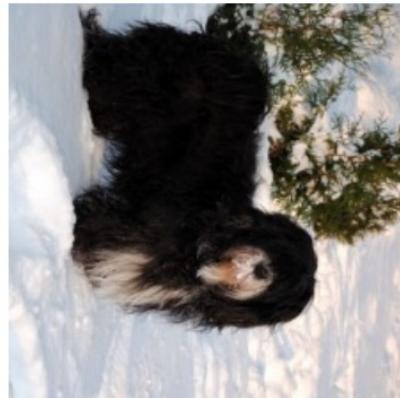
180

0

Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

RotNet: Predict Rotation

4-way classification task: How much was each image rotated? (0, 90, 180, or 270 degrees)



90

270

180

0

270

Gidaris et al, "Unsupervised representation learning by predicting image rotations", ICLR 2018

Which SSL Method is best?

Fair evaluation of SSL methods is very hard! No theory, so we need to rely on experiment

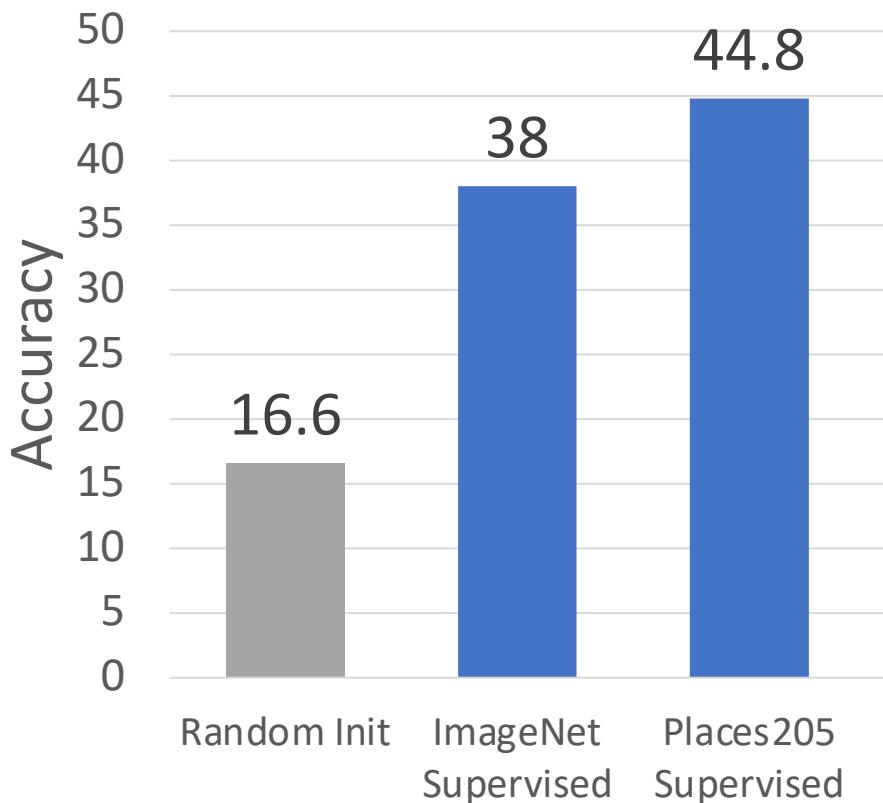
Many choices in experimental setup, huge variations from paper to paper:

- CNN architecture? AlexNet, ResNet50, something else?
- Pretraining dataset? ImageNet, or something else?
- Downstream task? ImageNet classification, detection, something else?
- Pretraining hyperparameters? Learning rates, training iterations, data augmentation?
- Transfer learning protocol?
 - Linear probe? From which layer? How to train linear models? SGD, something else?
 - Transfer learning hyperparameters? Data augmentation or BatchNorm during transfer learning?
 - Fine-tune? From which layer? Architecture of “head” you attach? Linear or nonlinear? Fine-tuning hyperparameters?
 - KNN? What value of K? Normalization on features?

Which SSL Method is best?

Some papers have tried to do fair comparisons of many SSL methods

Places205 Linear Classification from AlexNet conv5

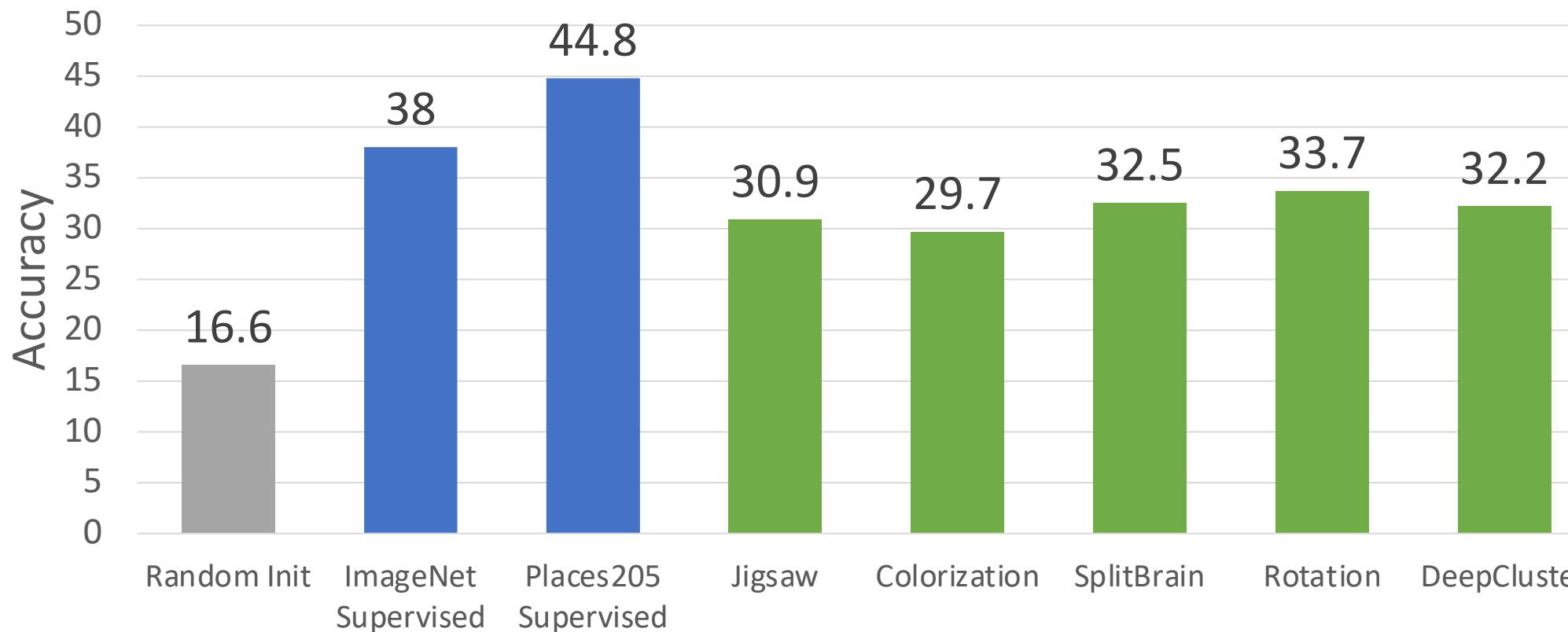


Goyal et al, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", ICCV 2019

Which SSL Method is best?

Some papers have tried to do fair comparisons of many SSL methods

Places205 Linear Classification from AlexNet conv5

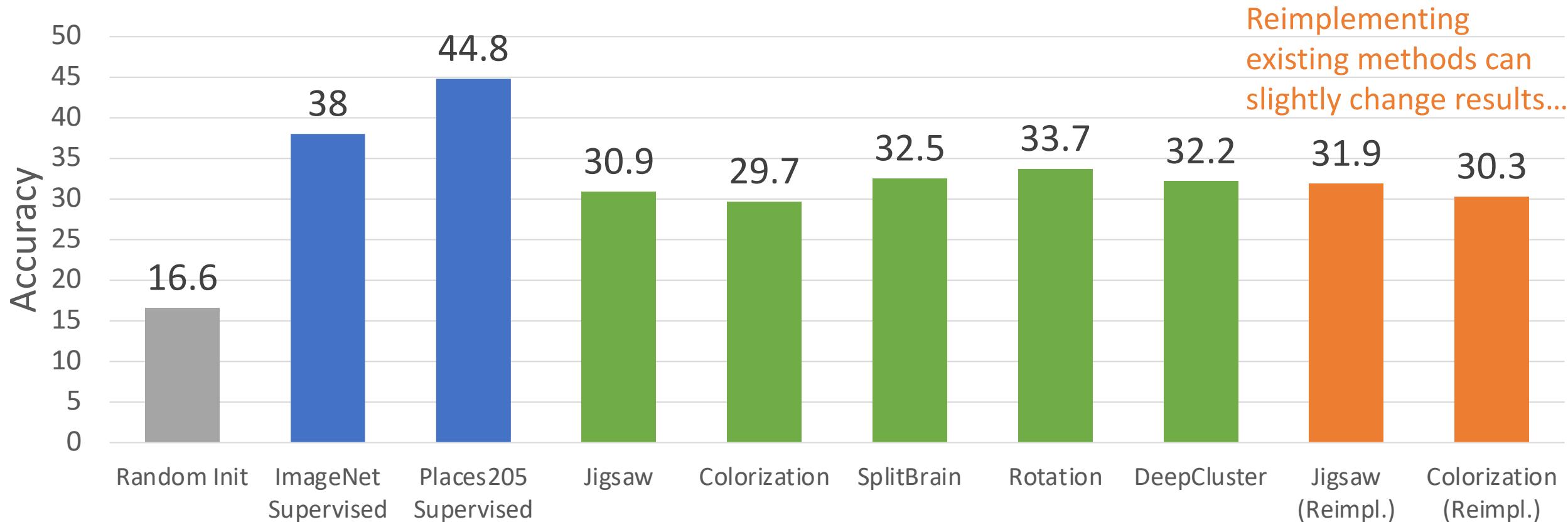


Goyal et al, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", ICCV 2019

Which SSL Method is best?

Some papers have tried to do fair comparisons of many SSL methods

Places205 Linear Classification from AlexNet conv5

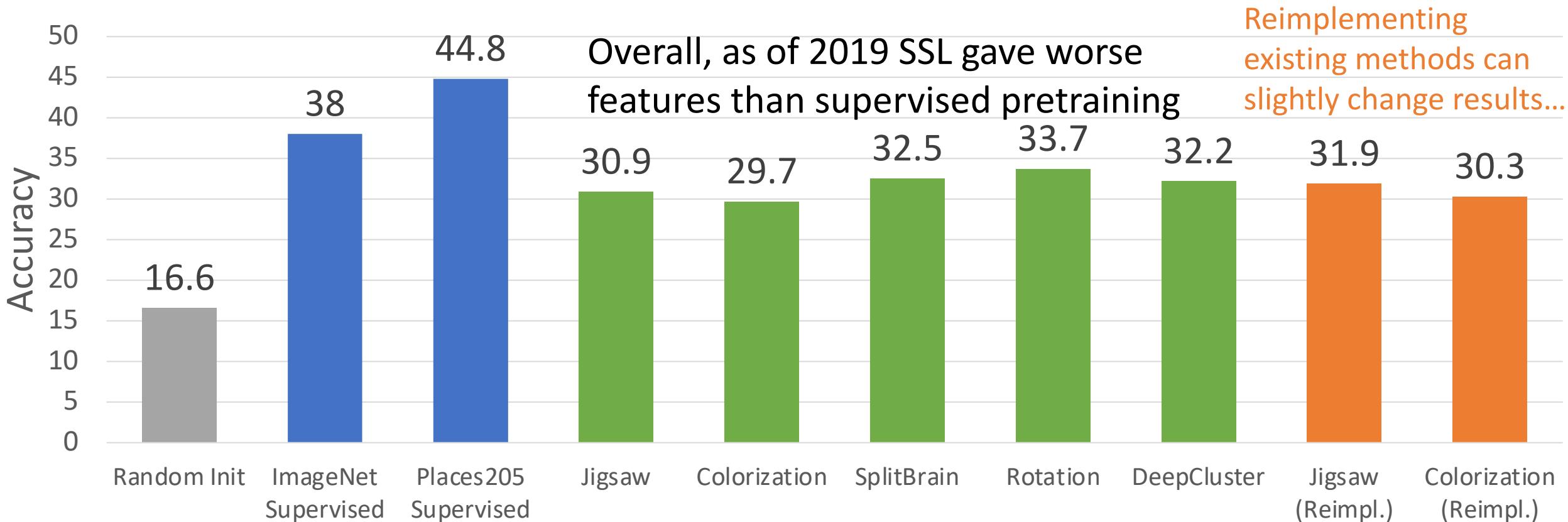


Goyal et al, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", ICCV 2019

Which SSL Method is best?

Some papers have tried to do fair comparisons of many SSL methods

Places205 Linear Classification from AlexNet conv5



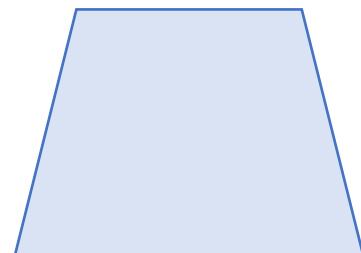
Goyal et al, "Scaling and Benchmarking Self-Supervised Visual Representation Learning", ICCV 2019

Self-Supervised Learning for Natural Language

Computer Vision

Image Features:

$H \times W \times C$

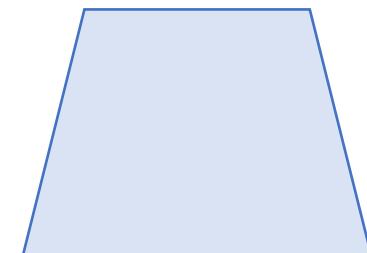


Input Image

Natural Language Processing

Word Features

$L \times C$



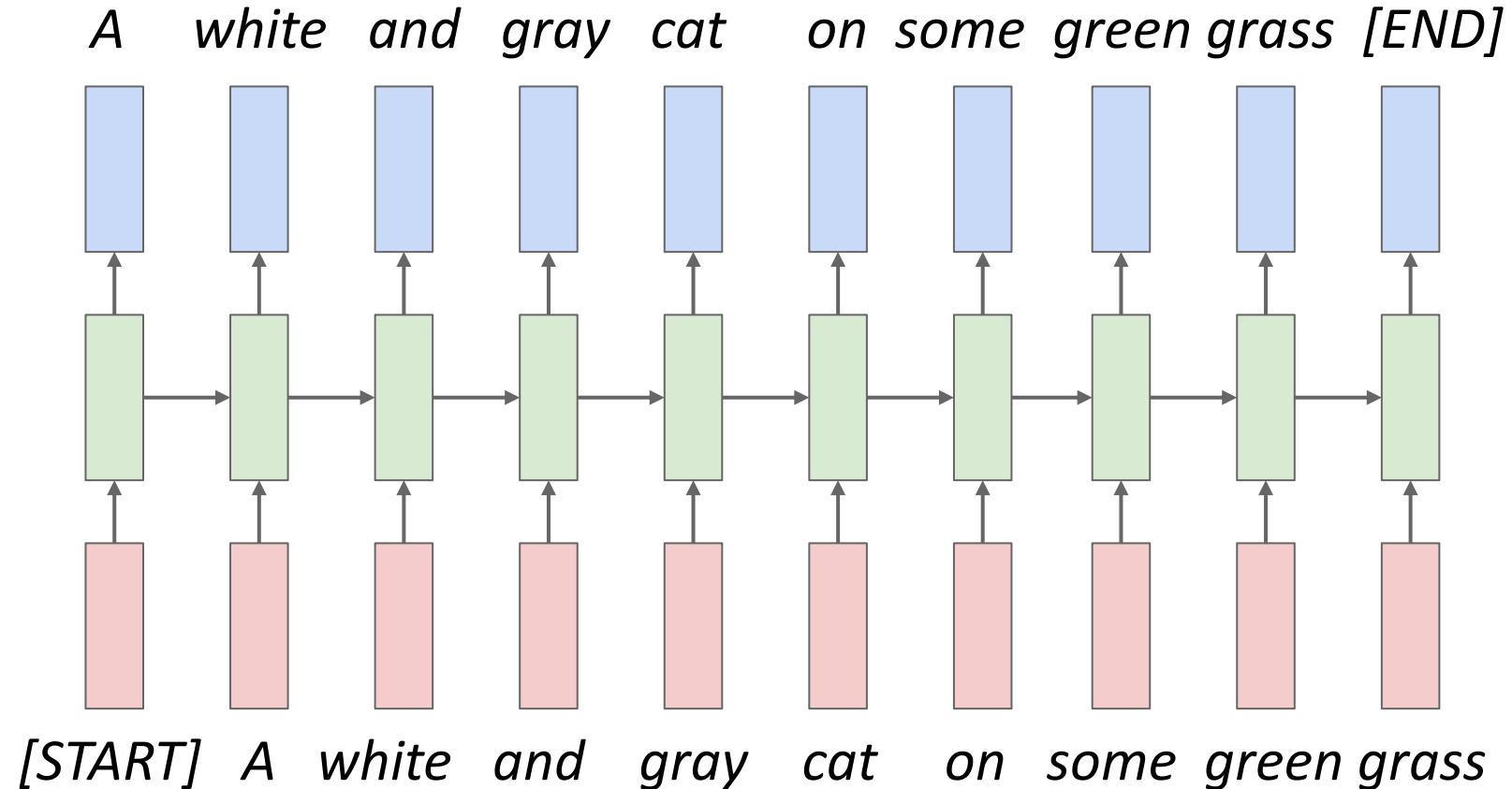
*A white and gray
cat standing outside
on the grass*

Input Sentence (L words)

Self-Supervised Learning for Natural Language

RNN language models train on raw text – no human labels required!

