Lecture 14: Representation learning and language
Announcements

- Project proposal info out
- Friday section: project office hours
- Ad: ECE Cider and Donuts: https://eecs.engin.umich.edu/event/ece-cider-and-donuts/ Tuesday at 9am in 3313 EECS
Supervised computer vision


Object segmentation [Gupta et al., “LVIS”, 2019]
Supervised computer vision

These methods need lots of labeled training examples.

[Lin et al., COCO dataset]
We still need *lots* of labeled examples

Mask R-CNN on COCO with Different Training Set Sizes

Object detection accuracy vs. dataset size

Image source: R. Girshick
We want models that can generalize
Transfer learning

Training

Object recognition

Testing

Scene recognition

Often, what we will be “tested” on is to learn to do something new.

Source: Isola, Freeman, Torralba
Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.

Source: Isola, Freeman, Torralba
Finetuning

Pretraining

Object recognition

dophins, cats, grizzly bears, angels, fish, chameleons, clown fish, iguanas, elephants

Finetuning

Scene recognition

Source: Isola, Freeman, Torralba
Finetuning

Pretraining
Object recognition

- dolphin
- cat
- grizzly bear
- angel fish
- chameleon
- clown fish
- iguana
- elephant

Finetuning
Place recognition

- bathroom
- kitchen
- bedroom
- living room
- hallway

Initialize the weights using the pretraining task!

Source: Isola, Freeman, Torralba
Finetuning

- Pretrain a network on task A (e.g., object recognition), resulting in parameters $\mathbf{W}$.

- Initialize a second network with some or all of $\mathbf{W}$.

- Train the second network on task B, resulting in parameters $\mathbf{W}'$.

- Why would we expect this to work?

Source: Isola, Freeman, Torralba
Visualizing representations
Deep net “electrophysiology”

[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]

Source: Isola, Freeman, Torralba
Visualizing and Understanding CNNs
[Zeiler and Fergus, 2014]

Gabor-like filters learned by layer 1

Image patches that activate each of the layer 1 filters most strongly

Source: Isola, Freeman, Torralba
Image patches that activate each of the **layer 2** neurons most strongly
Image patches that activate each of the **layer 3** neurons most strongly

Source: Isola, Freeman, Torralba
Image patches that activate each of the **layer 4** neurons most strongly

Source: Isola, Freeman, Torralba
Image patches that activate each of the layer 5 neurons most strongly.
CNNs learned the classical visual recognition pipeline

Edges
Texture
Colors

Segments
Parts

“clown fish”

Source: Isola, Freeman, Torralba
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 2

Source: Isola, Freeman, Torralba
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

conv 4

Source: Isola, Freeman, Torralba
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 5
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]
Linear probe

Object recognition net

“Dog”
Linear probe

Feature representation

\[ z \]

\[ \text{Linear Classifier} \]

\[ (W, b) \]

\[ y \]

"Rainforest"

Logistic regression:

\[ y = \sigma(Wz + b) \]
Transferring CNN features

Hand-crafted features

CNN features pretrained on ImageNet + linear classifier [Donahue et al. 2013]

[Xiao et al., CVPR 2010]
Case study: fine-tuning for object detection

Observations:
- ImageNet pretraining speeds up object detection training by 5x.
- No change in accuracy for this dataset — just training speed, perhaps because it is so large.
- Big performance gains for small/medium datasets (e.g. 1K examples per class).

