image captioning

a red truck is parked on a street lined with trees
visual question answering

- Is this truck considered “vintage”?
- Does the road look new?
- What kind of tree is behind the truck?
we’ve seen how to compute representations of words and sentences. what about images?
grayscale images are matrices

what range of values can each pixel take?
color images are tensors

Channels are usually RGB: Red, Green, and Blue
Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc
Convolution operator

\[ g(x, y) = \sum_{u} \sum_{v} k(u, v) f(x - u, y - v) \]

(filter, kernel)

Input image * Weights → Output image

\[
\begin{array}{cccc}
4 & 5 & 7 & 6 \\
3 & 2 & 8 & 0 \\
6 & 7 & 7 & 1 \\
3 & 0 & 1 & 1 \\
4 & 3 & 2 & 1 \\
\end{array}
\]

\[
\begin{array}{ccc}
0 & 0 & 0 \\
1 & 0 & 1 \\
0 & 0 & 0 \\
\end{array}
\]

\[
\begin{array}{ccc}
11 & 2 & 15 \\
13 & 8 & 12 \\
? & ? & ? \\
\end{array}
\]
demo:
http://setosa.io/ev/image-kernels/
Convolutional Layer (with 4 filters)

Input: 1x224x224

weights: 4x1x9x9

Output: 4x224x224

if zero padding, and stride = 1
Convolutional Layer (with 4 filters)

Input: 1x224x224

weights: 4x1x9x9

Output: 4x112x112

If zero padding, but stride = 2
pooling layers also used to reduce dimensionality

Convolutional Layers: slide a set of small filters over the image

Pooling Layers: reduce dimensionality of representation

why reduce dimensionality?

image: https://cs231n.github.io/convolutional-networks/
Alexnet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky  
University of Toronto  
kriz@cs.utoronto.ca

Ilya Sutskever  
University of Toronto  
ilya@cs.utoronto.ca

Geoffrey E. Hinton  
University of Toronto  
hinton@cs.utoronto.ca

the paper that started the deep learning revolution!
image classification

Classify an image into 1000 possible classes:
e.g. Abyssinian cat, Bulldog, French Terrier, Cormorant, Chickadee, red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.

- cat, tabby cat (0.71)
- Egyptian cat (0.22)
- red fox (0.11)
- ...

train on the ImageNet challenge dataset,
~1.2 million images
Alexnet

https://www.saagie.com/fr/blog/object-detection-part1
Alexnet

conv+pool

conv+pool

conv

conv

conv

linear

linear

linear+

softmax

https://www.saagie.com/fr/blog/object-detection-part1
What is happening?

https://www.saagie.com/fr/blog/object-detection-part1
Revolution of Depth

AlexNet, 8 layers  
(ILSVRC 2012)

VGG, 19 layers  
(ILSVRC 2014)

ResNet, **152 layers**  
(ILSVRC 2015)

**152 layers**

ILSVRC'15  
ResNet

ILSVRC'14  
GoogleNet

ILSVRC'14  
VGG

ILSVRC'13

ILSVRC'12  
AlexNet

ILSVRC'11  
shallow

ILSVRC'10

Slide by Mohammad Rastegari
ImageNet pretraining -> Instagram pretraining

Bigger models are saturated on ImageNet, but with more data bigger models do better

Biggest network was pretrained on 3.5B Instagram images

Trained on 336 GPUs for 22 days

Mahajan et al, “Exploring the Limits of Weakly Supervised Pretraining”, arXiv 2018
at the end of the day, we generate a fixed size vector from an image and run a classifier over it

\[
CNN(\text{image}) = \text{softmax: predict 'truck'}
\]
key insight: this vector is useful for many more tasks than just image classification!
we can use it for \textit{transfer learning}
simple visual QA

- \( i = \text{CNN}(\text{image}) \) > use an existing network trained for image classification and freeze weights
- \( q = \text{RNN}(\text{question}) \) > learn weights
- answer = \( \text{softmax}(\text{linear}([i; q])) \)

why isn’t this a good way of doing visual QA?
How many benches are shown?
visual attention

• Use the question representation $q$ to determine where in the image to look

How many benches are shown?
How many benches are shown?

attention over final convolutional layer in network: 196 boxes, captures color and positional information

softmax: predict answer
attention over final convolutional layer in network: 196 boxes, captures color and positional information

softmax: predict answer

how can we compute these attention scores?

How many benches are shown?
hard attention

attention over final convolutional layer in network: 196 boxes, captures color and positional information

How many benches are shown?

we can use reinforcement learning to focus on just one box
Is there a red shape above a circle? yes
Neural nets learn lexical groundings

Is there a red shape above a circle?