Their hidden states give features that transfer to many downstream tasks!



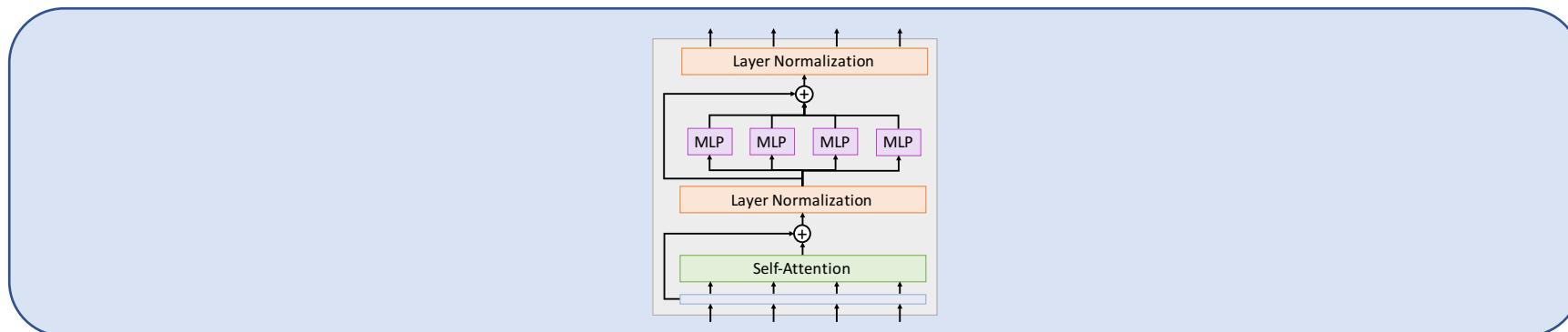
Peters et al, "Deep contextualized word representations", NAACL 2018

Self-Supervised Learning for Natural Language

Transformer-based language models work even better! Can scale up to very large datasets, and give extremely powerful features that transfer to downstream tasks

Wildly successful: larger models, larger datasets give better features that improve performance on many downstream NLP tasks. The dream of SSL made real!

A white and gray cat on some green grass [END]



[START] A white and gray cat on some green grass

Radford et al, "Language models are unsupervised multitask learners", 2019

Brown et al, "Language Models are Few-Shot Learners", arXiv 2020

Rae et al, "Scaling Language Models: Methods, Analysis, & Insights from Training Gopher", arXiv 2021

This Week: Pathways Language Model (PaLM)

Transformer with 118 layers, 48 heads, $d_{model}=18,432$, 540B parameters
Dataset: 780 billion tokens; trained on 6144 TPU-v4 chips

Bigger models
trained on more data
tend to give better
downstream task
performance

Model	Avg NLG	Avg NLU
GPT-3 175B	52.9	65.4
GLaM 64B/64E	58.4	68.7
PaLM 8B	41.5	59.2
PaLM 62B	57.7	67.3
PaLM 540B	63.9	74.7

NLG = Natural Language Generation (8 benchmarks)

NLU = Natural Language Understanding (21 benchmarks)

This Week: Pathways Language Model (PaLM)

Transformer with 118 layers, 48 heads, $d_{model}=18,432$, 540B parameters
Dataset: 780 billion tokens; trained on 6144 TPU-v4 chips

Bigger models
trained on more data
tend to give better
downstream task
performance

Model	Avg NLG	Avg NLU
GPT-3 175B	52.9	65.4
GLaM 64B/64E	58.4	68.7
PaLM 8B	41.5	59.2
PaLM 62B	57.7	67.3
PaLM 540B	63.9	74.7

NLG = Natural Language
Generation (8 benchmarks)

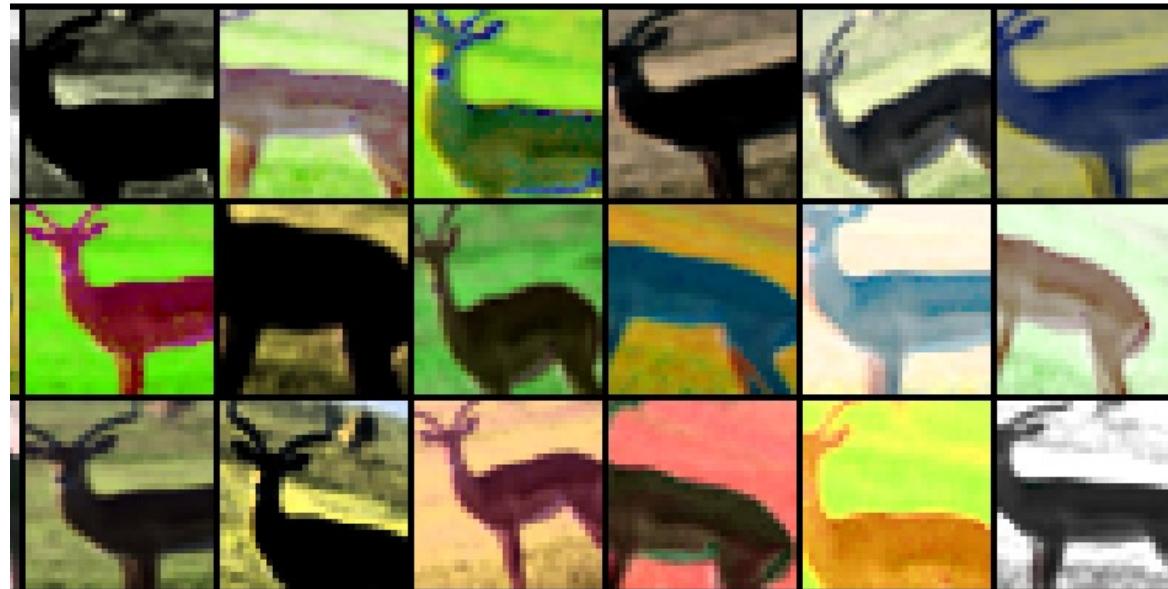
NLU = Natural Language
Understanding (21 benchmarks)

How can we achieve this success in vision?
Intensified interest in SSL since ~2018

Chowhery et al, "PaLM: Scaling Language Models with Pathways", 2022

Exemplar CNN: Invariance to Data Augmentation

Quiz: What is this?



Dosovitskiy et al, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks"

Exemplar CNN: Invariance to Data Augmentation

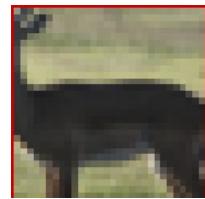
Quiz: What is this?



Answer: Deer!

Exemplar CNN: Invariance to Data Augmentation

Quiz: What is this?

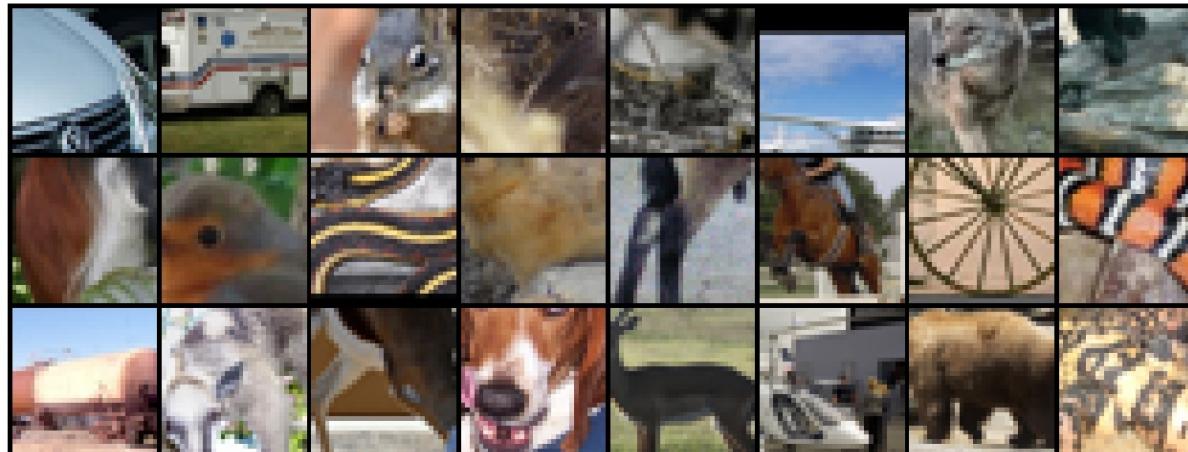


Different data augmentations (scale, shift, color jitter) of the same initial image patch

Answer: Deer!

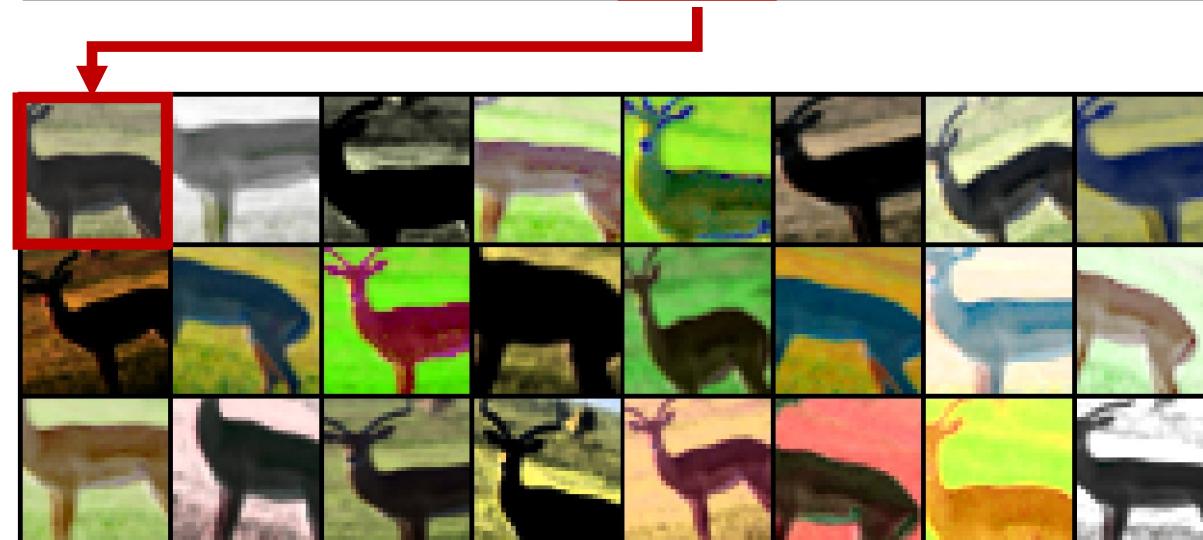
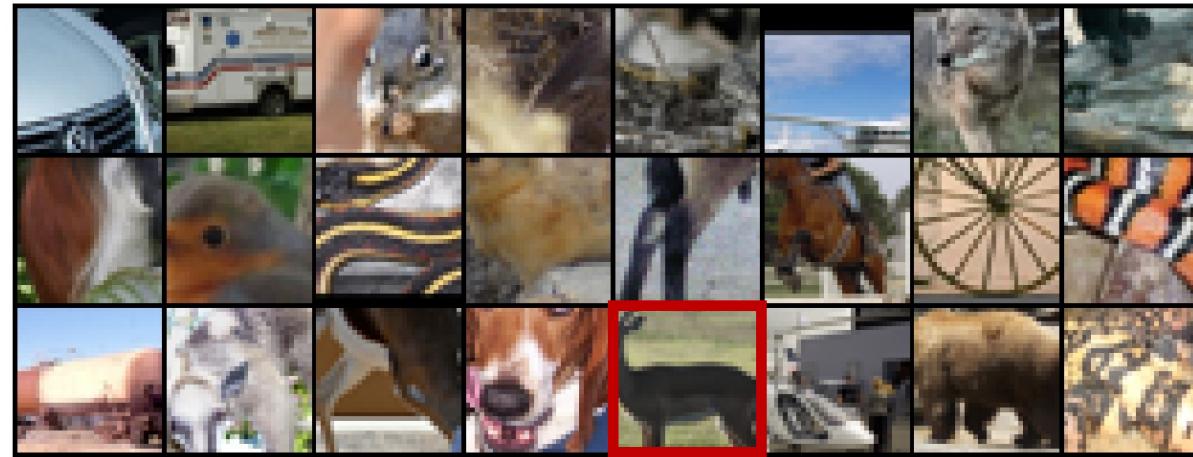
Exemplar CNN: Invariance to Data Augmentation

Given an initial dataset of N image patches



Exemplar CNN: Invariance to Data Augmentation

Given an initial dataset of N image patches

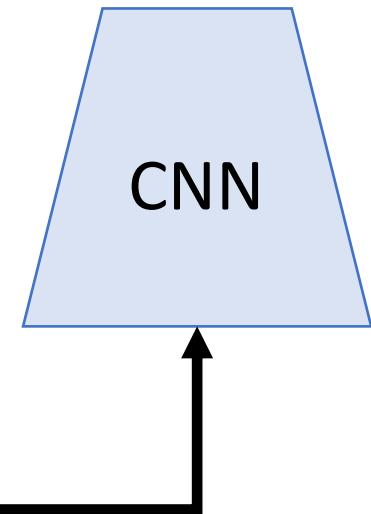
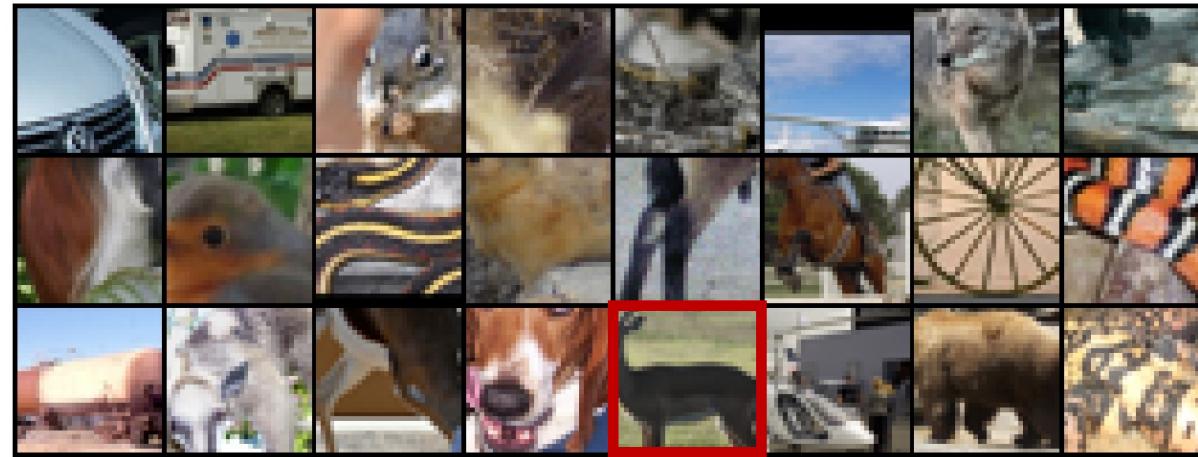


Sample K different augmentations for each; now have $K \cdot N$ total patches

Dosovitskiy et al, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks"

Exemplar CNN: Invariance to Data Augmentation

Given an initial dataset of N image patches

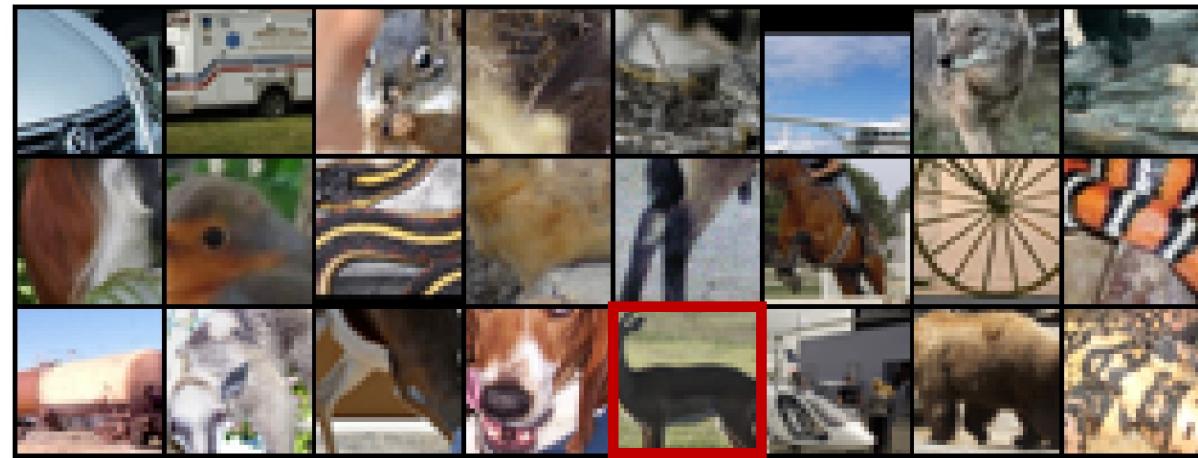


CNN inputs an augmented patch

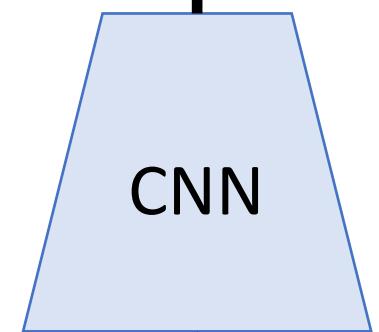
Dosovitskiy et al, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks"

