CS60010: Deep Learning
Spring 2023

Sudeshna Sarkar

Self-Supervised Learning

Sudeshna Sarkar

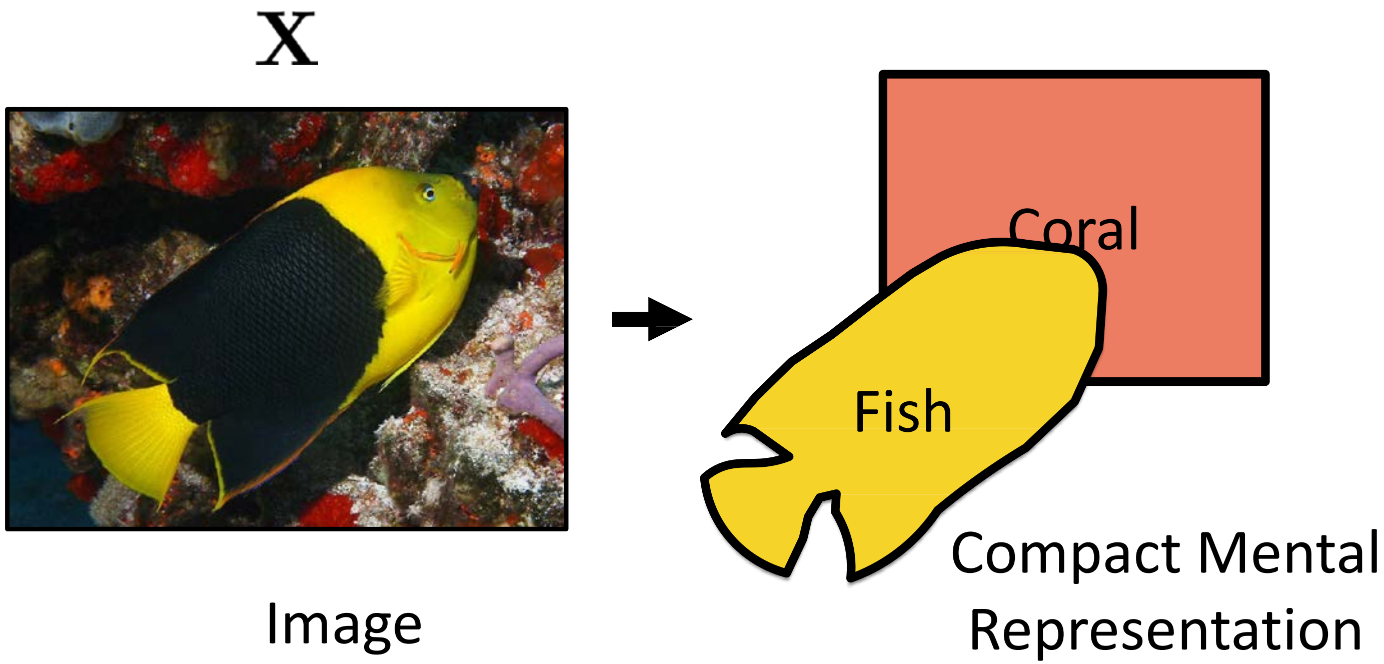
17 Mar 2023

Self-supervised Learning

- Self-supervised learning methods solve "pretext" tasks that produce good features for downstream tasks.
  - Learn with supervised learning objectives, e.g., classification, regression.
  - Labels of these pretext tasks are generated automatically

Representation learning

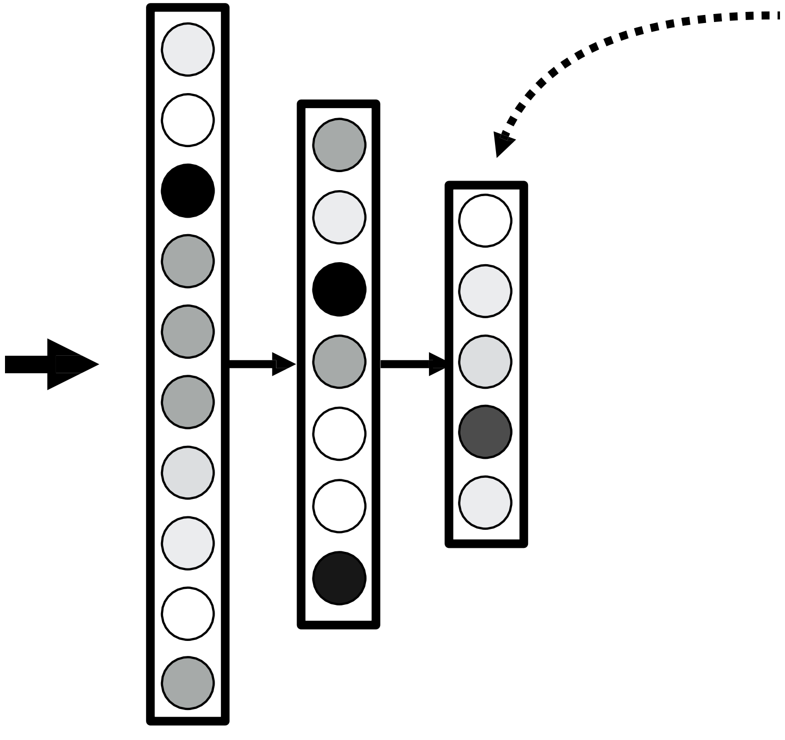
- Learn What?
- How to learn?
- Learn from what?



im2vec

x

Image



layer 3 representation of image

layer 1 representation of image

Represent image as a neural **embedding** — a vector/tensor of neural activations

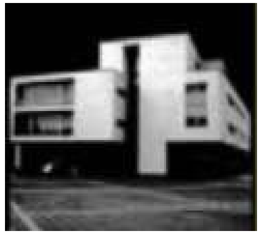
(perhaps representing a vector of detected texture patterns or object parts)

Slide credit: Phillip Isola

Investigating a representation via similarity analysis

How similar are these two images?

How about these two?



[Kriegeskorte et al. 2008]

Slide credit: Phillip Isola

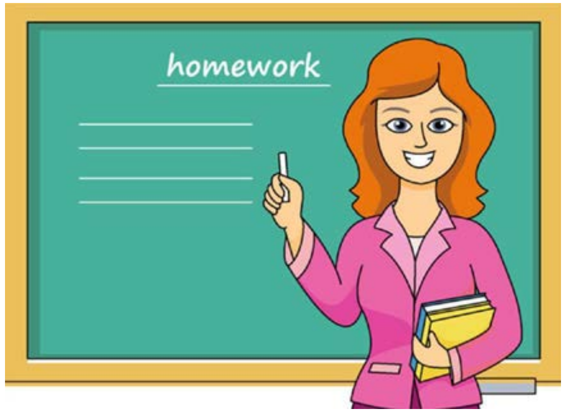
Problem: Supervised Learning is Expensive!

[slide credit: Justin Johnson]

Supervised computer vision

~~Hand-curated~~ training data

+ Informative
- Expensive
- Limited to teacher's knowledge



Vision in nature

Raw unlabeled training data

+ Cheap
- Noisy
- Harder to interpret

Slide credit: Phillip Isola

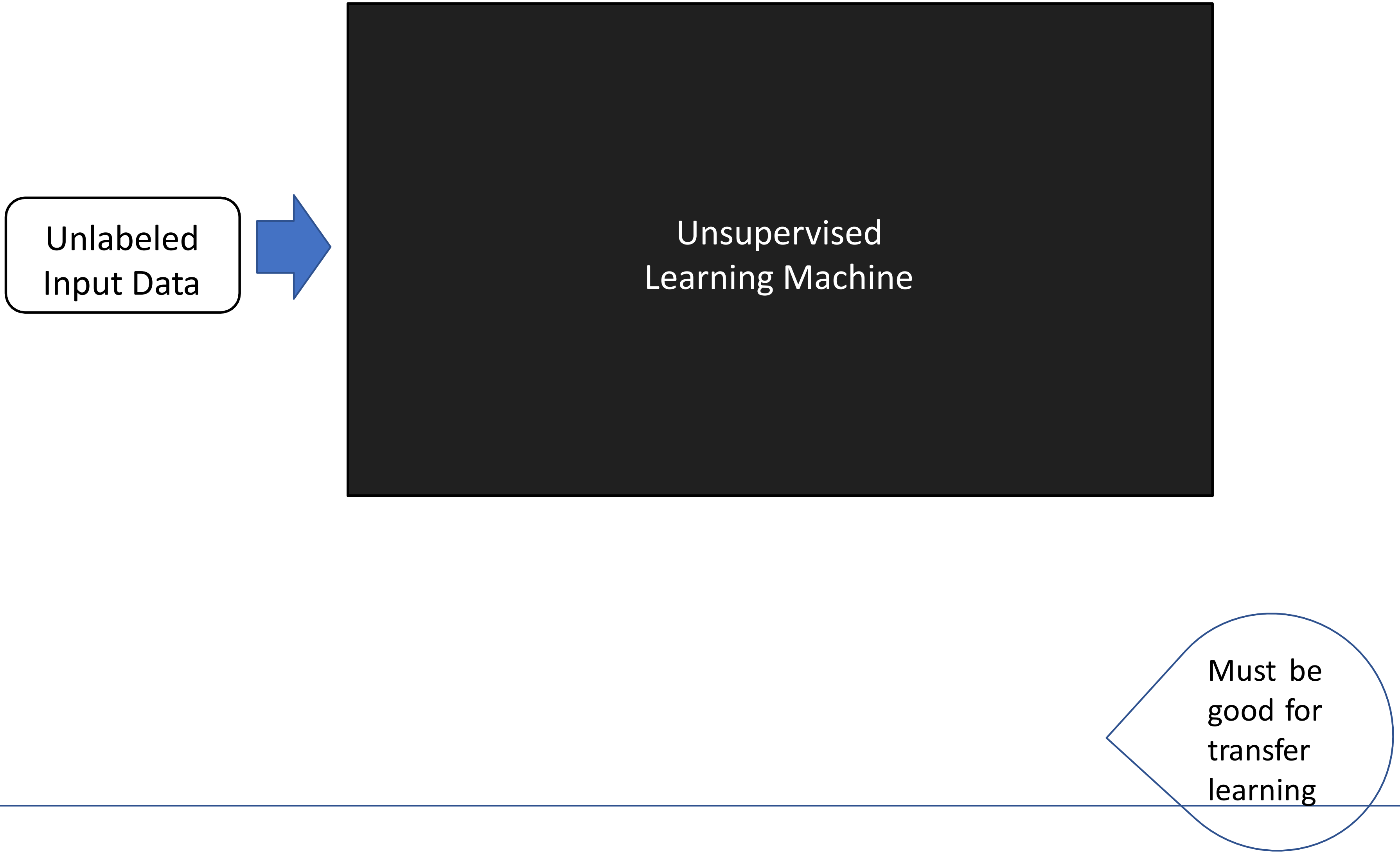
Representation Learning

Data

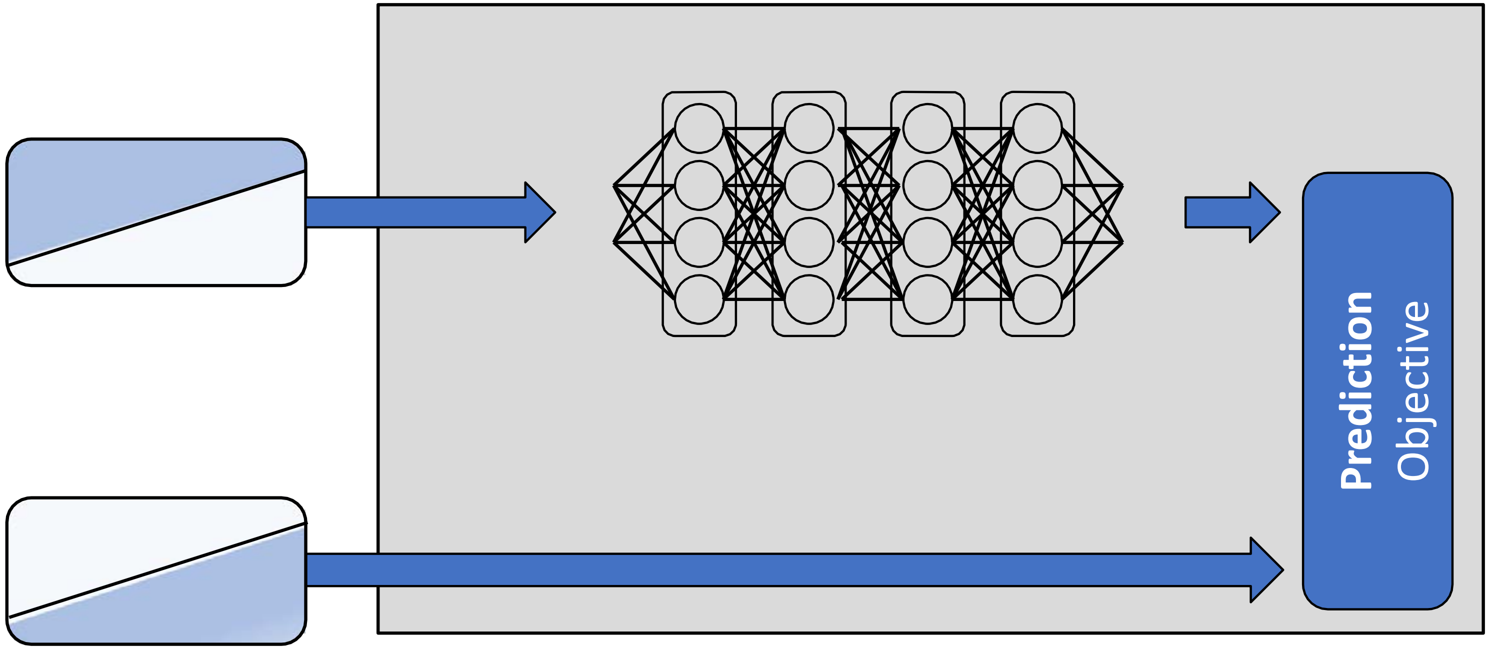
$$\{x^{(1)}\}$$
$$\{x^{(2)}\} \quad \rightarrow \quad \boxed{\text{Learner}} \quad \rightarrow \quad \text{Representations??}$$
$$\{x^{(3)}\}$$
$$\ldots$$

