Multimodal Learning

Danna Gurari
University of Colorado Boulder
Spring 2024

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2024/
Review

• Last lecture
  • Motivation
  • Foundation Models
  • NLP: Textual Prompting
  • CV: Visual Prompting (In-context Few-shot Learning)
  • Parameter-Efficient Tuning Methods
  • Latex Tutorial

• Assignments (Canvas)
  • Problem set 4 due earlier today
  • Final project outline due in 3 weeks

• Questions?
Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
Simultaneously Use 2+ Modalities

To date, most work focuses on the intersection of CV + NLP; e.g.,

**Caption:**
A bunch of small light brown mushrooms in a green field.

**Answer Visual Question:**

**Q:** Is it edible or poisonous?

**A:** Poisonous
Visual Assistance for People with Vision Loss; e.g.,

- Seeing AI
- Lookout by Google
- be my eyes
- TapTapSee
- ORCAM
- envision
Visual Assistance for People with Vision Loss; e.g., Microsoft Power Point (Office 365 demo)
Visual Assistance for People with Vision Loss

https://www.youtube.com/watch?v=cUSEFnZGlzY
Describing and Responding to Images Posted to Social Media with “Personality”

**Standard captioning output:** A plate with a sandwich and salad on it.

**Our model with different personality traits (215 possible traits, not all shown here):**

*Sweet* That is a lovely sandwich.

*Dramatic* This sandwich looks so delicious! My goodness!

*Anxious* I’m afraid this might make me sick if I eat it.

*Sympathetic* I feel so bad for that carrot, about to be consumed.

*Arrogant* I make better food than this

*Optimistic* It will taste positively wonderful!

*Money-minded* I would totally pay $100 for this plate.
Describing Products

Title: Stand Collar A-Line Dress

Fashion Caption: A pearly button accents the stand collar that gives this so-simple, yet so-chic A-line dress its retro flair

Color: Black and ivory

Meta: - 33" petite length (size 8P) - Hidden back-zip closure - Stand collar - Cap sleeves - Side-seam pockets – A-Lined - 63% polyester, 34% rayon, 3% spandex - Dry clean or hand wash, dry flat - Imported – Dress

Image Caption: A person in a dress

Yang et al. Fashion Captioning: Towards Generating Accurate Descriptions with Semantic Rewards. ECCV 2020
Medical VQA

(a) Q: what imaging method was used? A: us-d - doppler ultrasound

(b) Q: which plane is the image shown in? A: axial

(e) Q: what abnormality is seen in the image? A: nodular opacity on the left #metastastic melanoma
Education (e.g., for Preschoolers)

Answers questions about **quantity** and **colors** of detected objects

Audio Guide for Museums and Art Galleries

TEXTUAL DESCRIPTION
Bathers at Asnieres is an oil-on-canvas painting by the French artist Georges Pierre Seurat, the first of his two masterpieces on the monumental scale...

QUESTION
Who is the author of this painting?

IMAGE

Today’s Topics

• Multimodal applications

• **Image captioning dataset challenges**

• Image captioning algorithms

• **Visual question answering dataset challenges**

• Visual question answering algorithms

• Foundation models
### Sample of Existing Dataset Challenges

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conceptual Captions</th>
<th>Fashion Captioning</th>
<th>CUB-200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>COCO</strong></td>
<td>Woman on a horse jumping over a pole jump.</td>
<td>A glass bowl contains peeled tangerines and cut strawberries.</td>
<td>A person is holding a small container of cream upside down.</td>
</tr>
<tr>
<td><strong>VizWiz</strong></td>
<td>The billboard displays ‘Welcome to Yakima The Palm Springs of Washington’.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TextCaps</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Sample of Existing Dataset Challenges