Exemplar CNN: Invariance to Data Augmentation

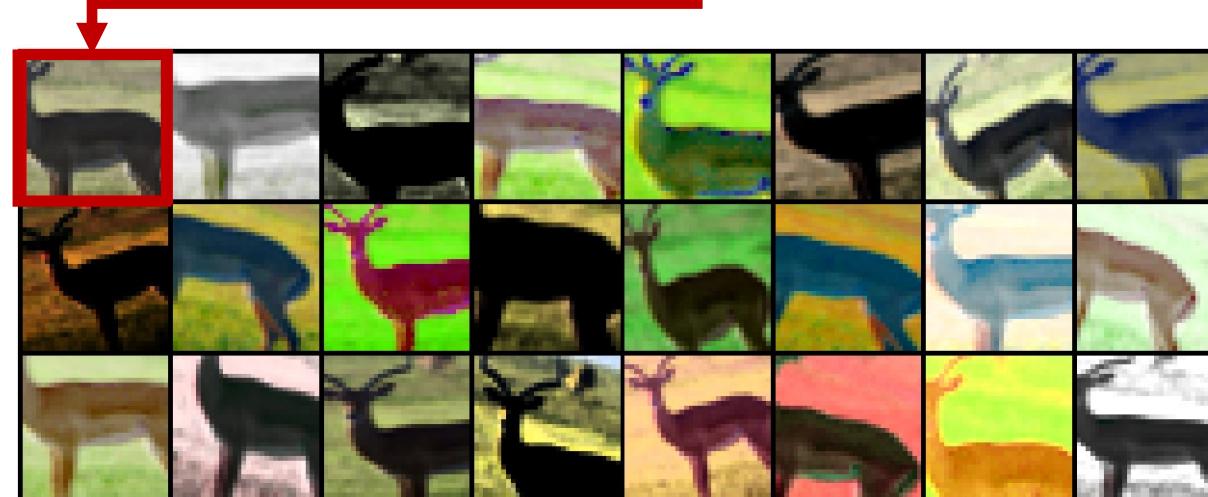
Given an initial dataset of N image patches



Predicts which of the N original images it came from (N -way classification)



Sample K different augmentations for each; now have $K \cdot N$ total patches



CNN inputs an augmented patch

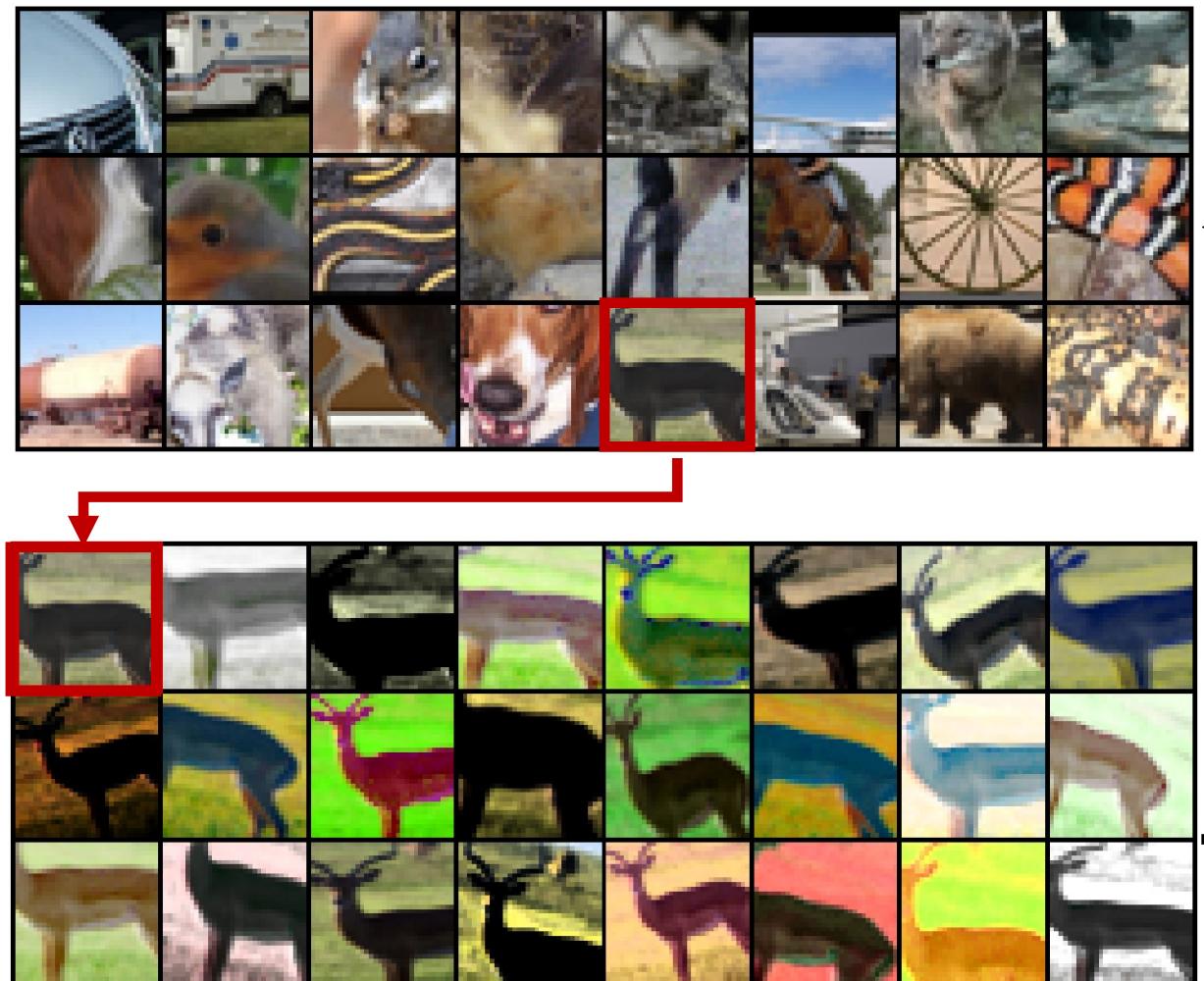
Dosovitskiy et al, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks"

Exemplar CNN: Invariance to Data Augmentation

Given an initial dataset of N image patches

Problem: number of parameters in final layer depends on N ; hard to scale

Sample K different augmentations for each; now have $K \cdot N$ total patches



Predicts which of the N original images it came from (N -way classification)

CNN

CNN inputs an augmented patch

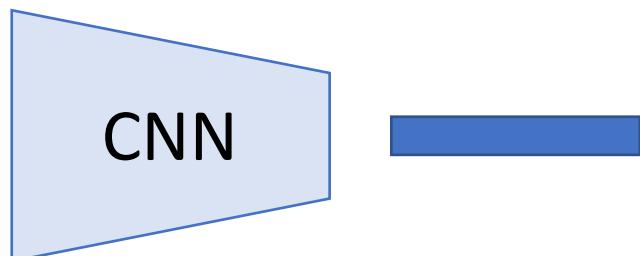
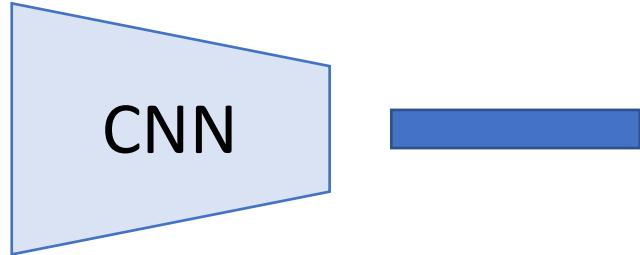
Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Similar images should have similar features



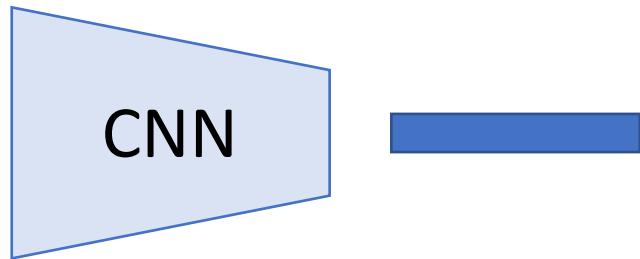
Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[White kitten image](#) is free for commercial use under the [Pixabay license](#)

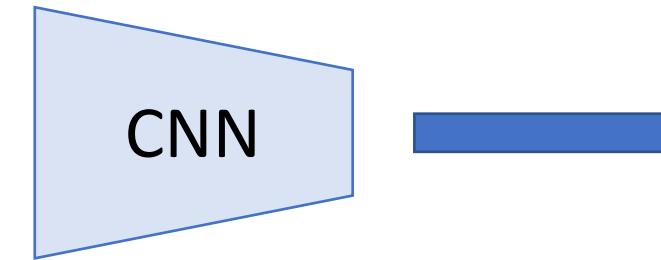
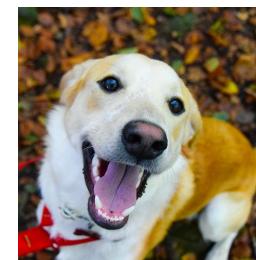
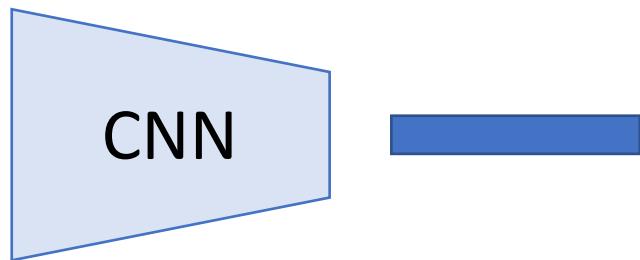
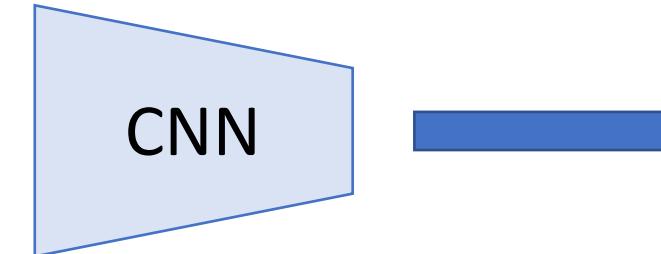
Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Similar images should have similar features



Dissimilar images should have dissimilar features



Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[White kitten image](#) is free for commercial use under the [Pixabay license](#)

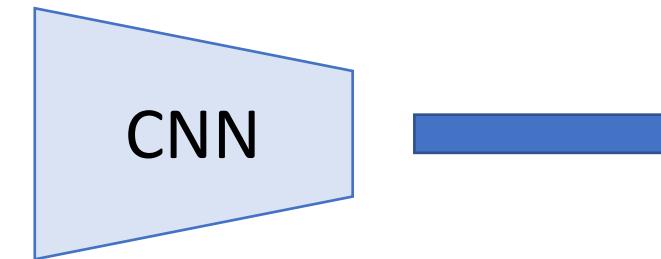
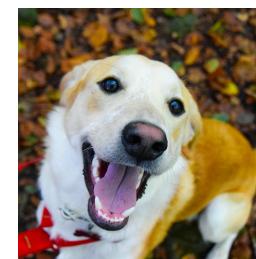
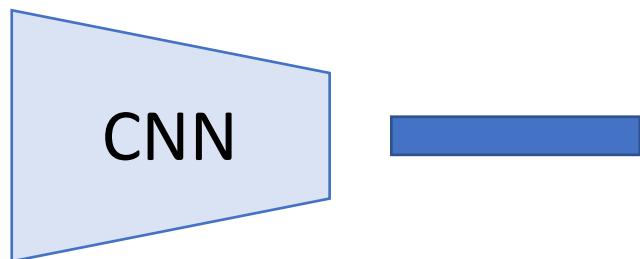
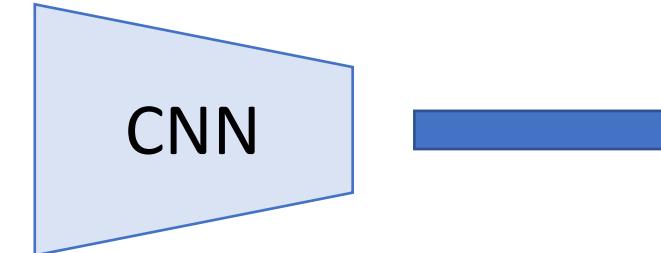
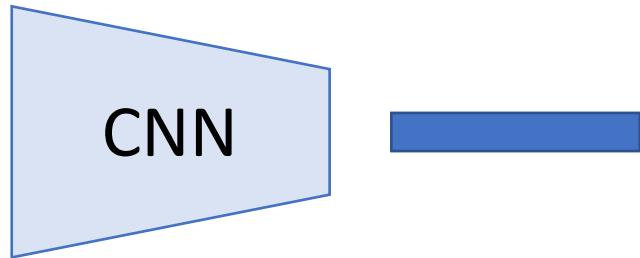
Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Let $d = \|\phi(x_1) - \phi(x_2)\|_2$ be the Euclidean distance between features for two images

Similar images should have similar features

Dissimilar images should have dissimilar features



Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[White kitten image](#) is free for commercial use under the [Pixabay license](#)

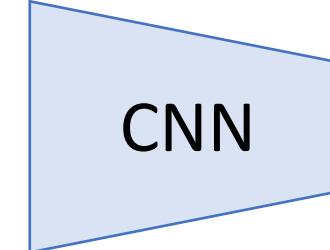
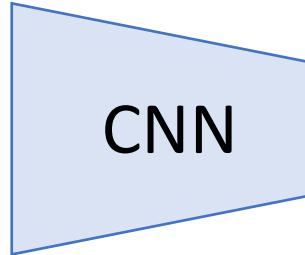
Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Let $d = \|\phi(x_1) - \phi(x_2)\|_2$ be the Euclidean distance between features for two images

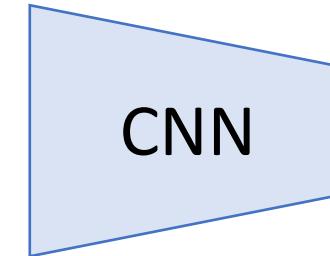
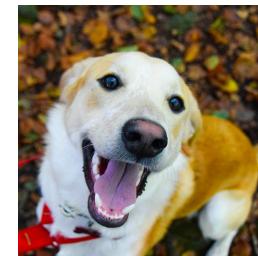
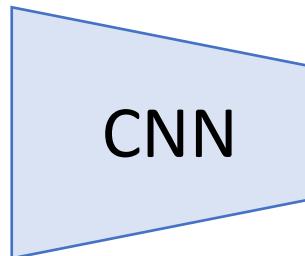
Similar images should have similar features

Dissimilar images should have dissimilar features



$$L_S(x_1, x_2) = d^2$$

Pull features together



Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[White kitten image](#) is free for commercial use under the [Pixabay license](#)

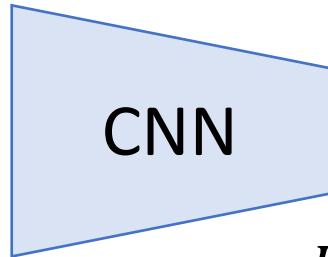
Contrastive Learning

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

Let $d = \|\phi(x_1) - \phi(x_2)\|_2$ be the Euclidean distance between features for two images

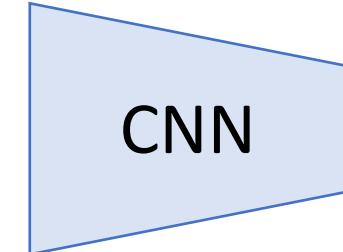
Similar images should have similar features

Dissimilar images should have dissimilar features



$$L_S(x_1, x_2) = d^2$$

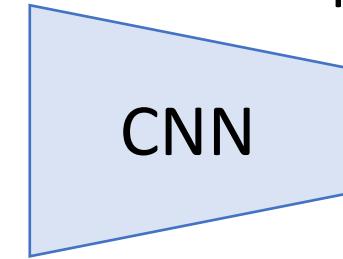
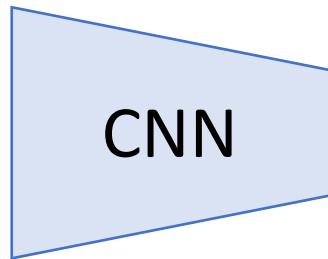
Pull features together



$$L_D(x_1, x_2)$$

$$= \max(0, m - d)^2$$

Push features apart
(up to margin m)



Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[White kitten image](#) is free for commercial use under the [Pixabay license](#)

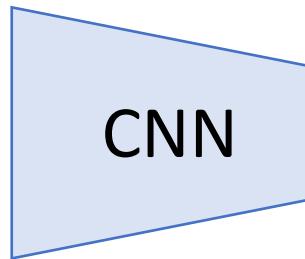
Contrastive Learning

Problem: Where to get positive and negative pairs?