Slide credit: Phillip Isola

Unsupervised + Deep Learning



Data Dropout Prediction

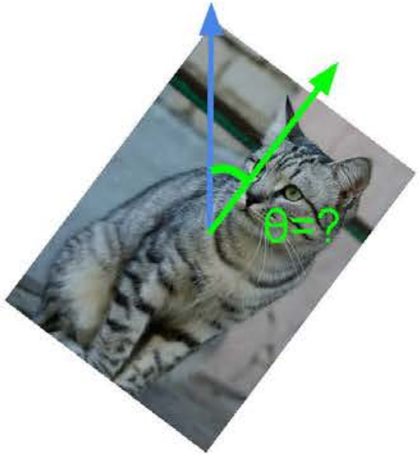


- Unsupervised / Self-supervised by predicting part of data from other part

Self-supervised pretext tasks

learn to predict image transformations / complete corrupted images.

image completion



rotation prediction

“jigsaw puzzle”

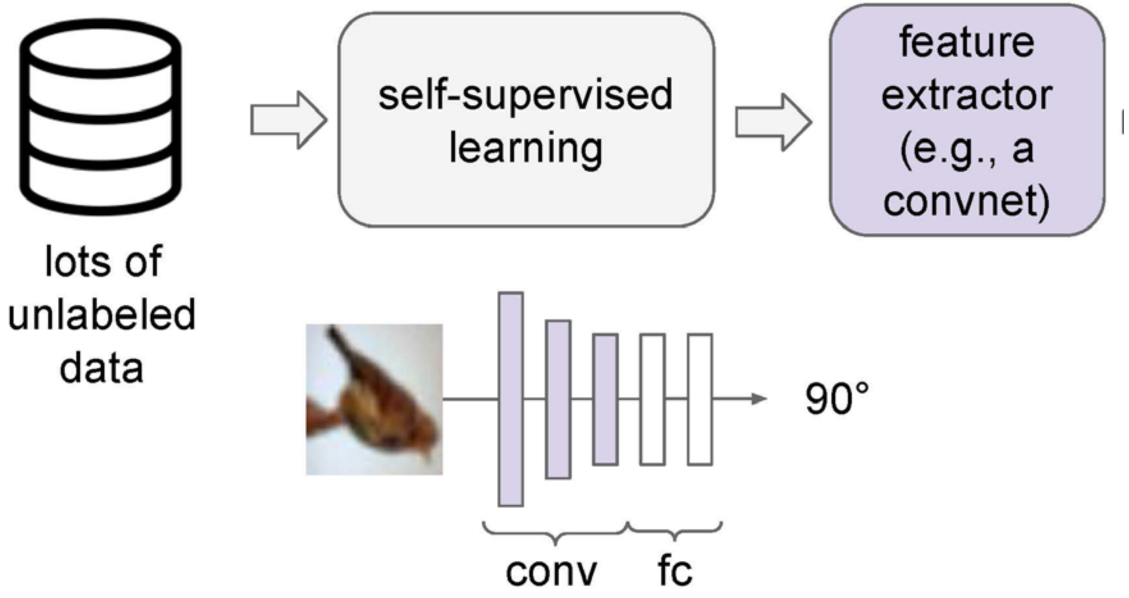
colorization

1. Solving the pretext tasks allow the model to learn good features.
2. We can automatically generate labels for the pretext tasks.

How to evaluate a self-supervised learning method?

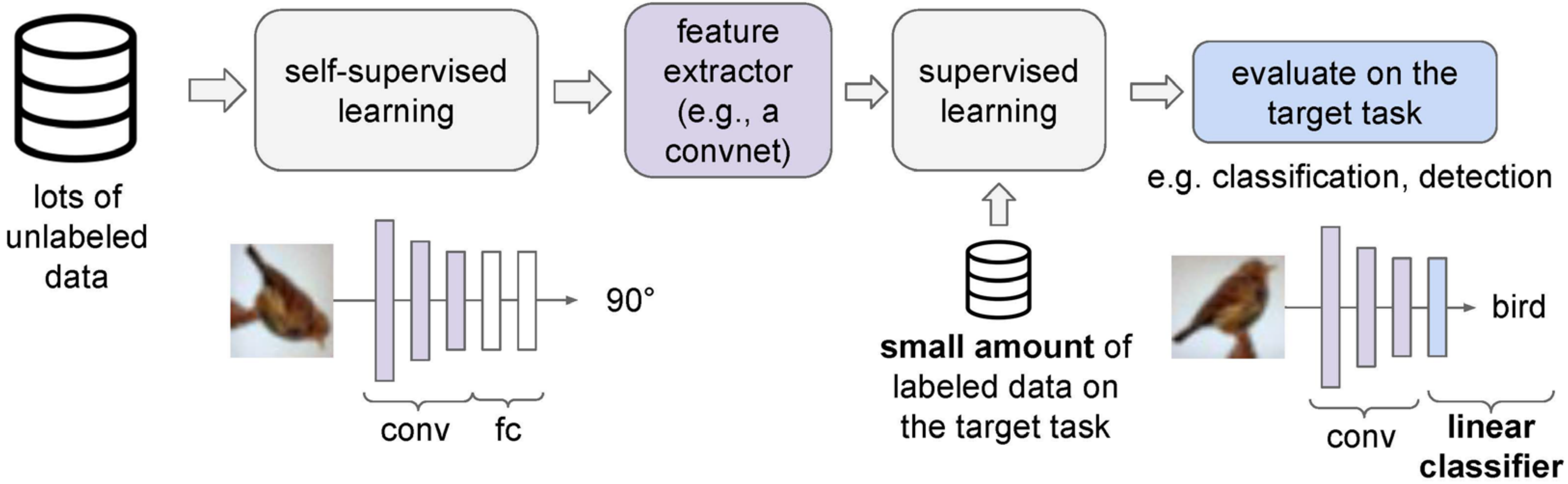
1. Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations
2. Evaluate the learned feature encoders on downstream target tasks
    - Attach a shallow network on the feature extractor;
    - train the shallow network on the target task with small amount of labeled data

How to evaluate a self-supervised learning method?



Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations

How to evaluate a self-supervised learning method?

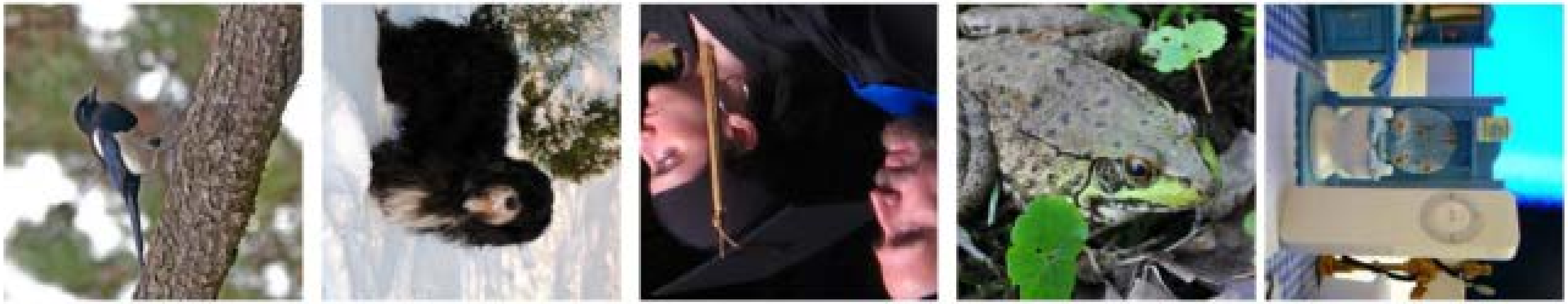


Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations

Evaluate the learned feature encoders on downstream target tasks

- Attach a shallow network on the feature extractor;
- train the shallow network on the target task with small amount of labeled data

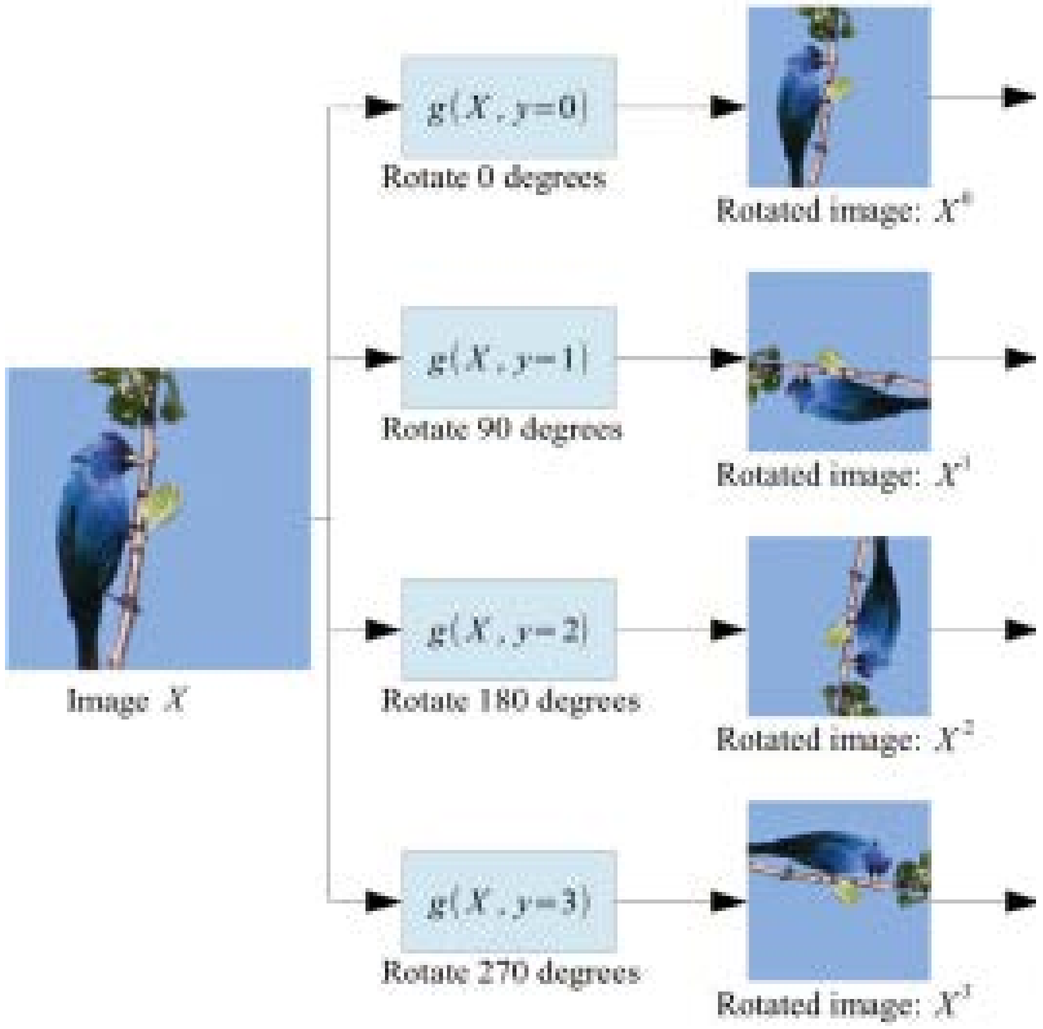
Pretext task: predict rotations



90° rotation 270° rotation 180° rotation 0° rotation 270° rotation

Hypothesis: a model could recognize the correct rotation of an object only if it has the “visual commonsense” of what the object should look like unperturbed.

