Image Captioning

Danna Gurari
University of Colorado Boulder
Spring 2022
Review

• Last week
  • Explosion of transformers
  • GPT
  • BERT
  • Limitations of transformer models
  • Programming tutorial

• Assignments (Canvas)
  • Problem set 4 due next week

• Questions?
Today’s Topics

• Image captioning applications

• Image captioning datasets

• Image captioning evaluation

• Challenge winner: encoder decoder pipeline with attention
Today’s Topics

• Image captioning applications

• Image captioning datasets

• Image captioning evaluation

• Challenge winner: encoder decoder pipeline with attention
A “Human-Like” Description

Captions: https://www.microsoft.com/cognitive-services/en-us/computer-vision-api
Visual Assistance for People with Visual Impairments

Facebook

https://www.youtube.com/watch?v=Tjuc8a836Q

Microsoft

https://www.youtube.com/watch?v=R2mC-NUAmMk
Alt Text for People with Visual Impairments

e.g., Microsoft Power Point (Office 365 demo)
Image Captioning for Newspaper Articles

Aiding Tourism with Captioned Images

Figure 7: Tourists from three different tour groups at the Salt Lake of Uyuni in Bolivia.

Figure 3: Examples for people shots (Peruvian Children, Korean Guards, Russian Singers).

Figure 8: The Cathedral of Cuzco, Peru, in different viewing angles (right, left and front).

Figure 4: Examples for animal photos (Humpback Whale, Kangaroos, Galapagos Giant Turtle).

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
What are other potential applications for image captioning?
Today’s Topics

• Image captioning applications

• Image captioning datasets

• Image captioning evaluation

• Challenge winner: encoder decoder pipeline with attention
Sample of Existing Dataset Challenges

COCO

- Woman on a horse jumping over a pole jump.
- A glass bowl contains peeled tangerines and cut strawberries.

VizWiz

- A person is holding a small container of cream upside down.
- The billboard displays ‘Welcome to Yakima The Palm Springs of Washington’.

TextCaps

- Conceptual Captions
  - Cars are on the streets.
  - Small stand of trees, just visible in the distance in the previous photo.
- Fashion Captioning
  - A decorative leather padlock on a compact bag with croc embossed leather.
- CUB-200
  - This bird is blue with white on its chest and has a very short beak.

## 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>

Challenge: What Instructions Should Be Provided When Collecting Captions from Human Annotators?
Class Task: How Would You Describe This Image?
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.
MSCOCO

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
VizWiz

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!"

Image

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.

Today’s Topics

• Image captioning applications

• Image captioning datasets

• Image captioning evaluation

• Challenge winner: encoder decoder pipeline with attention
Group Discussion: 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>

- *captions*: [ { "text": "a giraffe standing in the dirt" }, { "confidence": 0.982929349 } ]
Evaluation: Human Judgments

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Slightly Disagree</th>
<th>Slightly Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

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**

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?

Evaluation: Automated

• BLEU

• METEOR

• Rouge

• CIDEr

• SPICE

Do you think these two captions describe the same image?

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

Evaluation: Automated

• BLEU

Problem: n-gram methods scores these as very similar

• 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

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

- Image captioning applications
- Image captioning datasets
- Image captioning evaluation
- Challenge winner: encoder decoder pipeline with attention
Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu*
Jimmy Lei Ba†
Ryan Kiros†
Kyunghyun Cho*
Aaron Courville*
Ruslan Salakhutdinov†*
Richard S. Zemel†*
Joshua Bengio**

*Kelvin.Xu@umontreal.ca
*Jimmy@psi.utoronto.ca
*Ryankiros@cs.toronto.edu
*Kyu@umontreal.ca
*Aaron.courville@umontreal.ca
*Rsalakhu@cs.toronto.edu
*Zemel@cs.toronto.edu
*Yoshua.bengio@umontreal.ca

†University of Toronto, †University of Toronto, †CIFAR
Idea: Treat Problem Like Machine Translation
Idea: Treat Problem Like Machine Translation

Image

Encoder

Decoder

Caption

A bird flying over a body of water.
Recall Solution:

1. Encoder produces hidden state for every input

2. At each decoder time step, attention weights are computed that determine each input’s relevance for the prediction

3. At each decoder time step, a prediction is made based on the weighted sum of the inputs

https://towardsdatascience.com/attn-illustrated-attention-5ec4ad276ee3
Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input.

Model learns how to weight each input!
Approach: Key Difference

1. Input represents an image
Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input.

<table>
<thead>
<tr>
<th>Input</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>A bird is flying...</td>
<td>t = 1   t = 2   t = 3   t = 4</td>
</tr>
</tbody>
</table>
Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input
Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input.

<table>
<thead>
<tr>
<th>Input</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>bird</td>
</tr>
<tr>
<td>$t = 1$</td>
<td>$t = 2$</td>
</tr>
</tbody>
</table>
Intuition

Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input.

<table>
<thead>
<tr>
<th>Input</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>A bird is</td>
<td></td>
</tr>
<tr>
<td>t = 1 t = 2 t = 3</td>
<td></td>
</tr>
</tbody>
</table>
Decoder decides which inputs are needed for prediction at each time step; e.g., “soft attention” uses a weighted combination of the input

### Intuition

---

<table>
<thead>
<tr>
<th>Input</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Input Image" /></td>
<td><img src="image" alt="Target" /></td>
</tr>
</tbody>
</table>

A bird is flying...

t = 1  t = 2  t = 3  t = 4  ...
Input Representation: Idea

Use convolutional layer that map to regions of the input (e.g., pixel) space; e.g., 1rst layer with $h$ values

https://www.deeplearningbook.org/contents/convnets.html
Input Representation: Implementation

Each line represents a convolutional layer.

14×14×512 feature maps are reshaped to be inputs.

Grids reflect relative spatial coarseness at each layer.

VGG16:

Input Representation: Implementation

Each line represents a convolutional layer.

The size of each attended to region is therefore the size of the original image observed at this layer.

Grids reflect relative spatial coarseness at each layer.

VGG16:

Experimental Results

State-of-the-art performance on three dataset challenges (Flickr8k, Flicker30k, and MS COCO)
Examples where correct content was attended to when predicting the word:

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.
Examples where incorrect content was attended to when predicting the word:

- A large white **bird** standing in a forest.
- A woman holding a **clock** in her hand.
- A man wearing a hat and a hat on a **skateboard**.
- A person is standing on a beach with a **surfboard**.
- A woman is sitting at a table with a large **pizza**.
- A man is talking on his cell phone while another man watches.
Today’s Topics

• Image captioning applications

• Image captioning datasets

• Image captioning evaluation

• Challenge winner: encoder decoder pipeline with attention
The End