Assume we don't have labels for images, but we know whether some pairs of images are **similar** or **dissimilar**

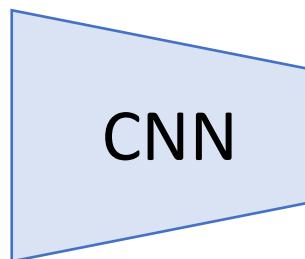
Let $d = \|\phi(x_1) - \phi(x_2)\|_2$ be the Euclidean distance between features for two images

Similar images should have similar features

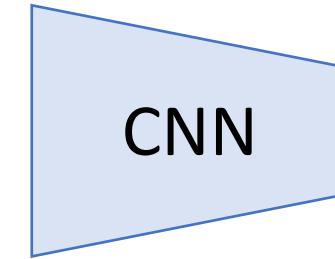


$$L_S(x_1, x_2) = d^2$$

Pull features together



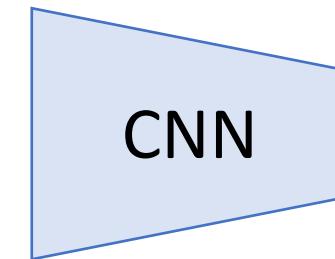
Dissimilar images should have dissimilar features



$$L_D(x_1, x_2)$$

$$= \max(0, m - d)^2$$

Push features apart
(up to margin m)



Contrastive Learning with Data Augmentation

Batch of
N images



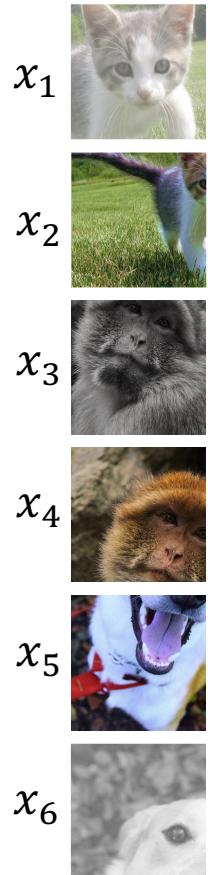
Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation

Batch of Two augmentations
N images for each image

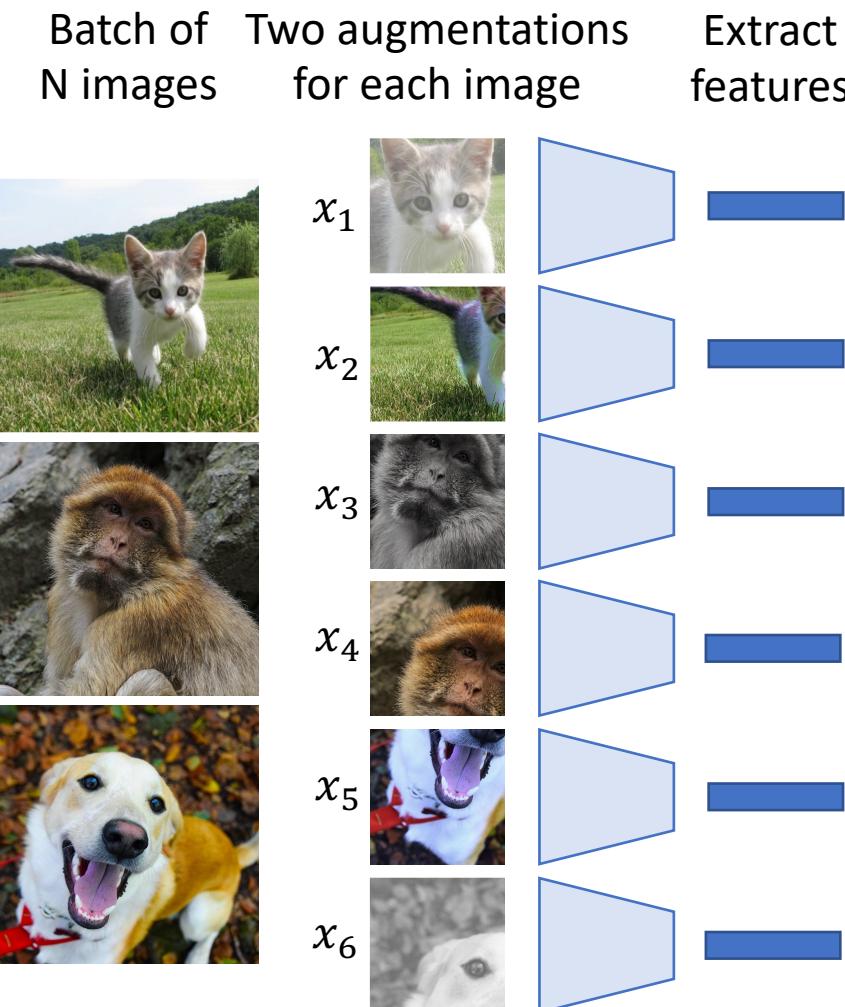


Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation



Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

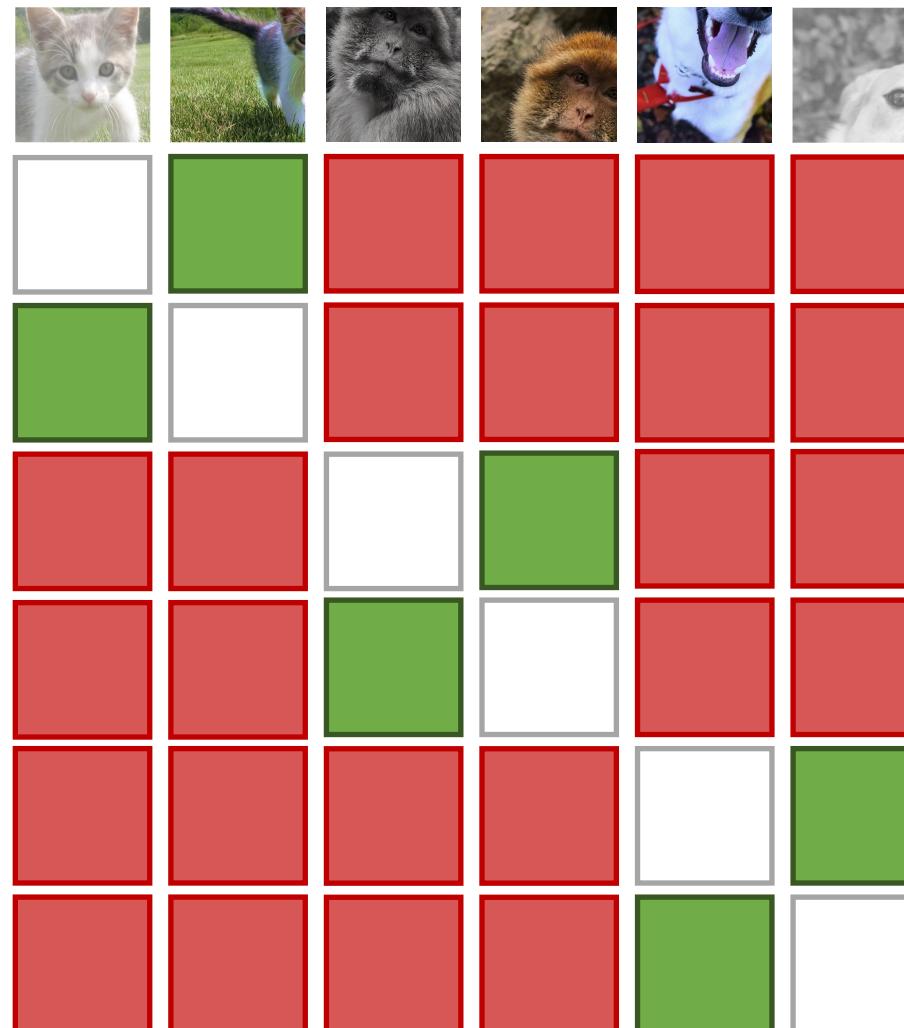
Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation

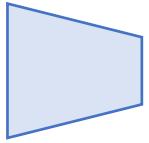
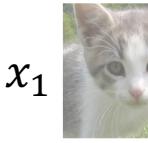
Batch of N images

Two augmentations for each image

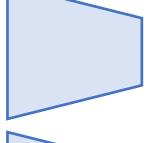
Extract features



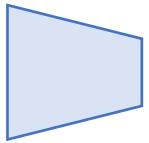
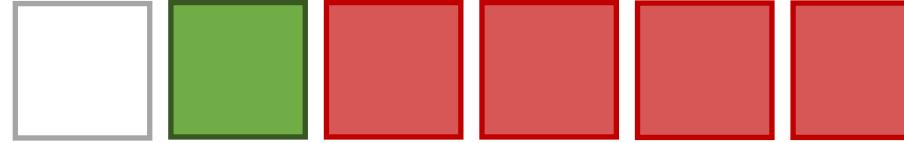
Each image tries to predict which of the *other* $2N-1$ images came from the same original image



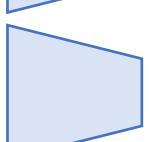
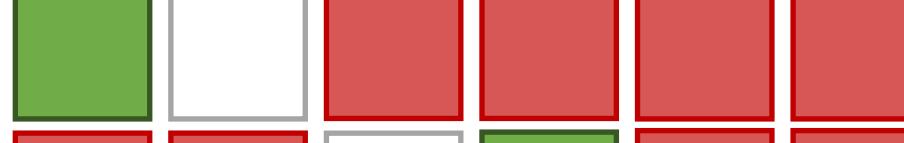
x_1



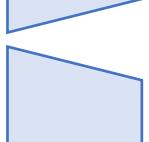
x_2



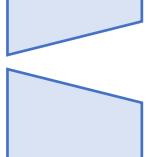
x_3



x_4



x_5



x_6

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018

Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019

Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019

Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020

He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020

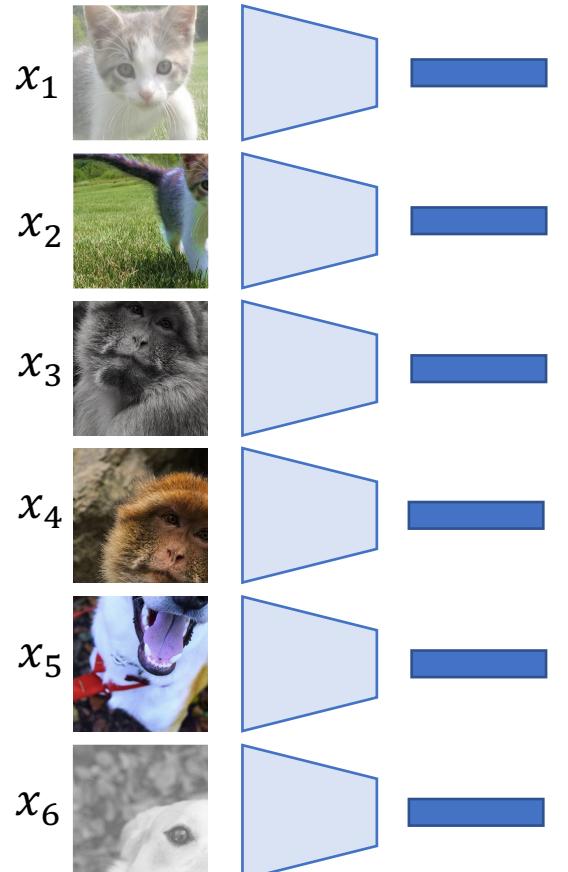
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation

Batch of N images



Two augmentations for each image



Extract features



Each image tries to predict which of the *other* $2N-1$ images came from the same original image

Similarity between x_i and x_j :

$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_j)\|}$$

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

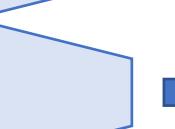
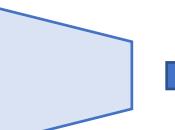
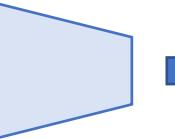
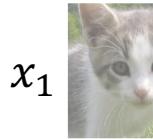
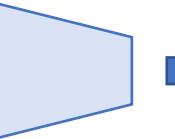
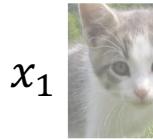
Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation

Batch of N images

Two augmentations for each image

Extract features



Each image tries to predict which of the *other* $2N-1$ images came from the same original image

Similarity between x_i and x_j :

$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_j)\|}$$

If (x_i, x_j) is a positive pair, then loss for x_i is:

$$L_i = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(s_{i,k}/\tau)}$$

(τ is a *temperature*)

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018

Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019

Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019

Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020

He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020

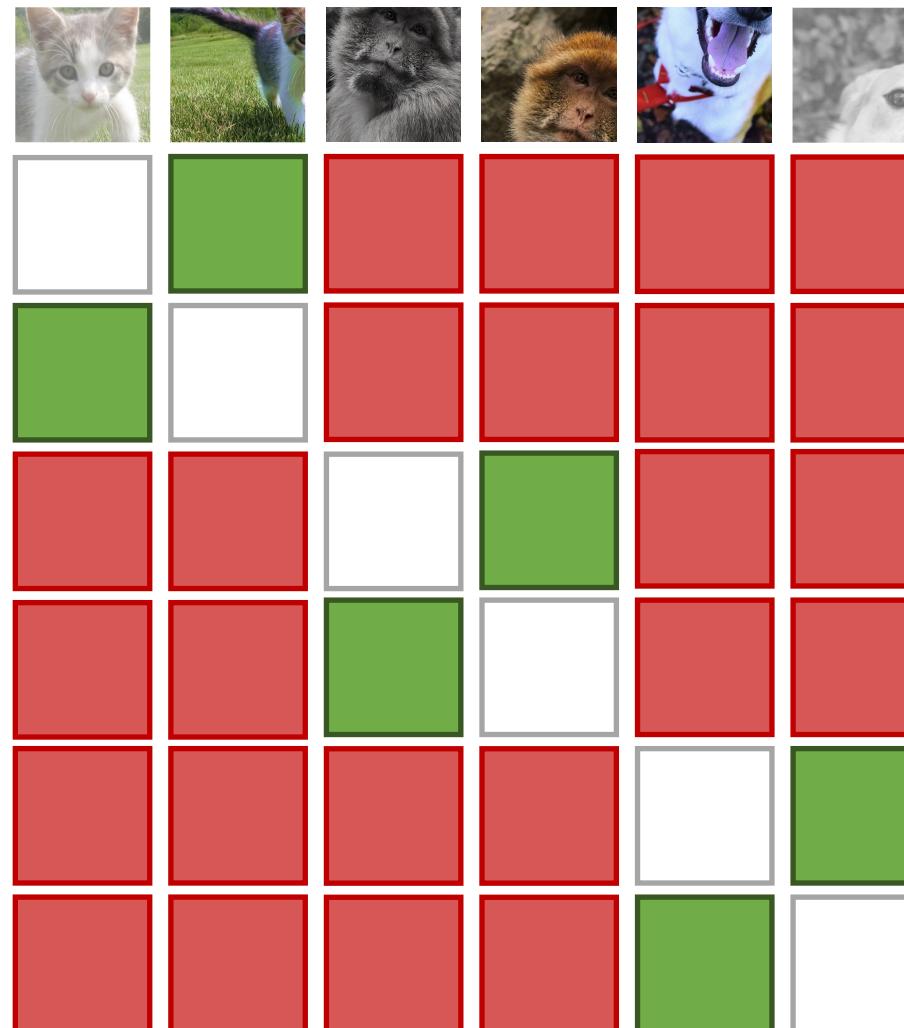
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning with Data Augmentation

Batch of N images

Two augmentations for each image

Extract features



Each image tries to predict which of the *other* $2N-1$ images came from the same original image

Similarity between x_i and x_j :

$$s_{i,j} = \frac{\phi(x_i)^T \phi(x_j)}{\|\phi(x_i)\| \cdot \|\phi(x_j)\|}$$

If (x_i, x_j) is a positive pair, then loss for x_i is:

$$L_i = -\log \frac{\exp(s_{i,i}/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(s_{i,k}/\tau)}$$

(τ is a *temperature*)