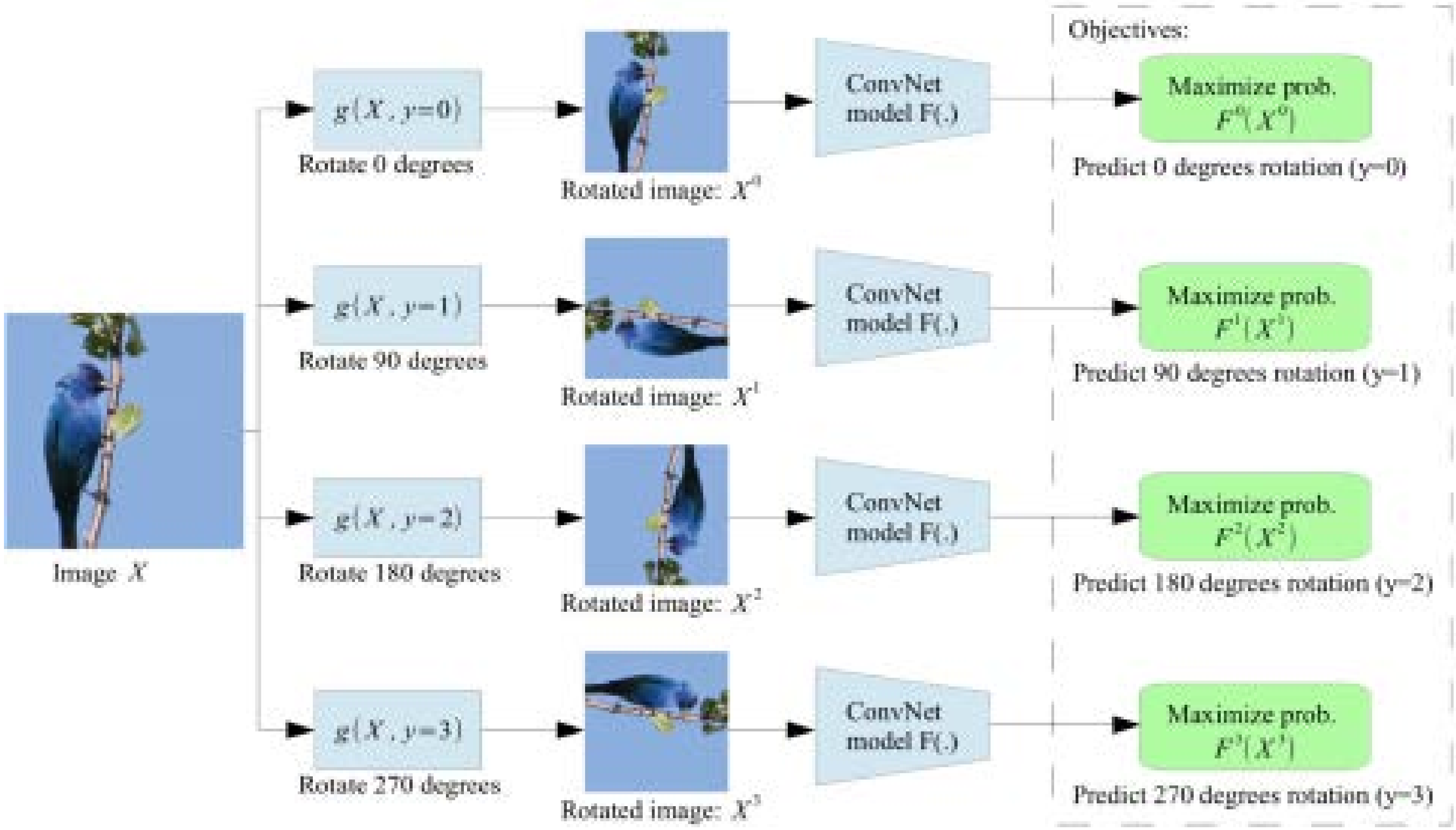
Pretext task: predict rotations



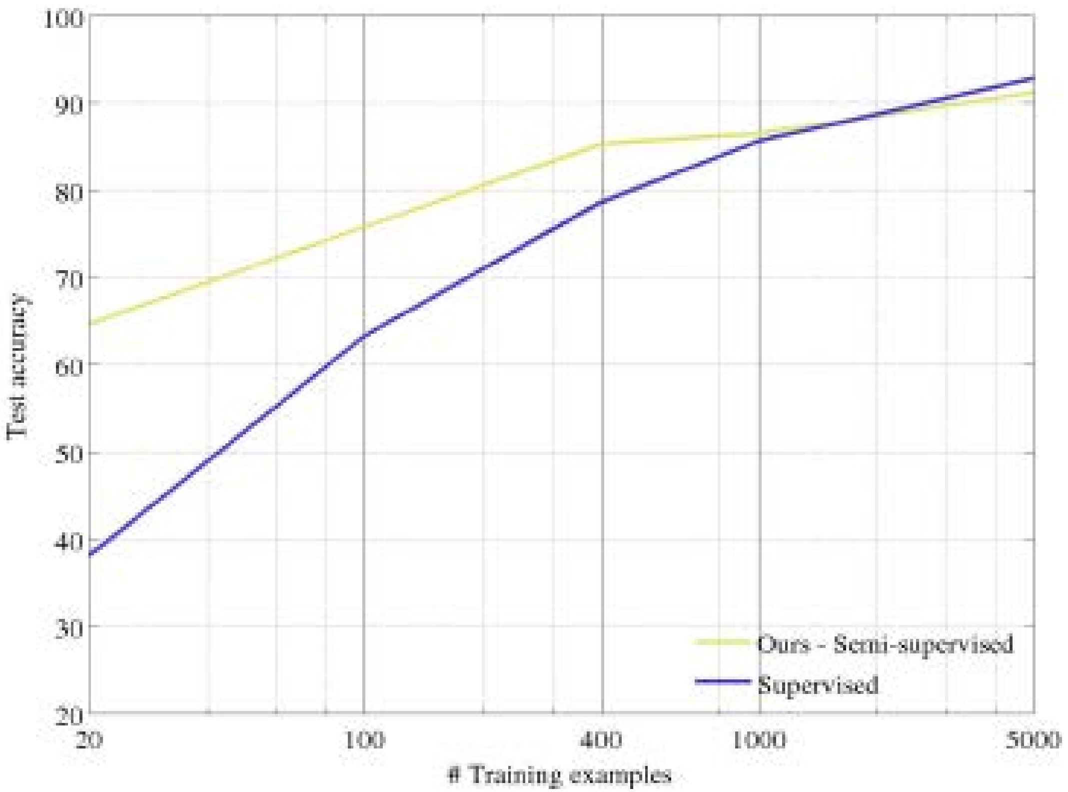
Self-supervised learning by rotating the entire input images.

The model learns to predict which rotation is applied (4-way classification)

Pretext task: predict rotations



Evaluation on semi-supervised learning



Self-supervised learning on CIFAR10 (entire training set)

Freeze conv1 + conv2 Learn conv3 + linear layers with subset of labeled CIFAR10 data (classification).

Transfer learned features to supervised learning

	Classification (%mAP)		Detection (%mAP)	Segmentation (%mIoU)
Trained layers	fc6-8	all	all	all
ImageNet labels	78.9	79.9	56.8	48.0
Random		53.3	43.4	19.8
Random rescaled Krähenbühl et al. (2015)	39.2	56.6	45.6	32.6
Egomotion (Agrawal et al., 2015)	31.0	54.2	43.9	
Context Encoders (Pathak et al., 2016b)	34.6	56.5	44.5	29.7
Tracking (Wang & Gupta, 2015)	55.6	63.1	47.4	
Context (Doersch et al., 2015)	55.1	65.3	51.1	
Colorization (Zhang et al., 2016a)	61.5	65.6	46.9	35.6
BIGAN (Donahue et al., 2016)	52.3	60.1	46.9	34.9
Jigsaw Puzzles (Noroozi & Favaro, 2016)	-	67.6	53.2	37.6
NAT (Bojanowski & Joulin, 2017)	56.7	65.3	49.4	
Split-Brain (Zhang et al., 2016b)	63.0	67.1	46.7	36.0
ColorProxy (Larsson et al., 2017)		65.9		38.4
Counting (Noroozi et al., 2017)	-	67.7	51.4	36.6
(Ours) RotNet	**70.87**	**72.97**	**54.4**	**39.1**

Pretrained with full ImageNet supervision

No pretraining

Self-supervised learning on ImageNet (entire training set) with AlexNet

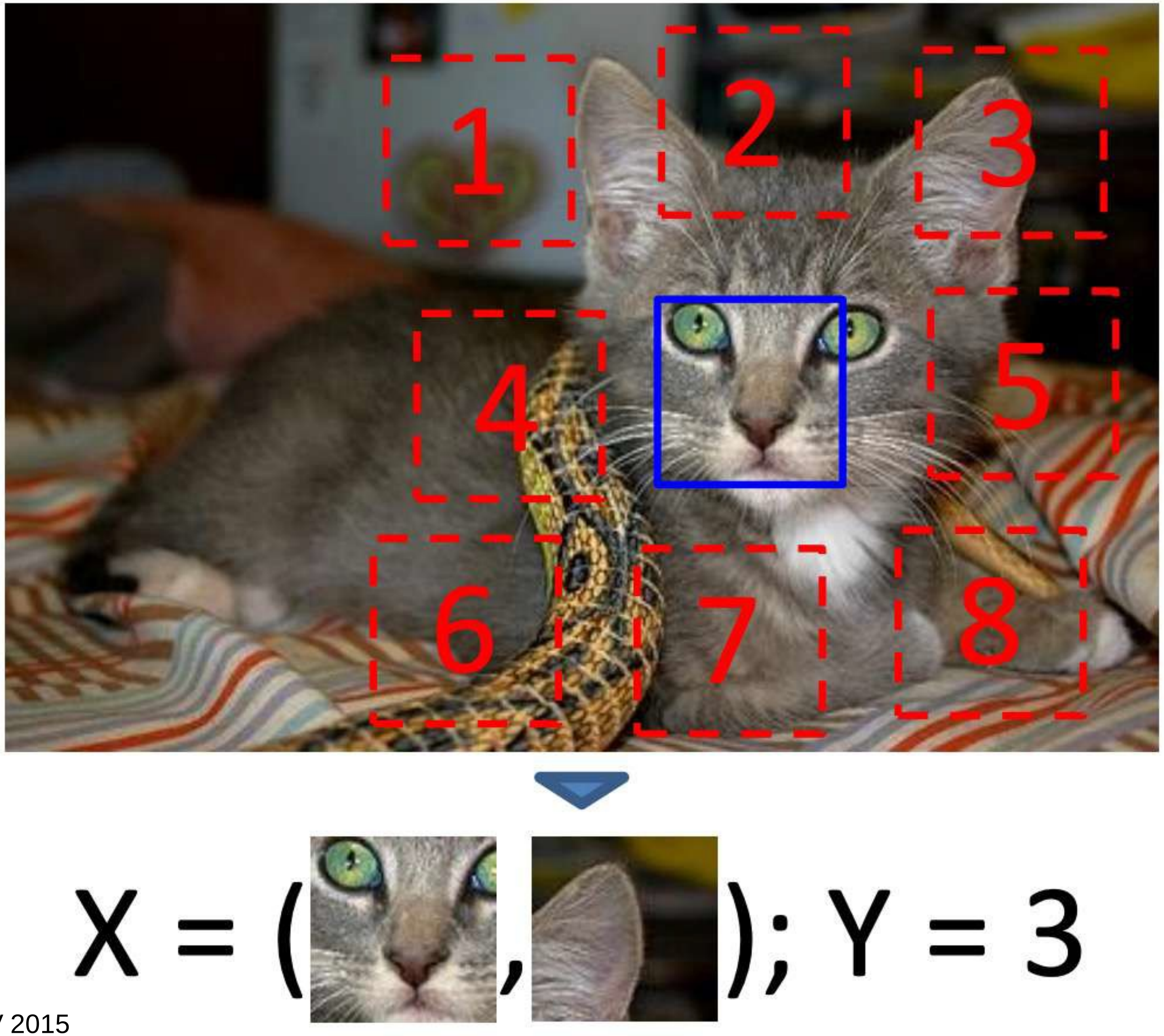
Finetune on labeled data from Pascal VOC 2007

Self-supervised learning with rotation prediction

Pretext task: predict relative patch locations

Model predicts relative location of two patches from the same image. <u>Discriminative</u> pretraining task

