Transfer Learning: Self-Supervised Learning

Danna Gurari

University of Colorado Boulder

Fall 2022

Review

- Last lecture topic:
  - Visual dialog applications
  - Visual dialog dataset
  - Visual dialog evaluation
  - Mainstream 2017 challenges: baseline approaches
  - LTMI: Transformer approach
  - Latex tutorial
- Assignments (Canvas)
  - Lab assignment 4 due earlier today
  - Final project proposal due in 1.5 weeks
- Questions?

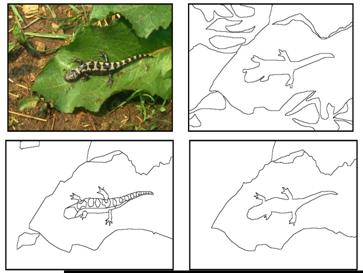
Today’s Topics

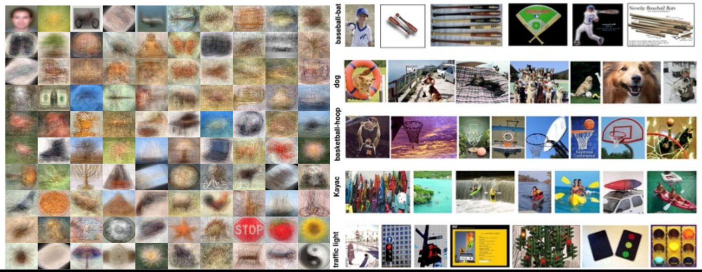
- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

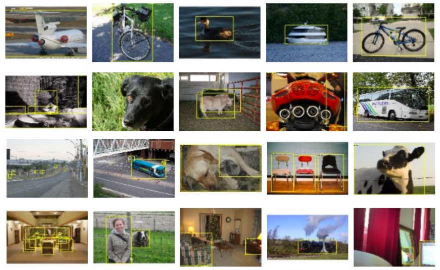
Today’s Topics

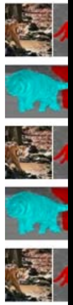
- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

Goal: Avoid Always Relying on Large Labeled Datasets

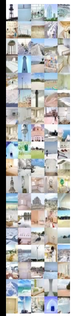


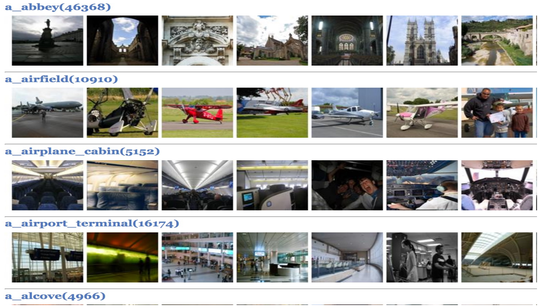






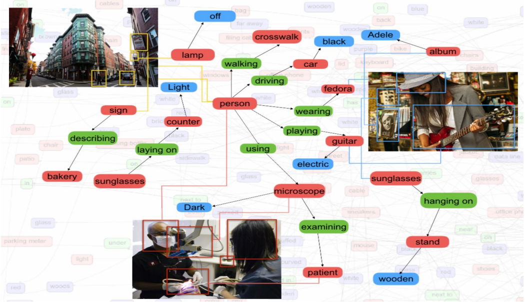
- Expensive
- Relatively Slow to Build Dataset





Places (2014)

MS COCO (2014)

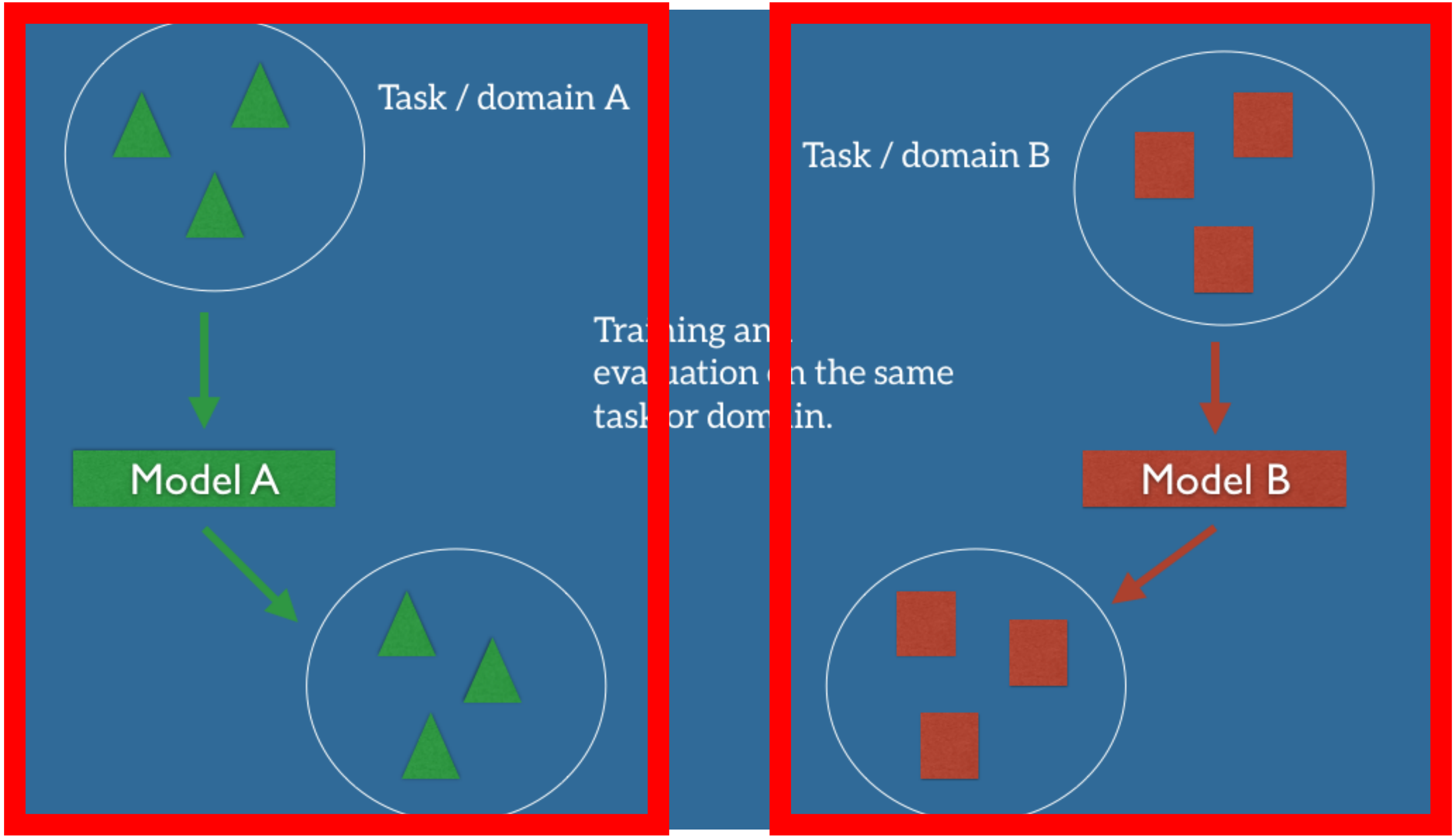


Visual Genome (2016)

Slide Credit: http://vision.cs.utexas.edu/slides/mit-ibm-august2018.pdf

Rather than Learn Solution from Scratch For Each Task/Domain Pair… (Problem for B)

Sufficient Labeled Data

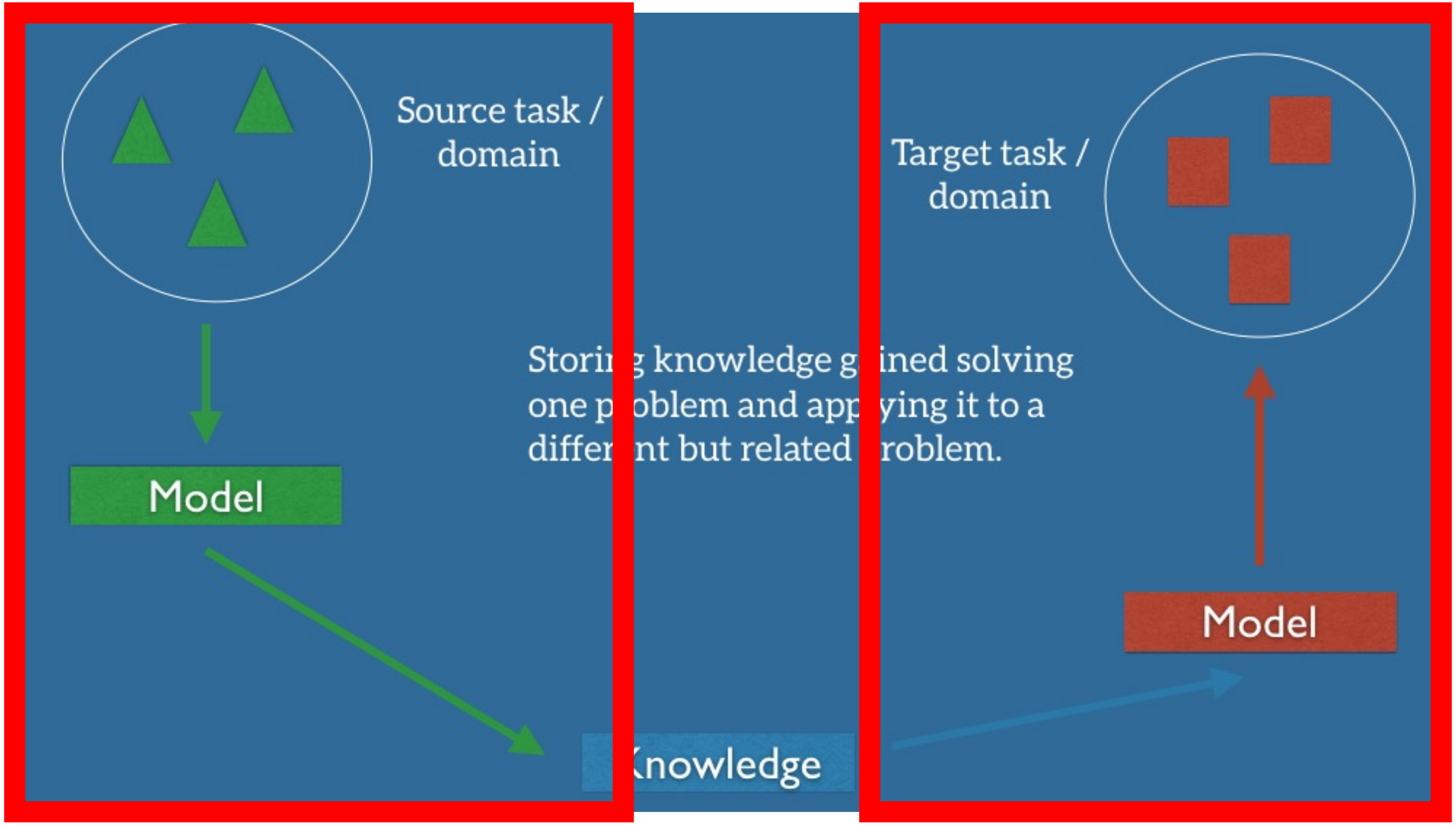


Insufficient Labeled Data

https://ruder.io/transfer-learning/

Idea: Improve the Learning for Conditions Not Observed During Training

Sufficient Labeled Data



Insufficient Labeled Data

https://ruder.io/transfer-learning/

Transfer Learning When Data Sampling Changes (e.g., Sentiment Classification)

News (formal and lengthy)

Tweets (informal and brief)

https://ruder.io/transfer-learning/

Transfer Learning When Feature Space Changes (e.g., Sentiment Classification in Different Language)

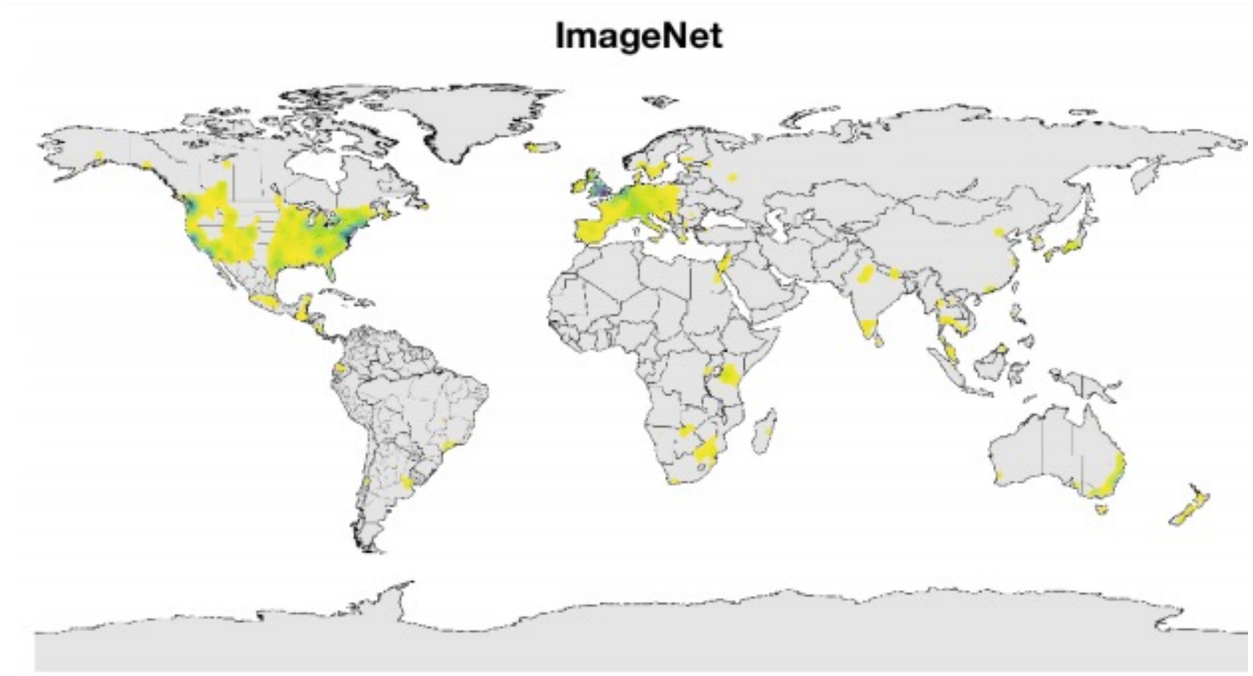
Cool charger
By Tiffany on March 30, 2015
Verified Purchase
Bought this for my Galaxy phone and I have to say, this is a pretty cool USB cord! :) I like the lights in the cord as it puts off a cool glowing effect in my room at night and it makes it much easier to see, thanks for the great product!

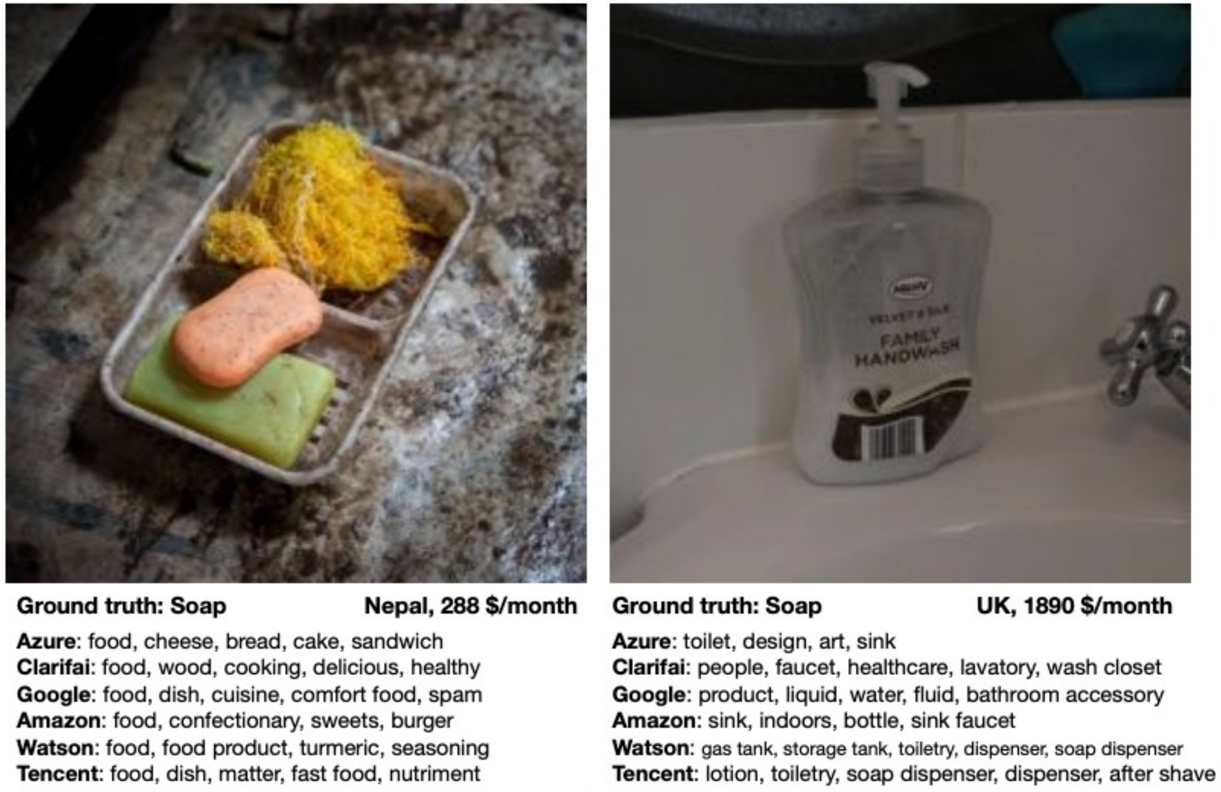
Definitely buying more.
By Krystal Willingham on March 28, 2015
Verified Purchase
I was impressed with how bright the lights on the cable are. It works amazing and as described. i received earlier than expected so that made me very happy. So far is working like a charm and I can't wait to buy a few more.

