Learning from vision and language
Zero-shot learning

• Question: can we teach a machine to recognize classes based on just textual descriptions?

An okapi is like a horse but with stripes on its legs

Gotcha

Okapi!
Zero-shot learning – Training the learner

Descr 1
Descr 2
Descr 3
Descr 4

Training

Zero Shot Learner
Zero-shot learning – training the learner

• General approach: align embeddings of images and words
Zero-shot learning

• Typically zero-shot learning techniques use attribute descriptions of classes
• But can use textual descriptions as well
• Need a text encoder
Weak supervision

• Zero-shot learning is typically performed on a particular domain
• Can we do “generic” zero-shot learning?
The world of internet images and captions

The world of internet images and captions

Vision-language pre-training

(1) Contrastive pre-training

Pepper the aussie pup

Text Encoder

<table>
<thead>
<tr>
<th>I₁</th>
<th>I₁·T₁</th>
<th>I₁·T₂</th>
<th>I₁·T₃</th>
<th>...</th>
<th>I₁·Tₜ</th>
</tr>
</thead>
<tbody>
<tr>
<td>I₂</td>
<td>I₂·T₁</td>
<td>I₂·T₂</td>
<td>I₂·T₃</td>
<td>...</td>
<td>I₂·Tₜ</td>
</tr>
<tr>
<td>I₃</td>
<td>I₃·T₁</td>
<td>I₃·T₂</td>
<td>I₃·T₃</td>
<td>...</td>
<td>I₃·Tₜ</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Iₙ</td>
<td>Iₙ·T₁</td>
<td>Iₙ·T₂</td>
<td>Iₙ·T₃</td>
<td>...</td>
<td>Iₙ·Tₜ</td>
</tr>
</tbody>
</table>

Image Encoder

(2) Create dataset classifier from label text

plane

<table>
<thead>
<tr>
<th>car</th>
</tr>
</thead>
<tbody>
<tr>
<td>dog</td>
</tr>
</tbody>
</table>

| A photo of an {object}. |

Text Encoder

(3) Use for zero-shot prediction

A photo of a dog.

Image Encoder

<table>
<thead>
<tr>
<th>I₁</th>
<th>I₁·T₁</th>
<th>I₁·T₂</th>
<th>I₁·T₃</th>
<th>...</th>
<th>I₁·Tₜ</th>
</tr>
</thead>
<tbody>
<tr>
<td>I₂</td>
<td>I₂·T₁</td>
<td>I₂·T₂</td>
<td>I₂·T₃</td>
<td>...</td>
<td>I₂·Tₜ</td>
</tr>
<tr>
<td>I₃</td>
<td>I₃·T₁</td>
<td>I₃·T₂</td>
<td>I₃·T₃</td>
<td>...</td>
<td>I₃·Tₜ</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>Iₙ</td>
<td>Iₙ·T₁</td>
<td>Iₙ·T₂</td>
<td>Iₙ·T₃</td>
<td>...</td>
<td>Iₙ·Tₜ</td>
</tr>
</tbody>
</table>

(1) Contrastive pre-training

- Vision-language pre-training - CLIP

- # image_encoder - ResNet or Vision Transformer
- # text_encoder - CBOW or Text Transformer
- # I[n, h, w, c] - minibatch of aligned images
- # T[n, l] - minibatch of aligned texts
- # W_i[d_i, d_e] - learned proj of image to embed
- # W_t[d_t, d_e] - learned proj of text to embed
- # t - learned temperature parameter

- # extract feature representations of each modality
  I_f = image_encoder(I) # [n, d_i]
  T_f = text_encoder(T) # [n, d_t]

- # joint multimodal embedding [n, d_e]
  I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
  T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

- # scaled pairwise cosine similarities [n, n]
  logits = np.dot(I_e, T_e.T) * np.exp(t)