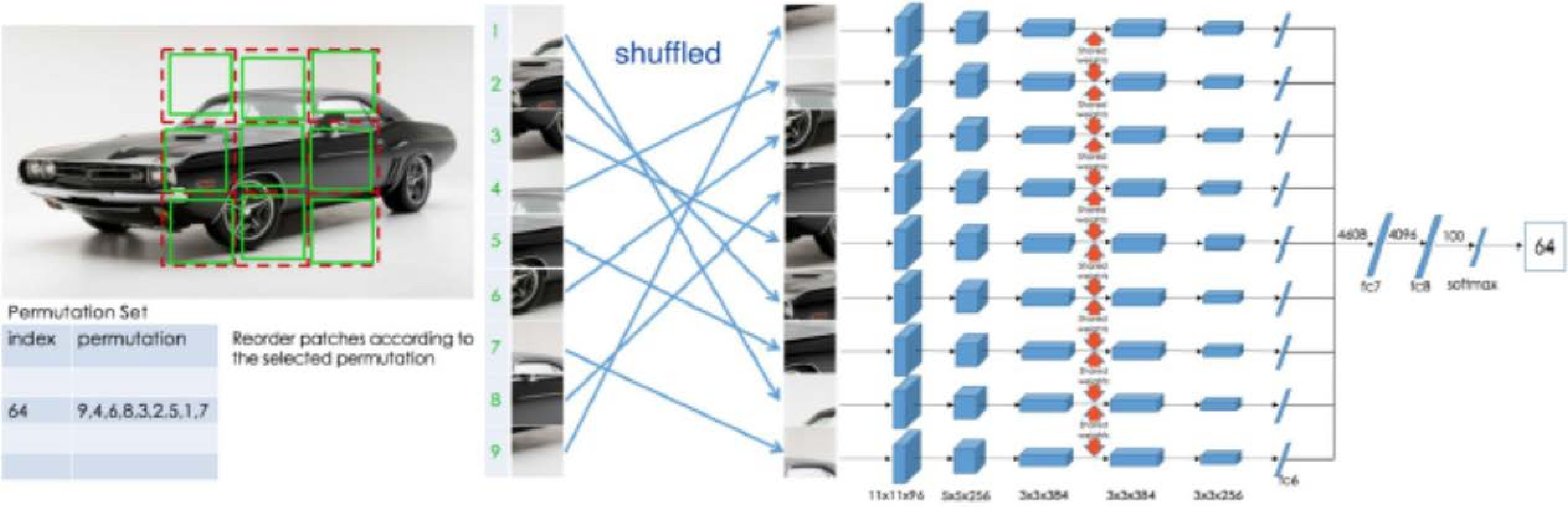
Intuition: Requires understanding objects and their parts



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

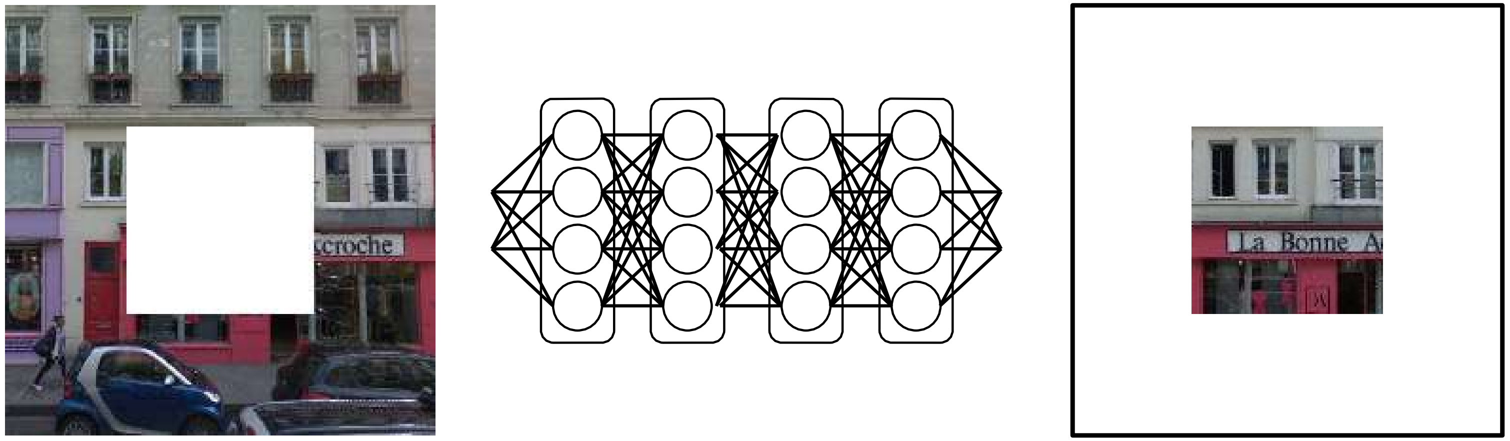
[slide credit: Justin Johnson]

Pretext task: solving "jigsaw puzzles"



Noroozi & Favaro, 2016)

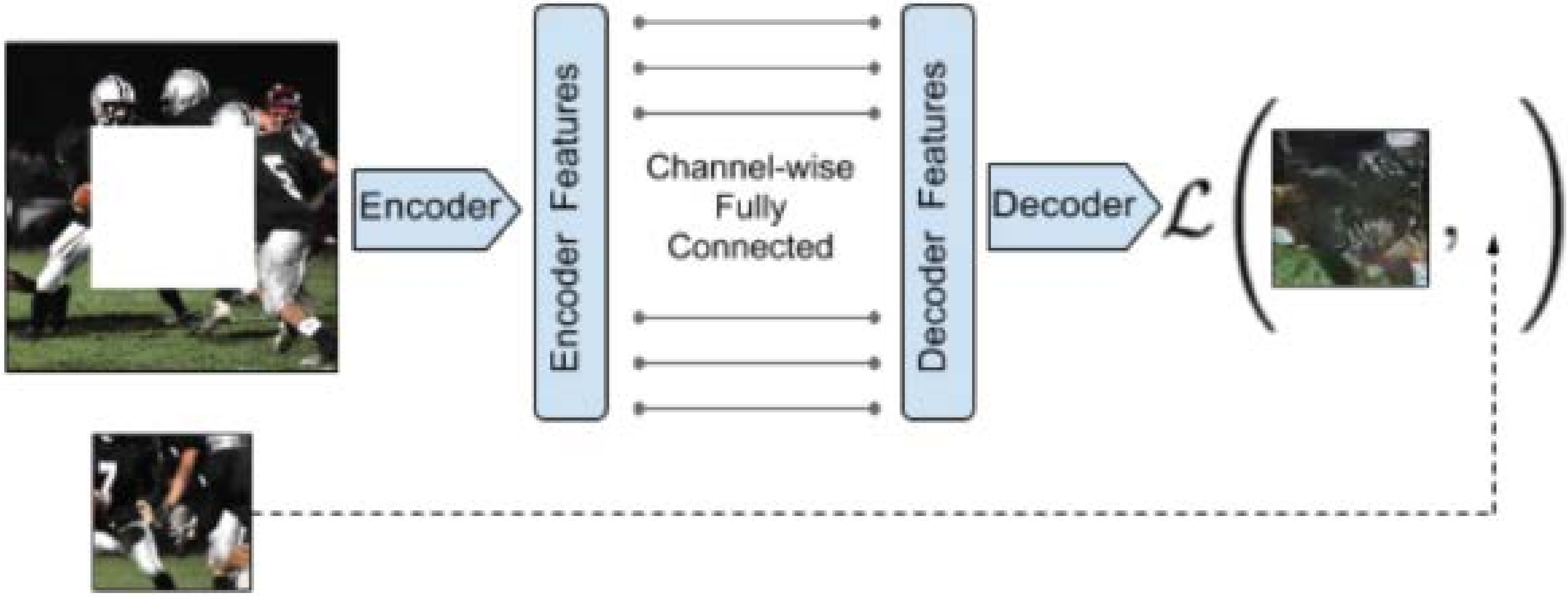
Pretext task: predict missing pixels (inpainting)



Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, Alexei Efros. CVPR 2016

Feature Learning by Inpainting

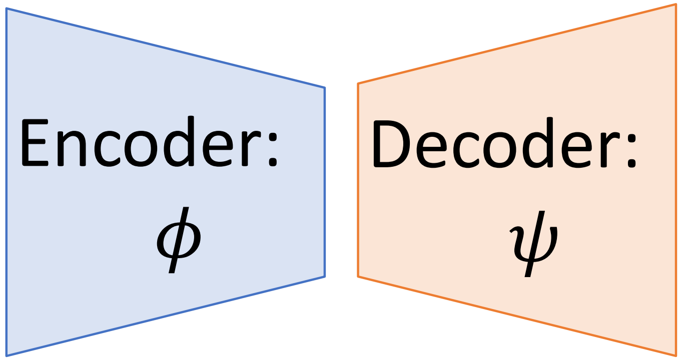
Learning to inpaint by reconstruction



Learning to reconstruct the missing pixels

Context Encoders: Learning by Inpainting

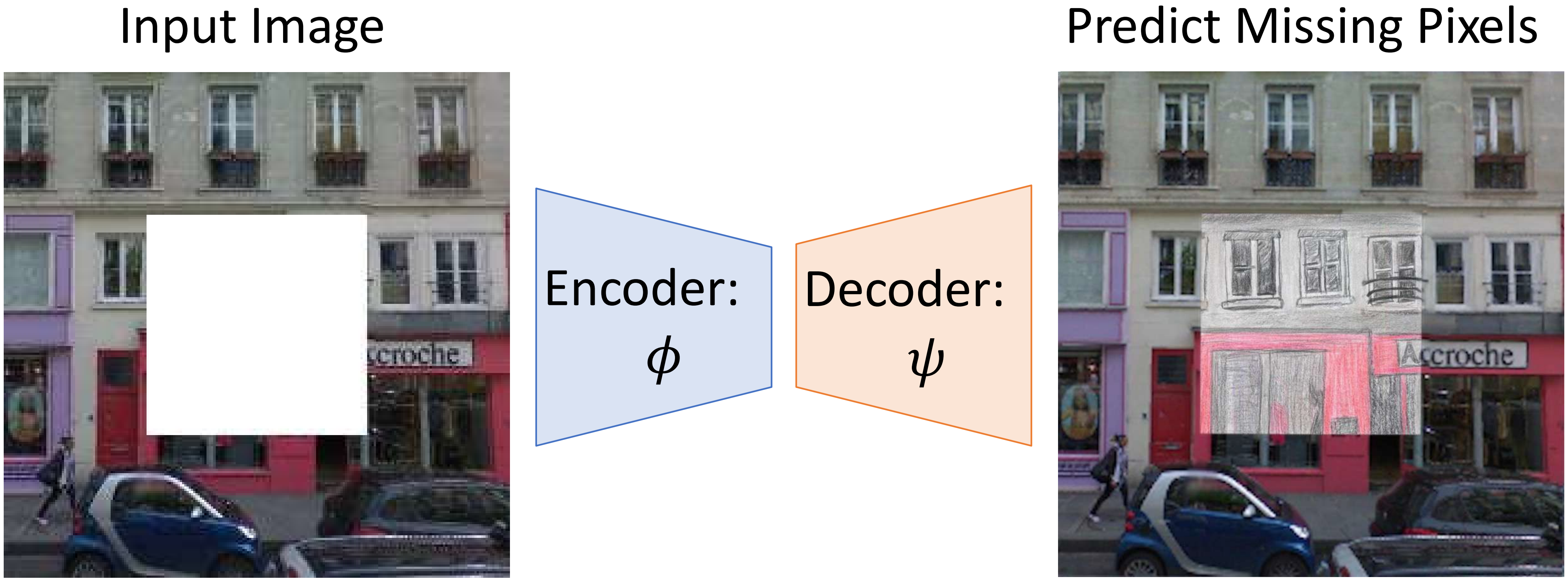
Input Image



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

[slide credit: Justin Johnson]