Spot It In the Crowd
By Heather-Joan Carls on March 29, 2015
Verified Purchase
Such a cool product. I was so happy with how bright the lights on the cable are. It shipped super fast. The light shuts off when the charging is complete, so that's super helpful. I don't have to keep checking.

https://www.nytimes.com/wirecutter/blog/lets-talk-about-amazon-reviews/

Transfer Learning When Target Categories Change (e.g., Items in Low Income Household vs ImageNet)

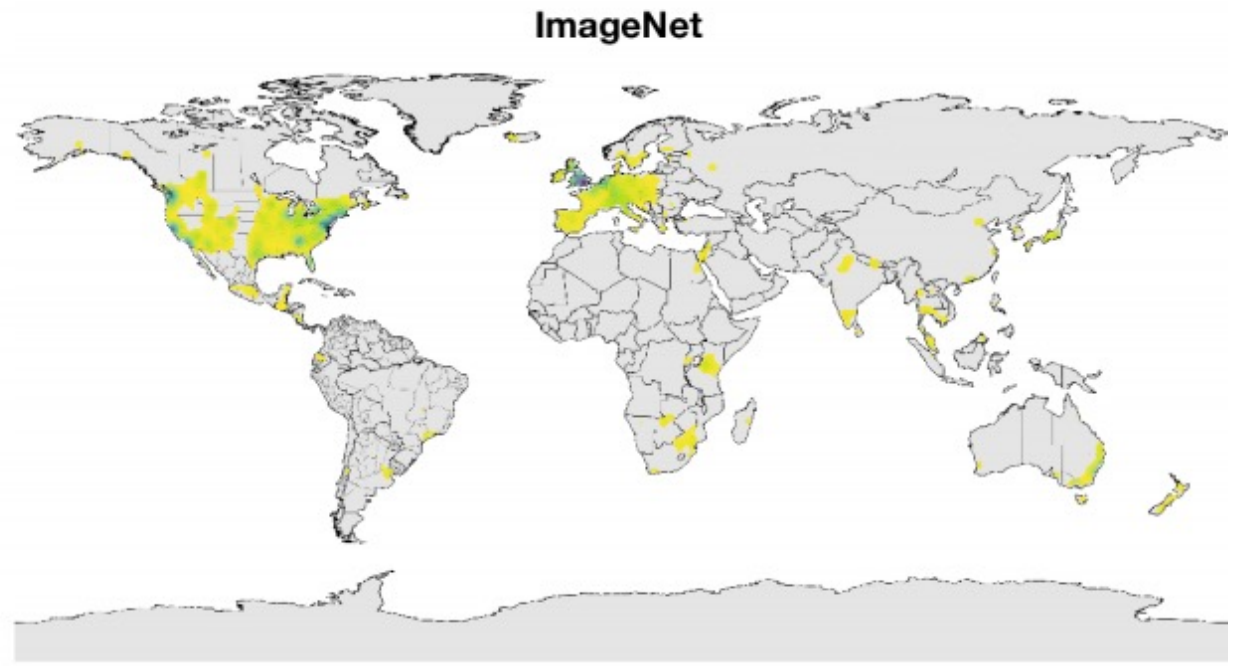


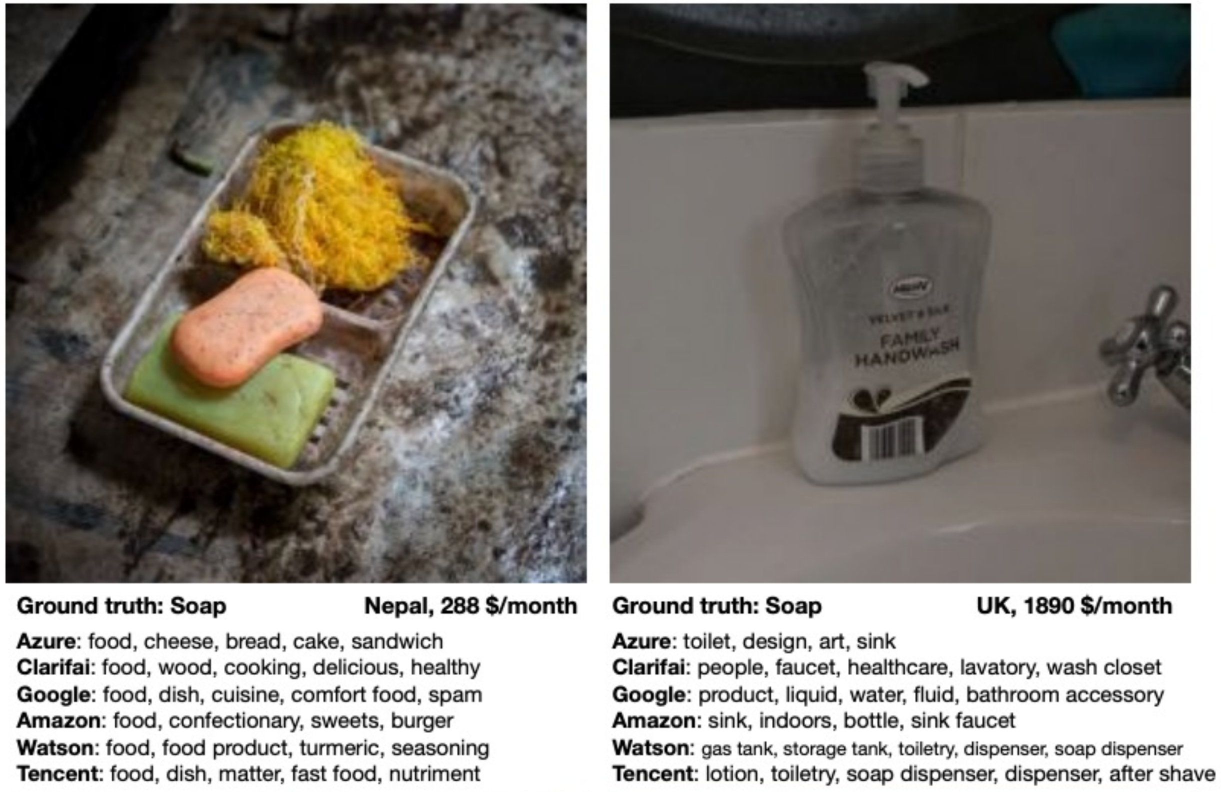


Zhao et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. 2017.

DeVries et al. Does object recognition work for everyone? CVPR workshops, 2019.

Transfer Learning When Limited Data Available (e.g., Items in Low Income Household vs ImageNet)



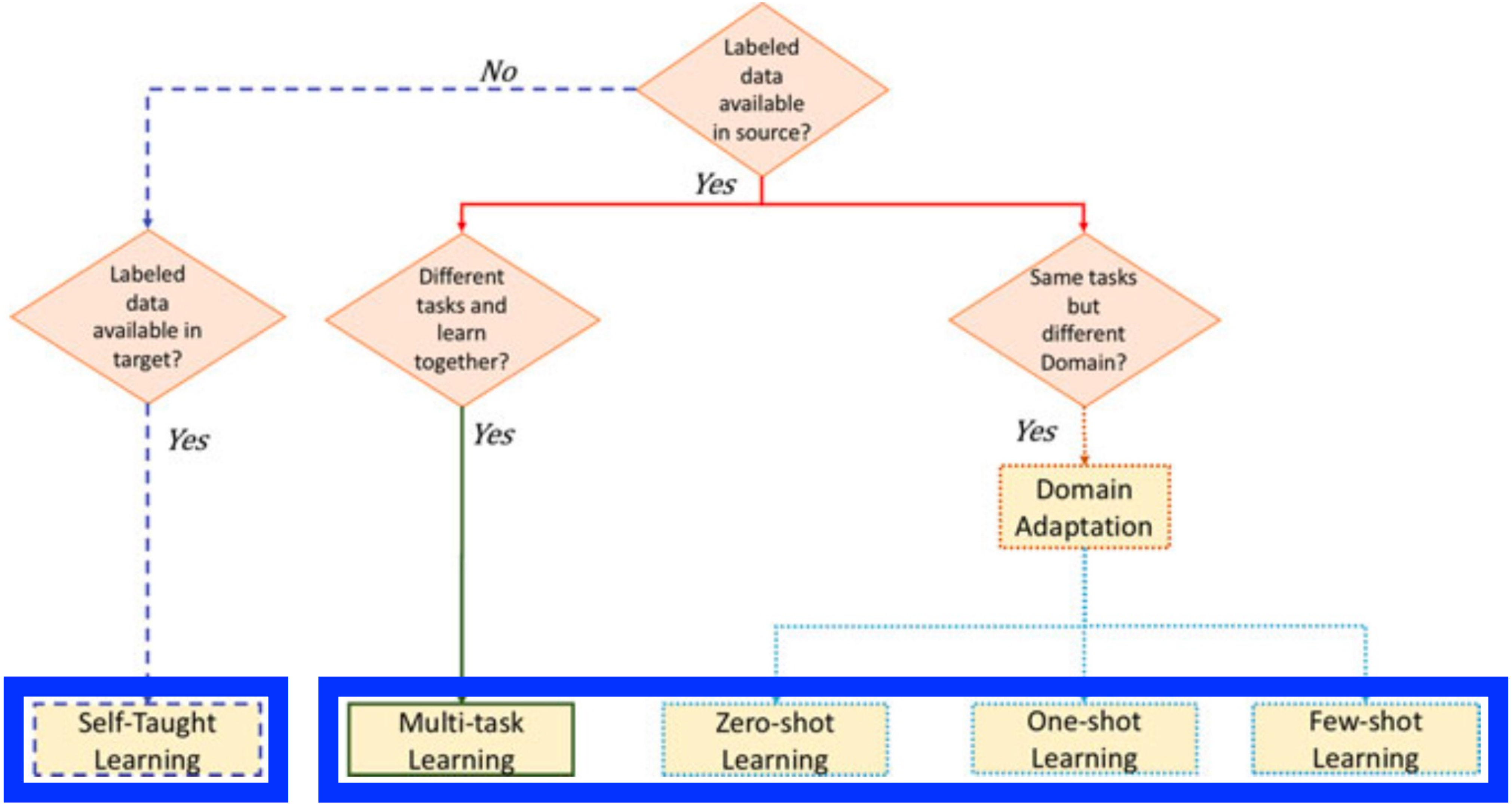


Zhao et al. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. 2017.

DeVries et al. Does object recognition work for everyone? CVPR workshops, 2019.

Transfer Learning Approaches

Today's focus



Next lecture

Kamath et al. Deep Learning for NLP and Speech Recognition. 2019.

Transfer Learning: Key Challenges

- What to transfer? i.e., what knowledge generalizes
- How to transfer?
- When to transfer? i.e., transferring knowledge can harm performance

Pan and Yang. A Survey on Transfer Learning. 2010.

Today’s Topics

- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

Goal: Create Generalizable Features

Key observation: features from a pretrained network can be useful for other datasets/tasks

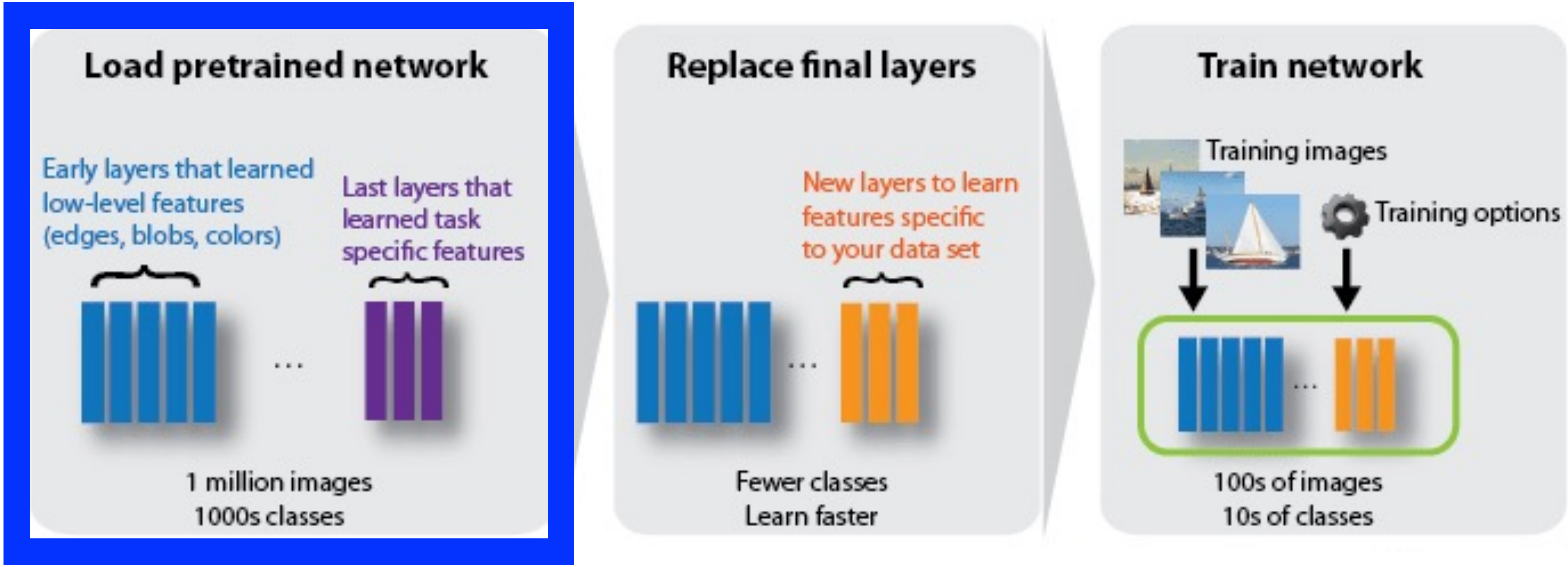


Image Source: https://www.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html

Intuition: How Do Humans Learn?

With Supervision

Learn from instruction

Unsupervised

Learn from experience

Today's scope

Self-Supervised Learning: Data Gives Supervision

- Relatively Cheap
- Can Collect Data Fast

https://lovevery.com/community/blog/child-development/the-surprising-learning-power-of-a-household-mirror/

https://www.rockettes.com/blog/how-to-use-the-mirror-in-dance-class/

Self-Supervised Learning: Data Gives Supervision

Approach: create features that are useful for other datasets/tasks

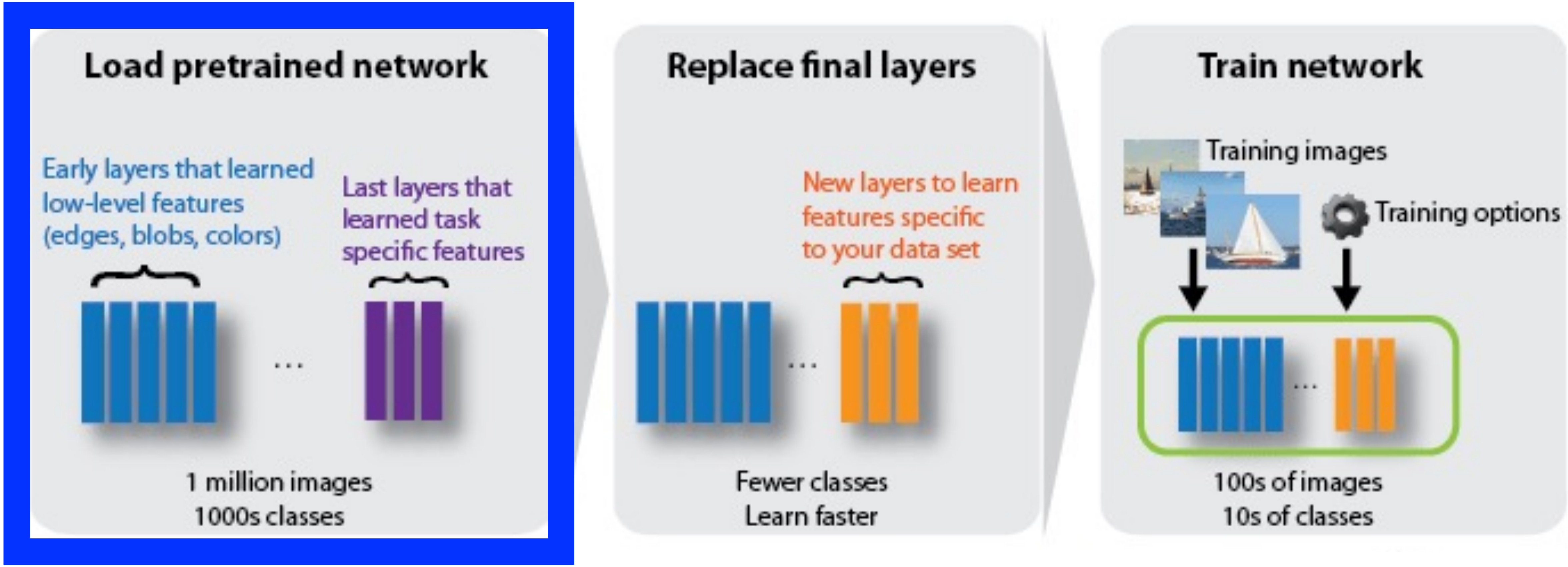
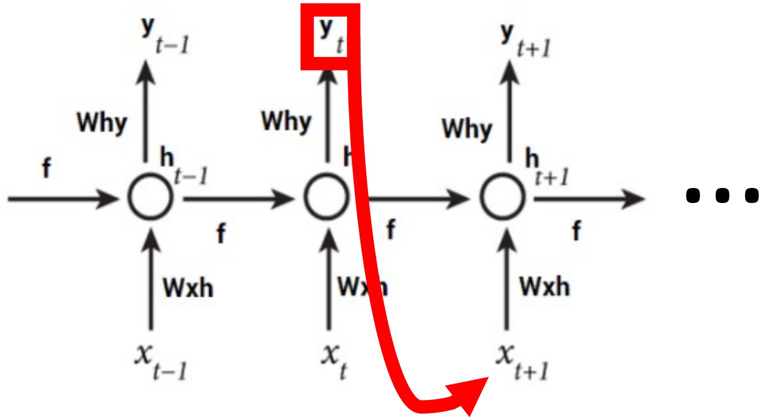