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Domain</th>
<th>Nb. Images</th>
<th>Nb. Caps (per Image)</th>
<th>Vocab Size</th>
<th>Nb. Words (per Cap.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO [128]</td>
<td>Generic</td>
<td>132K</td>
<td>5</td>
<td>27K (10K)</td>
<td>10.5</td>
</tr>
<tr>
<td>Flickr30K [129]</td>
<td>Generic</td>
<td>31K</td>
<td>5</td>
<td>18K (7K)</td>
<td>12.4</td>
</tr>
<tr>
<td>Flickr8K [19]</td>
<td>Generic</td>
<td>8K</td>
<td>5</td>
<td>8K (3K)</td>
<td>10.9</td>
</tr>
<tr>
<td>CC3M [130]</td>
<td>Generic</td>
<td>3.3M</td>
<td>1</td>
<td>48K (25K)</td>
<td>10.3</td>
</tr>
<tr>
<td>CC12M [131]</td>
<td>Generic</td>
<td>12.4M</td>
<td>1</td>
<td>523K (163K)</td>
<td>20.0</td>
</tr>
<tr>
<td>SBU Captions [4]</td>
<td>Generic</td>
<td>1M</td>
<td>1</td>
<td>238K (46K)</td>
<td>12.1</td>
</tr>
<tr>
<td>VizWiz [132]</td>
<td>Assistive</td>
<td>70K</td>
<td>5</td>
<td>20K (8K)</td>
<td>13.0</td>
</tr>
<tr>
<td>CUB-200 [133]</td>
<td>Birds</td>
<td>12K</td>
<td>10</td>
<td>6K (2K)</td>
<td>15.2</td>
</tr>
<tr>
<td>Oxford-102 [133]</td>
<td>Flowers</td>
<td>8K</td>
<td>10</td>
<td>5K (2K)</td>
<td>14.1</td>
</tr>
<tr>
<td>Fashion Cap. [134]</td>
<td>Fashion</td>
<td>130K</td>
<td>1</td>
<td>17K (16K)</td>
<td>21.0</td>
</tr>
<tr>
<td>BreakingNews [135]</td>
<td>News</td>
<td>115K</td>
<td>1</td>
<td>85K (10K)</td>
<td>28.1</td>
</tr>
<tr>
<td>GoodNews [136]</td>
<td>News</td>
<td>466K</td>
<td>1</td>
<td>192K (54K)</td>
<td>18.2</td>
</tr>
<tr>
<td>TextCaps [137]</td>
<td>OCR</td>
<td>28K</td>
<td>5/6</td>
<td>44K (13K)</td>
<td>12.4</td>
</tr>
<tr>
<td>Loc. Narratives [138]</td>
<td>Generic</td>
<td>849K</td>
<td>1/5</td>
<td>16K (7K)</td>
<td>41.8</td>
</tr>
</tbody>
</table>
Class Task: How Would You Describe This Image?

Form: https://forms.gle/Nbue5HcdP9Dib8Co8
Guidelines and Examples:

Read these guidelines carefully. You must write exactly two sentences.

1. Describe the action being performed and mention the person performing the action and all objects involved in the action.
2. Describe any objects in the image that are not directly involved in the action.

A man is reading a newspaper. It is cloudy and there are skyscrapers in the background.
A boy is typing on a laptop. There is a brown bookshelf behind him and a bright window.
A man is talking on the telephone. There is a red lampshade and three red chairs in the background.
Flickr8K and 30K

Guidelines:
- You must describe each of the following five images with one sentence.
- Please provide an accurate description of the activities, people, animals and objects you see depicted in the image.
- Each description must be a single sentence under 100 characters. Try to be concise.
- Please pay attention to grammar and spelling.
- We will accept your results if you provide a good description for all five images, leaving nothing blank.

Examples of good and bad descriptions.

(1) The dog is wearing a red sombrero.
   Very Good: This describes the two main objects concisely and accurately.

(2) White dog wearing a red hat.
   Good: Incomplete sentences like this are fine.

(3) The white dog is wearing a pink collar.
   Okay: This describes the dog, but it ignores the hat.

(4) The red hat is adorned with gold sequins.
   Bad: This ignores the dog.

(5) The dog is angry because he is hungry.
   Bad: This is speculation.

(6) The dog.
   Very Bad: This could describe any image of any dog.
Instructions:
- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Please describe the image:

Enter description here

Step 1: Please describe the image in one sentence.

- Describe all parts of the image that may be important to a person who is blind. E.g., imagine how you would describe this image on the phone to a friend.
- DO NOT speculate about what people in the image might be saying or thinking.
- DO NOT describe things that may have happened in the future or past.
- DO NOT use more than one sentence.
- If text is in the image, and is important, then you can summarize what it says. DO NOT use all the specific phrases that you see in the image.

- DO NOT describe the image quality issues. This is covered in Step 3. If the image quality issues make it impossible to recognize the visual content (e.g., image is totally black or white), then use the following description (you can copy-paste):

  Quality issues are too severe to recognize visual content. Copy to description

- Your description should contain at least 8 words.

Type here. Do not start the description with:
- "There is/are ..."
- "This is / These are ..."
- "The/This image/picture ...
- "It is/ It's ..."
Personality-Captions

215 personalities selected from this list: http://ideonomy.mit.edu/essays/traits.html

Comment on an Image

Description
In this task, you will be shown 5 images, and will write a comment about each image. The goal of this task is to write something about an image that someone else would find engaging.

STEP 1
With each new photo, you will be given a personality trait that you will try to emulate in your comment. For example, you might be given "snarky" or "glamorous". The personality describes YOU, not the picture. It is you who is snarky or glamorous, not the contents of the image.

STEP 2
You will then be shown an image, for which you will write a comment in the context of your given personality trait. Please make sure your comment has at least three words. Note that these are comments, not captions.

E.g., you may be shown an image of a tree. If you are "snarky", you might write "What a boring tree, I bet it has bad wood;" or, if you were "glamorous", you might write "What an absolutely beautiful tree! I would put this in my living room it’s so extravagant!"

Your assigned personality is:

Adventurous

Reminder - please do not write anything that involves any level of discrimination, racism, sexism and offensive religious/political comments, otherwise the submission will be rejected.

How Would You Evaluate Captions from an Algorithm?