Context Encoders: Learning by Inpainting



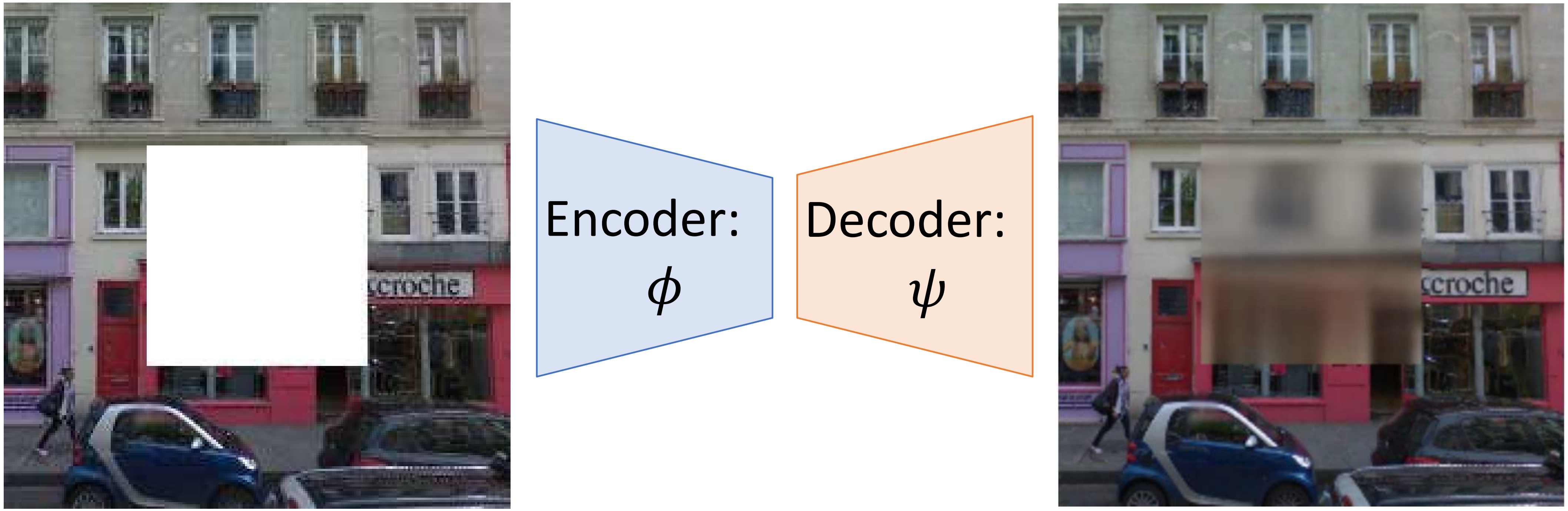
Pathak et al, “Context Encoders: Feature Learning by Inpainting”, CVPR 2016

[slide credit: Justin Johnson]

Context Encoders: Learning by Inpainting

Input Image

Predict Missing Pixels



L2 Loss

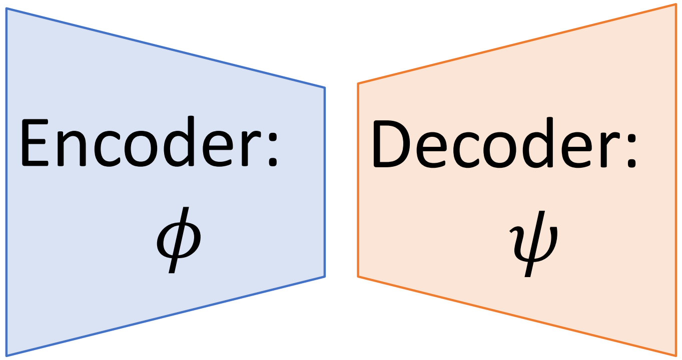
(Best for feature learning)

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

[slide credit: Justin Johnson]

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

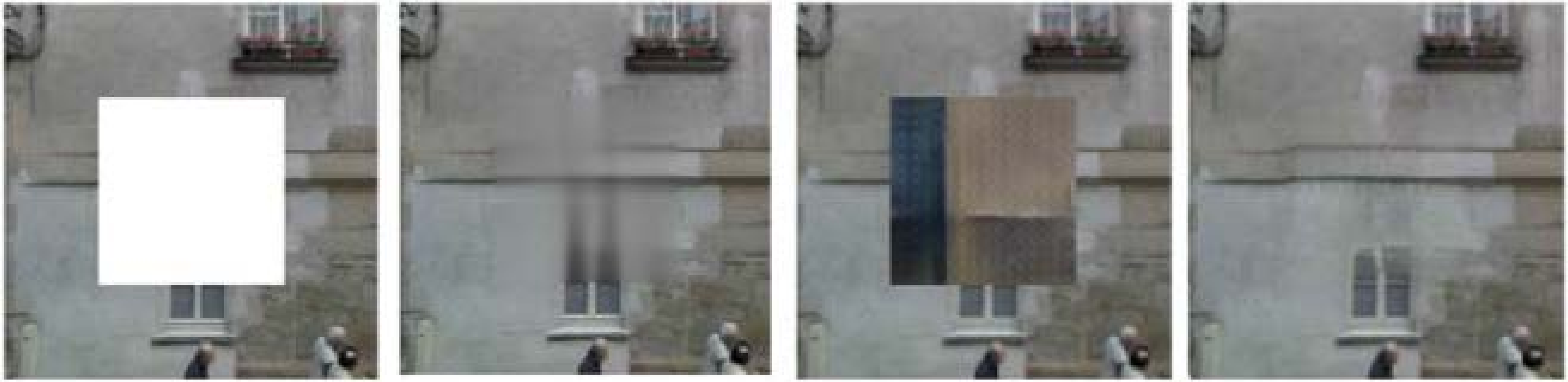
[slide credit: Justin Johnson]

Learning to inpaint by reconstruction

- Loss = reconstruction + adversarial learning

- Adversarial loss between “real” images and inpainted images

Inpainting evaluation



Input (context) reconstruction adversarial recon + adv

Pretext task: image coloring

Summary: pretext tasks from image transformations

- Pretext tasks focus on “visual common sense”, e.g., predict rotations, inpainting, rearrangement, and colorization.
- The models are forced learn good features about natural images, e.g., semantic representation of an object category, in order to solve the pretext tasks.
- We don’t care about the performance of these pretext tasks, but rather how useful the learned features are for downstream tasks (classification, detection, segmentation).
- Problems: 1) coming up with individual pretext tasks is tedious, and 2) the learned representations may not be general.

Pretext tasks from image transformations

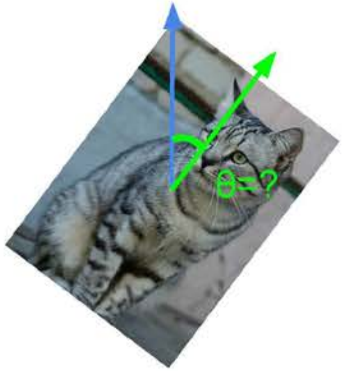
- Learned representations may be tied to a specific pretext task!Can we come up with a more general pretext task?

Contrastive representation learning

- Intuition and formulation
- Instance contrastive learning: SimCLR and MOCO
- Sequence contrastive learning: CPC

A more general pretext task?

image completion



rotation prediction

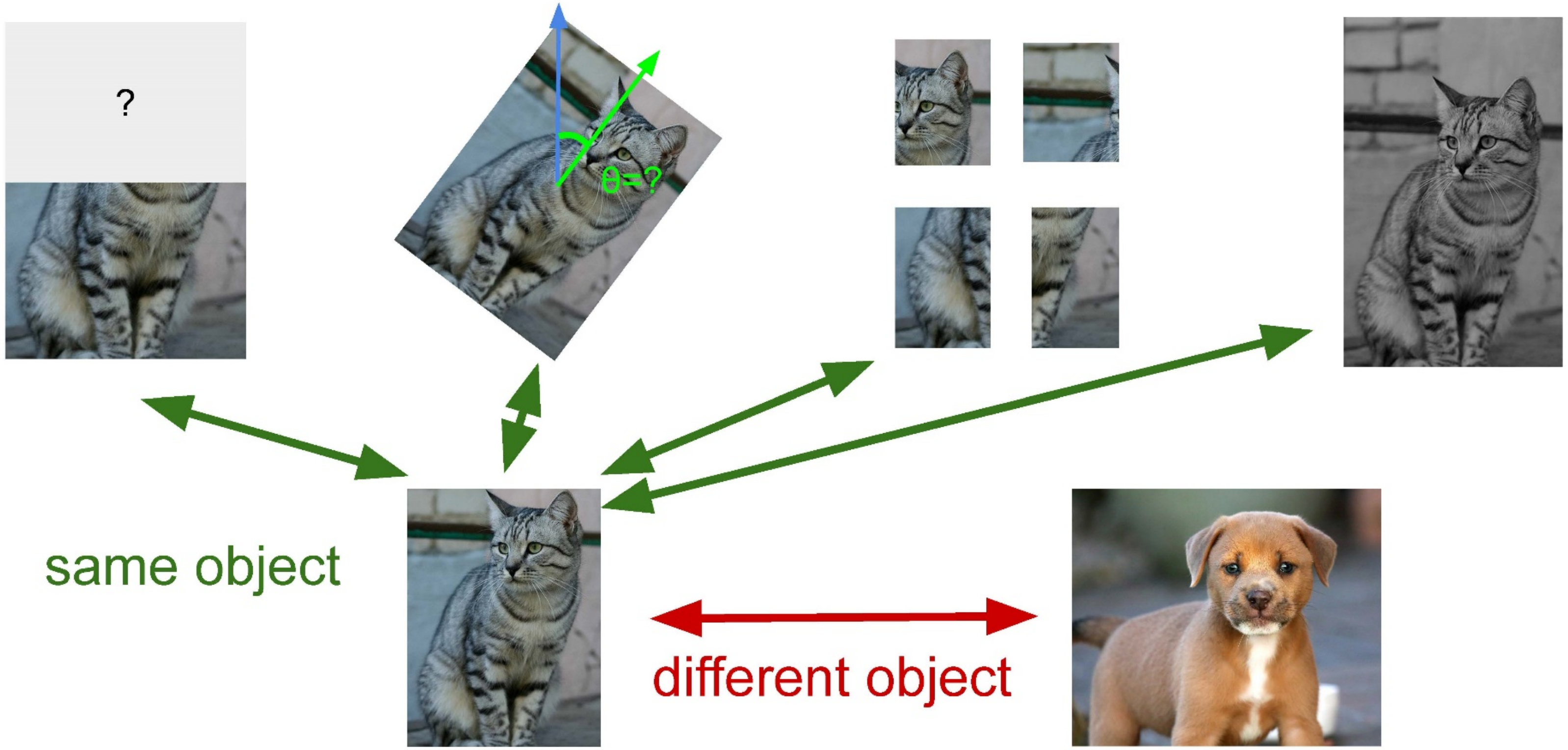
“jigsaw puzzle”

colorization

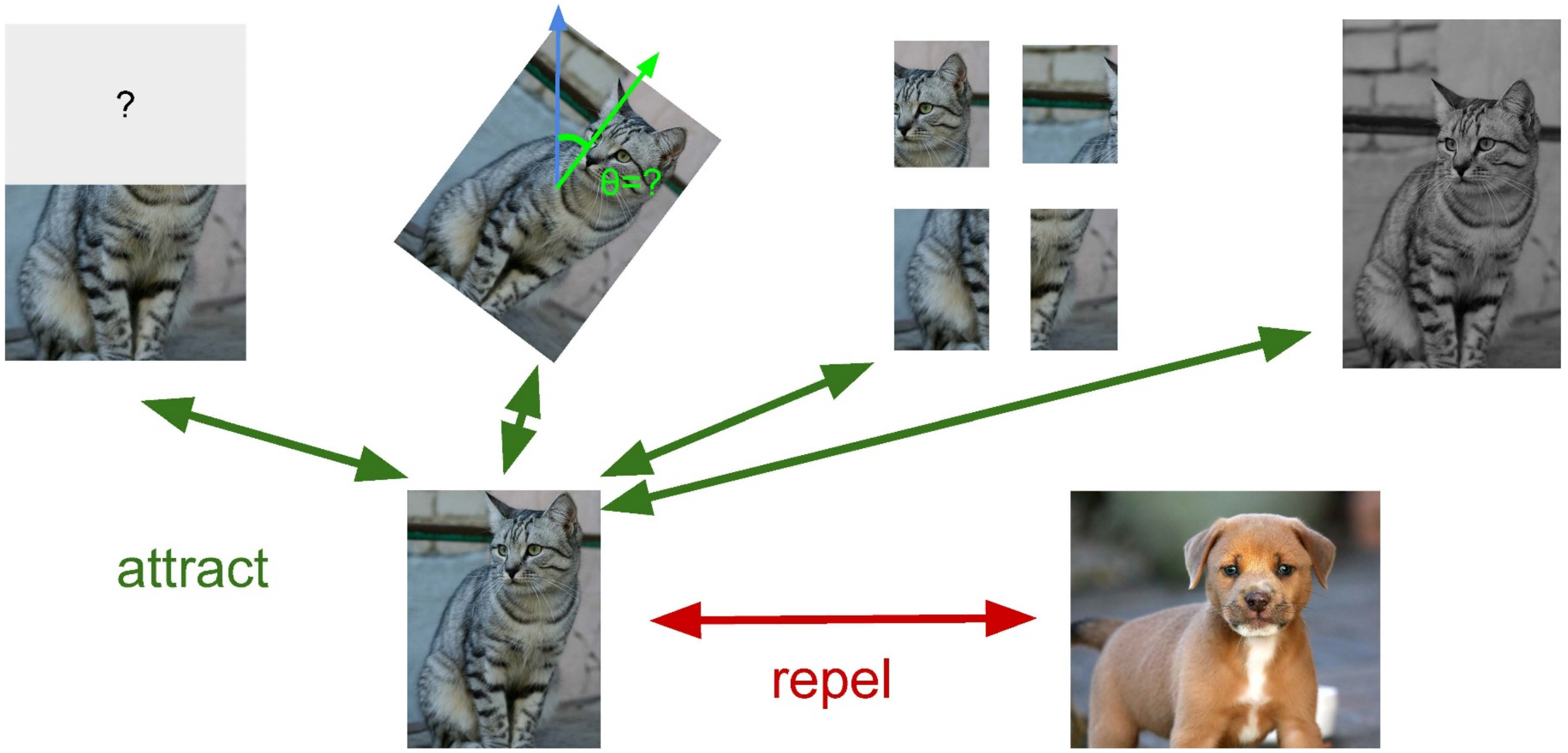
Learned representations may be tied to a specific pretext task!