Image Source: https://www.mathworks.com/help/deeplearning/ug/transfer-learning-using-alexnet.html

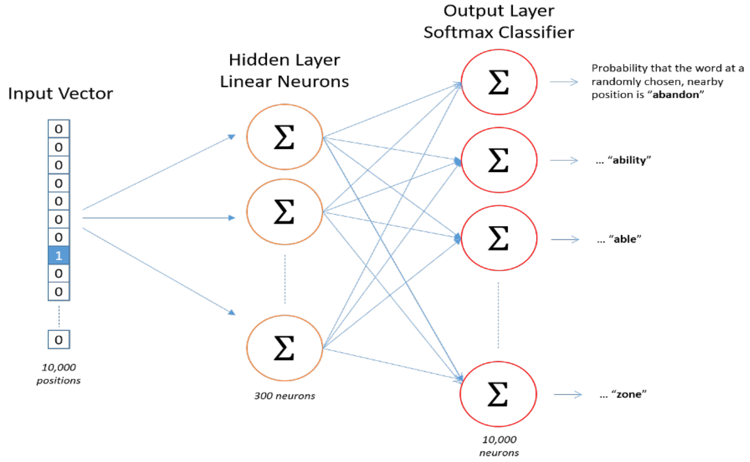
Self-Supervised Learning Methods Already Covered in This Course (Many NLP Methods)

Character prediction with RNNs



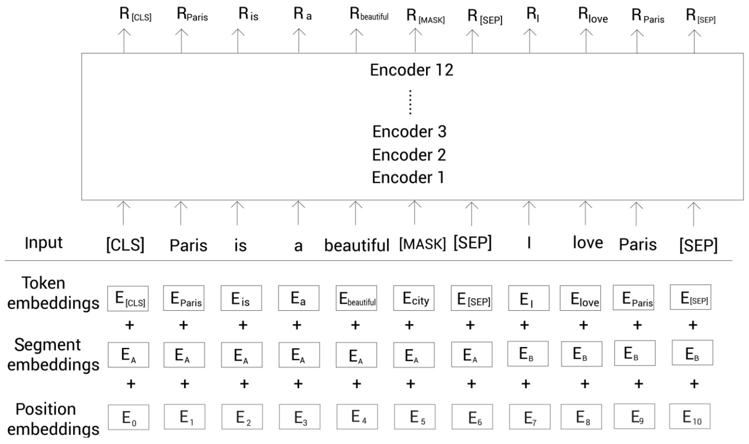
https://www.analyticsvidhya.com/blog/2017/12/introduction-to-recurrent-neural-networks/

Word embeddings (e.g., word2vec; predict nearby word for given word)



https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

Transformers (e.g., BERT and LXMERT with masking)



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Next: additional self-supervised learning methods explored in computer vision

Today's Topics

- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

Generative-based Methods

- Autoencoder: predict self
- Colorization: convert grayscale to color
- Video prediction: predict future frames

Generative-based Methods

- Autoencoder: predict self
- Colorization: convert grayscale to color
- Video prediction: predict future frames

Image Autoencoder Architecture

- Learn to copy the input to the output

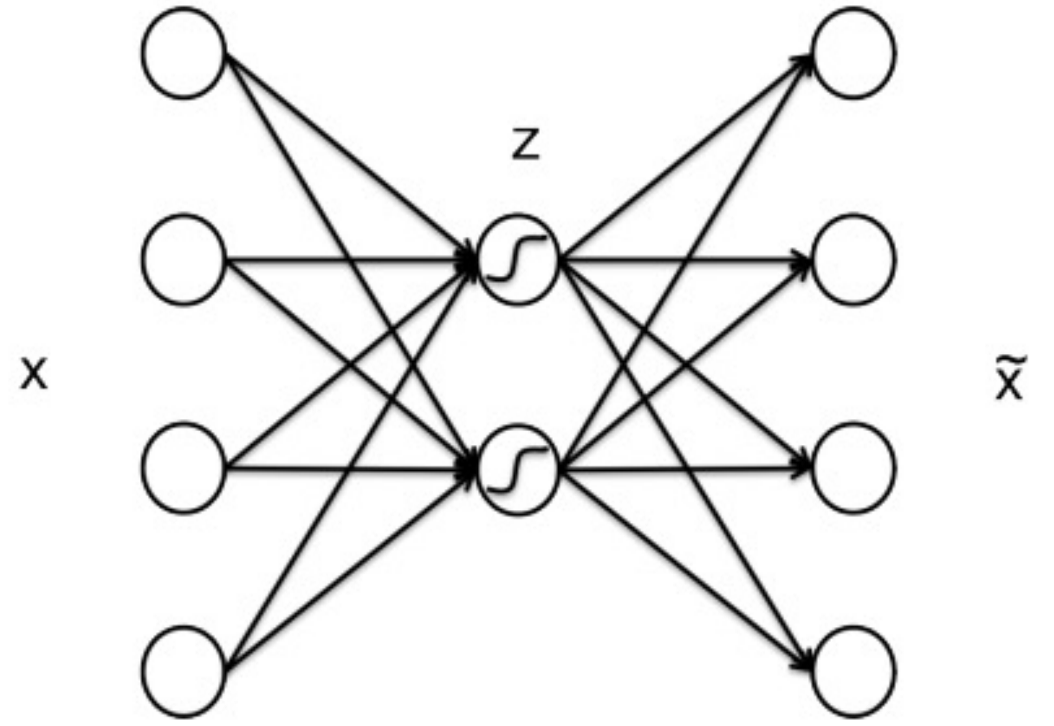


Figure Credit: https://lazyprogrammer.me/a-tutorial-on-autoencoders/

Image Autoencoder Architecture

- Consists of two parts:
  - **Encoder**: compresses inputs to an internal representation
  - **Decoder**: tries to reconstruct the input from the internal representation

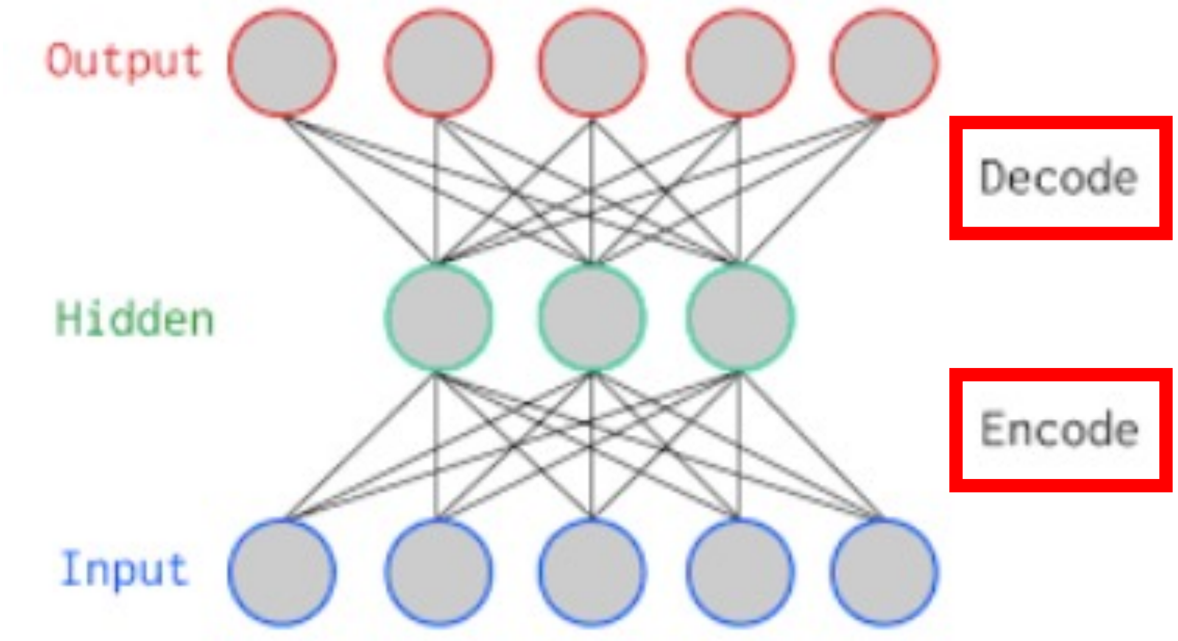
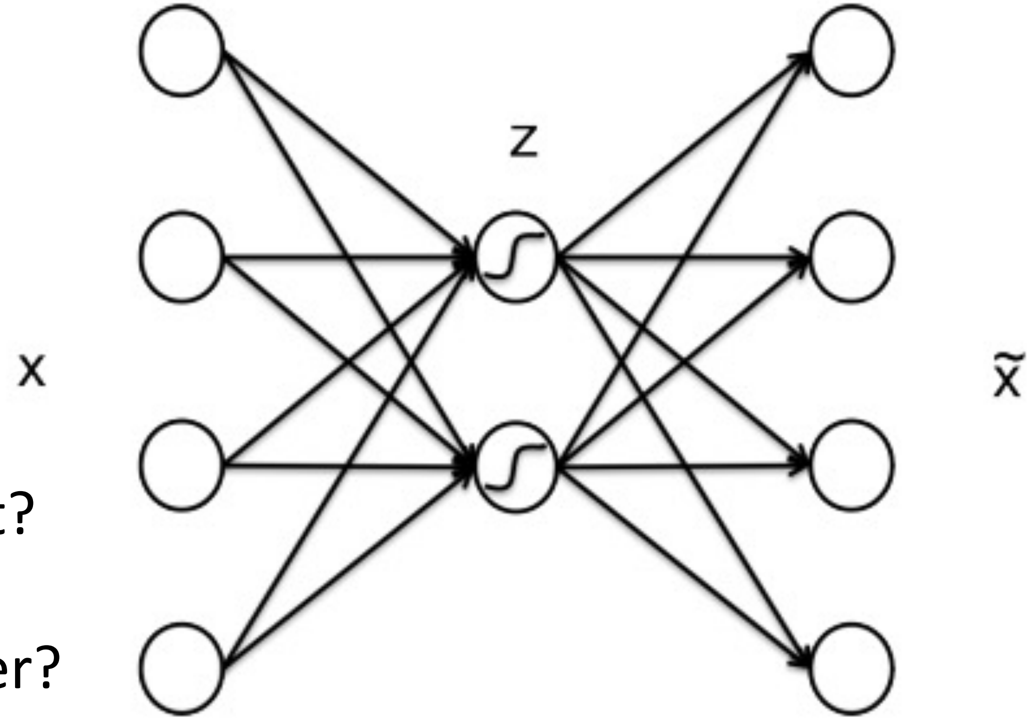


Figure Credit: https://www.datacamp.com/community/tutorials/autoencoder-keras-tutorial

Image Autoencoder Architecture

- Given this input 620 x 426 image (264,120 pixels):



  - What would a perfect autoencoder predict?
    - Itself
  - What number of nodes are in the final layer?
    - 264,120

Figure Credit: https://lazyprogrammer.me/a-tutorial-on-autoencoders/

Image Autoencoders

- Intuition: which number sequence is easier to remember?
  - **A:** 30, 27, 22, 11, 6, 8, 7, 2
  - **B:** 30, 15, 46, 23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5
- **B:** need learn only two rules
  - If even, divide by 2
  - If odd, multiply by 3 and add 1

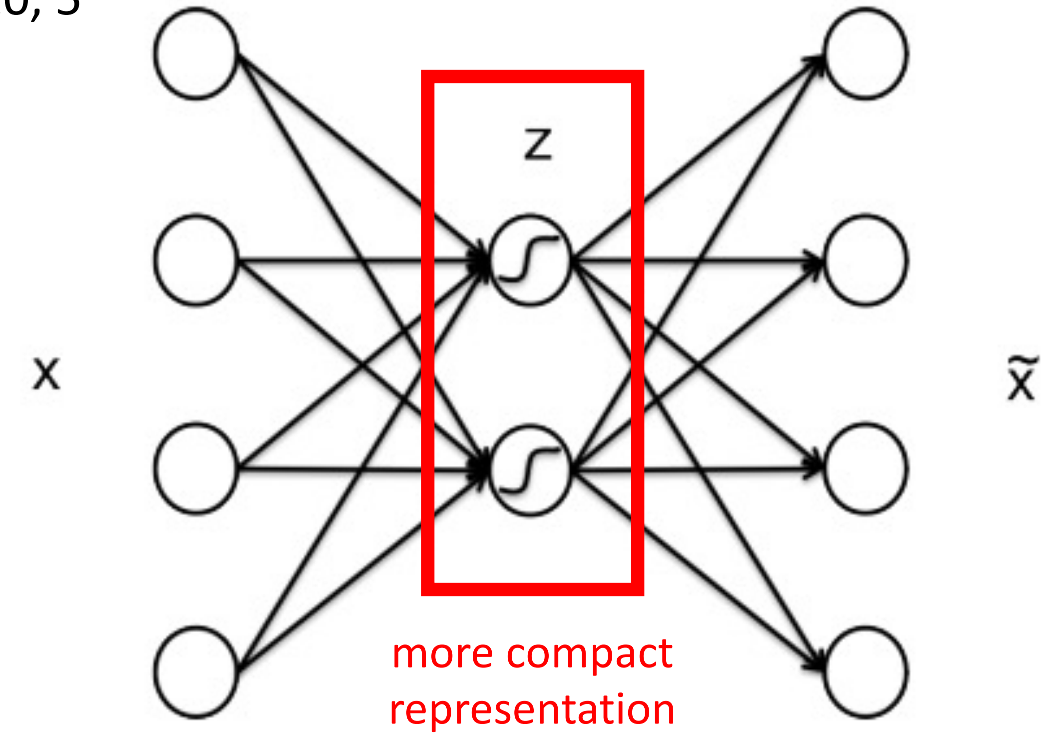


Figure Credit: https://lazyprogrammer.me/a-tutorial-on-autoencoders/

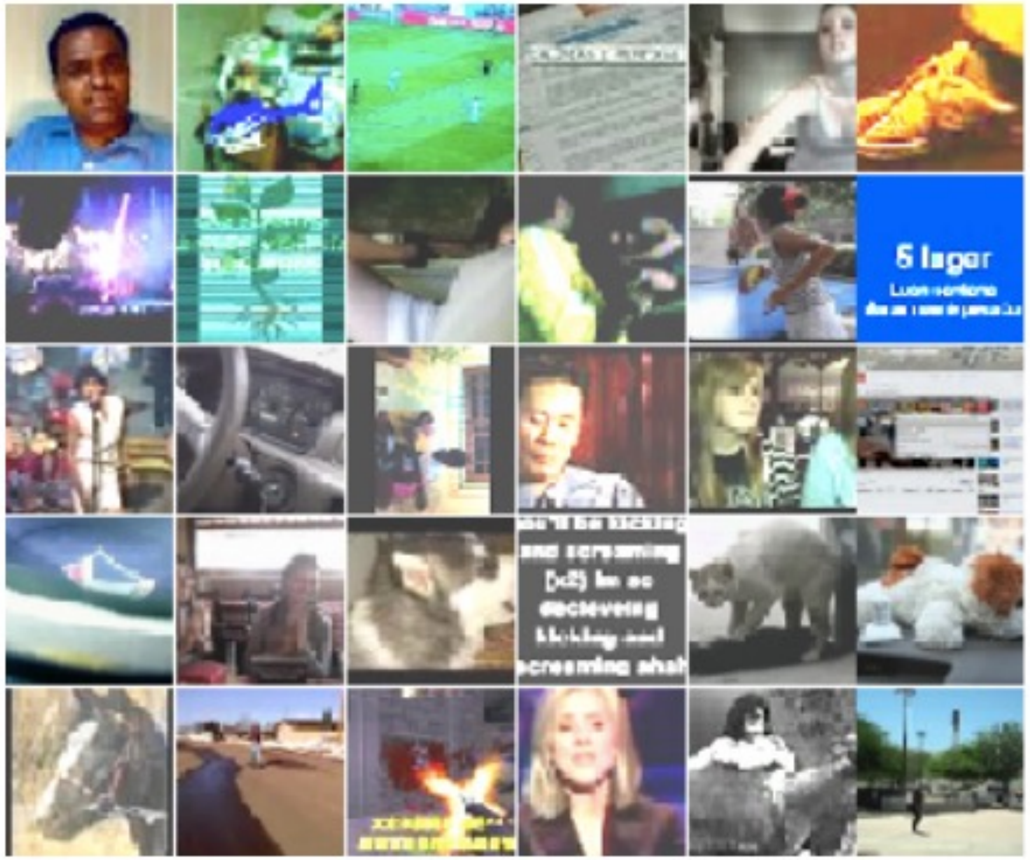
Image Autoencoder Training

Repeat until stopping criterion met:

1. **Forward pass**: propagate training data through network to make prediction
2. **Backward pass**: using predicted output, calculate error gradients backward
3. Update each weight using calculated gradients

Image Autoencoder Features

- e.g., training data:
  - 1 image taken from 10 million YouTube videos
  - Each image is in color and 200x200 pixels

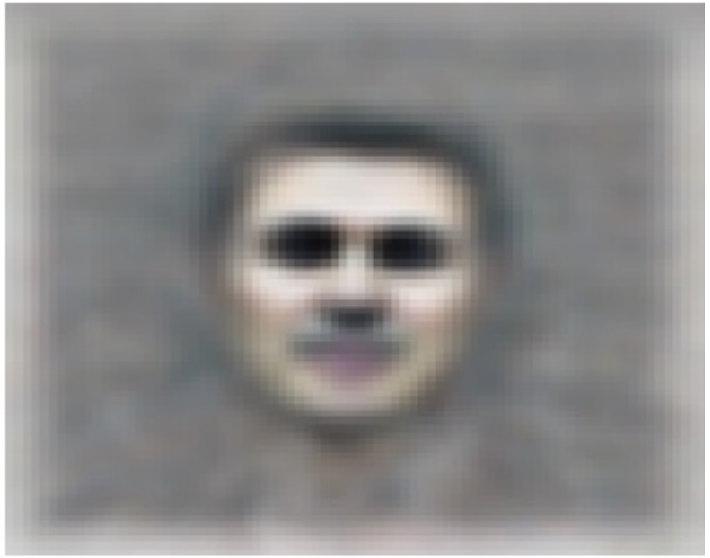


- What features do you think it learned?