[He et al. 2018]
Learning from examples

Training data

\[
\{x_1, y_1\} \quad \rightarrow \quad \{x_2, y_2\} \quad \rightarrow \quad \{x_3, y_3\} \quad \rightarrow \quad \cdots
\]

Learner

\[
\rightarrow \quad f : X \rightarrow Y
\]

Source: Isola, Freeman, Torralba
Representation Learning

Data

\[
\{x_1\}, \{x_2\}, \{x_3\}, \ldots
\]

\[
\xrightarrow{\text{Learner}}
\]

\[
\rightarrow \text{Representations}
\]

Source: Isola, Freeman, Torralba
How do we learn good representations?
Self-supervised learning methods
Recall: autoencoder

Compressed image code (vector $z$)

Image $X$ → $\hat{X}$

[e.g., Hinton & Salakhutdinov, Science 2006]

Source: Isola, Freeman, Torralba
\[
\text{arg min}_{\mathcal{F}} \mathbb{E}_X [||\mathcal{F}(X) - X||]
\]

Source: Isola, Freeman, Torralba
$X \hat{X} = \mathcal{F}(X)$

Source: Isola, Freeman, Torralba
Data compression

Source: Isola, Freeman, Torralba
Data prediction

Source: Isola, Freeman, Torralba
Denoising autoencoder

Noisy image → Reconstructed image

[Vincent et al., 2008]
Denoising autoencoder

Noisy image → Reconstruction process → Reconstructed image

Other types of “noise”.

[Vincent et al., 2008, Pathak et al. 2015, He, 2020]
Grayscale image: L channel

\[ X \in \mathbb{R}^{H \times W \times 1} \]

Color information: ab channels

\[ \hat{Y} \in \mathbb{R}^{H \times W \times 2} \]

Source: Isola, Freeman, Torralba

[Zhang, Isola, Efros, ECCV 2016]
Visualizing units

[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]

Source: Isola, Freeman, Torralba
Source: Isola, Torralba, Freeman

[“Colorful image colorization”, Zhang et al., ECCV 2016]
[“Colorful image colorization”, Zhang et al., ECCV 2016]
Source: Isola, Torralba, Freeman

[“Colorful image colorization”, Zhang et al., ECCV 2016]
Stimuli that drive selected neurons (conv5 layer)

- faces
- dog
- faces
- flowers
Classification performance
ImageNet Task [Russakovsky et al. 2015]

Accuracy

- autoencoder
- colorization

Layer:
- conv1, pool1, conv2, pool2, conv3, conv4, conv5, pool5

Source: Isola, Freeman, Torralba
Represent image as a vector of neural activations (perhaps representing a vector of detected texture patterns or object parts)
Example from language: word2vec

“Elephant” → dense vector representation of word, called a **word embedding**.

one-hot vector representation of word

Source: Isola, Freeman, Torralba
Words with similar meanings should be near each other

Source: Isola, Freeman, Torralba
word2vec

Words with similar meanings should be near each other

Proxy: words that are used in the same context tend to have similar meanings

words with similar contexts should be near each other
Next to the 'sofa' is a desk, and a 'person' is sitting behind it.
'sofa'
'armchair'
'bench'
'chair'
'deck chair'
'ottoman'
'seat'
'stool'
'swivel chair'
'loveseat'
'loveseat'...
word2vec

I parked the **car** in a nearby street. It is a red **car** with two doors, …

I parked the **vehicle** in a nearby street…

I parked the **car** in a nearby street. It is a red **car** with two doors, …
word2vec

word2vec, training

Output prob. That each word is in the context of the input word

\[ p = \frac{e^{x_i}}{\sum_j e^{x_j}} \]

Algebraic operations with the vector representation of words

\[ X = \text{Vector(“Paris”)} - \text{vector(“France”)} + \text{vector(“Italy”)} \]

Closest nearest neighbor to \( X \) is \( \text{vector(“Rome”)} \)
Context as Supervision
[Collobert & Weston 2008; Mikolov et al. 2013]

...
Context Prediction as Supervision

[Slide credit: Carl Doersch]
Semantics from a non-semantic task
Relative Position Task

8 possible locations

Randomly Sample Patch
Sample Second Patch

[Slide credit: Carl Doersch]
Patch Embedding (representation)

Classifier

CNN

Input | Nearest Neighbors

Note: connects *across* instances!