Can we come up with a more general pretext task?

A more general pretext task?



Contrastive Representation Learning



Contrastive Learning

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

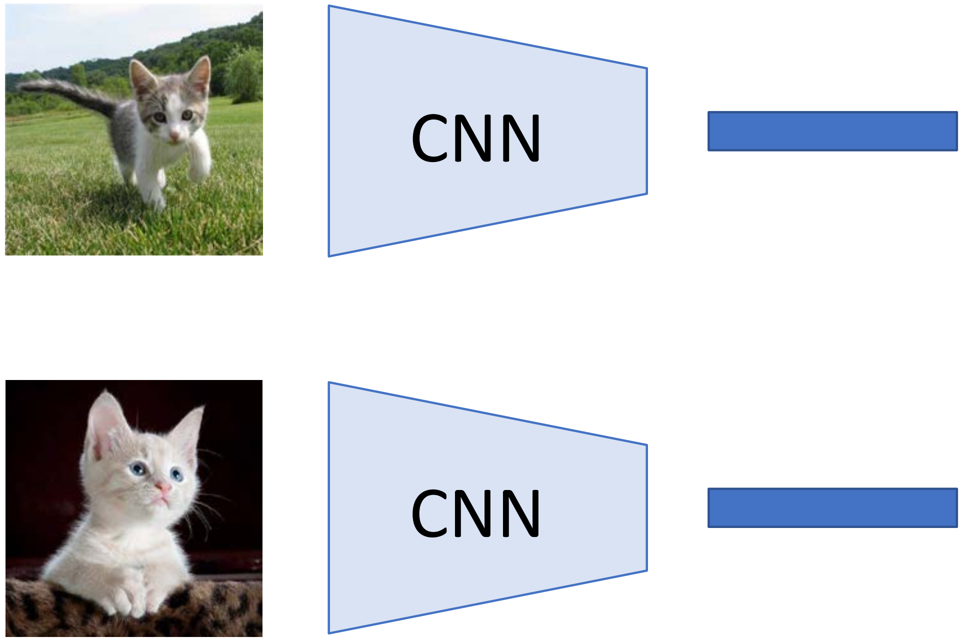
Hadsell et al, “Dimensionality Reduction by Learning and Invariant Mapping”, CVPR 2006

[slide credit: Justin Johnson]

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



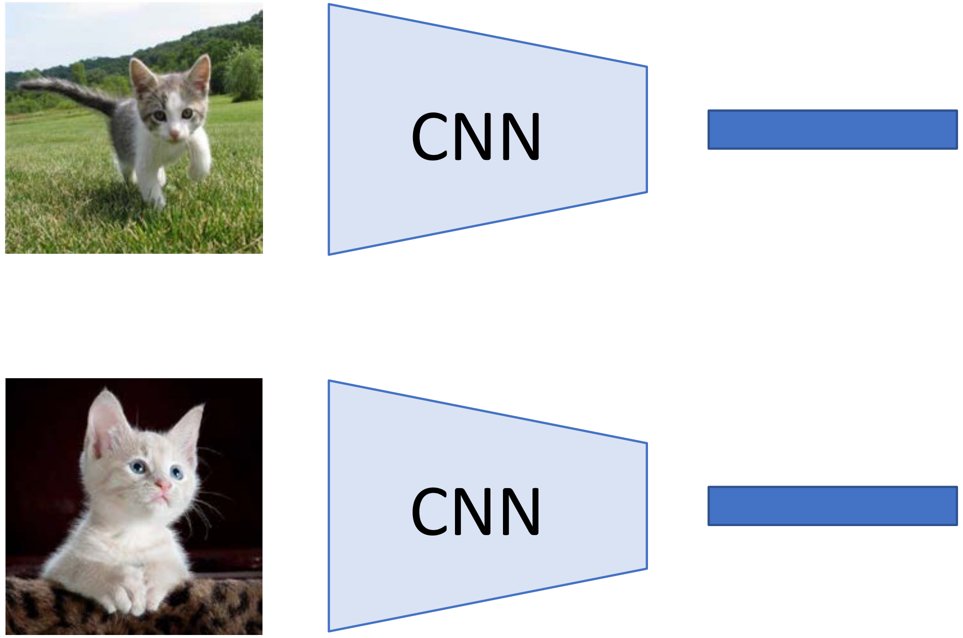
Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[slide credit: Justin Johnson]

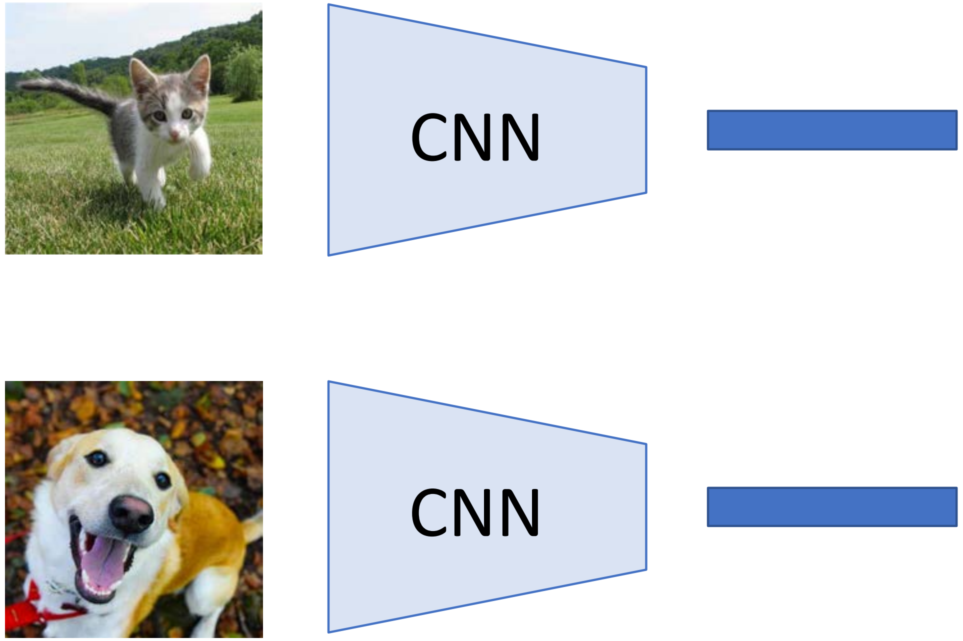
Contrastive Learning

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



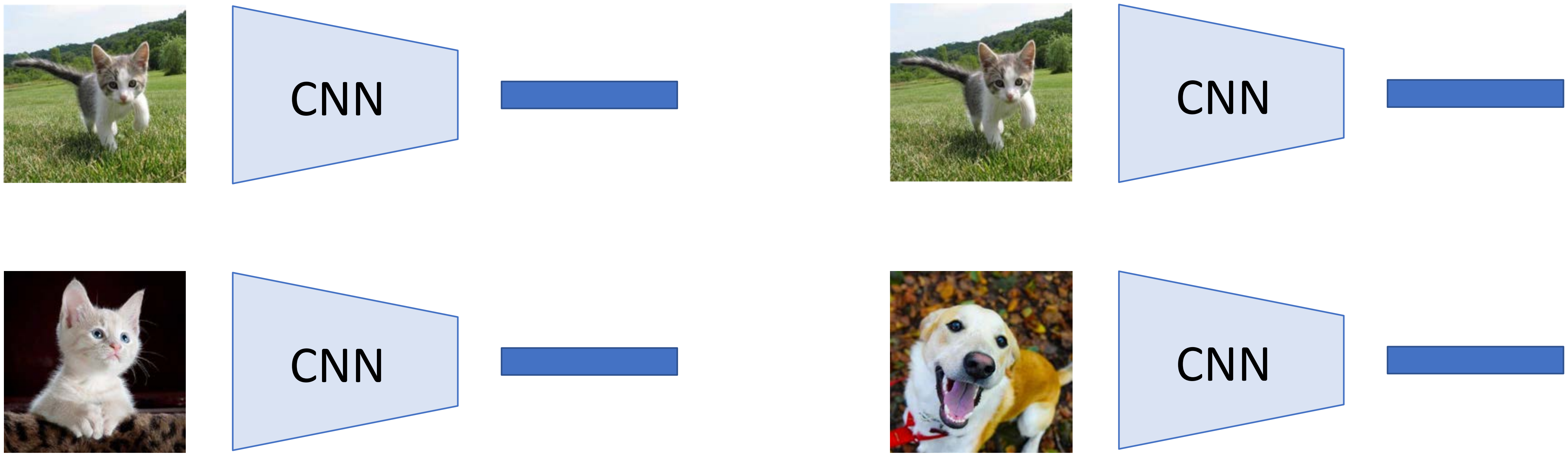
Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

Contrastive Learning

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

Let d be the Euclidean distance between features for two images

Similar images should have similar features   **Dissimilar** images should have dissimilar features



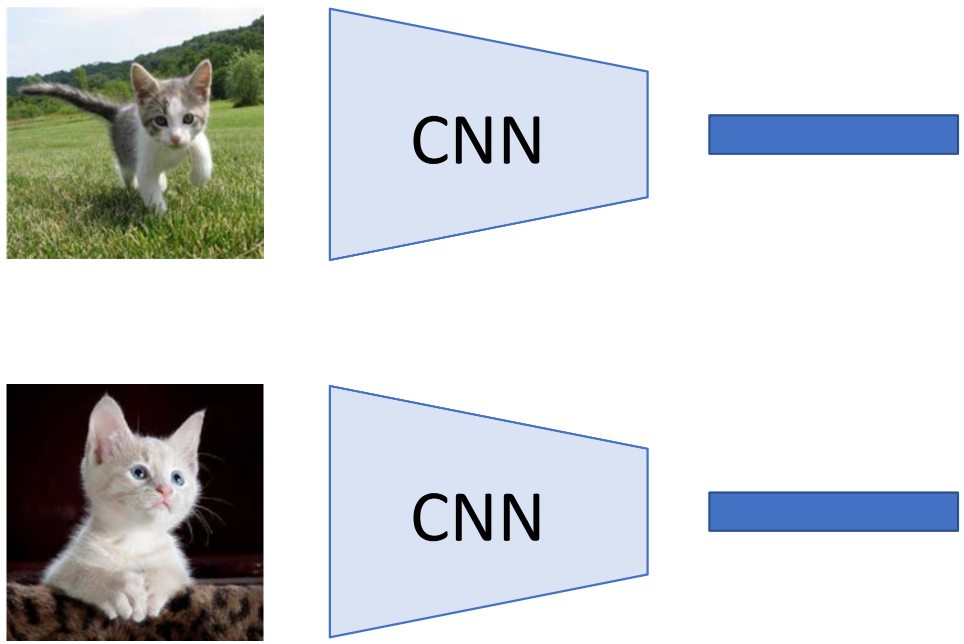
Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006

[slide credit: Justin Johnson]