<table>
<thead>
<tr>
<th>FEATURE NAME:</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>{ &quot;tags&quot;: [ &quot;outdoor&quot;, &quot;giraffe&quot;, &quot;animal&quot;, &quot;mammal&quot;, &quot;standing&quot;, &quot;field&quot;, &quot;top&quot;, &quot;branch&quot;, &quot;bird&quot;, &quot;eating&quot;, &quot;head&quot;, &quot;grazing&quot;, &quot;neck&quot;, &quot;water&quot;, &quot;large&quot;, &quot;man&quot;, &quot;grassy&quot;, &quot;tall&quot;, &quot;group&quot;, &quot;dirt&quot;, &quot;zoo&quot; ], &quot;captions&quot;: [ { &quot;text&quot;: &quot;a giraffe standing in the dirt&quot;, &quot;confidence&quot;: 0.982929349 } ] }</td>
</tr>
</tbody>
</table>
Evaluation: Human Judgments

Evaluation: Automated

• BLEU
• METEOR
• Rouge
• CIDEr
• SPICE
Evaluation: Automated

- **BLEU**
  Idea: compute similarities of n-grams between a predicted caption and each ground truth caption

- **METEOR**

- **Rouge**

- **CIDER**

- **SPICE**

  [Diagram of n-grams with examples for N=1, N=2, and N=3]

  http://recognize-speech.com/language-model/n-gram-model/comparison
Evaluation: Automated

- **BLEU**
- **METEOR**
- **Rouge**
- **CIDEr**
- **SPICE**

Idea: measure similarity of a predicted caption to how most people describe an image based on $n$-grams unique to the image

---

Evaluation: Automated

• BLEU
• METEOR
• Rouge
• CIDEr
• SPICE

What content do most people describe in this image?

A cow is standing in a field.
A cow with horns and long hair covering its face stands in a field.
A cow with hair over its eyes stands in a field.
This horned creature is getting his picture taken.
A furry animal with horns roams on the range.

Evaluation: Automated

• BLEU
  Do you think these two captions describe the same image?

• METEOR
  (a) A young girl standing on top of a tennis court.
  (b) A giraffe standing on top of a green field.

• Rouge

• CIDEr

• SPICE

Evaluation: Automated

- BLEU
  Problem: n-gram methods scores these as very similar
  (a) A young girl *standing on top of a tennis court.*
  (b) A giraffe *standing on top of a green field.*

- METEOR

- Rouge

- CIDEr

- SPICE

Evaluation: Automated

• BLEU

Do you think these two captions describe the same image?

• METEOR

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

• Rouge

• CIDEr

• SPICE

Evaluation: Automated

• BLEU

Problem: n-gram methods scores these as very different

• METEOR

(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

• Rouge

• CIDEr

• SPICE

Evaluation: Automated

Idea: compare scene graph of prediction to scene graph of ground truth

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

What is the meaningful semantic content in these captions?

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

Meaningful semantic content in these captions:

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
Historical Context

- **1945**: First programmable machine
- **1950**: Turing test
- **1956-1957**: AI
- **1959**: Perception
- **1960**: Machine learning

- **1980**: Neocognition
- **1986**: Effective learning strategy
- **1989**: Backpropagation for CNNs
- **1998**: MNIST, LeNet
- **2012**: Wave 3: rise of “deep learning”
- **2020**: Oscar
- **2021**: VinVL, CLIP
Oscar: Transformer Design

Pre-Training
- Large unlabelled datasets (e.g. Wikipedia, BookCorpus)
- Self-supervised training (hours to days)

Fine-Tuning
- Smaller labelled datasets (SQuAD, MNLI/CMNLI, Similarity)
- Task-specific fine tuning (minutes to hours)

Inference

Novelty: Adds **Explicit** Alignment Between Visual and Textual Concepts

- **Idea**: rather than have algorithm learn alignment between text and features describing image regions, **align them explicitly**

- **Motivating observations**: often, salient objects are mentioned in image descriptions and can be located by object detection algorithms
Each image is represented as description of objects detected with Faster R-CNN using features from Faster R-CNN.

Oscar: Architecture

Novelty is to incorporate tags predicted by Faster R-CNN

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Oscar: Architecture

Uses BERT architecture, initialized with pretrained BERT weights

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Oscar: 2 Pretraining Tasks
(Masked Token Loss and Contrastive Loss)

Like BERT, predict randomly masked tokens based on surrounding words, tags, and image information.

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Oscar: 2 Pretraining Tasks
(Masked Token Loss and **Contrastive Loss**)

Fully-connected layer added to enable binary classification based on the fused vision-language token representation

Goal is to determine whether tags are original when 50% of tags are replaced with randomly selected tag sequence in the dataset
Oscar: 2 Pretraining Dataset

6.5 million text-tag-image triplets derived from existing V+L datasets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Oscar: Transformer Design

Pre-Training
- Large unlabelled datasets (e.g. Wikipedia, BookCorpus)
- Self-supervised training (hours to days)

Fine-Tuning
- Smaller labelled datasets (SQuAD, MNLI/CMNLI, Similarity)
- Task-specific fine tuning (minutes to hours)

Inference

Oscar: 2 Fine-Tuning Task (Masked Token Loss)

Similar to pre-training, predict randomly masked tokens based on surrounding words, tags, and image information (on COCO dataset)

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Repeatedly predict a new [MASK] token, incorporating the predicted word into the sequence, until [STOP] is predicted.