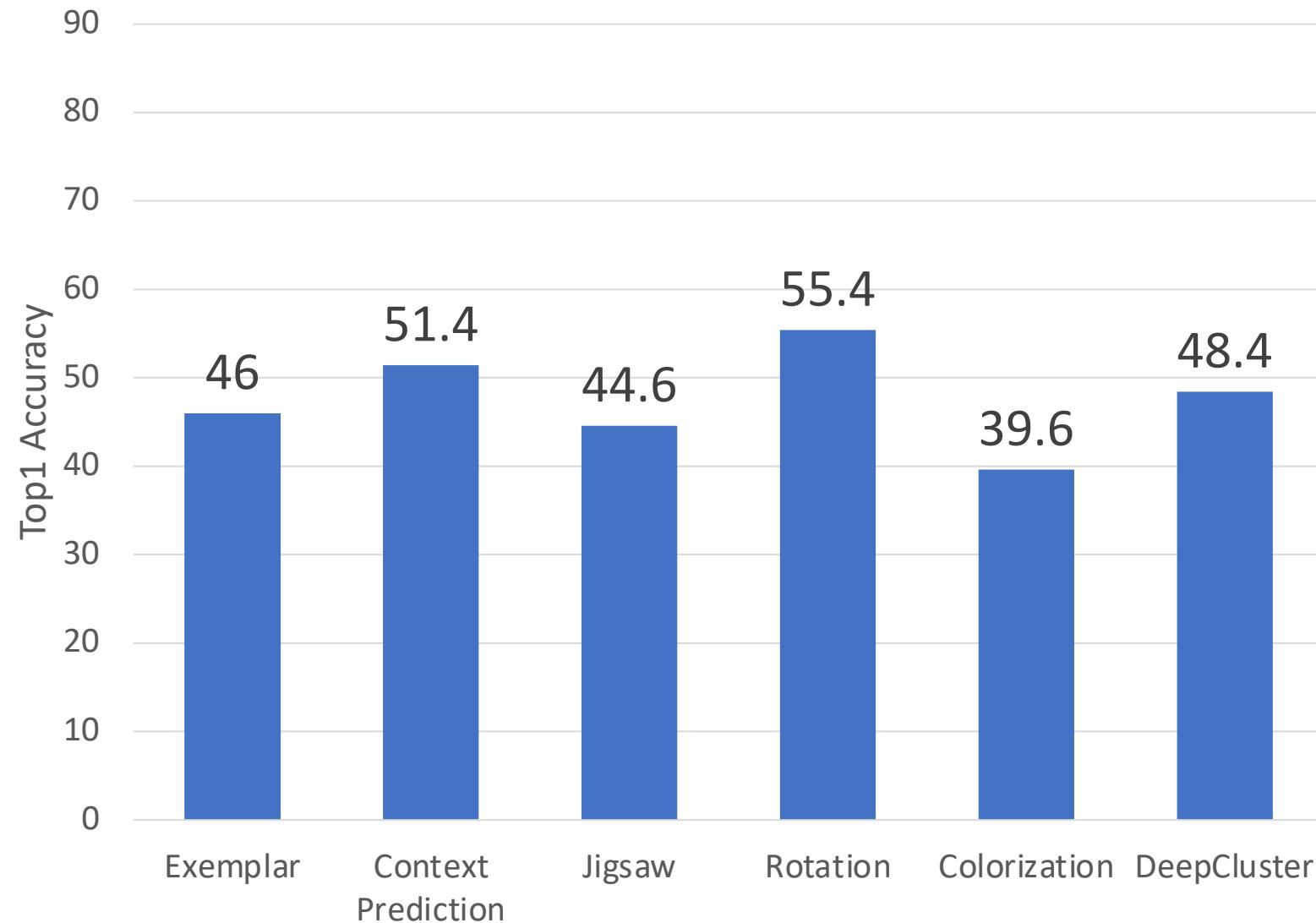
Interpretation: Cross-entropy loss over the other $2N-1$ elements in the batch!

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
 Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
 Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
 Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
 Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
 He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
 Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

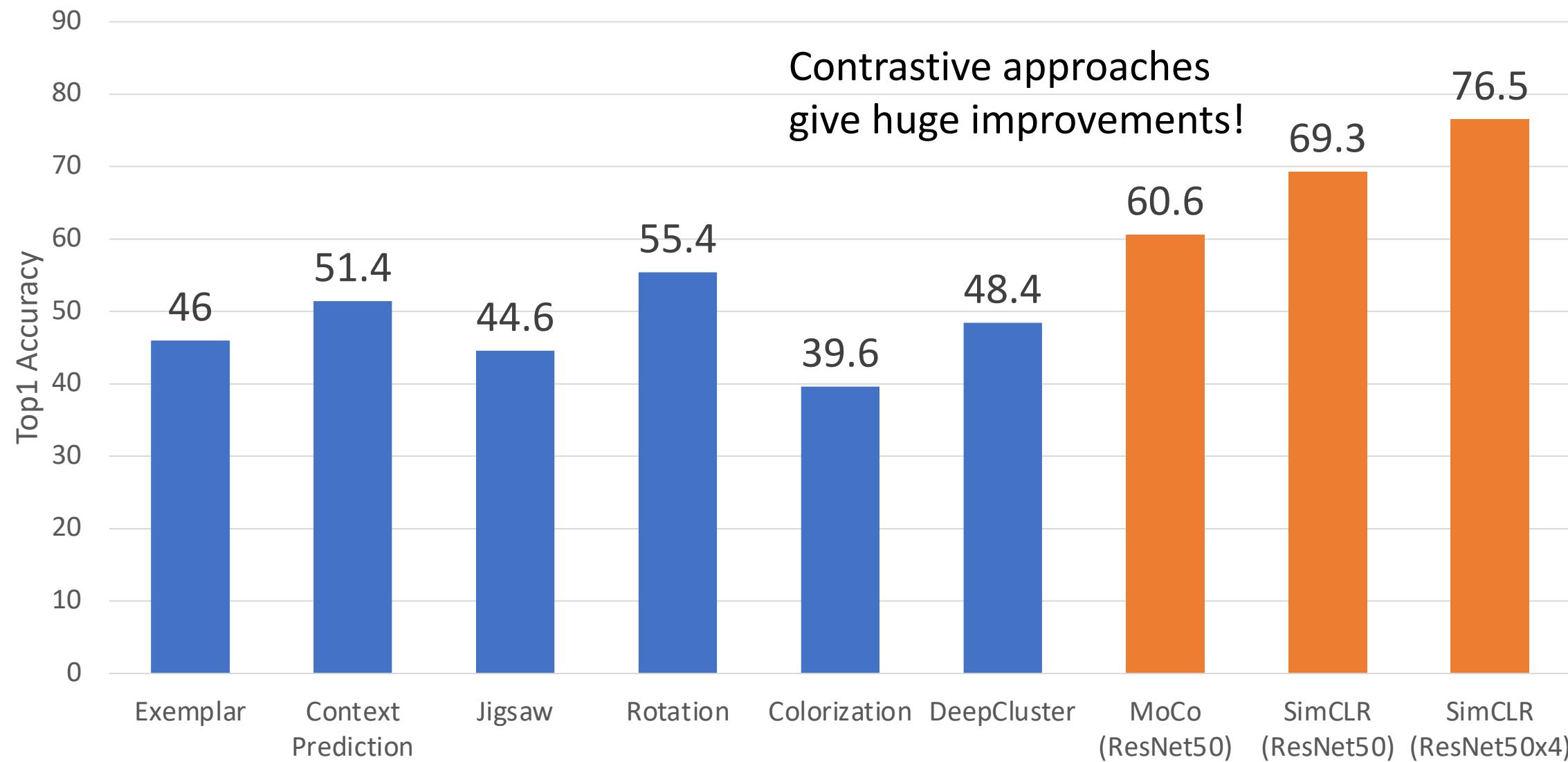
ImageNet Linear Classification from SSL Features



He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020
Chen et al, "An Empirical Study of Training Self-Supervised Vision Transformers", ICCV 2021

(Lots of caveats here ... different architectures, etc)

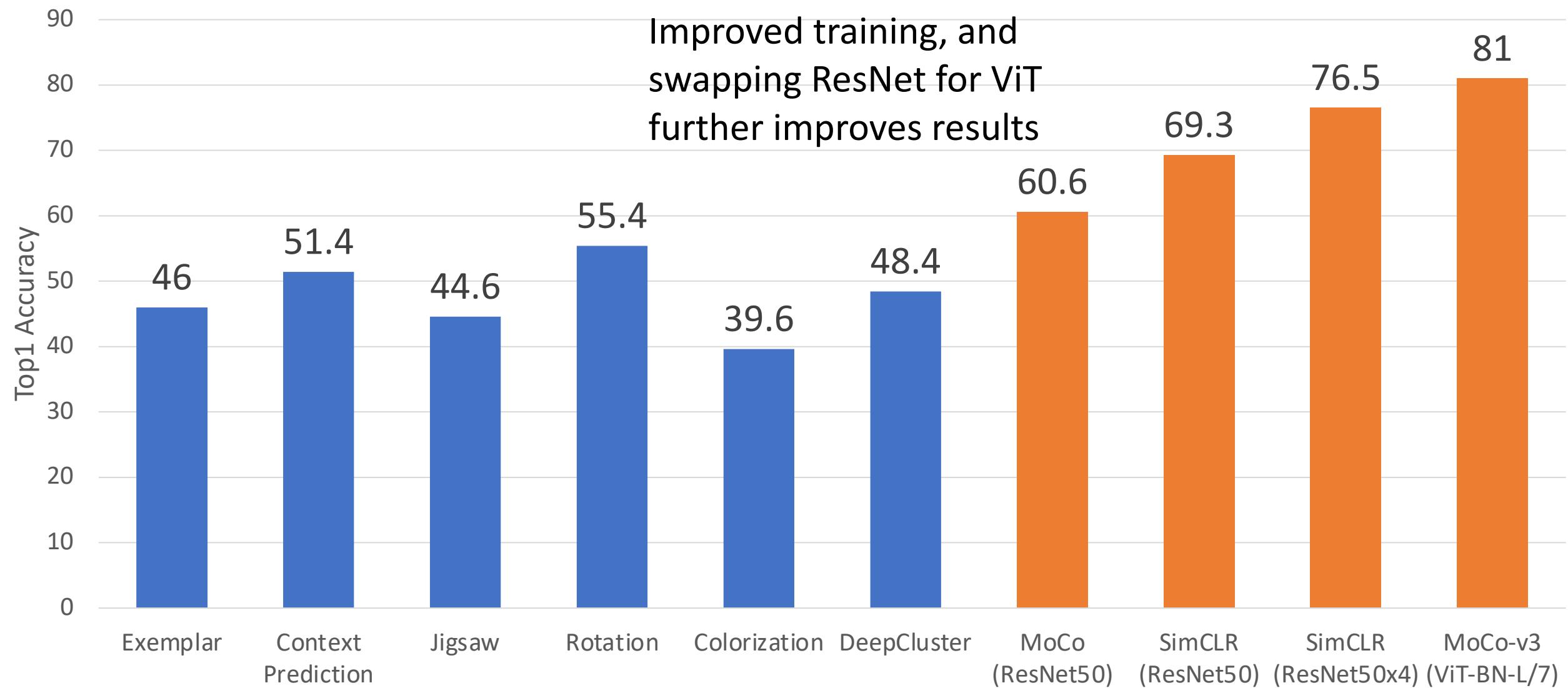
ImageNet Linear Classification from SSL Features



He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020
Chen et al, "An Empirical Study of Training Self-Supervised Vision Transformers", ICCV 2021

(Lots of caveats here ... different architectures, etc)

ImageNet Linear Classification from SSL Features

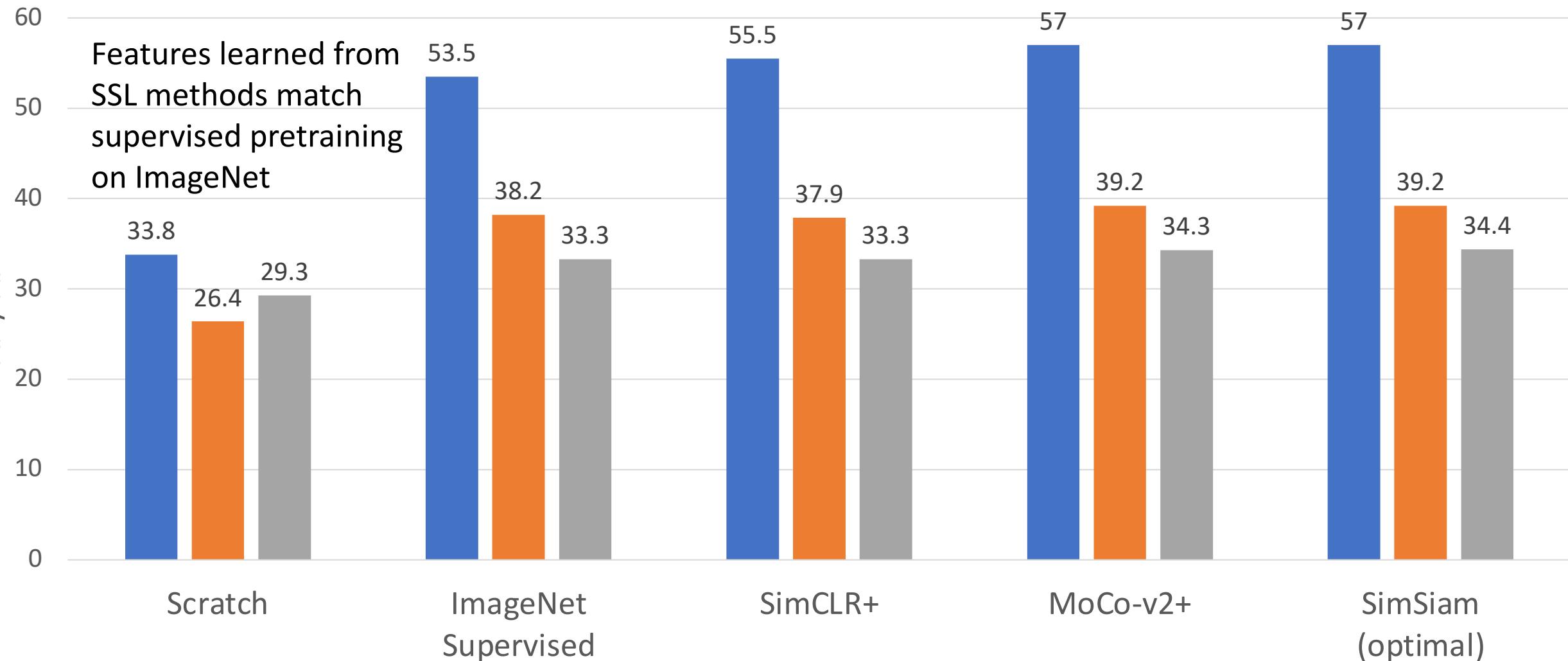


He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020
Chen et al, "An Empirical Study of Training Self-Supervised Vision Transformers", ICCV 2021

(Lots of caveats here ... different architectures, etc)

Contrastive SSL Pretraining then Finetuning on Detection

■ VOC 07+12 Detection ■ COCO Detection ■ COCO Instance Segmentation



He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Chen et al, "Improved Baselines with Momentum Contrastive Learning", arXiv 2020
Chen and He, "Exploring simple Siamese representation learning", CVPR 2021

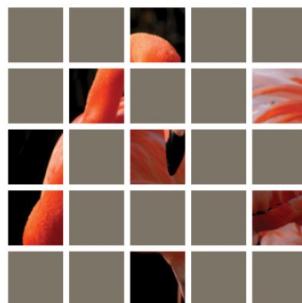
Masked Autoencoders (MAE)

A new old method dethrones contrastive learning? Denoising Autoencoder with Vision Transformer

Masked Autoencoders (MAE)

A new old method dethrones contrastive learning? Denoising Autoencoder with Vision Transformer