Quoc V. Le et al., Building High-level Features Using Large Scale Unsupervised Learning; ICML 2013.

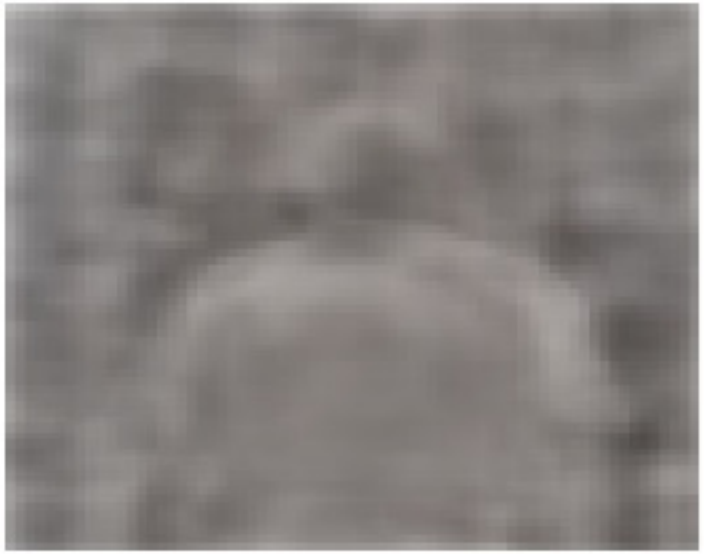
Image Autoencoder Features

- e.g., features learned include:



human face

cat face

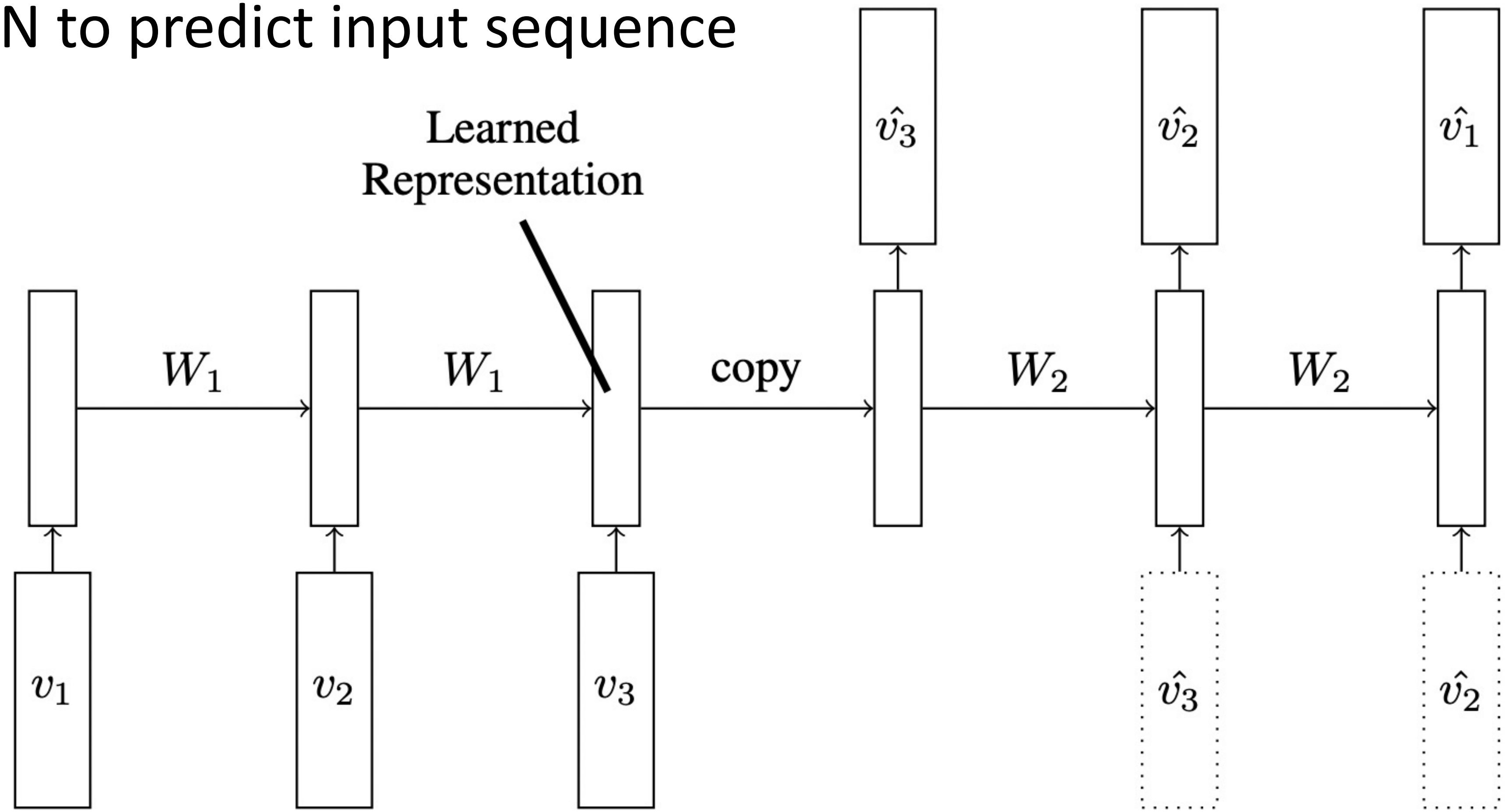


human body

Quoc V. Le et al., Building High-level Features Using Large Scale Unsupervised Learning; ICML 2013.

Video Autoencoder

- Train RNN to predict input sequence

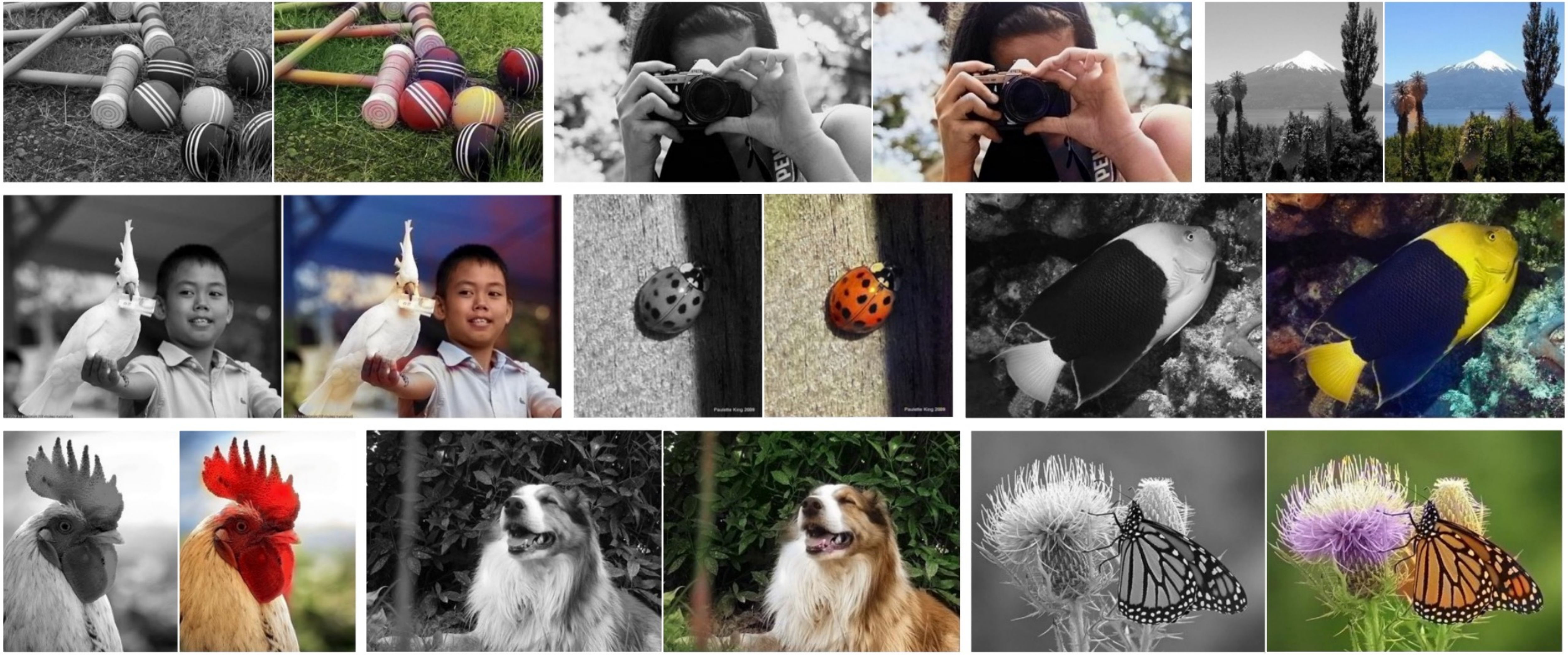


Srivastava et al., Unsupervised Learning of Video Representations using LSTMs; ICML 2015.

Generative-based Methods

- Autoencoder: predict self
- Colorization: convert grayscale to color
- Video prediction: predict future frames

Colorization: *Plausible* Coloring Results

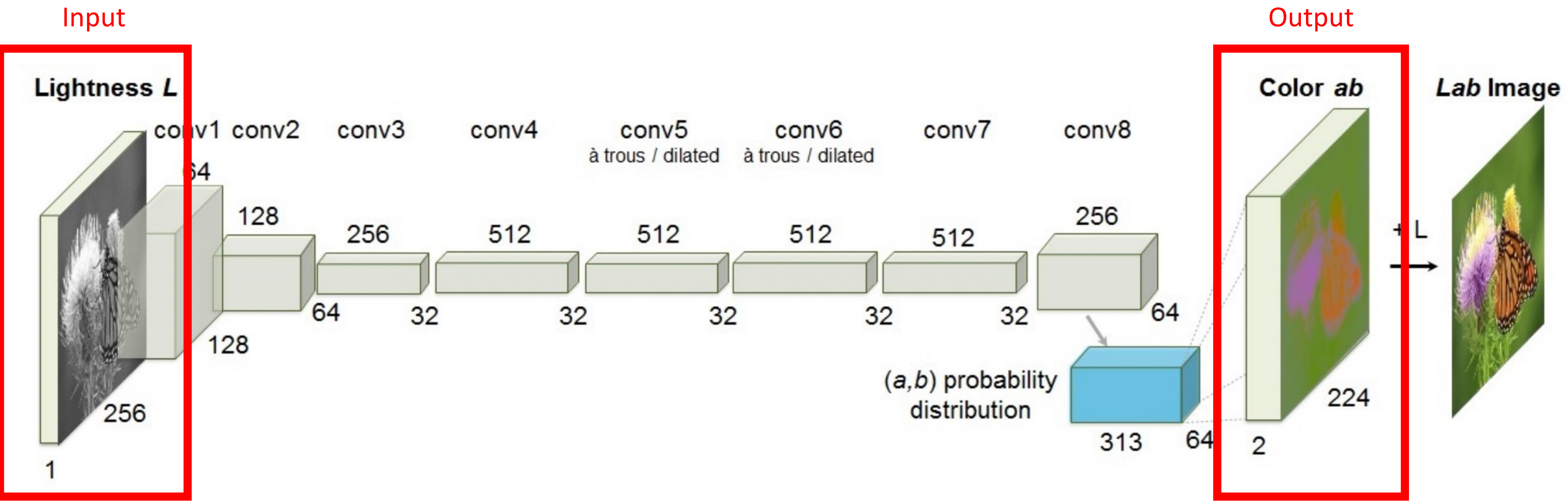


R. Zhang, P. Isoa, and A. A. Efros. Colorful Image Colorization. ECCV 2016.

Colorization: *Plausible* Coloring Results

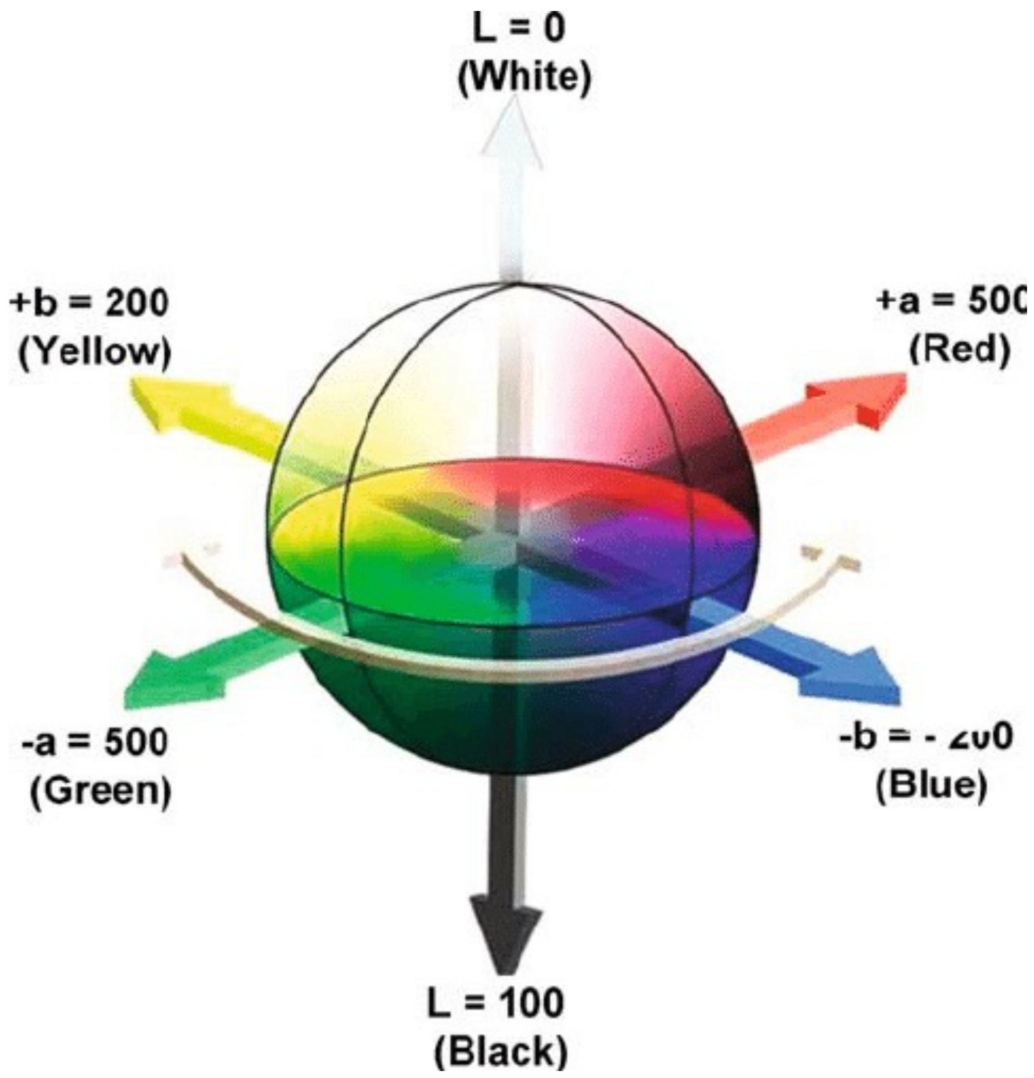
Figure Sources: https://www.flickr.com/photos/applesnpearsau/12197380673/in/photostream/; https://commons.wikimedia.org/wiki/File:JACQUES_VILET_-_1982,_Les_Fruits_du_Jardin.jpg

Image Colorization Architecture



R. Zhang, P. Isoa, and A. A. Efros. Colorful Image Colorization. ECCV 2016.

Image Colorization Architecture: CIE *Lab* Color

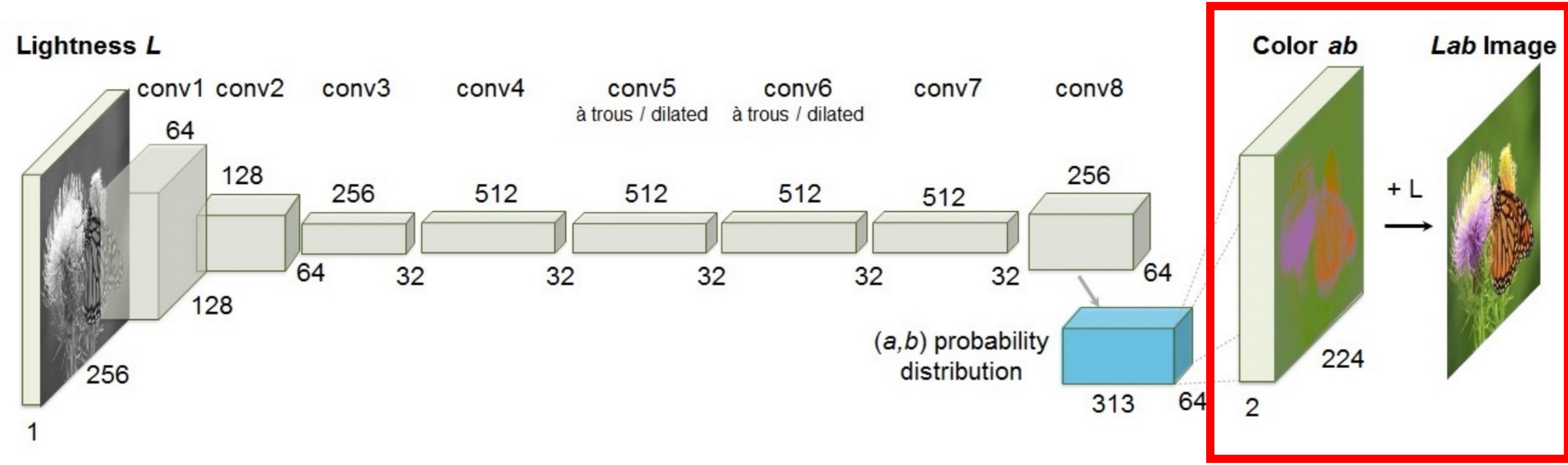


L indicates grayscale information whereas *a* and *b* represent colors

Figure source: https://www.researchgate.net/figure/The-cubical-CIE-Lab-color-space_fig3_23789543