Contrastive Learning

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

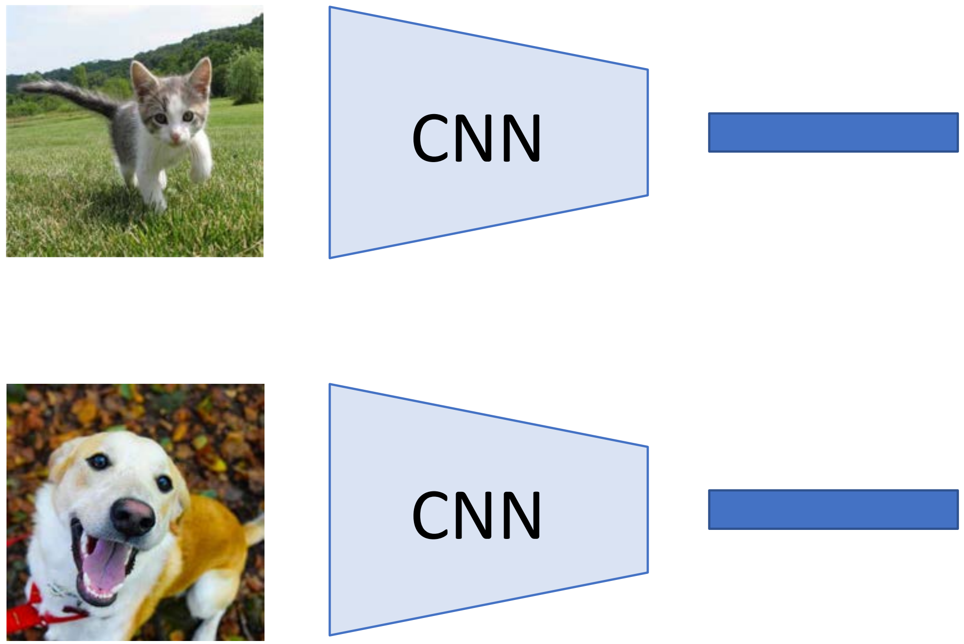
Similar images should have similar features



$$L_S(x1, x2) = d^2$$

Pull features together

Dissimilar images should have dissimilar features



$$L_D(x1, x2) = \max(0, m - d^2)$$

Push features apart

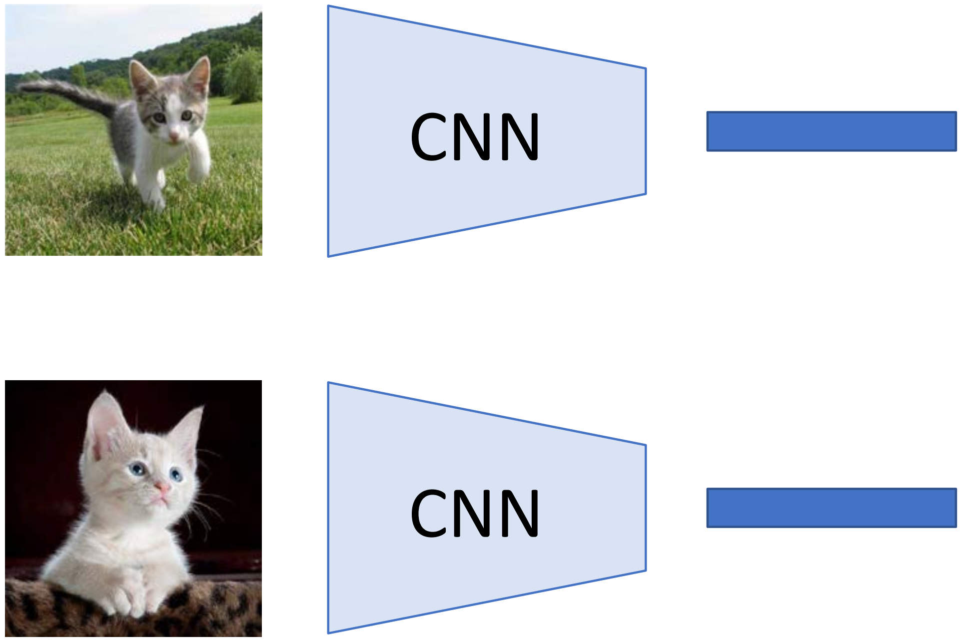
(upto margin m)

[slide credit: Justin Johnson]

Contrastive Learning

Problem: Where to get positive and negative pairs?

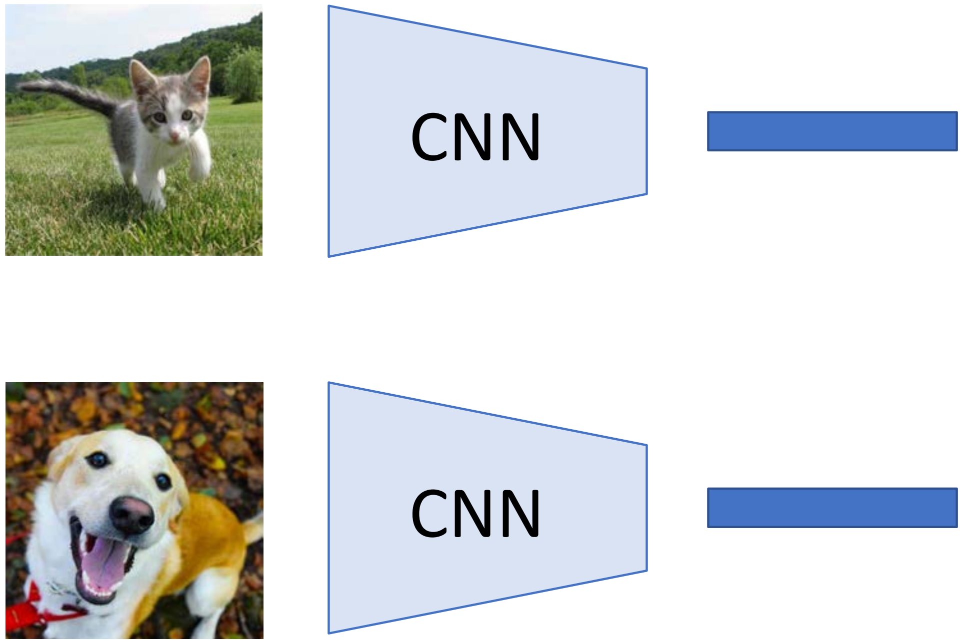
Similar images should have similar features



$$L_S(x1, x2) = d^2$$

Pull features together

Dissimilar images should have dissimilar features



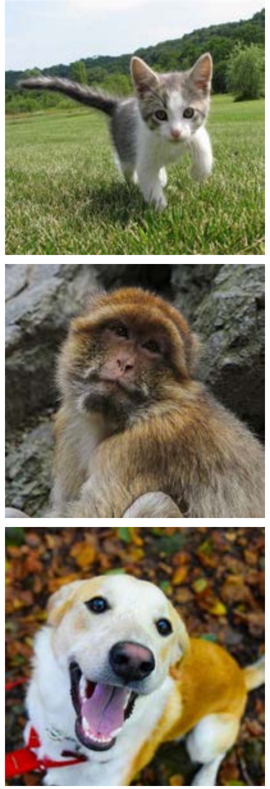
$$L_D(x1, x2) = \max(0, m - d^2)$$

Push features apart
(upto margin m)

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation

Batch of N images



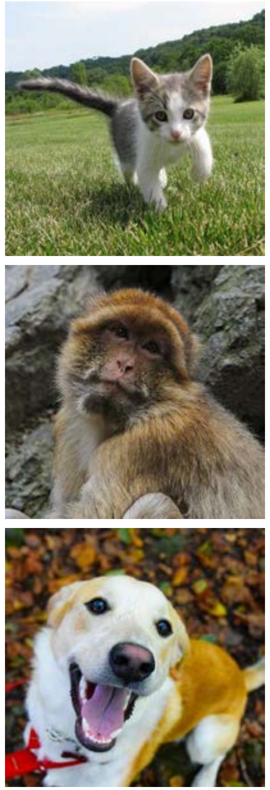
Hadsell et al, “Dimensionality Reduction by Learning and Invariant Mapping”, CVPR 2006
Wu et al, “Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination”, CVPR 2018
Van den Oord et al, “Representation Learning with Contrastive Predictive Coding”, NeurIPS 2018
Hjelm et al, “Learning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

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation

Batch of N images

Two augmentations for each image



$x_{!}$

$x_{"}$

$x_{\#}$

$x_{\$}$

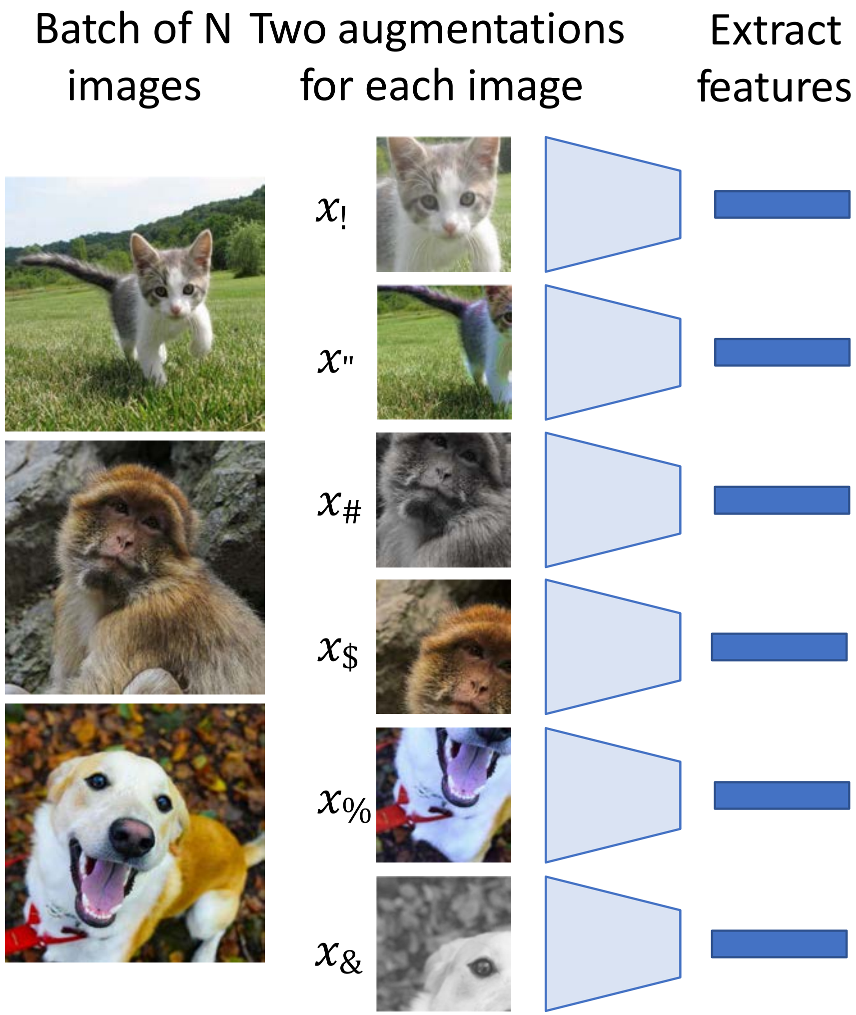
$x_{\%}$

$x_{\&}$

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

[slide credit: Justin Johnson]

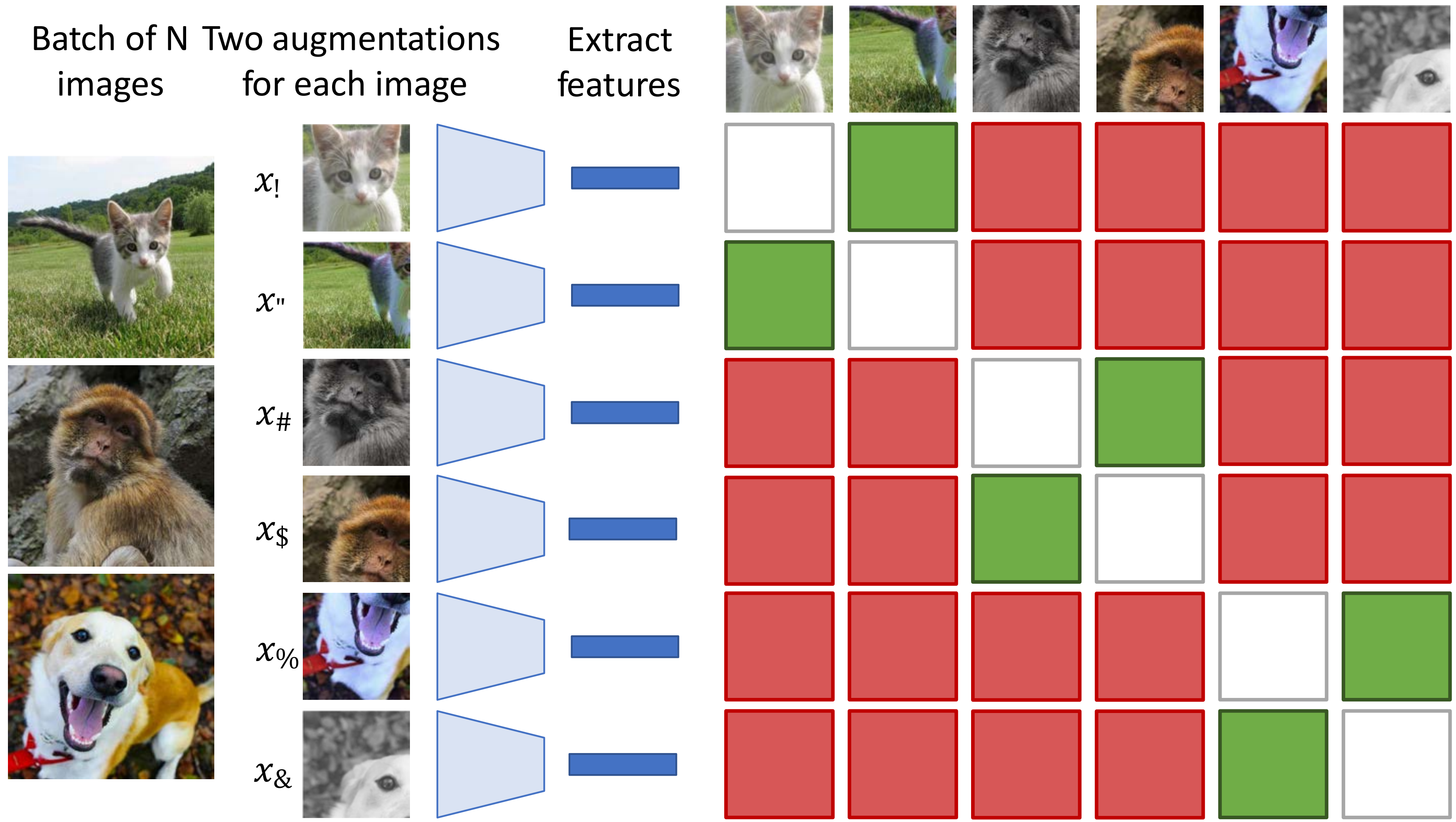
Contrastive Learning with Data Augmentation