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
Historical Context

1945  Turing test
1950  First programmable machine
1956-1957  AI
1959  Perceptron
1959  Machine learning
1980  Neocognition
1986  Neural networks with effective learning strategy
1989  Backpropagation for CNNs
1989  MNIST, LeNet
1998  Wave 3: rise of “deep learning”
2012  Oscar
2020-2021  VinVL, CLIP
Idea: Oscar + Improved Visual Representation
VinVL Architecture: Oscar + New Object Detector

Improved object detector to predict more diverse categories and train larger models on larger datasets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
VinVL: 2 Pretraining Tasks
(Masked Token Loss and **Contrastive Loss**)

- **Fully-connected layer added to enable 3-way classification**
- **based on the fused vision-language token representation**

---

**Features**

- [CLS]
- A **dog** is [MASK] on a **couch** [SEP]

**Network**

- Multi-Layer Transformers

**Embeddings**

- Word Tokens
- Object Tags
- Region Features

**Data**

- Trained on 8.85 million text-image pairs to decide whether either captions or answers are corrupted (50% are not) for caption-tags-image triplets and question-answer-image triplets

Li et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ECCV 2020
VinVL: Influence of Model and Dataset Sizes

200M images, each with 1 alt text description, collected from Internet

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022
VinVL: Influence of Model and Dataset Sizes

8 model sizes tested on COCO dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th>Width</th>
<th>MLP</th>
<th>Heads</th>
<th>Param (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tiny</td>
<td>6</td>
<td>256</td>
<td>1024</td>
<td>4</td>
<td>13.4</td>
</tr>
<tr>
<td>tiny12</td>
<td>12</td>
<td>256</td>
<td>1024</td>
<td>4</td>
<td>18.1</td>
</tr>
<tr>
<td>small</td>
<td>12</td>
<td>384</td>
<td>1536</td>
<td>6</td>
<td>34.3</td>
</tr>
<tr>
<td>small24</td>
<td>24</td>
<td>384</td>
<td>1536</td>
<td>6</td>
<td>55.3</td>
</tr>
<tr>
<td>base</td>
<td>12</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>111.7</td>
</tr>
<tr>
<td>base24</td>
<td>24</td>
<td>768</td>
<td>3072</td>
<td>12</td>
<td>196.7</td>
</tr>
<tr>
<td>large</td>
<td>24</td>
<td>1024</td>
<td>4096</td>
<td>16</td>
<td>338.3</td>
</tr>
<tr>
<td>huge</td>
<td>32</td>
<td>1280</td>
<td>5120</td>
<td>16</td>
<td>675.4</td>
</tr>
</tbody>
</table>

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022
VinVL: Influence of Model and Dataset Sizes

What trend(s) do you observe?

The trends of improved performance for large models and training datasets is generally observed for transformers.

Hu et al. Scaling Up Vision-Language Pre-training for Image Captioning. CVPR 2022
Historical Context

1945  Turing test
1950  First programmable machine
1956  AI
1957  Perceptron
1959  Machine learning
1980  Neural networks with effective learning strategy
1986  Backpropagation for CNNs
1989  MNIST, LeNet
1998  Wave 3: rise of “deep learning”
2012  Oscar
2020  VinVL
2021  CLIP
Why CLIP?

Named after the proposed technique: **Contrastive Language Image Pre-training**

CLIP Model: Novelty

• Train image analysis models with natural language supervision using the vast amounts of publicly available data on the Internet
CLIP Architecture

Tried 8 variants: 3 ViT & 5 ResNet
CLIP Training

Task: predict which image-text pairs match using 400 million image-text pairs from Internet containing any of 500,000 queries (e.g., words occurring 100+ times in English version of Wikipedia and all WordNet synonyms).

- Largest ResNet model took 18 days to train on 592 V100 GPUs and largest ViT took 12 days on 256 V100 GPUs.

- Experiments run with largest (“best”) ViT model.

Tried 8 variants: 3 ViT & 5 ResNet.

CLIP Training

- Learns feature embeddings for image and text encoders that push correct image-text pairs together and incorrect image-text pairs apart.

- Learns nouns, verbs, adjectives, and more!

Tried 8 variants: 3 ViT & 5 ResNet
Zero-Shot Performance Evaluated on Over 30 Datasets
CLIP Inference

**e.g., zero-shot classification:**

1. Compute feature embedding for names of all classes in the dataset by its encoder
2. Compute feature embedding of the image
3. Compute cosine similarity of each (image, text) pair embedding followed by softmax to identify most probable match

https://towardsdatascience.com/understanding-zero-shot-learning-making-ml-more-human-4653ac35ccab
Prompts “engineered” to mimic that training data often had sentences (instead of words):
- classification: “A photo of a {label}”
- fine-grained classification: “A photo of a {label}, a type of pet/food/aircraft/etc”
- satellite image classification: “A satellite photo of a {label}”
- ensembles: “A photo of a big/small/etc {label}”
## CLIP Evaluation