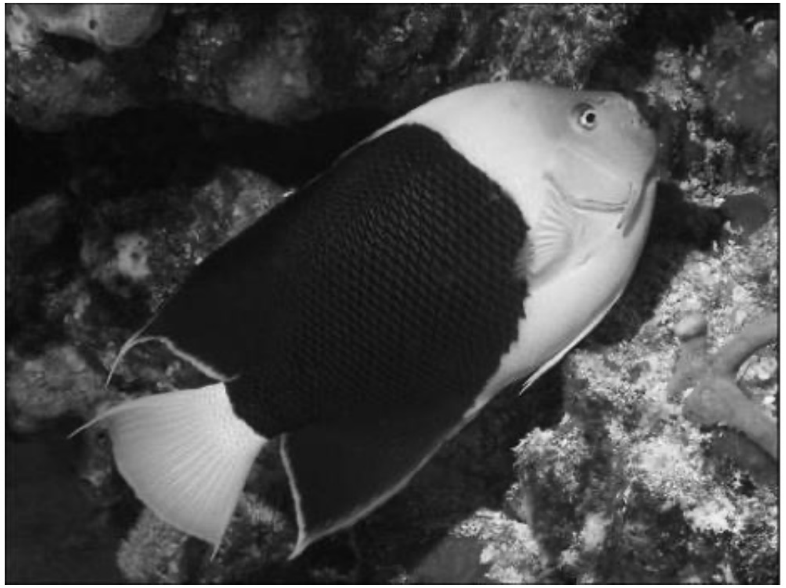
Image Colorization Architecture

Create image by combining predicted *a* and *b* channels with the *L* channel



R. Zhang, P. Isoa, and A. A. Efros. Colorful Image Colorization. ECCV 2016.

Image Colorization Architecture



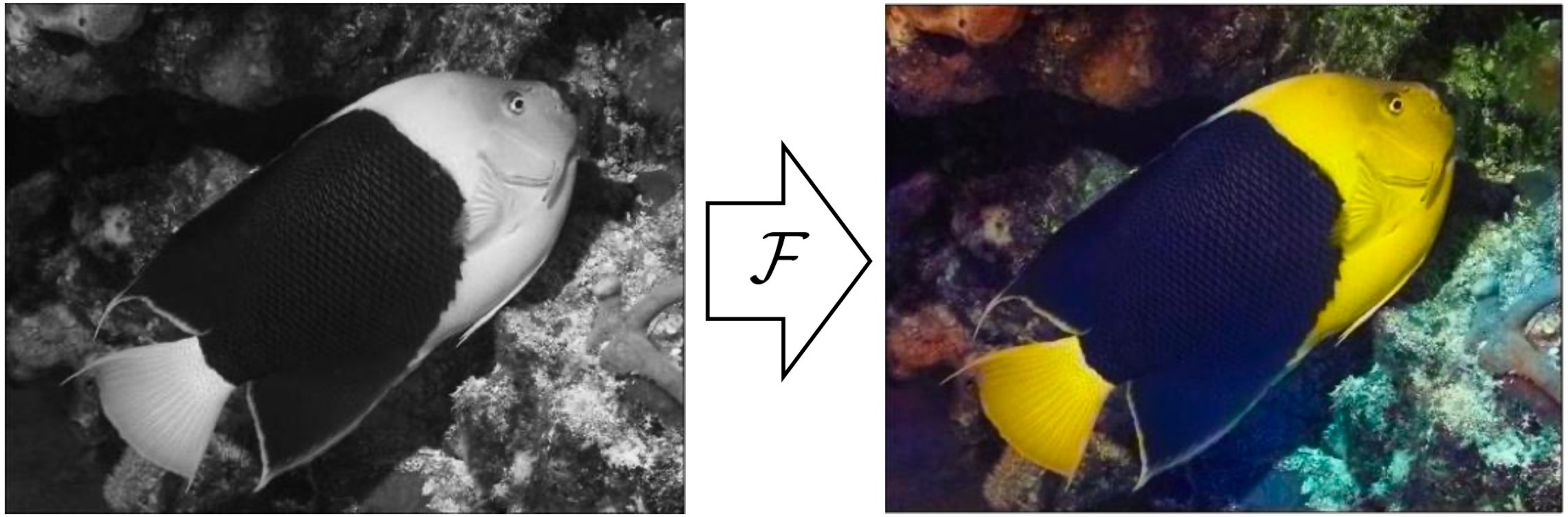
Grayscale image: *L* channel
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: *ab* channels
$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$

$\boxed{L}$ $\boxed{ab}$

Figure source: http://videolectures.net/eccv2016_zhang_image_colorization/

Image Colorization Architecture



Grayscale image: *L* channel

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Concatenate (L,ab)

$(\mathbf{X}, \widehat{\mathbf{Y}})$

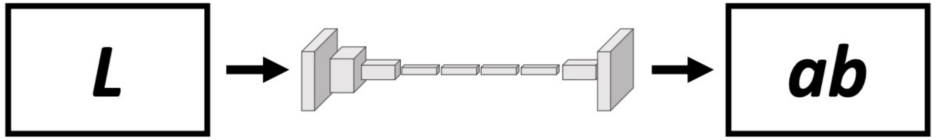


Figure source: http://videolectures.net/eccv2016_zhang_image_colorization/

Image Colorization Training

For 1.3 million ImageNet images, repeat until stopping criterion met:

1. **Forward pass**: propagate training data through network to make prediction
2. **Backward pass**: using predicted output, calculate error gradients backward
3. Update each weight using calculated gradients

R. Zhang, P. Isoa, and A. A. Efros. Colorful Image Colorization. ECCV 2016.

Image Colorization Features

Task requires understanding an image at the pixel and semantic-level

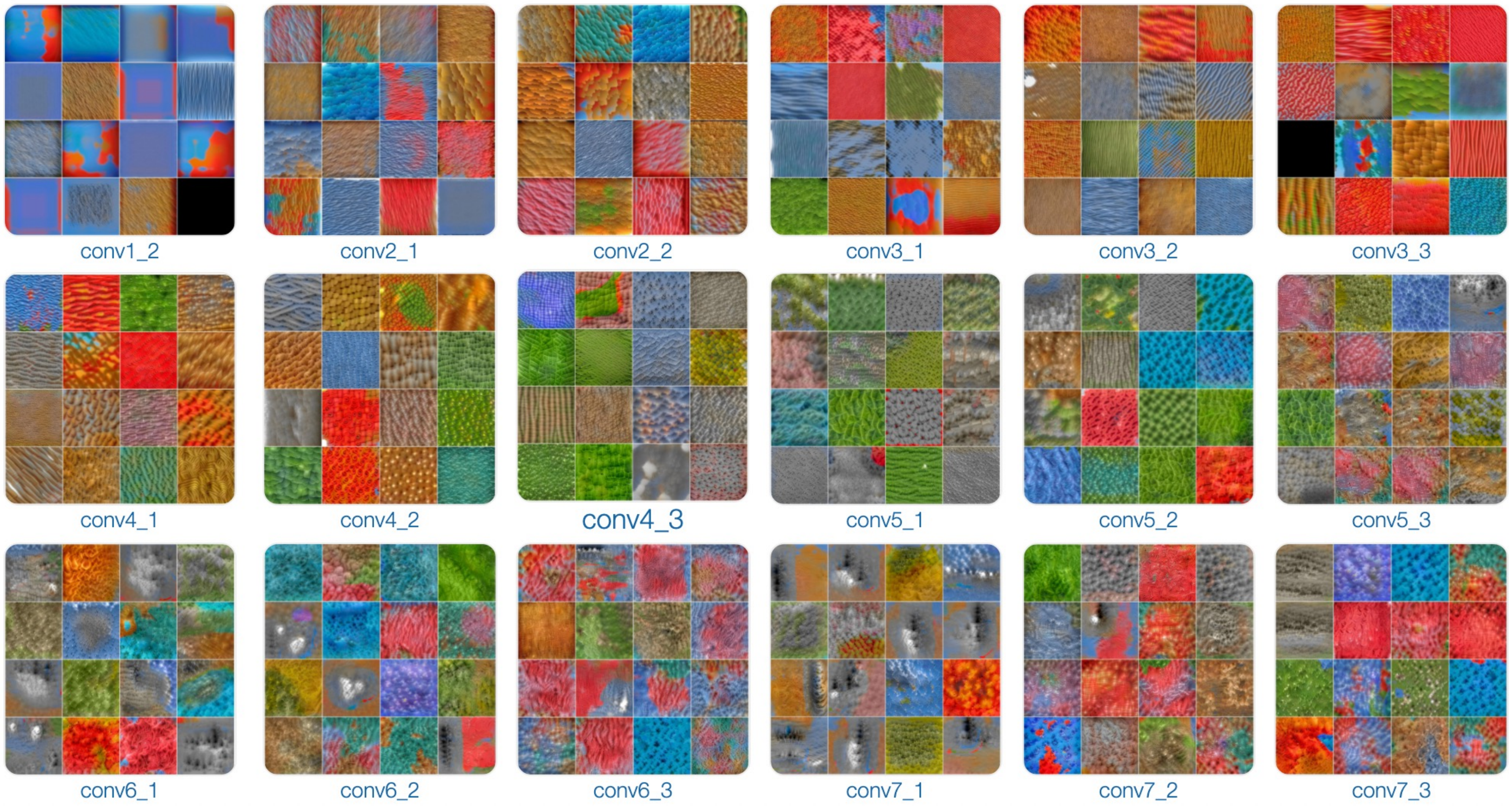


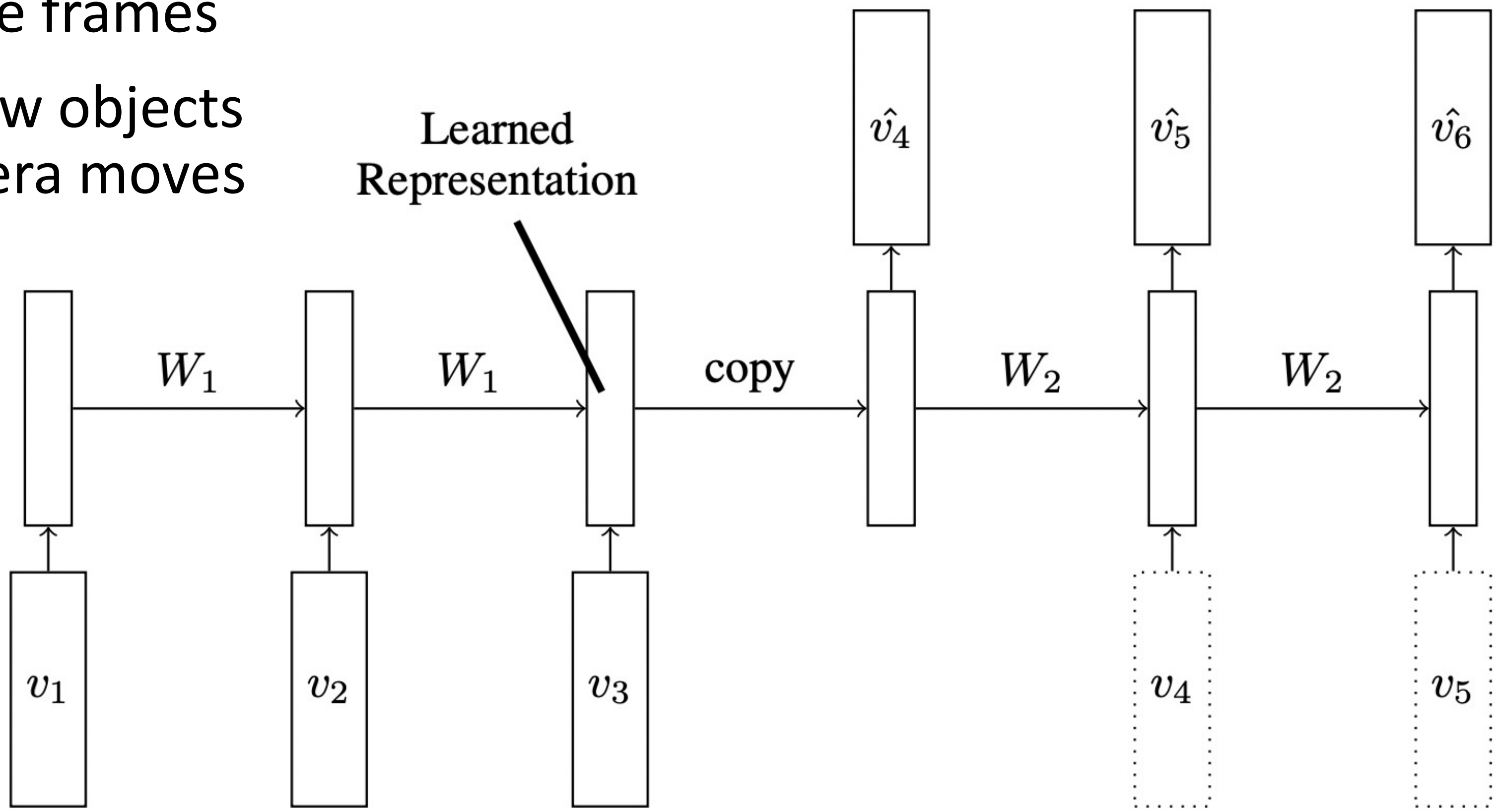
Figure source: http://richzhang.github.io/colorization/

Generative-based Methods

- Autoencoder: predict self
- Colorization: convert grayscale to color
- Video prediction: predict future frames

Video Prediction

- Train RNN to predict future frames
- Limitations: identifying new objects and background as a camera moves



What type of features might be learned?

Srivastava et al., Unsupervised Learning of Video Representations using LSTMs; ICML 2015.

Generative-based Methods

- Autoencoder: predict self
- Colorization: convert grayscale to color
- Video prediction: predict future frames