[Slide credit: Carl Doersch]
Revisiting autoencoders

Is prediction necessary?

\[ \hat{X} = F(X) \]
Contrastive learning

$z^T z \rightarrow $ High dot product with self

$z^T x_1 \rightarrow $ Low dot product with others

[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]
Contrastive learning

Minimize:

\[ \mathcal{L} = -\log \left( \frac{\exp(z^Tz)}{\sum_{i=1}^{n} \exp(z^T x_i)} \right) \]

Equivalent to softmax loss with each image as a category.

[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]
Contrastive learning

Build in invariance by comparing to \textbf{distorted} images.

\[
\frac{\exp\{z^\top \tilde{z}\}}{\exp\{z^\top \tilde{z} + \sum_i z^\top x_i\}}
\]

[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]
Data augmentation used in contrastive learning

(a) Original  (b) Crop and resize  (c) Crop, resize (and flip)  (d) Color distort. (drop)  (e) Color distort. (jitter)

(f) Rotate \{90^\circ, 180^\circ, 270^\circ\}  (g) Cutout  (h) Gaussian noise  (i) Gaussian blur  (j) Sobel filtering

[Chen et al., SimCLR, 2020]
Performance snapshot

ImageNet linear classification

Object detection finetuning

Comparable in many cases to supervised pretraining.

<table>
<thead>
<tr>
<th>pre-train</th>
<th>AP$_{50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>random init.</td>
<td>52.5</td>
</tr>
<tr>
<td>super. IN-1M</td>
<td>80.8</td>
</tr>
<tr>
<td>MoCo IN-1M</td>
<td>81.4 (+0.6)</td>
</tr>
<tr>
<td>MoCo IG-1B</td>
<td>82.1 (+1.3)</td>
</tr>
</tbody>
</table>

[He et al. Momentum contrastive learning 2019]
Goal: Set up a pre-training scheme to induce a “useful” representation

Source: Richard Zhang
Language
Language-based supervision

e.g., [Radford et al., “CLIP”, 2021]
Language-based supervision

e.g., [Radford et al., “CLIP”, 2021]
What about language?

“A giraffe standing in the grass next to a tree”
How to represent words as numbers?

Prediction \( \hat{y} \)

\[ f_\theta : X \rightarrow \mathbb{R}^K \]

- dolphin
- cat
- grizzly bear
- angel fish
- chameleon
- clown fish
- iguana
- elephant

Source: Isola, Torralba, Freeman
How to represent words as numbers?

Or, represent each character as a class (e.g., $K=26$ for English letters), and represent words as a sequence of characters.

Source: Isola, Torralba, Freeman
How to represent words as numbers?

Rather than having just a handful of possible object classes, we can represent all words in a large vocabulary using a very large $K$ (e.g., $K=100,000$).

Use “chunks” of characters instead.

Source: Isola, Torralba, Freeman
This problem is called **image captioning**.
**Teacher forcing**: provide previous ground truth word. Makes training easier. Reduces need to model long-range dependencies.

Source: Isola, Torralba, Freeman
Maximum likelihood: maximize probability of all ground truth words.
Testing

Samples

Outputs $p_\theta(\cdot)$

Hidden

Input

```
  Sample from predicted distribution over words.
  Alternatively, sample most likely word.
```

Source: Isola, Torralba, Freeman
Captioning: popular topic circa 2015

Vinyals et al., 2015
Donahue et al., 2015
Karpathy and Fei-Fei, 2015
Hodosh et al., 2013

Fang et al., 2015
Mao et al., 2015
Ordonez et al., 2011
Kulkarni et al., 2011

Chen and Zitnick, 2015
Farhadi et al., 2010
Mitchell et al., 2012
Kiros et al., 2015

... and many more
Show and Tell: A Neural Image Caption Generator

[Vinyals et. al., CVPR 2015]

Source: Isola, Torralba, Freeman
Show and Tell: A Neural Image Caption Generator
[Vinyals et. al., CVPR 2015]
Transformer-based captioning

Desai and Johnson, VirTex, 2020
Good source of features

Language Supervised Pretraining

A brown and white puppy lying on green lawn looking at apples.