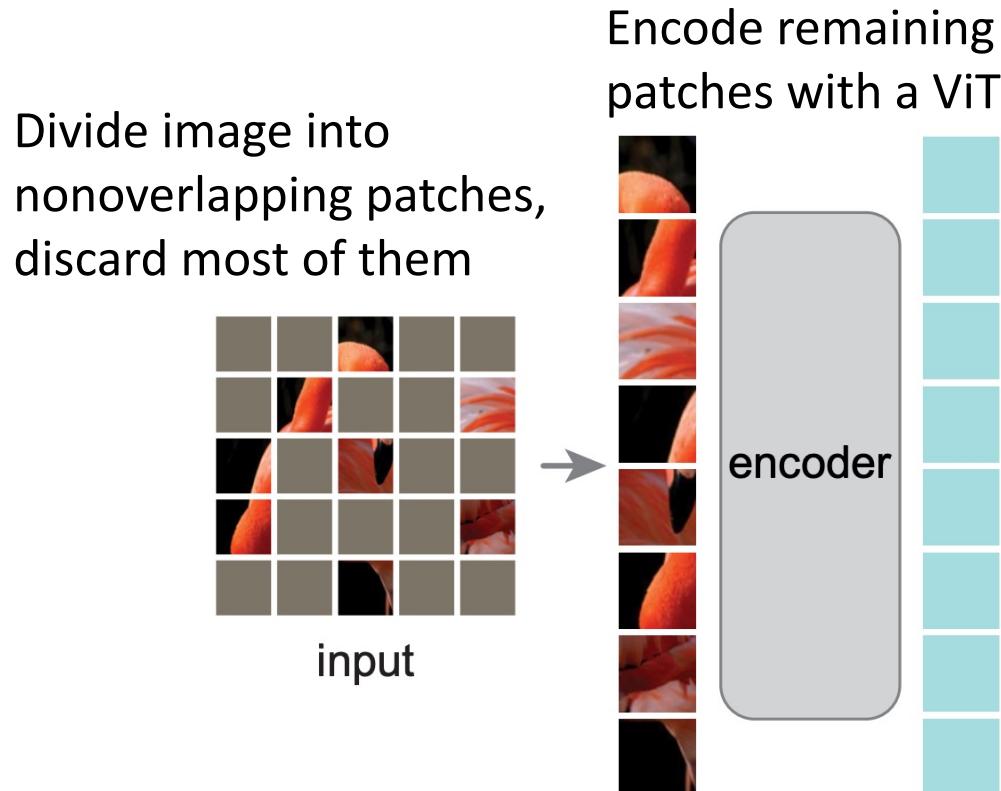
Divide image into
nonoverlapping patches,
discard most of them



input

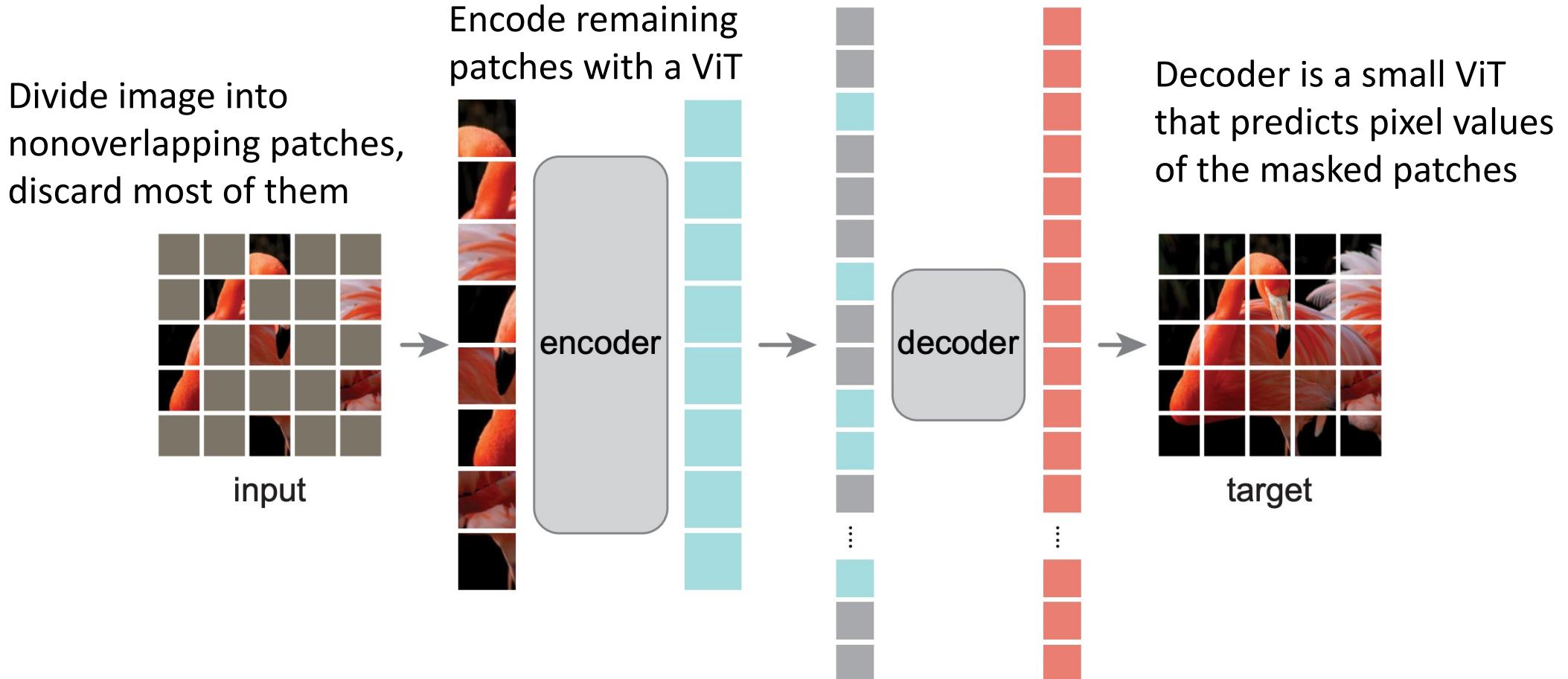
Masked Autoencoders (MAE)

A new old method dethrones contrastive learning? Denoising Autoencoder with Vision Transformer



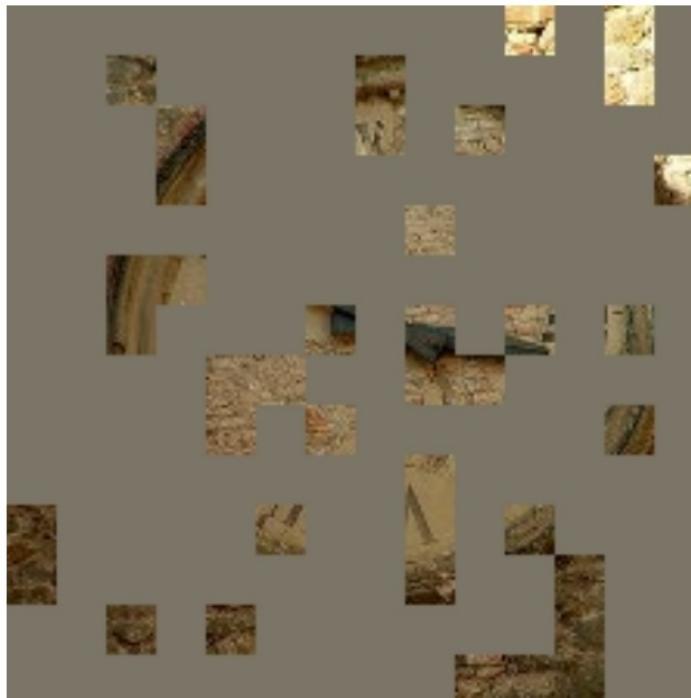
Masked Autoencoders (MAE)

A new old method dethrones contrastive learning? Denoising Autoencoder with Vision Transformer



Masked Autoencoders (MAE): Reconstructions

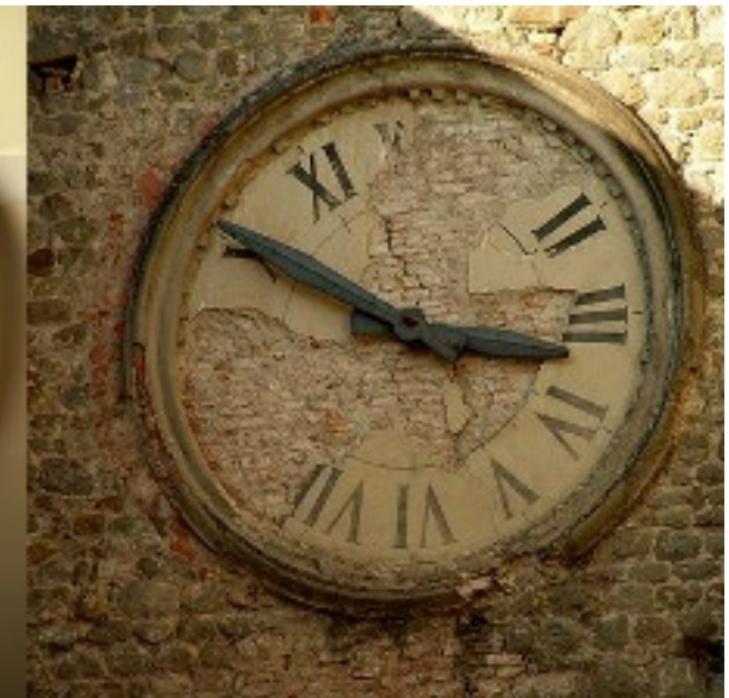
Input Patches



Prediction



Actual Image



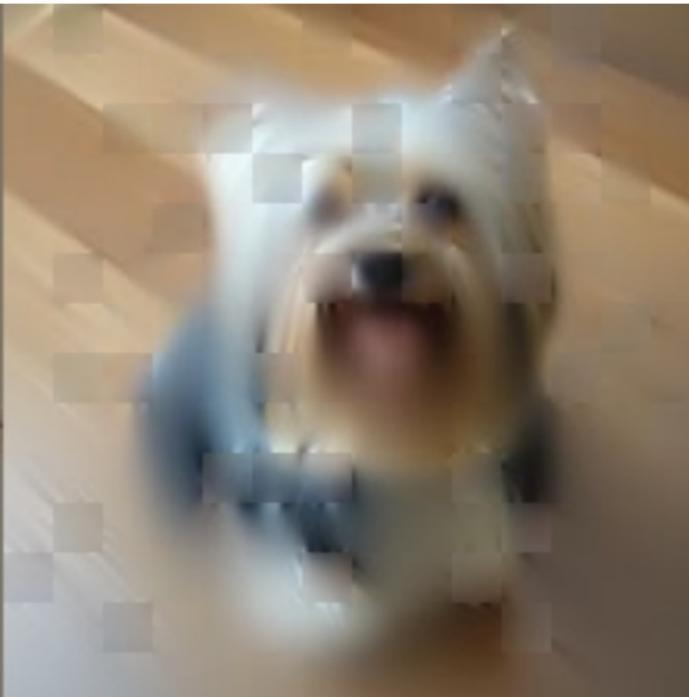
He et al, "Masked Autoencoders are Scalable Vision Learners", CVPR 2022

Masked Autoencoders (MAE): Reconstructions

Input Patches



Prediction



Actual Image

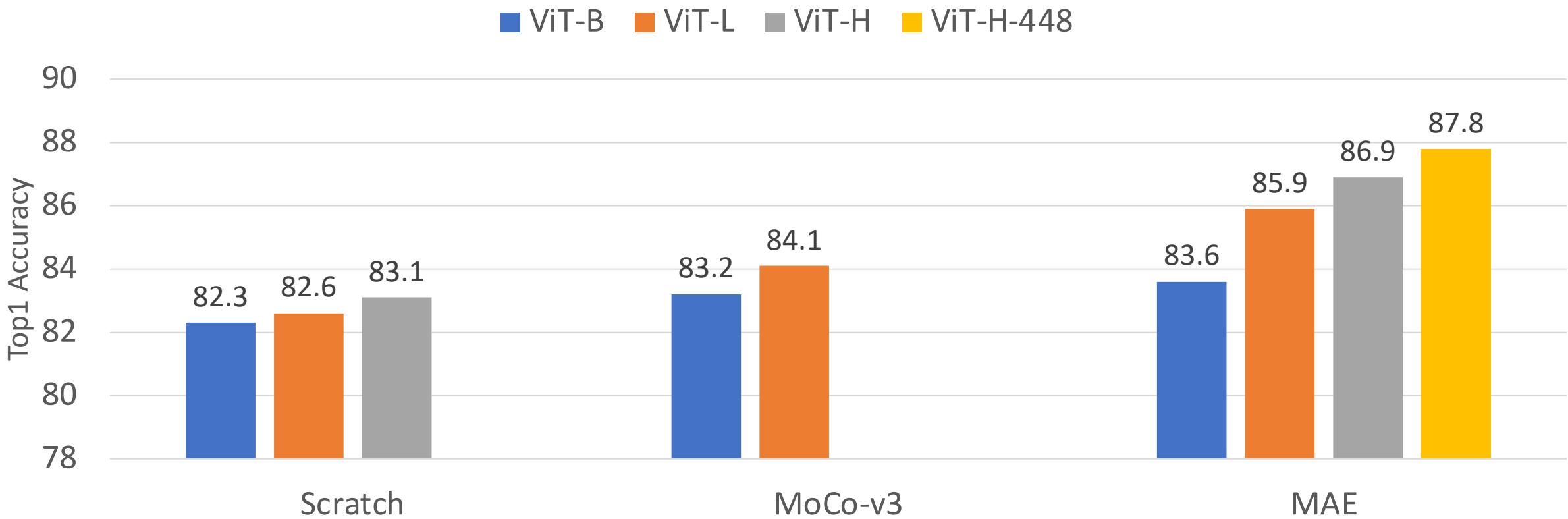


Masked Autoencoders (MAE): Reconstructions



He et al, "Masked Autoencoders are Scalable Vision Learners", CVPR 2022

SSL Pretraining, then finetuning for ImageNet Classification



MAE Pretraining outperforms training from scratch, and allows scaling to larger ViT models

A rant about data...

The motivation of SSL is scaling to large data
that can't be labeled

A rant about data...

The motivation of SSL is scaling to large data
that can't be labeled

Most papers pretrain on (unlabeled)
ImageNet, then evaluate on ImageNet!

Unlabeled ImageNet is still curated: single
object per image, balanced classes

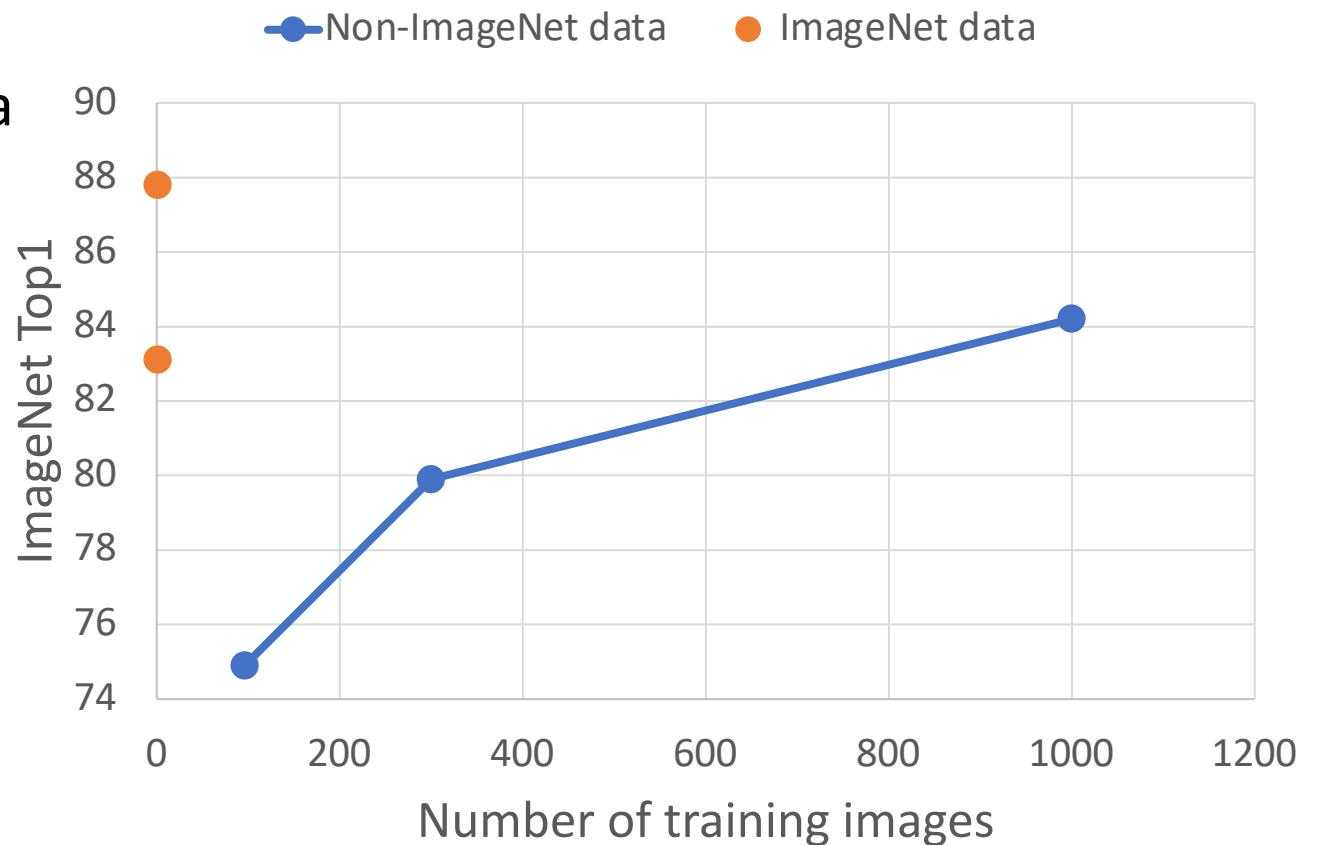
A rant about data...