[lyyer et al. 2014, Bordes et al. 2014, Yang et al. 2015, Malinowski et al., 2015]

Slide credit: Jacob Andreas
Semantic parsers learn composition

Is there a red shape above a circle?

[Yes]


Slide credit: Jacob Andreas
Neural module networks learn both!

Is there a red shape above a circle?

yes

Slide credit: Jacob Andreas
Neural module networks

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Neural module networks

Is there a red shape above a circle?

Slide credit: Jacob Andreas
Neural module networks

Is there a red shape above a circle?

red
exists
above

true

yes

Slide credit: Jacob Andreas
Sentence meanings are computations

Is there a red shape above a circle?

Slide credit: Jacob Andreas
NLVR$^2$: natural language for visual reasoning! (Suhr et al., 2018)

TRUE OR FALSE: the left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.
Figure 1: A snapshot from an interaction in CEREALBAR. The current instruction is in bold. The large image shows the entire environment. This overhead view is available only to the leader. The follower sees a first-person view only (bottom right). The zoom boxes (top) show the leader and follower.
"Okay, pick up yellow hearts and run past me toward the bush sticking out, on the opposite side is 3 green stars"
Image Captioning

Explain Images with Multimodal Recurrent Neural Networks, Mao et al.
Deep Visual-Semantic Alignments for Generating Image Descriptions, Karpathy and Fei-Fei
Show and Tell: A Neural Image Caption Generator, Vinyals et al.
Long-term Recurrent Convolutional Networks for Visual Recognition and Description, Donahue et al.
Learning a Recurrent Visual Representation for Image Caption Generation, Chen and Zitnick
this is our ImageNet CNN, now used as a feature extractor
this is our ImageNet CNN, now used as a feature extractor
before:

\[ h = \tanh(Wxh \times x + Whh \times h) \]

now:

\[ h = \tanh(Wxh \times x + Whh \times h + Wih \times v) \]

Let's use the image features to create a conditional LM.
test image

sample!

<START>

x0 ← START

straw

hat
<START>

<START>

x0

<STRAT>

straw

hat

y0

y1

y2

h0

h1

h2

sample

<END> token

=> finish.
Image Captioning: Failure Cases

A woman is holding a cat in her hand

A bird is perched on a tree branch

A woman standing on a beach holding a surfboard

A man in a baseball uniform throwing a ball

A person holding a computer mouse on a desk
Image Captioning with Attention

RNN focuses its attention at a different spatial location when generating each word

Image Captioning with Attention

Image Captioning with Attention

A woman is throwing a **frisbee** in a park.

A **dog** is standing on a hardwood floor.

A **stop** sign is on a road with a mountain in the background.

A little **girl** sitting on a bed with a teddy bear.

A group of **people** sitting on a boat in the water.

A giraffe standing in a forest with **trees** in the background.

---

VilBERT (vision and language BERT)

(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

Lu et al., 2019 ("VilBERT")
OpenAI’s CLIP: Contrastive language-image pretraining

- VilBERT and similar methods (e.g. LXMERT) rely on small labeled datasets like MSCOCO and Visual Genome (~100K images each)

- OpenAI collect 400 million (image, text) pairs from the web

- Then, they train an image encoder and a text encoder with a simple contrastive loss: given a collection of images and text, predict which (image, text) pairs actually occurred in the dataset

Radford et al., 2021 ("CLIP")
1. Contrastive pre-training

pepper the aussie pup

Text Encoder

Image Encoder

I_1, I_2, I_3, ..., I_N

T_1, T_2, T_3, ..., T_N

https://openai.com/blog/clip/
Similar to GPT-3, you can use CLIP for zero-shot learning.
<table>
<thead>
<tr>
<th>DATASET</th>
<th>IMAGENET RESNET101</th>
<th>CLIP VIT-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>76.2%</td>
<td>76.2%</td>
</tr>
<tr>
<td>ImageNet V2</td>
<td>64.3%</td>
<td>70.1%</td>
</tr>
<tr>
<td>ImageNet Rendition</td>
<td>37.7%</td>
<td>88.9%</td>
</tr>
<tr>
<td>ObjectNet</td>
<td>32.6%</td>
<td>72.3%</td>
</tr>
<tr>
<td>ImageNet Sketch</td>
<td>25.2%</td>
<td>60.2%</td>
</tr>
<tr>
<td>ImageNet Adversarial</td>
<td>2.7%</td>
<td>77.1%</td>
</tr>
</tbody>
</table>