Today's Topics

- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

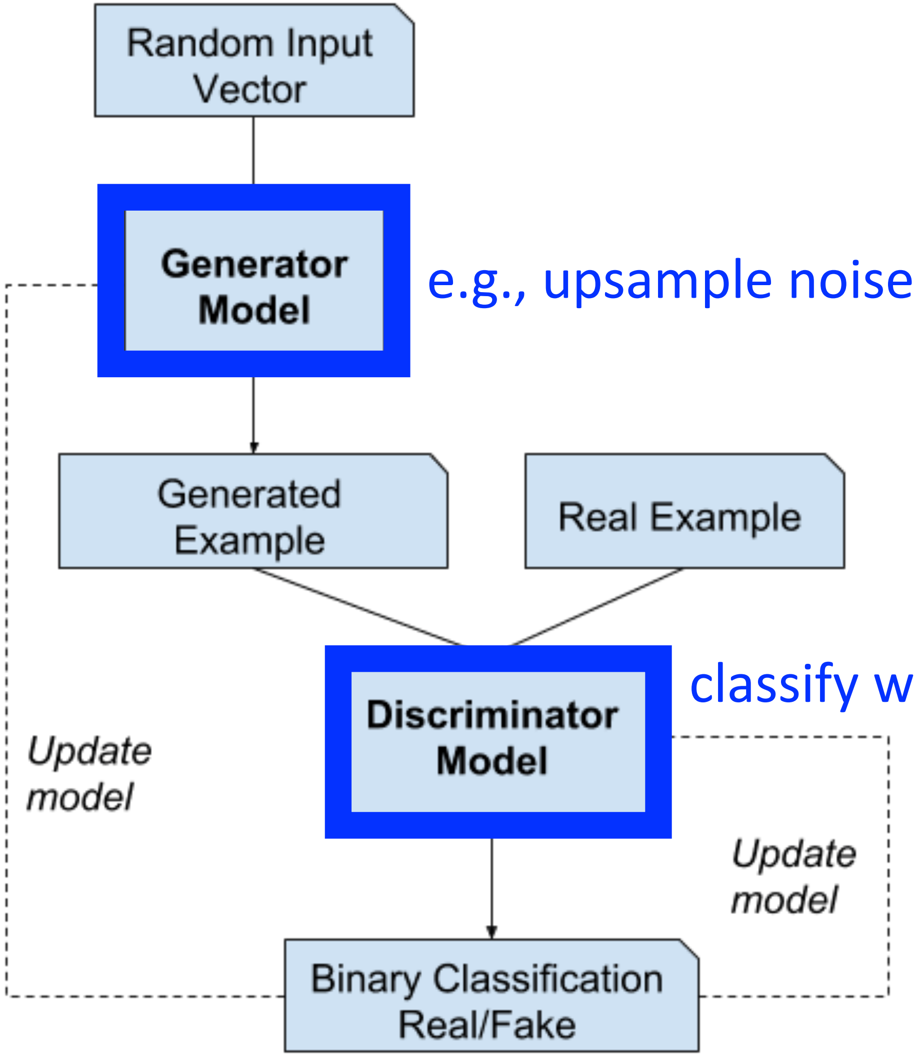
Generative adversarial networks

- Generative adversarial networks (GANs)
- Context encoder

Generative adversarial networks

- Generative adversarial networks (GANs)
- Context encoder

GAN: Basic Architecture



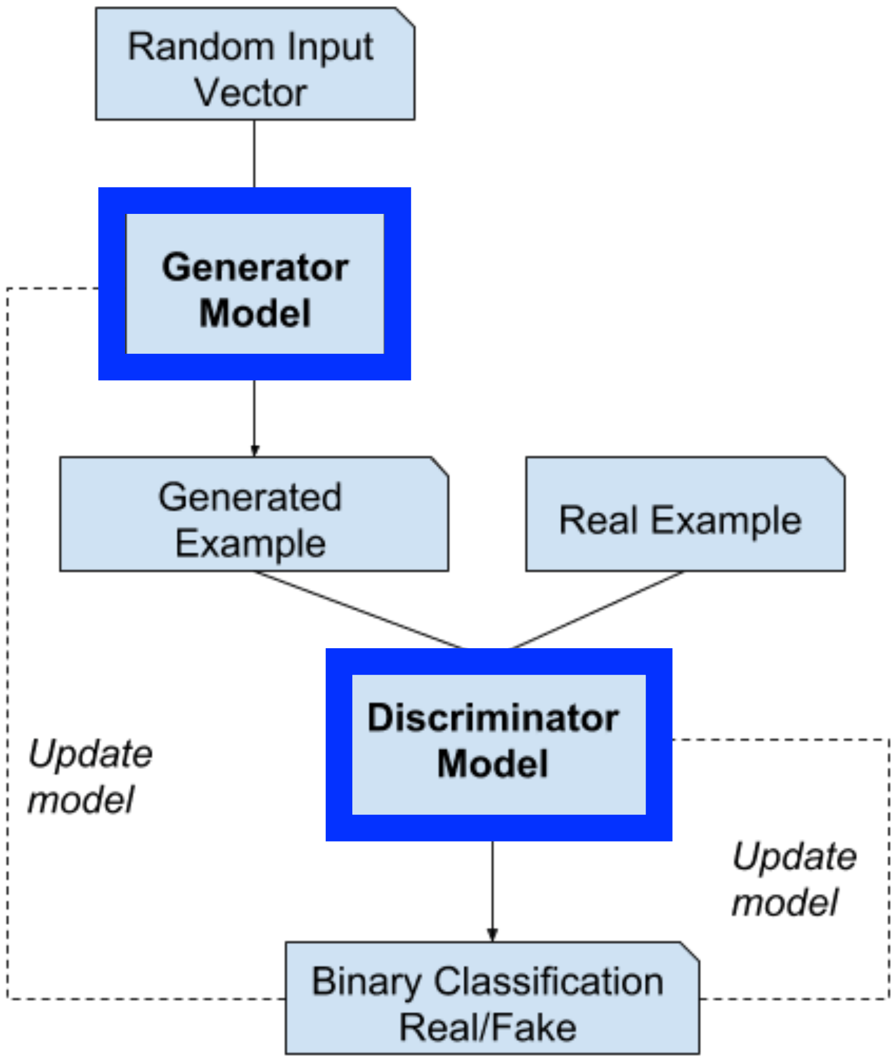
e.g., upsample noise to image of desired size

classify which examples are real

Consists of two models that compete against each other

https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

GAN: Training



The two models are iteratively trained separately

- Train discriminator using fake and real images
- Train generator using just fake images and penalize it when the discriminator recognizes images are fake

https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/

GAN: Discriminator Loss Function

Discriminator tries to minimize classification error

Discriminator wants a value of 1 for real images

Discriminator wants a value of 0 for fake images

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{x}\sim p_{\text{data}}}} \log \boxed{D(\boldsymbol{x})} - \frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{z}}} \log\left(1 - \boxed{D\left(G(\boldsymbol{z})\right)}\right)$$

Real image

Input noise

https://arxiv.org/pdf/1701.00160.pdf

GAN: Generator Loss Function

Generator tries to maximize classification error

$$J^{(G)} = -J^{(D)}$$

Want the discriminator to mistakenly arrive at a value of 1 for fake images

$$J^{(G)} = -\frac{1}{2}\mathbb{E}_{\boxed{\boldsymbol{z}}} \log \boxed{D\left(G(\boldsymbol{z})\right)}$$

Input noise

https://arxiv.org/pdf/1701.00160.pdf

DGANs: GANs that Use Convolutional Layers

Bedrooms generated by observing over 3M bedroom images

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

DGANs: GANs that Use Convolutional Layers

What objects does it learn to generate?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

DGANs: GANs that Use Convolutional Layers

What objects may it not have learned to generate?

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

DGANs: GANs that Use Convolutional Layers

Faces generated by observing over 3M images of 10K people

Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

DGANs: GANs that Use Convolutional Layers

What does it generate poorly or not all?

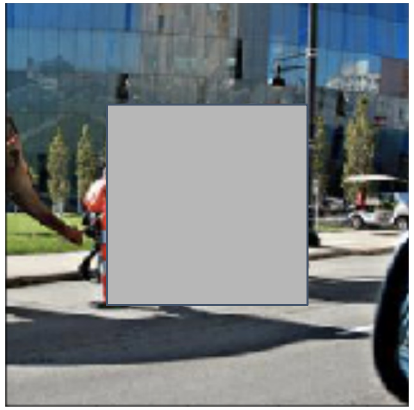
Radford et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv 2015.

Generative adversarial networks

- Generative adversarial networks (GANs)
- Context encoder

Task: Hole Filling

- What might fit into this hole?



- Many items may plausibly fit into the hole:

- Challenge: have up to 1 known ground truth region per hole

Architecture

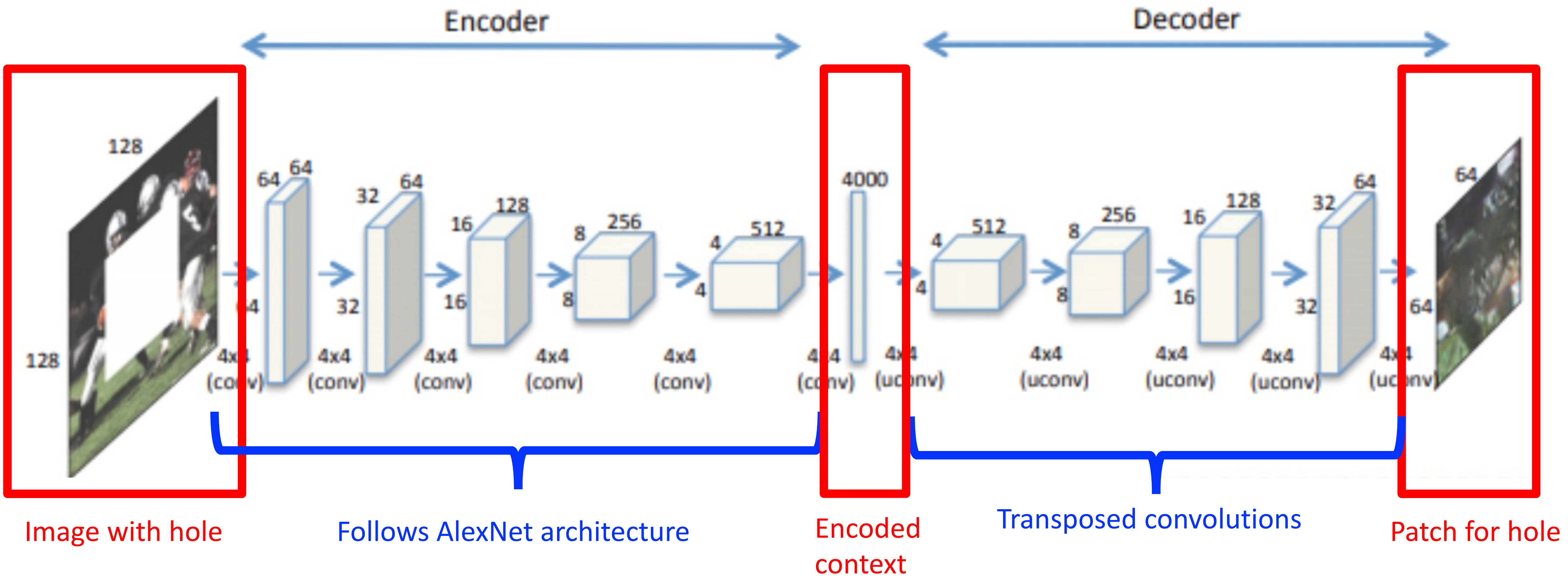


Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

Training: Loss Functions ($\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\boxed{\mathcal{L}_{rec}}$)

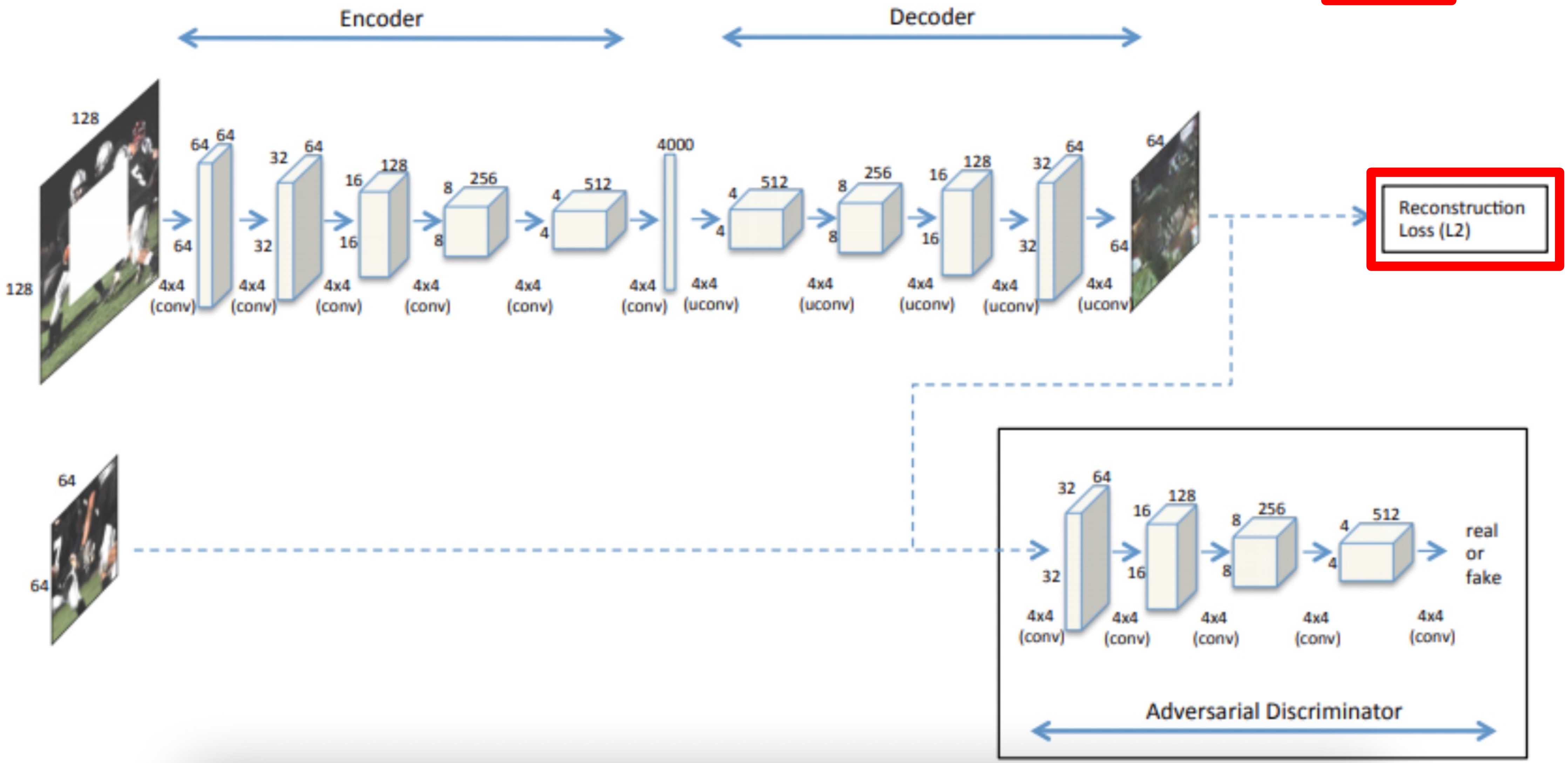
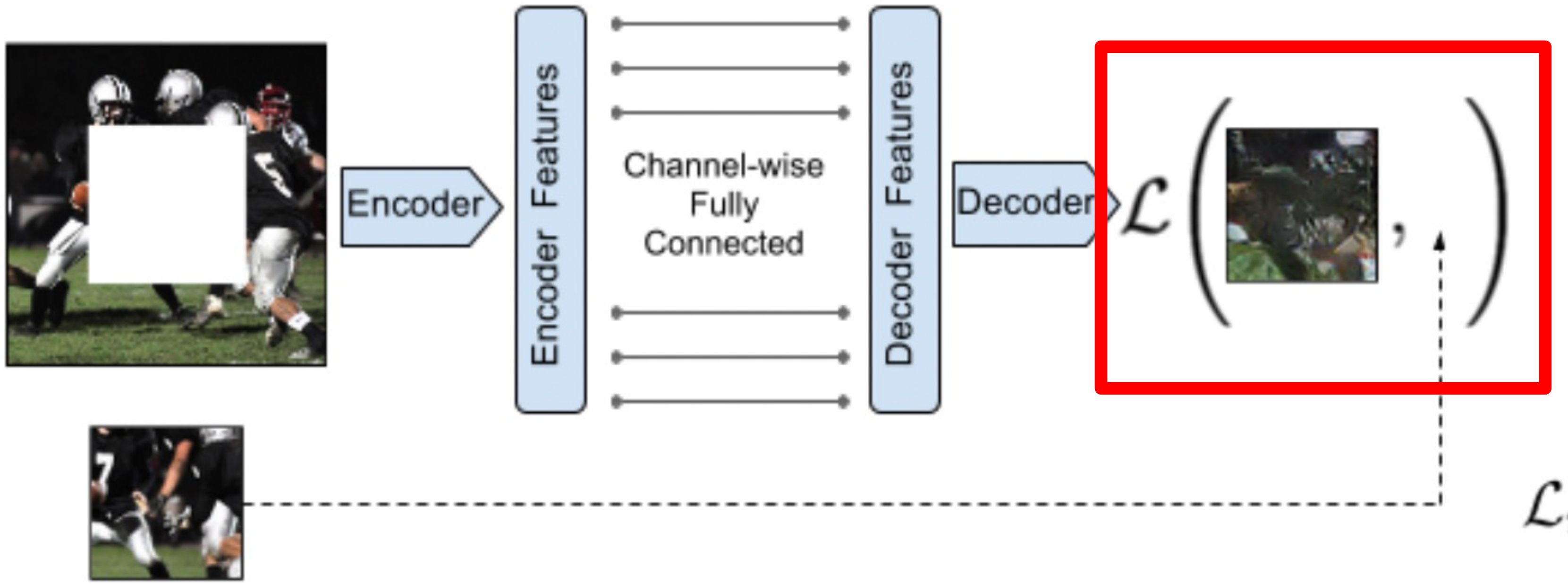


Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

Training: Reconstruction Loss (i.e., Self-Supervised Learning Approach)



Learns to fit into the context by computing the L2 loss to compare the original patch content (P) to the predicted patch content created by the model when given the image with hole (CE(X')).

$$\mathcal{L}_{rec} = \|\mathbf{P} - CE(\mathbf{X}')\|_2^2$$

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

Training: Reconstruction Loss (i.e., Self-Supervised Learning Approach)

(a) Input context

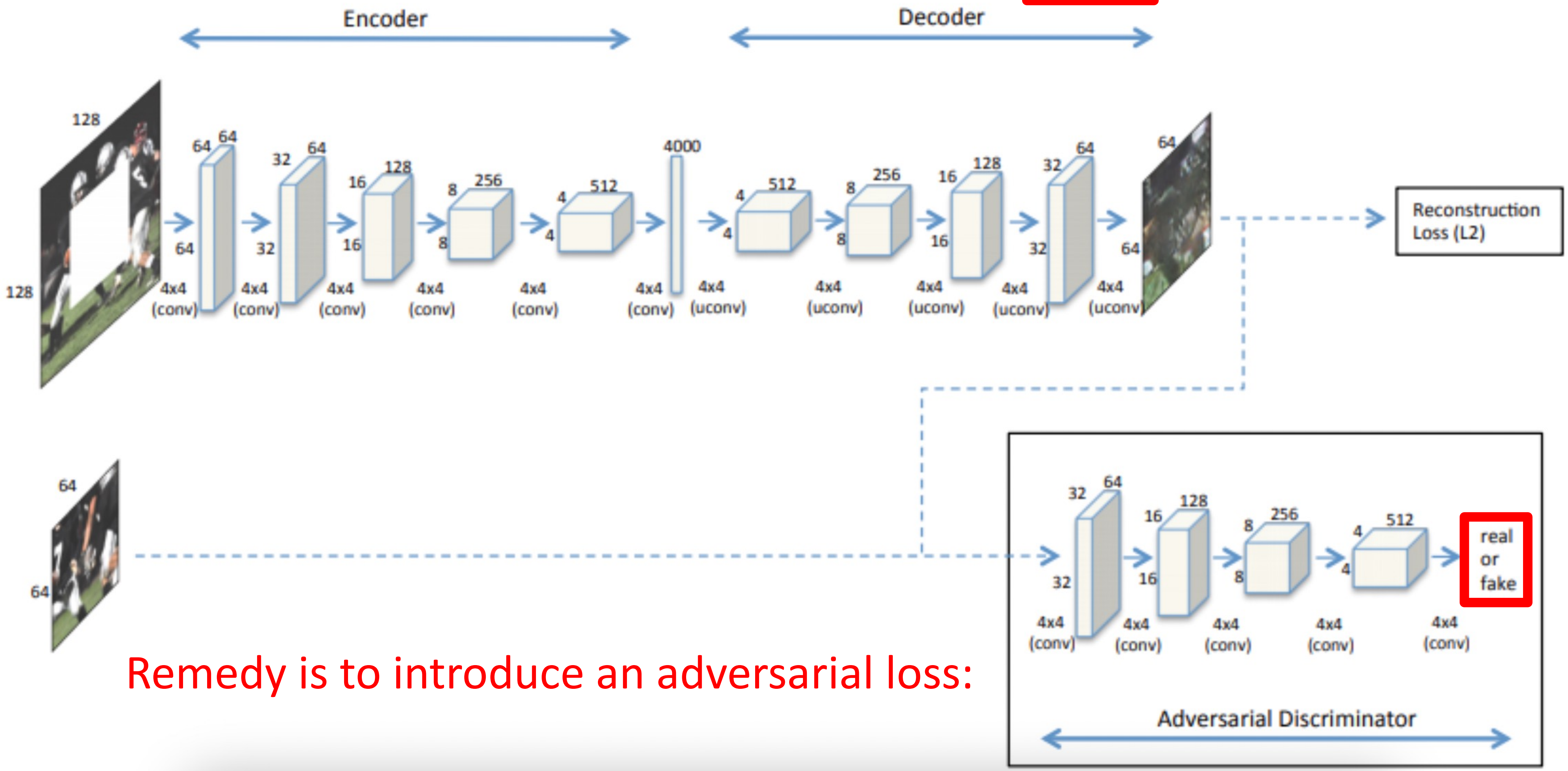
(c) Context Encoder ($L2$ loss)

Why might training with this loss function alone lead to blurry results?