Task: Image Captioning

Example: Object Detection

Downstream Transfer

Faster R-CNN

Transformers

PASCAL VOC Linear Clf. (mAP)

- VirTex (5 caption)
- VirTex (1 caption)
- ImageNet-sup

[Desai and Johnson, VirTex, 2020]
VQA: Visual Question Answering
www.visualqa.org


Abstract—We propose the task of free-form and open-ended Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ~0.25M images, ~0.76M questions, and ~10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance.

2016

What is the mustache made of?

[http://www.visualqa.org/challenge.html]
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.
### Questions and answers collected with Amazon Mechanical Turk

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is something under the sink broken?</td>
<td>yes</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>What number do you see?</td>
<td>33</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Does this man have children?</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Is this man crying?</td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Can you park here?</td>
<td>no</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>What color is the hydrant?</td>
<td>white and orange</td>
<td>red</td>
<td></td>
</tr>
<tr>
<td>Has the pizza been baked?</td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>What kind of cheese is topped on this pizza?</td>
<td>feta</td>
<td>feta</td>
<td>mozzarella</td>
</tr>
<tr>
<td>What kind of store is this?</td>
<td>bakery</td>
<td>bakery</td>
<td>art supplies</td>
</tr>
<tr>
<td>Is the display case as full as it could be?</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>How many pickles are on the plate?</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>What is the shape of the plate?</td>
<td>circle</td>
<td>round</td>
<td>circle</td>
</tr>
</tbody>
</table>

Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.
Architecture

There are 1000 possible answers in this system. Questions are unlimited.

[Agrawal et al., "VQA: Visual Question Answering" 2016]
what is on the ground?

Submit

Predicted top-5 answers with confidence:
sand 90.748%
snow 2.656%
beach 4.16%
surfboards 0.677%
water 0.526%
what color is the umbrella?

<table>
<thead>
<tr>
<th>Color</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>yellow</td>
<td>96.090%</td>
</tr>
<tr>
<td>white</td>
<td>4.11%</td>
</tr>
<tr>
<td>black</td>
<td>0.663%</td>
</tr>
<tr>
<td>blue</td>
<td>0.541%</td>
</tr>
<tr>
<td>gray</td>
<td>0.362%</td>
</tr>
</tbody>
</table>
are we alone in the universe?

Submit

Predicted top-5 answers with confidence:

- no [78.23%]
- yes [21.76%]
- people [0.001%]
- birds [0.000%]
- out [0.000%]
what is the meaning of life?

Submit

Predicted top-5 answers with confidence:

beach
15.26%
sand
0.53%
seagull
4.70%
tower
2.93%
rocks
1.74%
what is the yellow thing?

Predicted top-5 answers with confidence:

frisbee 79.644%
surfboard 7.319%
banana 2.644%
lemon 2.438%
surfboards 2.438%
how many trains are in the picture?

Submit

Predicted top-5 answers with confidence:

3  30.233%
5  18.270%
4  17.000%
2  11.343%
6  7.806%
What’s going on?
The Giraffe-Tree problem

A giraffe standing in the grass next to a tree.

[“Measuring Machine Intelligence Through Visual Question Answering”, Zitnick et al., 2016]
Nearest neighbor baseline

Test

Train

Source: L. Zitnick
Nearest Neighbor

A black and white cat sitting in a bathroom sink.

Two zebras and a giraffe in a field.

See mscoco.org for image information

Source: L. Zitnick
A man riding a motorcycle on a beach.

An airplane is parked on the tarmac at an airport.