### Subset of datasets shown here:

Classification evaluation spanned fine-grained classification (e.g., food, bird, aircraft, and car categories), distribution shifts for ImageNet categories (e.g., corrupted images), and more

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classes</th>
<th>Train size</th>
<th>Test size</th>
<th>Evaluation metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food-101</td>
<td>102</td>
<td>75,750</td>
<td>25,250</td>
<td>accuracy</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>10</td>
<td>50,000</td>
<td>10,000</td>
<td>accuracy</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>100</td>
<td>50,000</td>
<td>10,000</td>
<td>accuracy</td>
</tr>
<tr>
<td>Birdsnap</td>
<td>500</td>
<td>42,283</td>
<td>2,149</td>
<td>accuracy</td>
</tr>
<tr>
<td>SUN397</td>
<td>397</td>
<td>19,850</td>
<td>19,850</td>
<td>accuracy</td>
</tr>
<tr>
<td>Stanford Cars</td>
<td>196</td>
<td>8,144</td>
<td>8,041</td>
<td>accuracy</td>
</tr>
<tr>
<td>FGVC Aircraft</td>
<td>100</td>
<td>6,667</td>
<td>3,333</td>
<td>mean per class</td>
</tr>
<tr>
<td>Pascal VOC 2007 Classification</td>
<td>20</td>
<td>5,011</td>
<td>4,952</td>
<td>11-point mAP</td>
</tr>
<tr>
<td>Describable Textures</td>
<td>47</td>
<td>3,760</td>
<td>1,880</td>
<td>accuracy</td>
</tr>
<tr>
<td>Oxford-IIIIT Pets</td>
<td>37</td>
<td>3,680</td>
<td>3,669</td>
<td>mean per class</td>
</tr>
<tr>
<td>Caltech-101</td>
<td>102</td>
<td>3,060</td>
<td>6,085</td>
<td>mean-per-class</td>
</tr>
<tr>
<td>Oxford Flowers 102</td>
<td>102</td>
<td>2,040</td>
<td>6,149</td>
<td>mean per class</td>
</tr>
<tr>
<td>MNIST</td>
<td>10</td>
<td>60,000</td>
<td>10,000</td>
<td>accuracy</td>
</tr>
<tr>
<td>Facial Emotion Recognition 2013</td>
<td>8</td>
<td>32,140</td>
<td>3,574</td>
<td>accuracy</td>
</tr>
<tr>
<td>STL-10</td>
<td>10</td>
<td>1000</td>
<td>8000</td>
<td>accuracy</td>
</tr>
<tr>
<td>EuroSAT</td>
<td>10</td>
<td>10,000</td>
<td>5,000</td>
<td>accuracy</td>
</tr>
<tr>
<td>RESISC45</td>
<td>45</td>
<td>3,150</td>
<td>25,200</td>
<td>accuracy</td>
</tr>
<tr>
<td>GTSRB</td>
<td>43</td>
<td>26,640</td>
<td>12,630</td>
<td>accuracy</td>
</tr>
<tr>
<td>KITTI</td>
<td>4</td>
<td>6,770</td>
<td>711</td>
<td>accuracy</td>
</tr>
<tr>
<td>Country211</td>
<td>211</td>
<td>43,200</td>
<td>21,100</td>
<td>accuracy</td>
</tr>
<tr>
<td>PatchCamelyon</td>
<td>2</td>
<td>294,912</td>
<td>32,768</td>
<td>accuracy</td>
</tr>
<tr>
<td>UCF101</td>
<td>101</td>
<td>9,537</td>
<td>1,794</td>
<td>accuracy</td>
</tr>
<tr>
<td>Kinetics700</td>
<td>700</td>
<td>494,801</td>
<td>31,669</td>
<td>mean(top1, top5)</td>
</tr>
<tr>
<td>CLEVR Counts</td>
<td>8</td>
<td>2,000</td>
<td>500</td>
<td>accuracy</td>
</tr>
<tr>
<td>Hateful Memes</td>
<td>2</td>
<td>8,500</td>
<td>500</td>
<td>ROC AUC</td>
</tr>
<tr>
<td>Rendered SST2</td>
<td>2</td>
<td>7,792</td>
<td>1,821</td>
<td>accuracy</td>
</tr>
<tr>
<td>ImageNet</td>
<td>1000</td>
<td>1,281,167</td>
<td>50,000</td>
<td>accuracy</td>
</tr>
</tbody>
</table>
CLIP: Qualitative Results

**Food101**
- Correct label: guacamole
- Correct rank: 1/101
- Correct probability: 90.15%

- A photo of guacamole, a type of food.
- A photo of ceviche, a type of food.
- A photo of edamame, a type of food.
- A photo of tuna tartare, a type of food.
- A photo of hummus, a type of food.

**SUN397**
- Correct label: television studio
- Correct rank: 1/397
- Correct probability: 90.22%

- A photo of a television studio.
- A photo of a podium indoor.
- A photo of a conference room.
- A photo of a lecture room.
- A photo of a control room.