The motivation of SSL is scaling to large data that can't be labeled

Most papers pretrain on (unlabeled) ImageNet, then evaluate on ImageNet!

Unlabeled ImageNet is still curated: single object per image, balanced classes

Self-Supervised Learning on larger datasets hasn't been as successful as NLP



- Caron et al, "Unsupervised pre-training of images features on non-curated data", ICCV 2019
Chen et al, "Big self-supervised models are strong semi-supervised learners", NeurIPS 2020
Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021
Goyal et al, "Self-supervised Pretraining of Visual Features in the Wild", arXiv 2021
He et al, "Masked Autoencoders are Scalable Vision Learners", arXiv 2021

A rant about data...

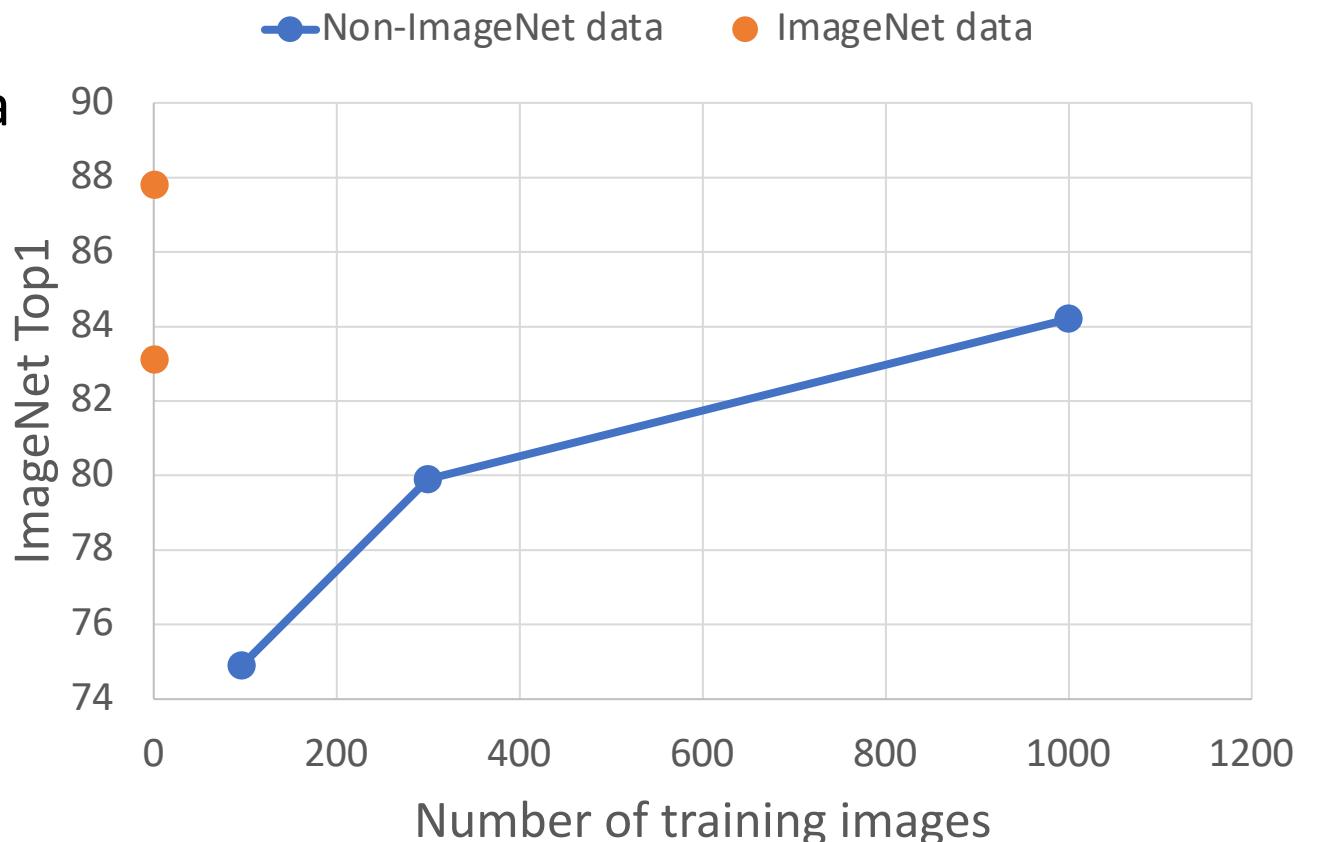
The motivation of SSL is scaling to large data that can't be labeled

Most papers pretrain on (unlabeled) ImageNet, then evaluate on ImageNet!

Unlabeled ImageNet is still curated: single object per image, balanced classes

Self-Supervised Learning on larger datasets hasn't been as successful as NLP

Idea: What if we go beyond isolated images?



Caron et al, "Unsupervised pre-training of images features on non-curated data", ICCV 2019
Chen et al, "Big self-supervised models are strong semi-supervised learners", NeurIPS 2020
Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021
Goyal et al, "Self-supervised Pretraining of Visual Features in the Wild", arXiv 2021
He et al, "Masked Autoencoders are Scalable Vision Learners", arXiv 2021

Multimodal Self-Supervised Learning

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015

Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015

Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Multimodal Self-Supervised Learning

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015

Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015

Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Sound: Image with audio track from video

Owens et al, "Ambient Sound Provides Supervision for Visual Learning", ECCV 2016

Arandjelovic and Zisserman, "Look, Listen and Learn", ICCV 2017

Multimodal Self-Supervised Learning

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015

Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015

Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Sound: Image with audio track from video

Owens et al, "Ambient Sound Provides Supervision for Visual Learning", ECCV 2016

Arandjelovic and Zisserman, "Look, Listen and Learn", ICCV 2017

3D: Image with depth map or point cloud

Xie et al, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding", ECCV 2020

Zhang et al, "Self-supervised pretraining of 3D features on any point-cloud", CVPR 2021

Multimodal Self-Supervised Learning

Don't learn from isolated images -- take images together with some **context**

Video: Image together with adjacent video frames

Agrawal et al, "Learning to See by Moving", ICCV 2015

Wang et al, "Unsupervised Learning of Visual Representations using Videos", ICCV 2015

Pathak et al, "Learning Features by Watching Objects Move", CVPR 2017

Sound: Image with audio track from video

Owens et al, "Ambient Sound Provides Supervision for Visual Learning", ECCV 2016

Arandjelovic and Zisserman, "Look, Listen and Learn", ICCV 2017

3D: Image with depth map or point cloud

Xie et al, "PointContrast: Unsupervised Pre-training for 3D Point Cloud Understanding", ECCV 2020

Zhang et al, "Self-supervised pretraining of 3D features on any point-cloud", CVPR 2021

Language: Image with natural-language text

Sariyildiz et al, "Learning Visual Representations with Caption Annotations", ECCV 2020

Desai and Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR 2021

Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021

Jia et al, "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision", ICLR 2021

Desai et al, "RedCaps: Web-curated Image-Text data created by the people, for the people", NeurIPS 2021

Why Language?

Large dataset of
(image, caption)



a dog with his
head out the
window of the car



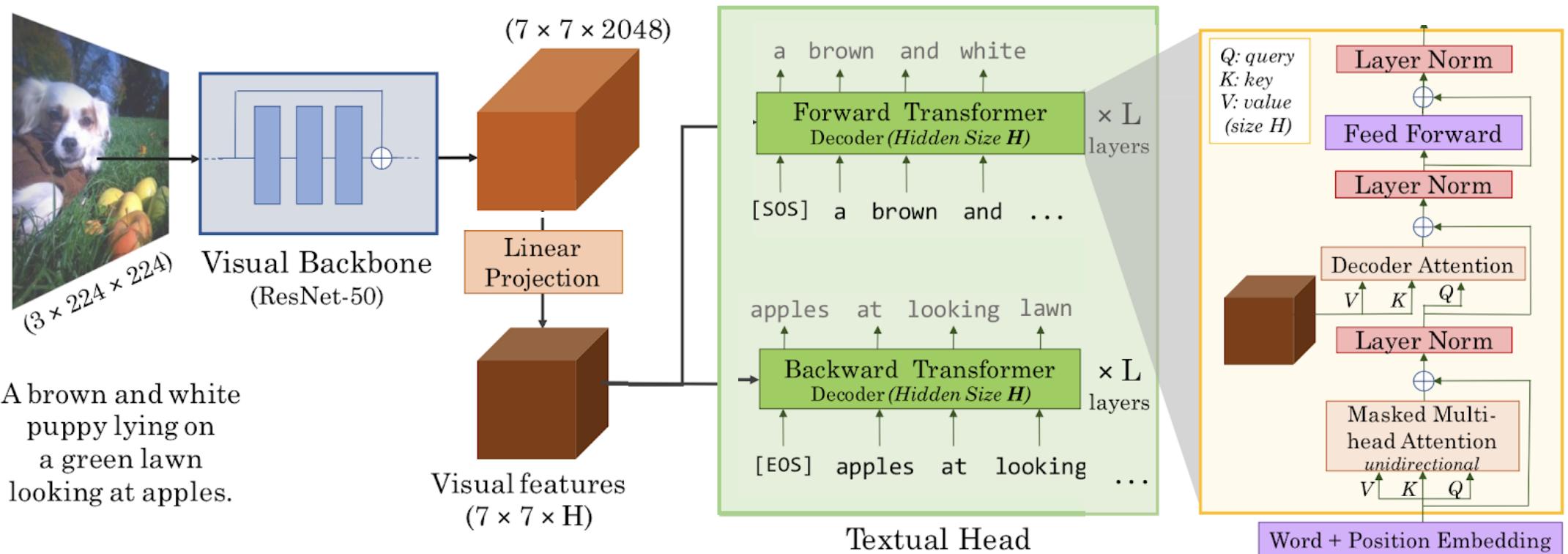
a black and orange
cat is resting on a
keyboard and yellow
back scratcher

1. Semantic density: Just a few words give rich information

2. Universality: Language
can describe any concept

3. Scalability: Non-experts
can easily caption images;
data can also be collected
from the web at scale

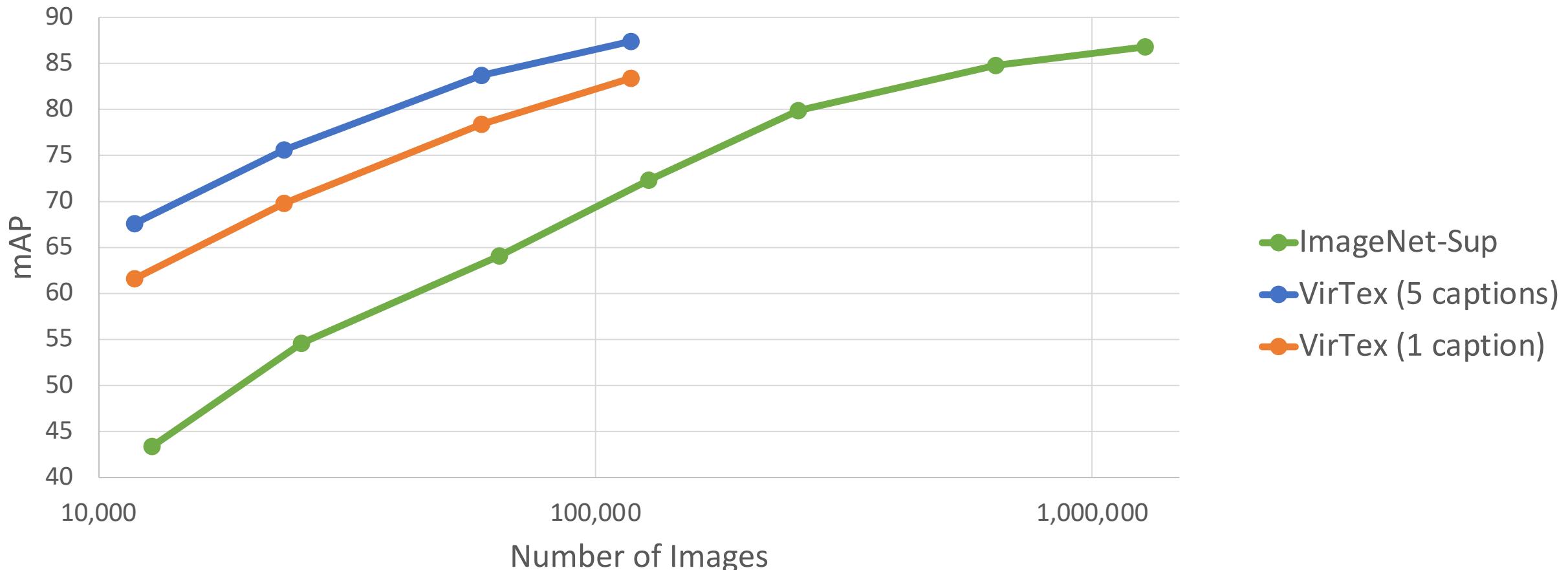
Generating Captions



Desai and Johnson, "Desai and Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR 2021

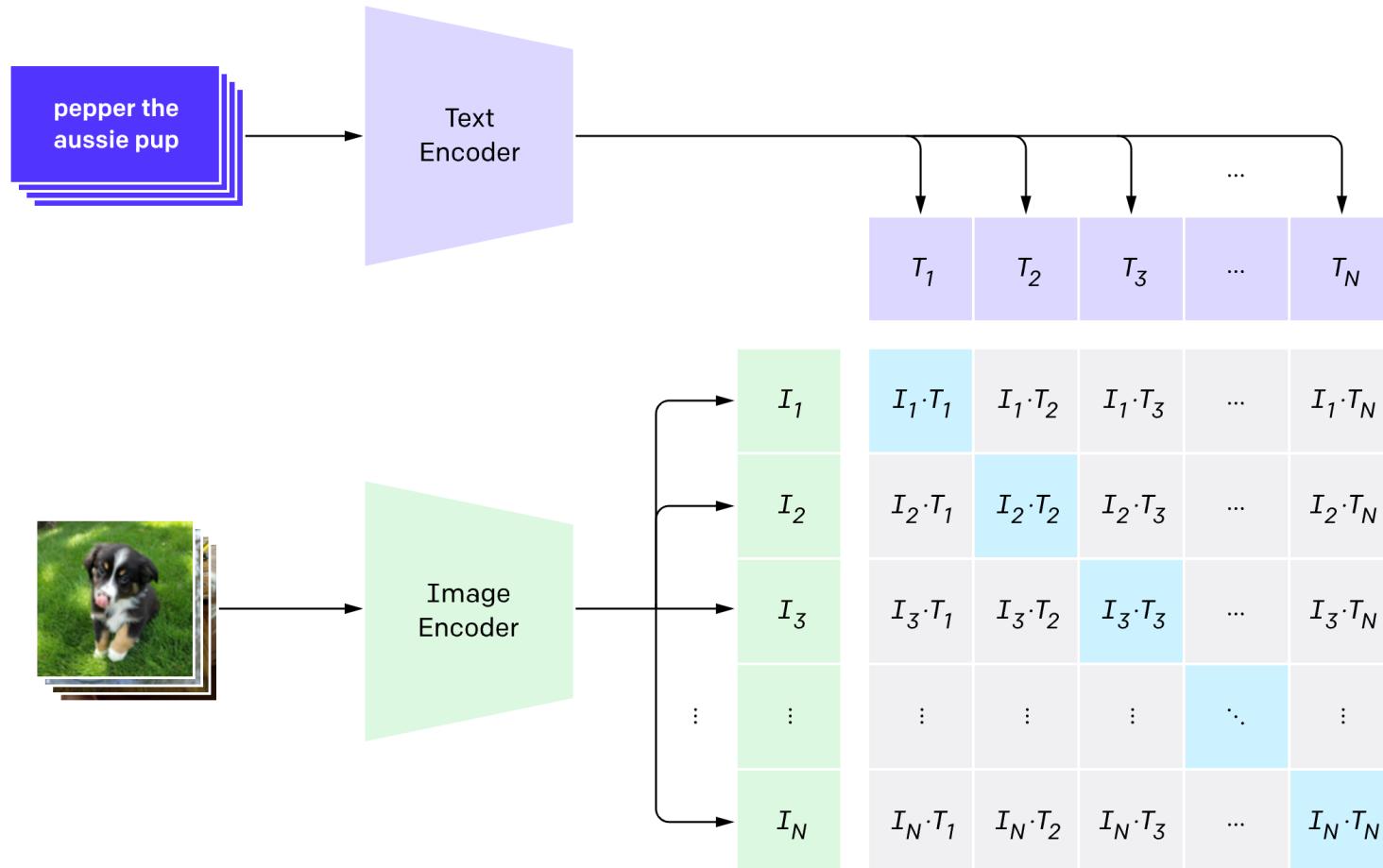
Generating Captions

PASCAL VOC Linear Classification



Desai and Johnson, "Desai and Johnson, "VirTex: Learning Visual Representations from Textual Annotations", CVPR 2021