- It averages the multiple plausible inpaintings for a hole

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

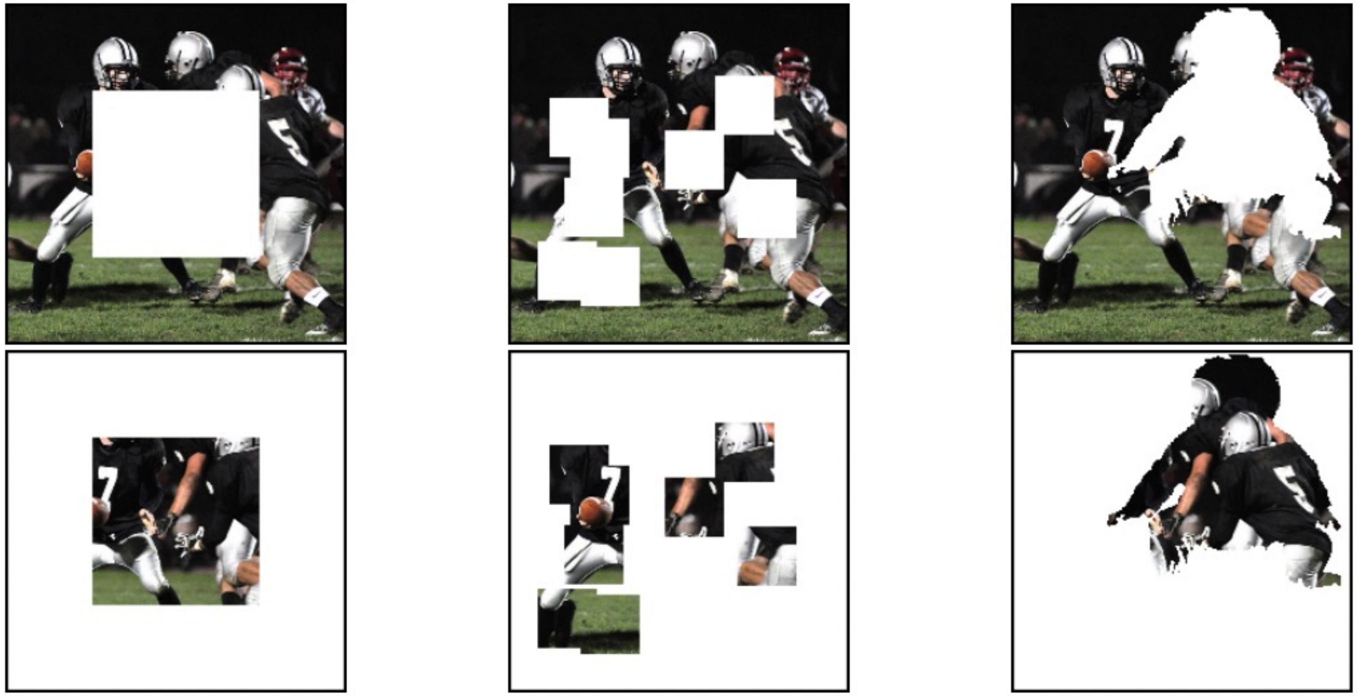
Training: Loss Functions ($\mathcal{L} = \lambda_{adv}\boxed{\mathcal{L}_{adv}} + \lambda_{rec}\mathcal{L}_{rec}$)



Remedy is to introduce an adversarial loss:

Figure source: https://medium.com/knowledge-engineering-seminar/context-encoder-image-inpainting-using-gan-ccd6a1ea5fb7

Training: Datasets

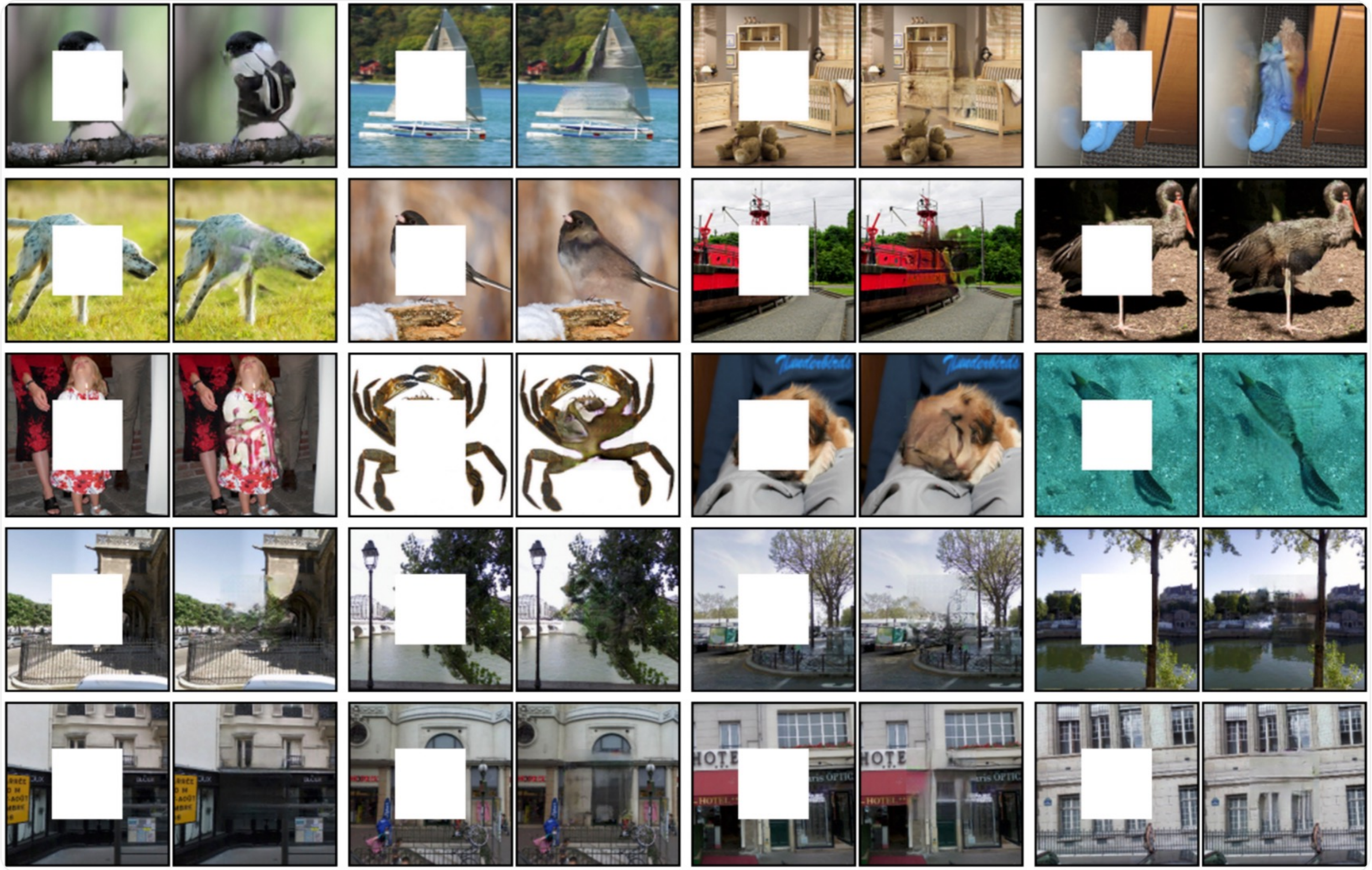


(a) Central region (b) Random block (c) Random region

Training completed on ImageNet (all 1.2M and a 100K subset) for three hole types

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

Results: https://www.cs.cmu.edu/~dpathak/context_encoder/



What type of features might be learned?

Pathak et al., Context encoders: Feature learning by inpainting; CVPR 2016

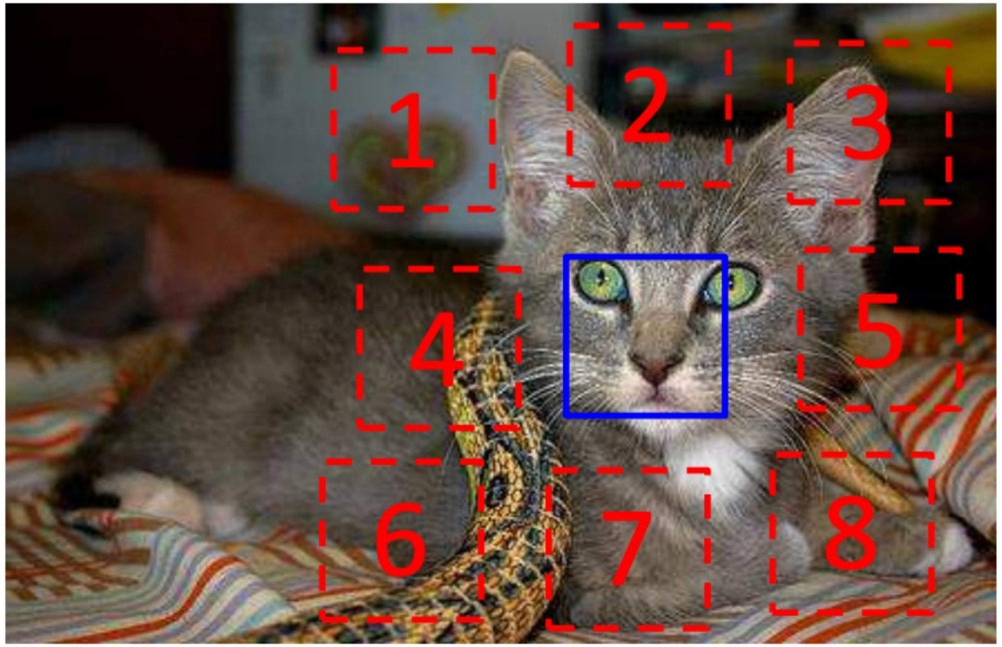
Today’s Topics

- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods

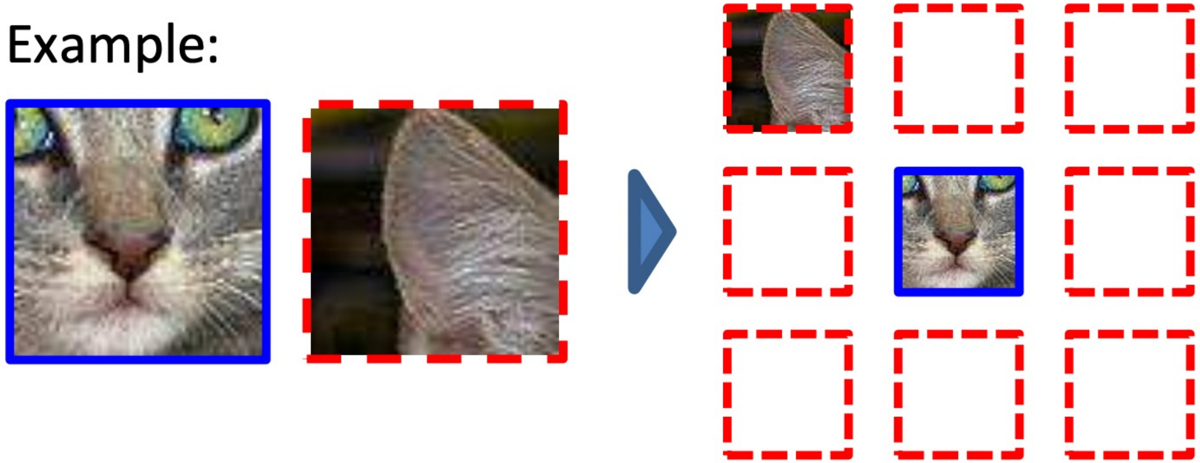
Context-based Methods

- Spatial context: predict relative positions of image patches
- Timing context: predict relative positions of video frames
- Similarity context: clustering

Spatial Context: Predict Image Index Per Patch



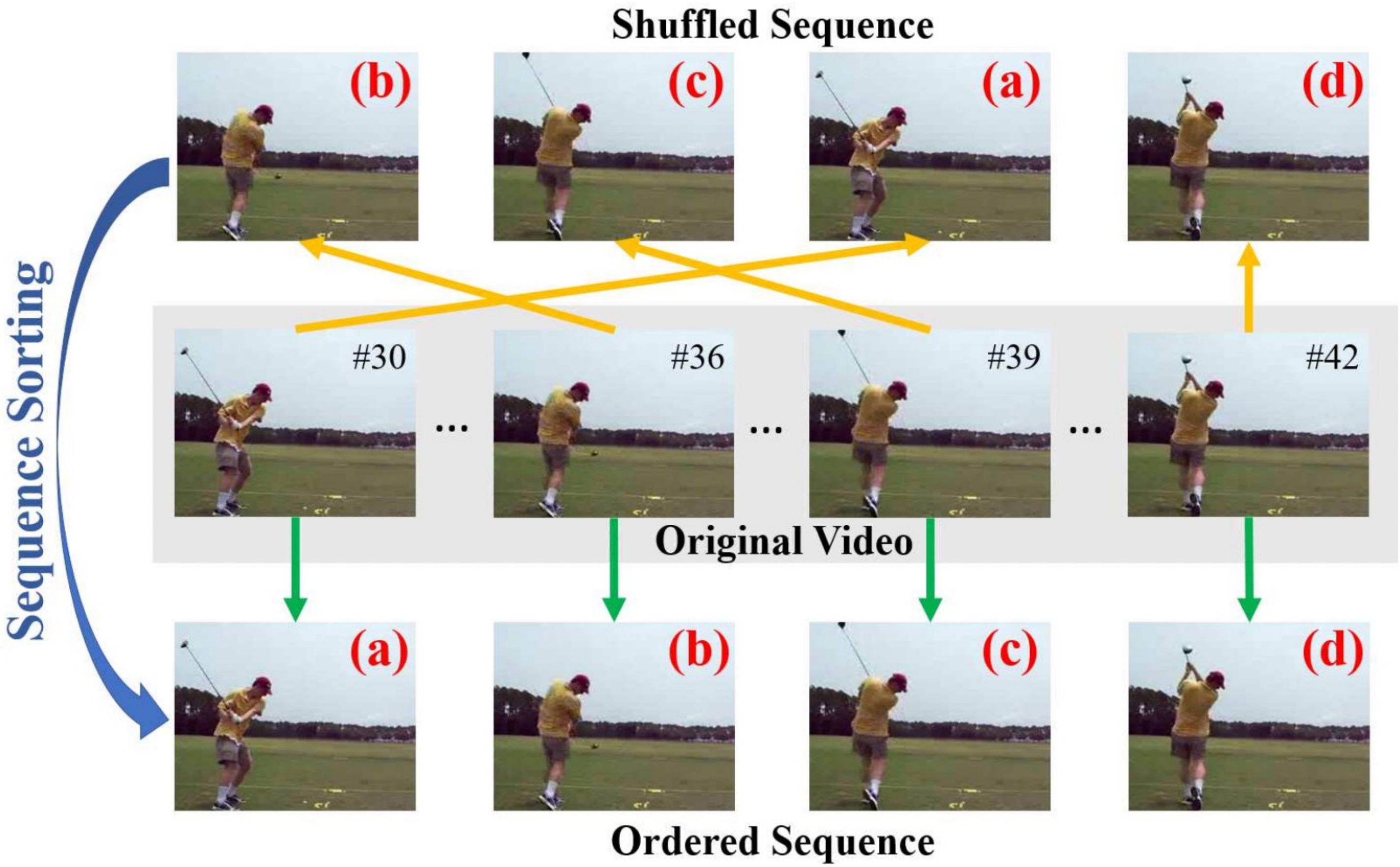
Example:



What type of features might be learned?

Carl Doersch, Abhinav Gupta, and Alexei A. Efros, Unsupervised Visual Representation Learning by Context Prediction; ICCV 2015.

Timing Context : Predict Order of Video Frames

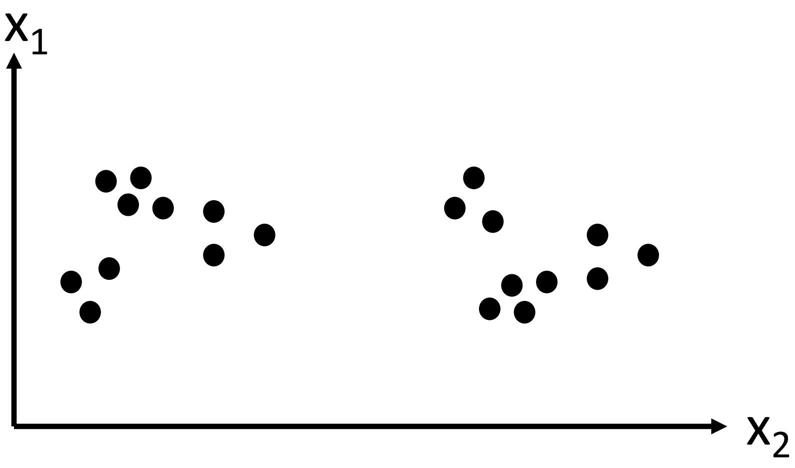


What type of features might be learned?