CLIP: Qualitative Results

CLIP: Qualitative Results

Oxford-IIIT Pets

- correct label: Maine Coon
- correct rank: 1/37
- correct probability: 99.99%

- a photo of a maine coon, a type of pet.
- a photo of a persian, a type of pet.
- a photo of a ragdoll, a type of pet.
- a photo of a birman, a type of pet.
- a photo of a siamese, a type of pet.

FGVC Aircraft

- correct label: Boeing 717
- correct rank: 2/100
- correct probability: 9.91%

- a photo of a mcdonnell douglas md-90, a type of aircraft.
- a photo of a boeing 717, a type of aircraft.
- a photo of a fokker 100, a type of aircraft.
- a photo of a mcdonnell douglas dc-9-30, a type of aircraft.
- a photo of a boeing 727-200, a type of aircraft.

CLIP: Qualitative Results

CLIP: Qualitative Results

Hateful Memes

Correct label: meme
Correct rank: 1/2  Correct probability: 99.20%

Get the jumper cables

German Traffic Sign Recognition Benchmark (GTSRB)

Correct label: red and white triangle with exclamation mark warning
Correct rank: 1/43  Correct probability: 45.75%

Image of a red and white triangle with exclamation mark warning traffic sign.
Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
Status Quo (Approach to Create 14+ Datasets)

1. Category Selection
   - Person
   - Car
   - N. Apple

2. Image Collection
   - flickr™
   - Google™

3. Question Collection
   - OR
   - Contrived Questions

Constrained set of concepts
Pre-qualified images (quality, privacy)
Contrived Questions
e.g., Question Generation
e.g., Answer Generation

10 answers collected from 10 crowdworkers

Mainstream VQA Challenge (held for 6 years)

Visual Question Answering and Dialog Workshop

Location: Seaside Ballroom B, Long Beach Convention & Entertainment Center

at CVPR 2019, June 17, Long Beach, California, USA

What is the mustache made of?

AI System

bananas

Two zebras are walking around their pen at the zoo.

Q1: Any people in the shot?
A1: No, there aren’t any.

Q2: Any other animal?
A2: No, just zebras.

https://visualqa.org/workshop.html
VizWiz: Authentic Use Case
VizWiz: Authentic Use Case

Image

VizWiz: Authentic Use Case

VizWiz: Authentic Use Case

Image + Question

VizWiz: Authentic Use Case
VizWiz: Authentic Use Case

Users agreed to share 44,799 (62%) of requests for dataset creation

VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions (removes voice)

2. Re-save images (removes metadata)
VizWiz: Authentic Use Case

**Anonymization**

1. Transcribe questions
   ![Microphone](image1.png) ➔ ![Text](image2.png)

2. Re-save images
   ![Floppy Disk](image3.png)

**In-House Filtering**

(personally identifying information)

![Images](image4.png)
VizWiz: Authentic Use Case

Anonymization
1. Transcribe questions
2. Re-save images

In-House Filtering

Data Labeling
(high quality answers)
VizWiz: Authentic Use Case

**VQA:** 32,842 image/question pairs $\rightarrow$ 328,420 answers
VizWiz-VQA Grand Challenge (5th year in 2023)

2022 VizWiz Grand Challenge Workshop

Visual Question Answering

- Q: Does this foundation have any sunscreen? A: yes
- Q: What is this? A: 10 euros
- Q: What color is this? A: green
- Q: What type of pills are these? A: unsuitable image
- Q: What type of soap is this? A: unsuitable image
- Q: Who is this mail for? A: unsuitable image

Answering Grounding

- Q: What is this? A: dog
- Q: What does the package say? A: burrito
- Q: What is this? A: crystal
- Q: How many tablets in this box? A: 8

Few-Shot Object Recognition

- 67 Blind and low-vision collectors
- 486 Objects
- 3,822 Videos
- 2,687,934 Frames

https://vizwiz.org
Key Difference of Real-World Use Case from Status Quo: VQs Can Be Unanswerable!

**Q:** What is the expiration date?
**A:** unanswerable

**Q:** What is this a gift card for?
**A:** unanswerable

**Q:** What temperature is the dial set to?
**A:** unanswerable

[Gurari et al. CVPR 2018]
Does this picture look scary?

Hi there can you please tell me what flavor this is?

Which side of the room is the toilet on?

Is my monitor on?
Crowdsourced Answers

Is my monitor on?
(1) yes
(2) yes
(3) yes
(4) yes
(5) yes
(6) yes
(7) yes
(8) yes
(9) yes
(10) yes

Hi there can you please tell me what flavor this is?
(1) sweet pepper
(2) sweet pepper
(3) sweet pepper
(4) sweet pepper
(5) sweet pepper
(6) sweet pepper
(7) sweet pepper
(8) sweet pepper
(9) sweet pepper
(10) sweet pepper

Does this picture look scary?
(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

Which side of the room is the toilet on?
(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right
Evaluating Automated Predictions: Basic Equation

\[
\text{accuracy} = \min\left(\frac{\# \text{humans that provided that answer}}{3}, 1\right)
\]
Evaluating Automated Predictions: Example

Does this picture look scary?
(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

What is the accuracy of an algorithm prediction of
- “yes”? 
- “no”? 
- “maybe”? 

\[
\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)
\]
Evaluating Automated Predictions: Example

What is the accuracy of an algorithm prediction of
- “right”?
- “left”?
- “right side”?
- “center”?
- “bottom”?

