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

4 5 7 6 6
3 2 8 0 7
6 7 7 1 5
3 0 1 1 1
4 3 2 1 7

0 0 0 0 0
1 0 1
0 0 0

11 2 15
13 8 12

?
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
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

conv+pool

conv+pool

conv

conv

conv

linear

linear

linear+

softmax

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

Deep Neural Network

Input Layer

Hidden Layer 1  Hidden Layer 2  Hidden Layer 3

Output Layer

edges  combinations of edges  object models

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
Transformer encoders for vision
Self-attention on pixels
Self-attention on pixels
Self-attention on pixels

I

(2) (a) Autoregressive

(b) BERT

Issues?
Justin Johnson
March 23, 2022
Lecture 18

Idea #4: Standard Transformer on Patches

Dosovitskiy et al, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021

An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021
Slides from Justin Johnson
N input patches, each of shape 3x16x16
Linear projection to D-dimensional vector

N input patches, each of shape 3x16x16
Add positional embedding: learned D-dim vector per position

Linear projection to D-dimensional vector

N input patches, each of shape 3x16x16
Output vectors

Exact same as NLP Transformer!

Add positional embedding: learned D-dim vector per position

Linear projection to D-dimensional vector

N input patches, each of shape 3x16x16
Transformer

Output vectors

Exact same as NLP Transformer!

Add positional embedding: learned D-dim vector per position

Linear projection to D-dimensional vector

$N$ input patches, each of shape 3x16x16

Linear projection to C-dim vector of predicted class scores

Special extra input: classification token (D dims, learned)
16x16 patches = 16*16*3 = 768d embedding
Vision Transformers (ViT) outperform ResNets with larger datasets.
Okay, so we can encode text with Transformers, and we can encode images with Transformers....

Since the architectures are now basically the same, can we train a single model on both modalities?
OpenAI’s CLIP: Contrastive language-image pretraining

- 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

\[
\begin{array}{cccccc}
T_1 & T_2 & T_3 & \cdots & T_N \\
\end{array}
\]

\[
\begin{array}{cccccc}
I_1 & T_1 & I_1 & T_2 & I_1 & T_3 & \cdots & I_1 & T_N \\
I_2 & T_1 & I_2 & T_2 & I_2 & T_3 & \cdots & I_2 & T_N \\
I_3 & T_1 & I_3 & T_2 & I_3 & T_3 & \cdots & I_3 & T_N \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
I_N & T_1 & I_N & T_2 & I_N & T_3 & \cdots & I_N & T_N \\
\end{array}
\]
Similar to GPT-3, you can use CLIP for zero-shot learning

2. Create dataset classifier from label text

3. Use for zero-shot prediction
<table>
<thead>
<tr>
<th>DATASET</th>
<th>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>
SimVLM: prefix LM for image/text

Wang et al., ICLR 2022 ("SimVLM")
VisualGPT: cross-attention between text decoder and image features

Chen et al., CVPR 2022 ("VisualGPT")
LAION-5B: a dataset of 5 billion image/text pairs!

1. Feed in Common Crawl
2. Webpage Filtering
3. Download Image-Text Pairs
4. Content Filtering
5. Store Data

Schuhmann et al., 2022
Major copyright issues…

Stable Diffusion and other image-generating AI products could not exist without the work of painters, illustrators, photographers, sculptors, and other artists. Stable Diffusion was trained on the LAION-5B dataset. LAION-5B contains 5.85 billion image-text pairs. Most of the images contained in the dataset are copyrighted, and LAION claims no ownership in them. As it notes, “The images are under their copyright.”

On January 13, 2023, the Joseph Saveri Law Firm, LLP filed a complaint in the U.S. District Court for the Northern District of California on behalf of Sarah Andersen, Kelly McKernan, Karla Ortiz, and a class of other artists and stakeholders against Stability AI Ltd.; Stability AI, Inc.; DeviantArt, Inc.; and Midjourney, Inc. This suit alleges copyright infringement, DMCA violations, right of publicity violations, breach of the DeviantArt Terms of Service, unfair competition, and unjust enrichment. It likewise seeks damages and injunctive relief to compensate the class for harms already incurred and to prevent future harms.
Relatedly…

Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems

The internet site has long been a forum for discussion on a huge variety of topics, and companies like Google and OpenAI have been using it in their A.I. projects.

PaLM-E: An Embodied Multimodal Language Model

Given \texttt{<emb>} ... \texttt{<img>} Q: How to grasp blue block? A: First, grasp yellow block

Large Language Model (PaLM)

Control

A: First, grasp yellow block and ...
PaLM-E

https://palm-e.github.io/#demo