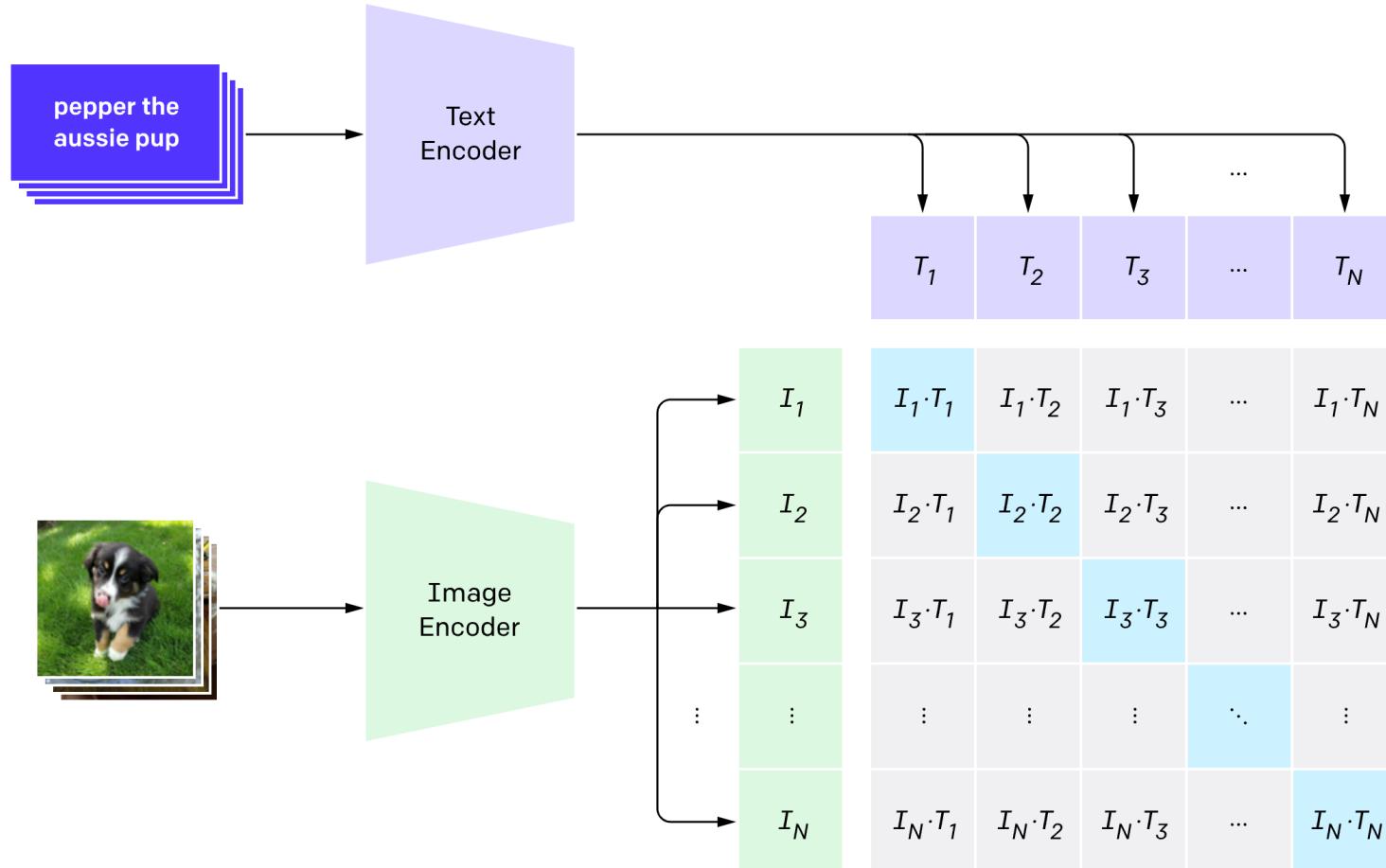
Matching Images and Text



Contrastive loss: Each image predicts which caption matches

Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021
Jia et al, "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision", ICLR 2021

Matching Images and Text: CLIP



Contrastive loss: Each image predicts which caption matches

Large-scale training on 400M (image, text) pairs from the internet

Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021

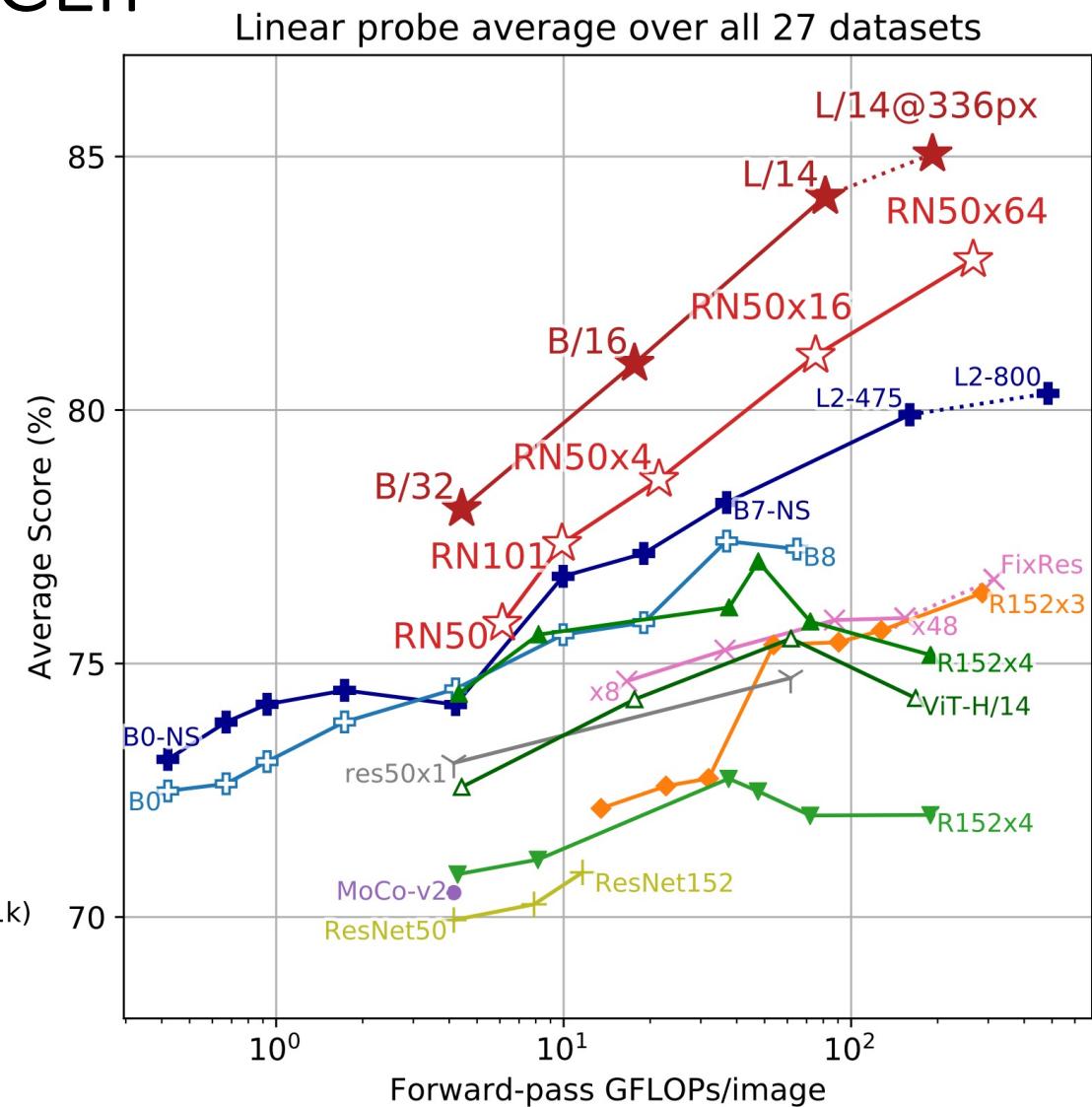
Jia et al, "Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision", ICLR 2021

Matching Images and Text: CLIP

Very strong performance on many downstream vision problems!

Performance continues to improve with larger models

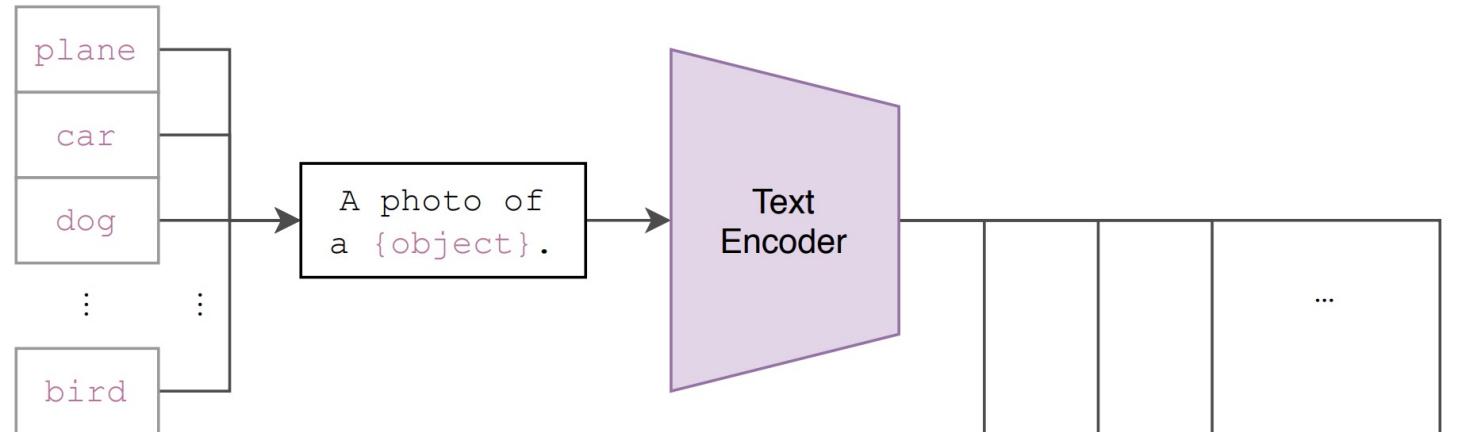
- | | | |
|-----------------------------|------------------------|----------------------|
| ★ CLIP-ViT | ✖ Instagram-pretrained | △ ViT (ImageNet-21k) |
| ★ CLIP-ResNet | ◆ SimCLRv2 | ▲ BiT-M |
| — EfficientNet-NoisyStudent | — BYOL | ▼ BiT-S |
| + EfficientNet | ● MoCo | + |



Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021

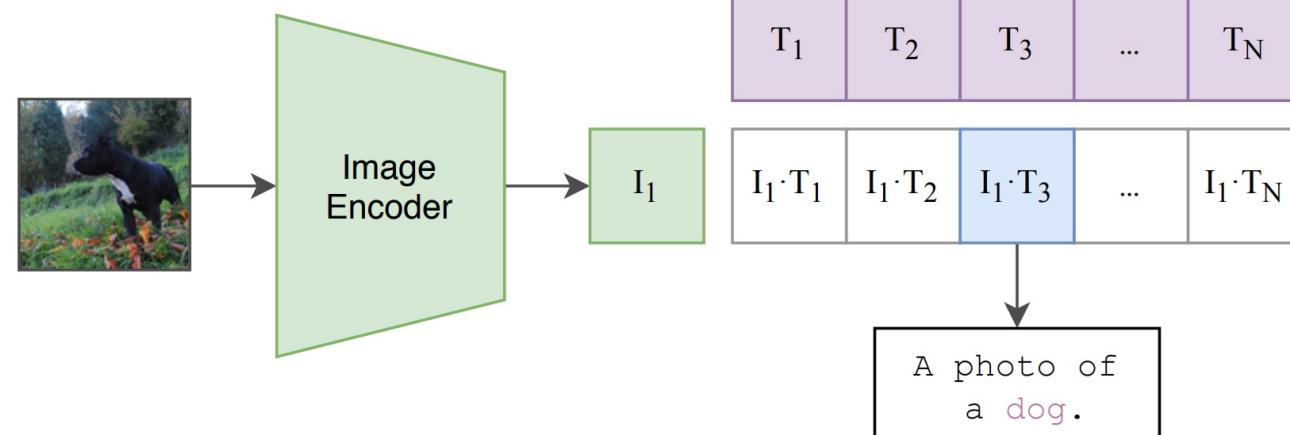
CLIP: Zero-Shot Classification

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

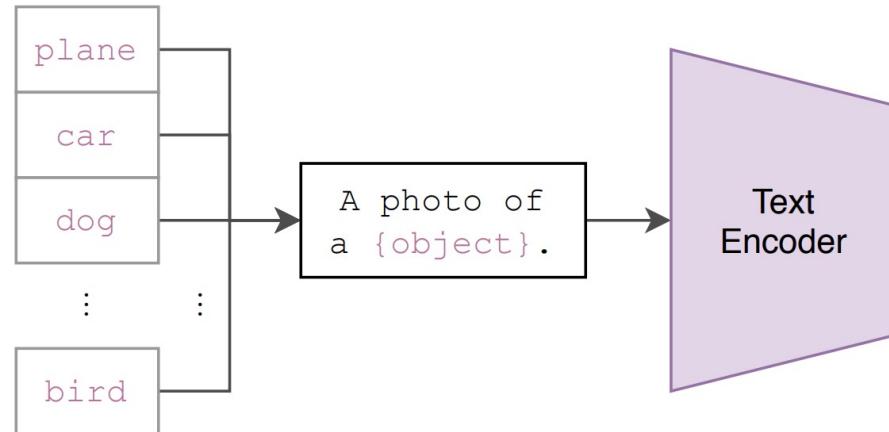
Language enables **zero-shot classification**:
Classify images into categories without any additional training data!



Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021

CLIP: Zero-Shot Classification

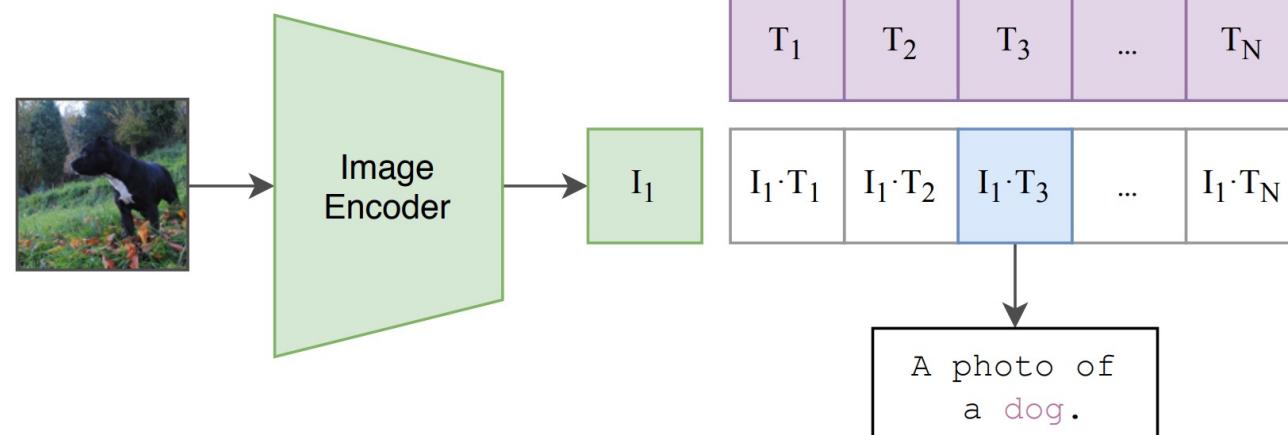
(2) Create dataset classifier from label text



Problem: CLIP training dataset is private; can't reproduce results

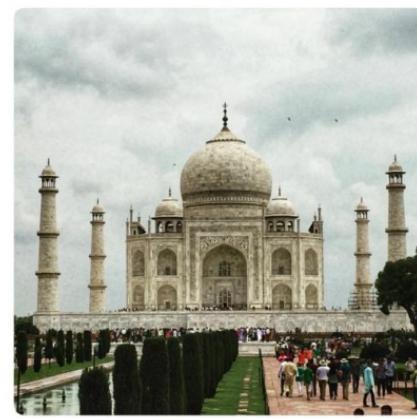
(3) Use for zero-shot prediction

Language enables **zero-shot classification**:
Classify images into categories without any additional training data!



Radford et al, "Learning Transferable Visual Models from Natural Language Supervision", ICML 2021

RedCaps: Images and Captions from Reddit



r/birdpics: male
northern cardinal

r/crafts: my mom
tied this mouse

r/itookapicture: itap of the taj mahal

r/perfectfit: this
lemon in my drink

r/shiba: mlem!

Data from 350 manually-chosen subreddits
12M high-quality (image, caption) pairs

Summary

Self-Supervised Learning (SSL) aims to scale up to larger datasets without human annotation

First train for a **pretext** task, then **transfer to downstream tasks**

Many pretext tasks: context prediction, jigsaw, colorization, clustering, rotation

SSL has been wildly successful for language

Intense research on SSL in vision; current best are contrastive, masked autoencoding

Multimodal SSL uses images together with additional context

Multimodal SSL with vision + language has been very successful; seems very promising!