Which side of the room is the toilet on?
(1) right
(2) left
(3) right
(4) right
(5) right
(6) right
(7) right side
(8) right
(9) center
(10) right

accuracy = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)

Implementation detail: to enable fair comparison to humans, compute 10 rounds of comparing a prediction with each possible set of 9 human-supplied answers.
Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
Key Idea: Multimodal Representation Rather Than Single Modality Representations

- e.g., vision representation with AlexNet
- e.g., language representation with BERT

[Diagram of AlexNet architecture]

https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Pretrains using language and vision input

Generates representations for image and vision separately as well as jointly

Tan and Bansal, EMNLP 2019.
[CLS] is added to the start of the sequence
LXMERT: Language Input

Each word is represented as sum of its word embedding and position encoding.
Transformer encoder (i.e., BERT); what does its output represent?
LXMERT: Language Input

Transformer encoder (i.e., BERT); represents words with their relationships to all words

Tan and Bansal, EMNLP 2019.
LXMERT: Vision Input

Each image is represented as a description of $m$ objects detected with Faster R-CNN using features from Faster R-CNN and position encodings.
LXMERT: Architecture

Transformer encoder (i.e., BERT); what does its output represent?
LXMERT: Architecture

Transformer encoder (i.e., BERT); represents objects with their relationships to all objects
LXMERT: Architecture

Learns cross-modality representations by aligning entities in the two modalities

Tan and Bansal, EMNLP 2019.
LXMERT: Architecture

Two cross-attention layers are functions of the “query” with “keys” and “values” from the opposite modalities’ features.
LXMERT: Architecture

Two more layers of transformer encoders (i.e., BERT)
LXMERT: Output

New representation of input detected objects

New representation of input language sequence
LXMERT: Output

Cross-modality representation is the [CLS] token appended at the start of the sentence.

Tan and Bansal, EMNLP 2019.
Pretrained Faster R-CNN can locate 1,600 categories and only 36 object detections are kept per image.

Number of layers mimics the size of BERT base of 12 layers; i.e., (5+9)/2 + 5
LXMERT: Architecture

What might be strengths and limitations of the resulting feature representations based on the architecture used?

A woman riding a bike with a dog in a basket.
LXMERT: Summary of Architecture

Tan and Bansal, EMNLP 2019.
LXMERT: Pretraining Task 1 (Language)

Task used for BERT: mask 15% of input words and then predict them.
LXMERT: Pretraining Task 1 (Language)

Unlike BERT, vision modality can resolve language ambiguity; e.g., shows what is being eaten

Tan and Bansal, EMNLP 2019.
LXMERT: Pretraining Tasks 2 & 3 (Vision)

Mask 15% of input objects and then predict their original feature values and categories

Who is eating the carrot?

RoI-Feature Regression
Detected-Label Classification

(Goal: predict same category predicted by Faster R-CNN)
Knowledge about other objects and the language should help predict masked objects.

Who is eating the carrot?

RoI-Feature Regression
Detected-Label Classification
(Goal: predict same category predicted by Faster R-CNN)
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)

Task 4: predict if caption and image match, where 50% of the captions are random

A dog watching a rabbit eat a carrot
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)

Task 5: perform VQA
LXMERT: 5 Pretraining Tasks

Tan and Bansal, EMNLP 2019.
LXMERT: All Pretraining Task Losses Are Summed During Training

What might be strengths and limitations of the resulting feature representations based on the type of pretraining tasks used?

Tan and Bansal, EMNLP 2019.
LXMERT: Training Data for Pretraining

<table>
<thead>
<tr>
<th>Image Split</th>
<th>Images</th>
<th>Sentences (or Questions)</th>
<th>COCO-Cap</th>
<th>VG-Cap</th>
<th>VQA</th>
<th>GQA</th>
<th>VG-QA</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS COCO - VG</td>
<td>72K</td>
<td>361K</td>
<td>-</td>
<td>387K</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.75M</td>
</tr>
<tr>
<td>MS COCO intersection VG</td>
<td>51K</td>
<td>256K</td>
<td>2.54M</td>
<td>271K</td>
<td>515K</td>
<td>724K</td>
<td>4.30M</td>
<td></td>
</tr>
<tr>
<td>VG - MS COCO</td>
<td>57K</td>
<td>-</td>
<td>2.85M</td>
<td>-</td>
<td>556K</td>
<td>718K</td>
<td>4.13M</td>
<td></td>
</tr>
</tbody>
</table>

All images are from two image sets, MS COCO and Visual Genome, which were collected by scraping images from the photo-sharing website Flickr

(Visual Genome includes the MS COCO images)
## LXMERT: Training Data for Pretraining

<table>
<thead>
<tr>
<th>Image Split</th>
<th>Images</th>
<th>Sentences (or Questions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>COCO-Cap</td>
</tr>
<tr>
<td>MS COCO - VG</td>
<td>72K</td>
<td>361K</td>
</tr>
<tr>
<td>MS COCO ∩ VG</td>
<td>51K</td>
<td>256K</td>
</tr>
<tr>
<td>VG - MS COCO</td>
<td>57K</td>
<td>-</td>
</tr>
</tbody>
</table>

Language annotations came from 2 image captioning and 3 VQA datasets, authored by crowdworkers paid to create captions, questions, and answers.