Source: L. Zitnick
### COCO Caption Challenge Results

<table>
<thead>
<tr>
<th>Model</th>
<th>CIDEr-D</th>
<th>Meteor</th>
<th>ROUGE-L</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google[4]</td>
<td>0.943</td>
<td>0.254</td>
<td>0.53</td>
<td>0.309</td>
</tr>
<tr>
<td>MSR Captivator[9]</td>
<td>0.931</td>
<td>0.248</td>
<td>0.526</td>
<td>0.308</td>
</tr>
<tr>
<td>m-RNN[15]</td>
<td>0.917</td>
<td>0.242</td>
<td>0.521</td>
<td>0.299</td>
</tr>
<tr>
<td>MSR[8]</td>
<td>0.912</td>
<td>0.247</td>
<td>0.519</td>
<td>0.291</td>
</tr>
<tr>
<td>Nearest Neighbor[11]</td>
<td>0.886</td>
<td>0.237</td>
<td>0.507</td>
<td>0.280</td>
</tr>
<tr>
<td>m-RNN (Baidu/ UCLA)[16]</td>
<td>0.886</td>
<td>0.238</td>
<td>0.524</td>
<td>0.302</td>
</tr>
<tr>
<td>Berkeley LRCN[2]</td>
<td>0.869</td>
<td>0.242</td>
<td>0.517</td>
<td>0.277</td>
</tr>
<tr>
<td>Human[5]</td>
<td>0.854</td>
<td>0.252</td>
<td>0.484</td>
<td>0.217</td>
</tr>
<tr>
<td>Montreal/Toronto[10]</td>
<td>0.85</td>
<td>0.243</td>
<td>0.513</td>
<td>0.268</td>
</tr>
<tr>
<td>PicSOM[13]</td>
<td>0.833</td>
<td>0.231</td>
<td>0.505</td>
<td>0.281</td>
</tr>
<tr>
<td>MLBL[7]</td>
<td>0.74</td>
<td>0.219</td>
<td>0.499</td>
<td>0.26</td>
</tr>
<tr>
<td>ACVT[1]</td>
<td>0.709</td>
<td>0.213</td>
<td>0.483</td>
<td>0.246</td>
</tr>
<tr>
<td>NeuralTalk[12]</td>
<td>0.674</td>
<td>0.21</td>
<td>0.475</td>
<td>0.224</td>
</tr>
<tr>
<td>Tsinghua Bigeye[14]</td>
<td>0.673</td>
<td>0.207</td>
<td>0.49</td>
<td>0.241</td>
</tr>
<tr>
<td>MIL[6]</td>
<td>0.666</td>
<td>0.214</td>
<td>0.468</td>
<td>0.216</td>
</tr>
<tr>
<td>Brno University[3]</td>
<td>0.517</td>
<td>0.195</td>
<td>0.403</td>
<td>0.134</td>
</tr>
</tbody>
</table>

Source: L. Zitnick
Aside: biases in data collection
Getting more humans in the annotation loop

Labeling to get a Ph.D.

Labeling for money
(Sorokin, Forsyth, 2008)

Amazon Mechanical Turk
Artificial Artificial Intelligence

Labeling for fun
Luis Von Ahn and Laura Dabbish 2004

Labeling because it gives you added value

Visipedia
(Belongie, Perona, et al)

Labeling to prove you’re human

Source: Isola, Torralba, Freeman
Beware of the human in your loop

• What do you know about them?
• Will they do the work you pay for?

Let’s check a few simple experiments

Source: Isola, Torralba, Freeman
People have biases…

Turkers were offered 1 cent to pick a number from 1 to 10.

~850 turkers

Experiment by Greg Little

Source: Isola, Torralba, Freeman

From http://groups.csail.mit.edu/uid/deneme/
Do humans have consistent biases?

Experiment by Greg Little

Source: Isola, Torralba, Freeman

From http://groups.csail.mit.edu/uid/deneme/
Are humans reliable even in simple tasks?

Experiment by Greg Little

Results of 100 HITS:
A: 2
B: 96
C: 2

Source: Isola, Torralba, Freeman

From http://groups.csail.mit.edu/uid/deneme/
Do humans do what you ask for?

Experiment by Rob Miller
From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman
So we can sometimes collect good training data.