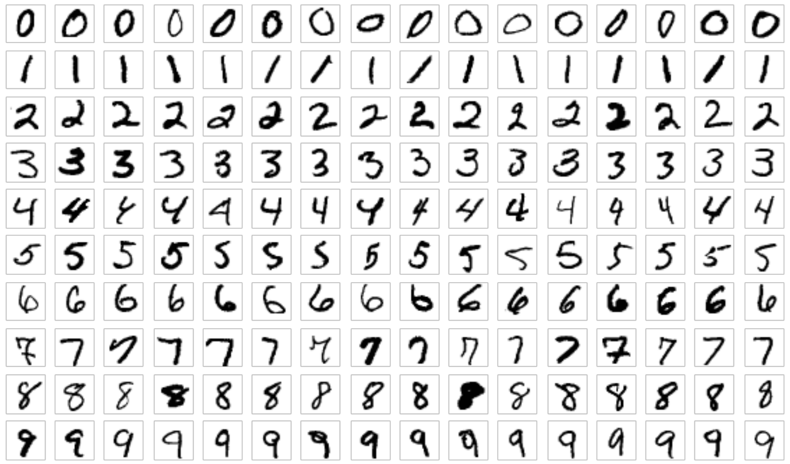
Lee et al., Unsupervised Representation Learning by Sorting Sequences; ICCV 2017.

Similarity Context: Predict Clusters

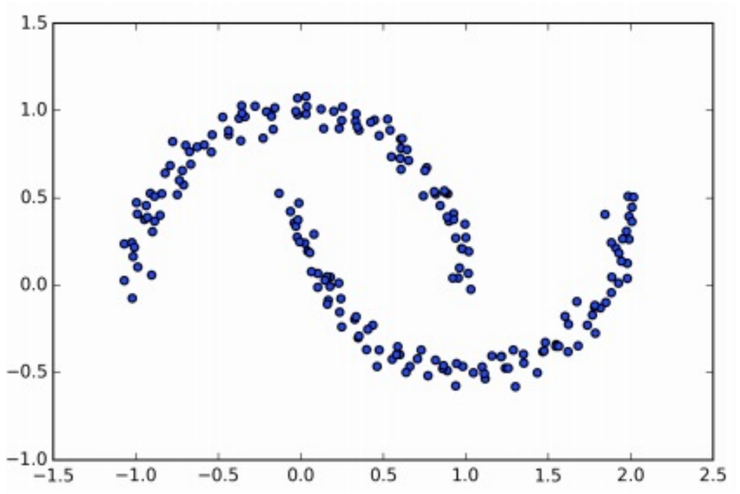
A.



B.



C.

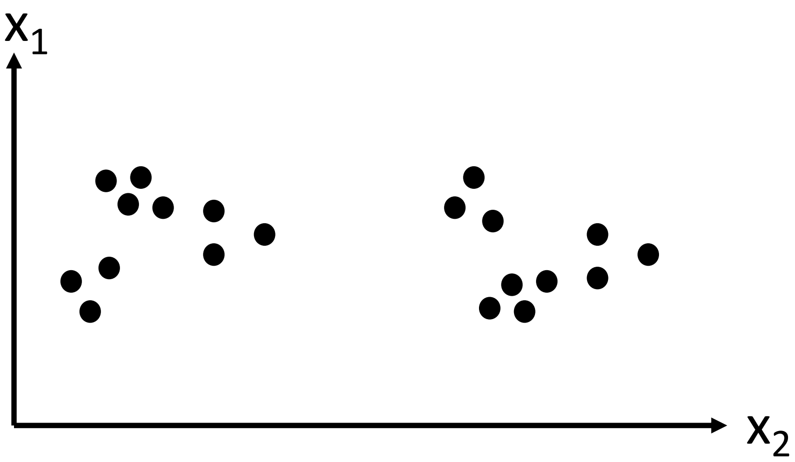


CNNS are trained to identify cluster assignments OR to recognize whether images belong to the same cluster

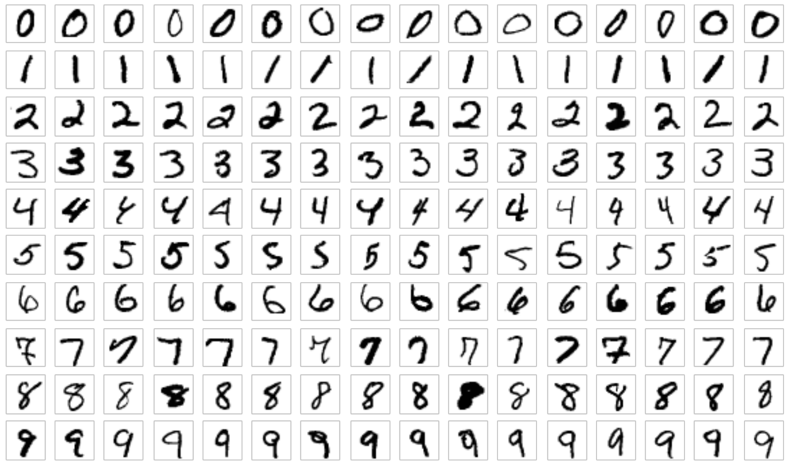
Raschka and Mirjalili; Python Machine Learning

Clustering

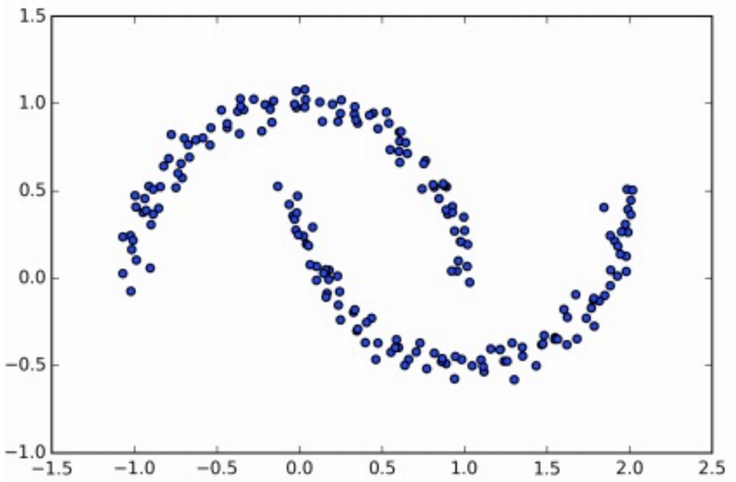
A.



B.



C.

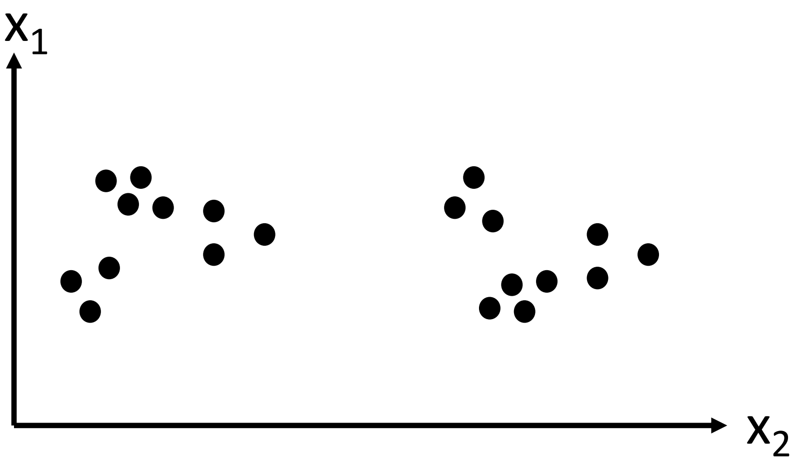


Create groupings so entities in a group will be similar to each other and different from the entities in other groups.

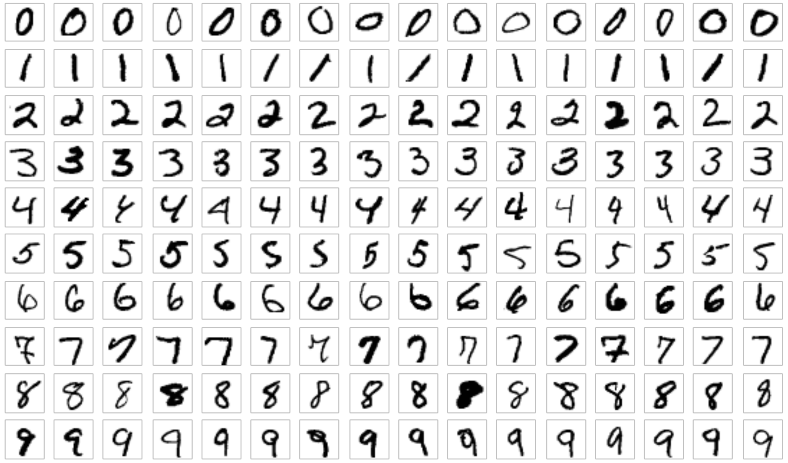
Raschka and Mirjalili; Python Machine Learning

Clustering: Key Questions

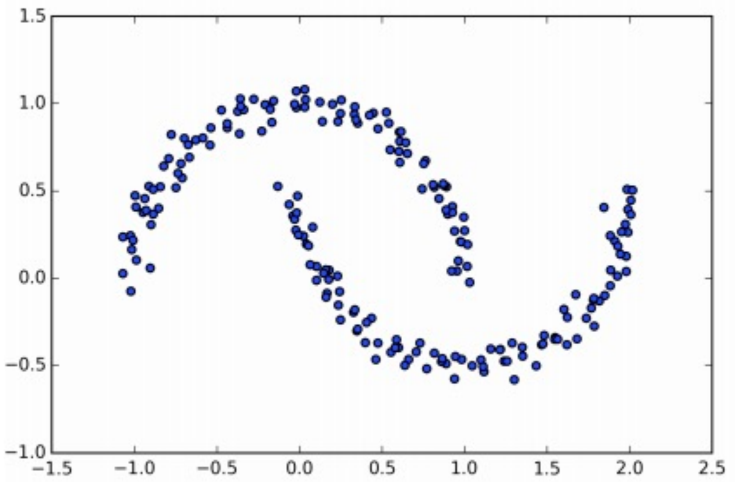
A.



B.



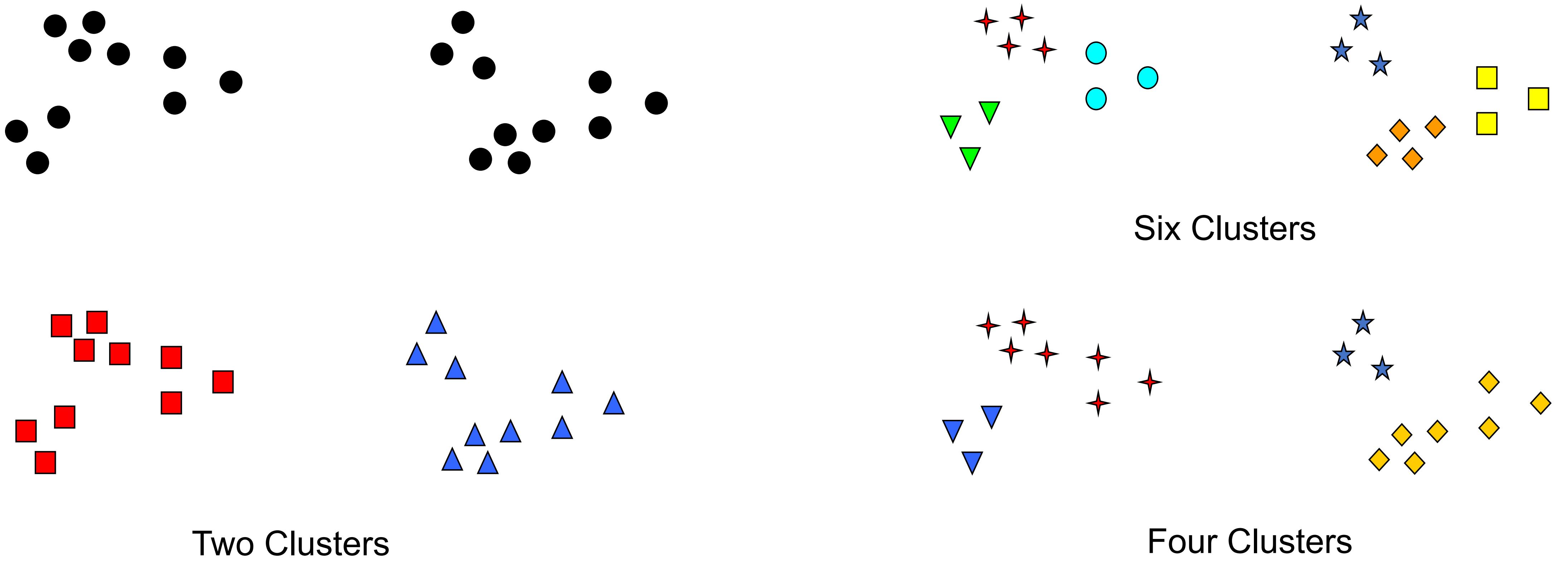
C.



- How many data clusters to create?
- What "algorithm" to use to partition the data?

Raschka and Mirjalili; Python Machine Learning

Clustering: How Many Clusters to Create?

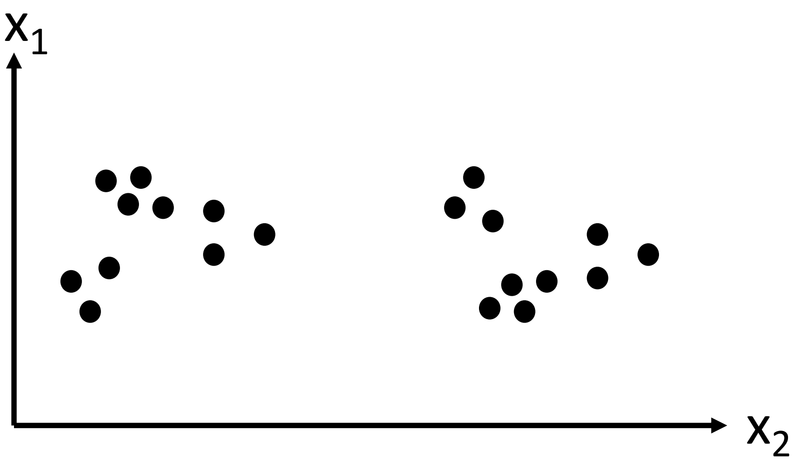


Number of clusters can be ambiguous.

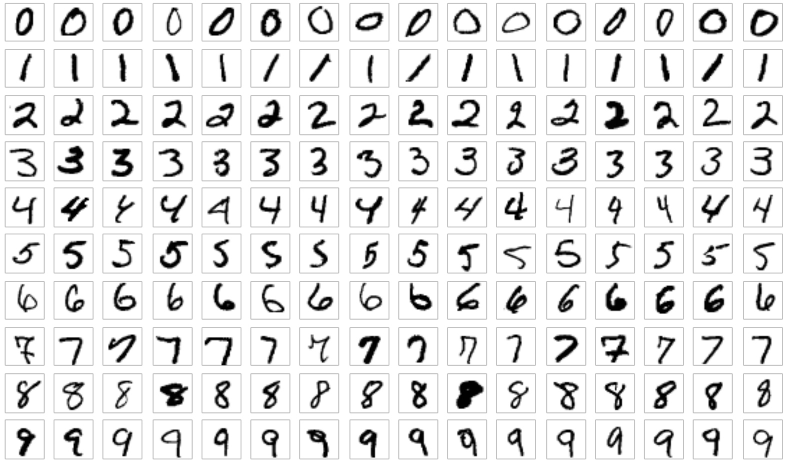
Slide adapted from: https://www-users.cs.umn.edu/~kumar001/dmbook/slides/chap7_basic_cluster_analysis.pdf

Clustering

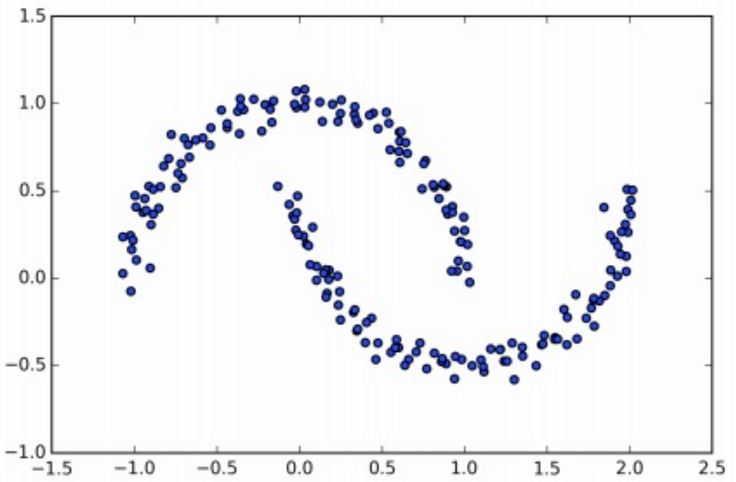
A.



B.



C.



Create groupings so entities in a group will be similar to each other and different from the entities in other groups.

What type of features might be learned?

Raschka and Mirjalili; Python Machine Learning

Context-based Methods: How Might Such Methods Be Used in the NLP Field?

- Spatial context: predict relative positions of image patches
- Timing context: predict relative positions of video frames
- Similarity context: clustering

Today's Topics

- Transfer learning definition
- Overview of self-supervised learning
- Generative-based methods
- Generative adversarial networks
- Context-based methods


The End