Tan and Bansal, EMNLP 2019.
### LXMERT: Training Data for Pretraining

A total of 9.18M image-sentence pairs are included for 180,000 images (questions in VQA datasets are used for the image-sentence pairs).

<table>
<thead>
<tr>
<th>Image Split</th>
<th>Images</th>
<th>Sentences (or Questions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>COCO-Cap</td>
</tr>
<tr>
<td>MS COCO - VG</td>
<td>72K</td>
<td>361K</td>
</tr>
<tr>
<td>MS COCO ∩ VG</td>
<td>51K</td>
<td>256K</td>
</tr>
<tr>
<td>VG - MS COCO</td>
<td>57K</td>
<td>-</td>
</tr>
<tr>
<td>All</td>
<td>180K</td>
<td>617K</td>
</tr>
</tbody>
</table>

Tan and Bansal, EMNLP 2019.
LXMERT: Training Data for Pretraining

<table>
<thead>
<tr>
<th>Image Split</th>
<th>Images</th>
<th>Sentences (or Questions)</th>
<th>COCO-Cap</th>
<th>VG-Cap</th>
<th>VQA</th>
<th>GQA</th>
<th>VG-QA</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS COCO - VG</td>
<td>72K</td>
<td>361K</td>
<td>-</td>
<td>387K</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.75M</td>
</tr>
<tr>
<td>MS COCO ∩ VG</td>
<td>51K</td>
<td>256K</td>
<td>2.54M</td>
<td>271K</td>
<td>515K</td>
<td>724K</td>
<td>4.30M</td>
<td></td>
</tr>
<tr>
<td>VG - MS COCO</td>
<td>57K</td>
<td>-</td>
<td>2.85M</td>
<td>-</td>
<td>556K</td>
<td>718K</td>
<td>4.13M</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>180K</td>
<td>617K</td>
<td>5.39M</td>
<td>658K</td>
<td>1.07M</td>
<td>1.44M</td>
<td>9.18M</td>
<td></td>
</tr>
</tbody>
</table>

What might be strengths and limitations of the resulting feature representations based on the type of training data that is used?
LXMERT: Fine-Tuning Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary</td>
</tr>
<tr>
<td>Human</td>
<td></td>
</tr>
<tr>
<td>Image Only</td>
<td></td>
</tr>
<tr>
<td>Language Only</td>
<td>66.8</td>
</tr>
<tr>
<td>State-of-the-Art</td>
<td>85.8</td>
</tr>
<tr>
<td>LXMERT</td>
<td>88.2</td>
</tr>
</tbody>
</table>

Achieved the best performance, with stronger gains over prior work for questions that lead to “binary” and “other” answers.

Tan and Bansal, EMNLP 2019.
### LXMERT: Fine-Tuning Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary</td>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Image Only</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Language Only</td>
<td>66.8</td>
<td>31.8</td>
<td>27.6</td>
</tr>
<tr>
<td>State-of-the-Art</td>
<td>85.8</td>
<td>53.7</td>
<td>60.7</td>
</tr>
<tr>
<td>LXMERT</td>
<td><strong>88.2</strong></td>
<td><strong>54.2</strong></td>
<td><strong>63.1</strong></td>
</tr>
</tbody>
</table>

The representations also led to the best performance for an additional VQA dataset and a visual reasoning task (i.e., does statement describe two images or not).

Tan and Bansal, EMNLP 2019.
Today’s Topics

• Multimodal applications
• Image captioning dataset challenges
• Image captioning algorithms
• Visual question answering dataset challenges
• Visual question answering algorithms
• Foundation models
Contributors from academia & industry!

Zhang et al. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv 2024.
Many convert visual inputs to text descriptions that LLMs understand (i.e., prompts) or finetune the vision encoder to align with a frozen pre-trained LLM

Zhang et al. MM-LLMs: Recent Advances in MultiModal Large Language Models. arXiv 2024.
Famous Models in Industry (Proprietary)

• Next frontier: models that are multimodal from design

• OpenAI’s GPT-4(Vision): vision + language model
  • Results: https://openai.com/contributions/gpt-4v

• Google’s Gemini: vision + language model
  • Testing results: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
Prompt Types: Textual vs Visual

• When Might One Choose A Visual Prompt Versus a Textual Prompt?

  • Greater equity for different languages as non-English languages often are poorly supported if at all

  • Empowering people appropriately based on their (dis)abilities: e.g., blind and deaf users
Today’s Topics

• Multimodal applications

• Image captioning dataset challenges

• Image captioning algorithms

• Visual question answering dataset challenges

• Visual question answering algorithms

• Foundation models
The End