- # symmetric loss function
  labels = np.arange(n)
  loss_i = cross_entropy_loss(logits, labels, axis=0)
  loss_t = cross_entropy_loss(logits, labels, axis=1)
  loss = (loss_i + loss_t)/2
Vision-language pre-training

Using vision-language embeddings

Trained on COCO Segmentation

GLIP: Vision-language for object detection

Vision and Communication with Language
Vision, Language and Embodiment

• Embodied agents have to communicate
• Humans must be able to communicate with machines
• Multimodal input provides strong supervisory signal
The many flavors of vision and language tasks

Grounding
- Zero-shot learning
- Visual question answering

Image captioning
Image captioning - The task

A group of young men playing soccer.
Image captioning - why?

• Alt-text for visually impaired
• Test for true understanding?
Image captioning - evaluation

• Given computer-generated caption and human caption, compute match
• BLEU from machine translation community
• Computes (modified) n-gram precision

Reference: A group of people playing soccer
Candidate: People playing baseball.
BLEU: 1/3
Image captioning


Attention (Transformers)

• Comes from the NLP community
• Is an approach for processing sets

Attention (Transformers)

- Comes from the NLP community
- Is an approach for processing sets
Attention (Transformers)

- Comes from the NLP community
- Is an approach for processing sets
Attention (Transformers)

Input

Queries

Keys

Values

Output

Attention
Attention (Transformers) for Encoding sequences

Attention for Outputing Sequences

Attention for Outputing Sequences

Attention for Outputing Sequences

Modern image captioning - BLIP

Evaluation Metrics

Slide credit: Larry Zitnick
Evaluation Metrics

Human captions

Slide credit: Larry Zitnick
A man riding a wave on a surfboard in the water.

Slide credit: Larry Zitnick
A man riding a wave on a surfboard in the water.

“surfboard”

Slide credit: Larry Zitnick
The post-captioning world

- Captioning is hard to evaluate!
  - Frame task so that it is easy to evaluate objectively

- Datasets are biased!
  - Control dataset bias
I'm going to crush the rebellion... but first, let me take a selfie. #captionbot

I am not really confident, but I think it's a man taking a selfie in front of a building.
Reasoning

• Want vision systems to reason about what is going on
  • Identify objects and scenes
  • Identify relationships between objects
  • Understand physics of the world
  • Understand social interactions, intent etc.
  • Incorporate knowledge: common sense, pop culture, ...
Visual Question Answering

- Direct motivation: assistive technology
- Indirect motivation: sandbox for reasoning

“We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (e.g., kitchen, beach), people’s expressions and poses, and properties of objects (e.g., color of objects, their texture). Your task is to stump this smart robot! Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image.”
Methods for VQA

How many bikes are there?
Methods for VQA

How many bikes are there?

Compositional reasoning

What is the color of the kitten to the left of the blue kitten?
Compositional reasoning

What is the color of the kitten to the left of the blue kitten?
Compositional reasoning

- Look left
- Triangle detector
- Blue detector
- And
- Get color

Look left
- Triangle detector
- Get color

Red
What is the color of the kitten to the left of the blue kitten?
What is the color of the kitten to the left of the blue kitten?
Compositional reasoning

• How do we learn a mapping from language to trees?
  • Problem: semantic parsing
  • Option 1: Syntactic parsing
  • Option 2: Use supervision

Neural module networks. Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Dan Klein. CVPR 2016
Learning to reason: End-to-end module networks for visual question answering. Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko. ICCV 2017
Inferring and Executing Programs for Visual Reasoning
Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick. ICCV, 2017
Compositional reasoning

Question: Are there more cubes than yellow things?

Answer: Yes

Program Generator

Predicted Program

Execution Engine

Classifier
The problem with VQA

• Dataset biases allow cheating
  • Only-question Bag-of-Words: 53.7% (vs ~65% for state-of-the-art)

• Require common sense to answer
  • “What is the moustache made of?”

• Hard to diagnose error
  • Is the problem understanding the question?
  • Or understanding the image?
Clever Hans