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

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation

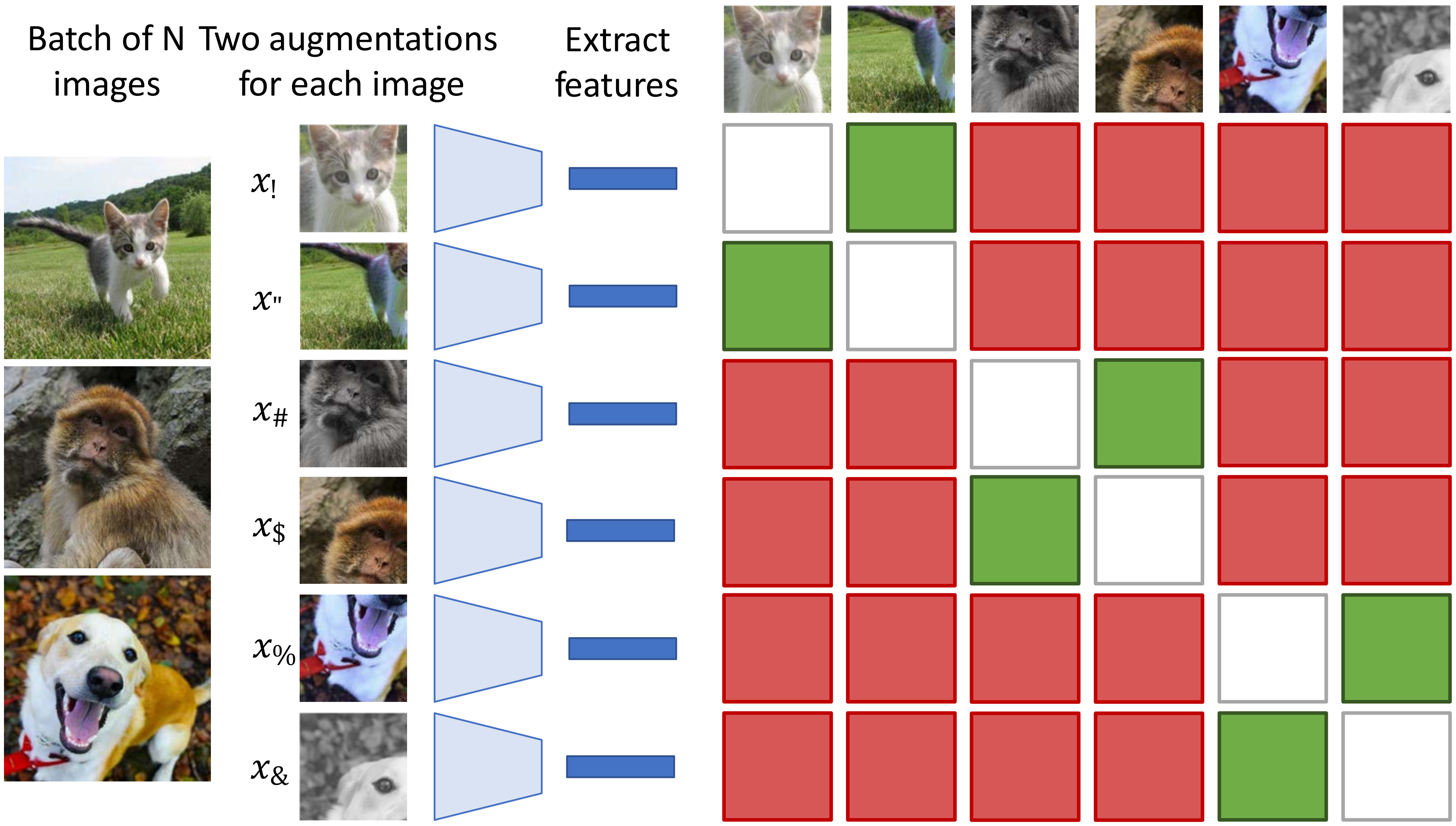


Each image tries to predict which o
the *other* 2N-1 images came from
the same original image

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

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation



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

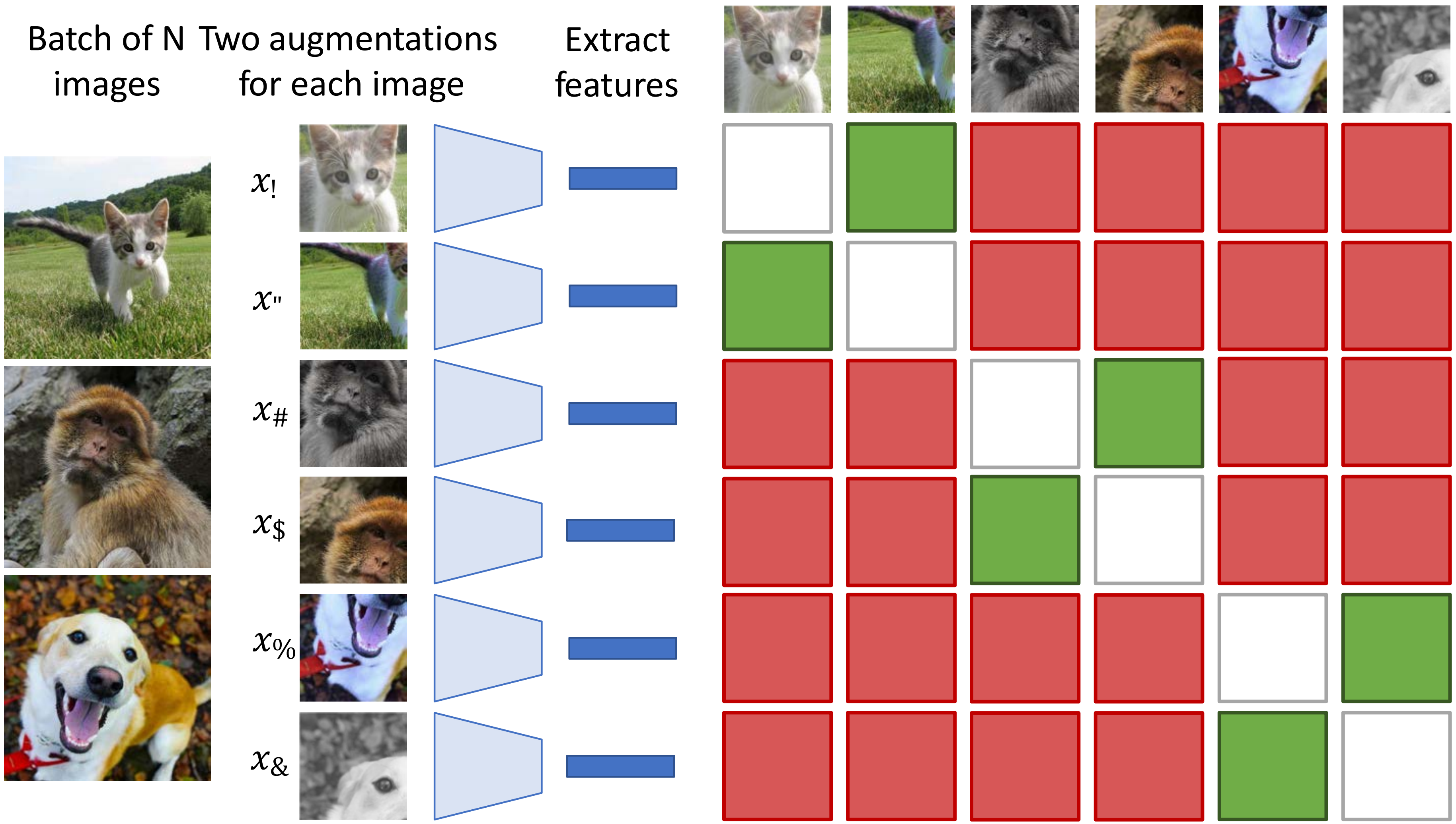
Hadsell et al, “Dimensionality Reduction by Learning and Invariant Mapping”, CVPR 2006
Wu et al, “Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination”, CVPR 2018
Van den Oord et al, “Representation Learning with Contrastive Predictive Coding”, NeurIPS 2018

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

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation



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:

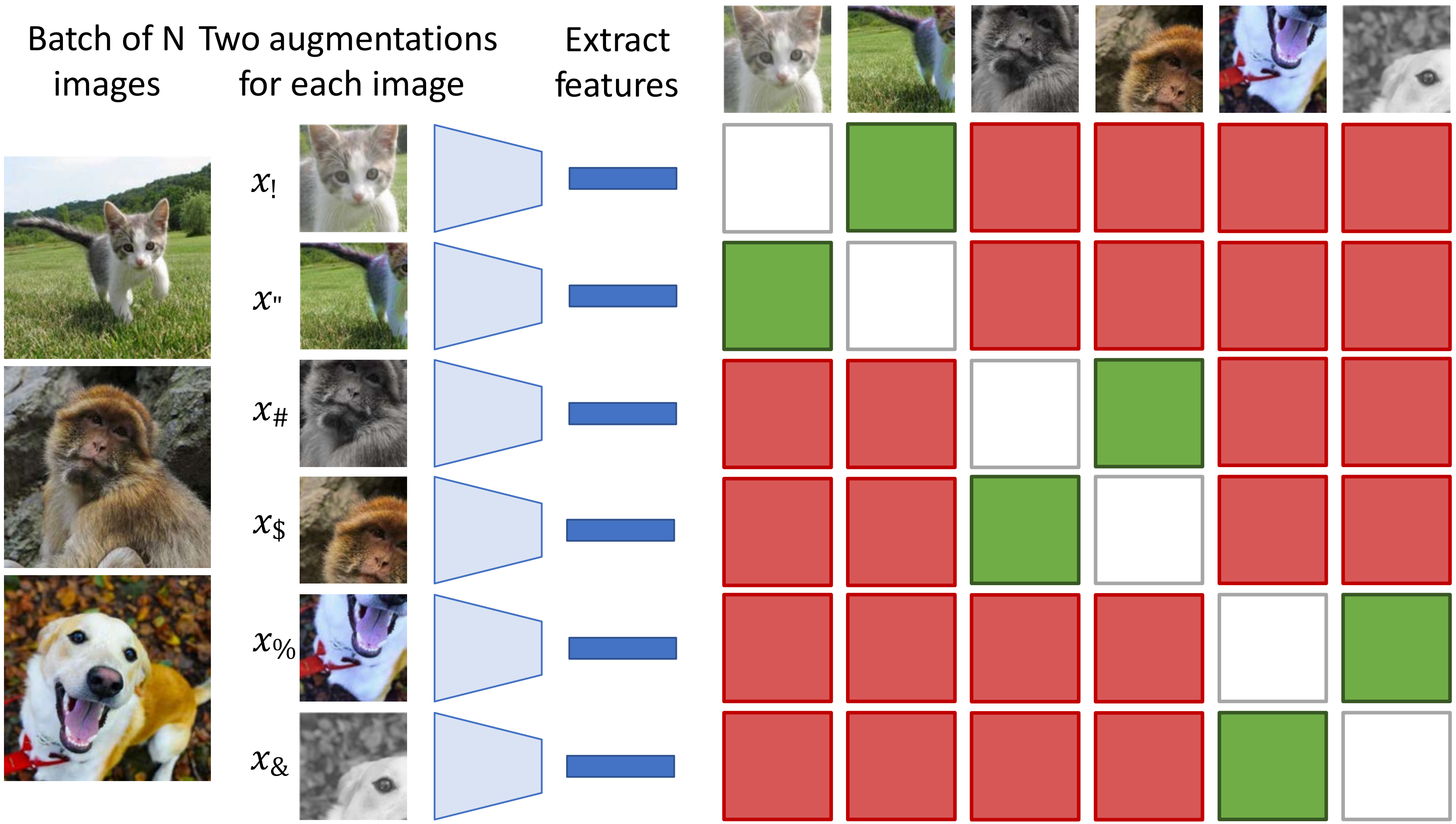
$$L_i = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1, k \neq i}^{2N} \exp(s_{i,k}/\tau)}$$

(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

[slide credit: Justin Johnson]

Contrastive Learning with Data Augmentation



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:

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

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

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

(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

[slide credit: Justin Johnson]