But suppose we messed up. Our test setting doesn’t look like the training data!

How can we bridge the domain gap?

Source: Isola, Torralba, Freeman
Finding more representative images

Places365 Kitchen

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]
Finding more representative images

VLOG Kitchen

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]
Name that dataset game

1. Caltech101
2. Tiny
3. MSRC
4. Corel
5. UIUC
6. PASCAL 07
7. LabelMe
8. 15 Scenes
9. COIL-100
10. Caltech256
11. ImageNet
12. SUN09

Some recent directions
Learning representations from language

Contrastive learning

\[
\log \left( \frac{\exp(I_i \cdot T_i)}{\sum_j \exp(I_i \cdot T_j)} \right)
\]

[Radford et al., "CLIP", 2021]
Learning representations from language

maximize:

$$\log \left( \frac{\exp(I_i \cdot T_i)}{\sum_j \exp(I_i \cdot T_j)} \right)$$

[Radford et al., "CLIP", 2021]
Learning representations from language

\[
\log \left( \frac{\exp(I_i \cdot T_i)}{\sum_j \exp(I_i \cdot T_j)} \right)
\]

Contrastive learning

[Radford et al., "CLIP", 2021]
“Zero-shot” classification

(1) Create classifier from label text

(2) Test how well each prompt fits an image

[Radford et al., "CLIP", 2021]
“Zero-shot” classification

[Radford et al., "CLIP", 2021]
“Zero-shot” classification

Radford et al., "CLIP", 2021
“Zero-shot” classification

[Radford et al., "CLIP", 2021]
A giraffe standing in the grass next to a tree
“A giraffe standing in the grass next to a tree”
(a) a tapir made of accordion.  
(b) an illustration of a baby hedgehog in a christmas sweater walking a dog  
(c) a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign  
(d) the exact same cat on the top as a sketch on the bottom

[Ramesh et al., “Zero-Shot Text-to-Image Generation”, 2021]
Text-to-image

[Image with various text-to-image creations for a capybara sitting in a field at sunrise, showing different styles and mediums like pencil sketch, pixel art illustration, illustration, sepia tone photograph, black and white photograph, cartoon, ukiyo-e print, wood engraving, postage stamp, and poster.

[Ramesh et al., “Zero-Shot Text-to-Image Generation”, 2021]
Diffusion text-to-image synthesis

- a teddy bear on a skateboard in Times Square
- A photo of Michelangelo’s sculpture of David wearing headphones DJing
- “A sea otter with a pearl earring” by Johannes Vermeer
- 3D render of a cute tropical fish in an aquarium on a dark blue background, digital art

[“DALL-E 2”, Ramesh et al., 2022]
Diffusion text-to-image synthesis

1. Train CLIP

“a corgi playing a flame throwing trumpet”

[Clip objective] [Image encoder]

prior

decoder

[Ramesh et al., DALL-E 2, 2021]
Diffusion text-to-image synthesis

2. Estimate image embedding from text embedding
Diffusion text-to-image synthesis

3. Conditional model

“a corgi playing a flame throwing trumpet”

[Image: Diagram of a conditional model for diffusion text-to-image synthesis]

[Ramesh et al., DALL-E 2, 2021]
Conditional diffusion

Basic idea

- Unconditional diffusion: predict noise at step $t$ with neural net: $\epsilon_\theta(x_t, t)$.
- Conditional diffusion: predict noise with: $\epsilon_\theta(x_t, c, t)$, where $c$ conditional input.
Summary

1. Deep nets learn representations

2. This is useful because representations transfer — they act as prior knowledge that enables quick learning on new tasks

3. Representations can also be learned without labels

4. Without labels there are many ways to learn representations. We saw:
   1. representations as compressed codes
   2. representations that are predictive of their context

5. Language is a powerful form of supervision

6. Language is a natural “user interface” for computer vision systems

Adapted from Isola, Freeman, Torralba
Next class: sound and touch
Next class: vision and language