AUTOMATED IMAGE CAPTIONING AND IMAGE-TEXT ALIGNMENT

Natalie Parde
parde@uic.edu

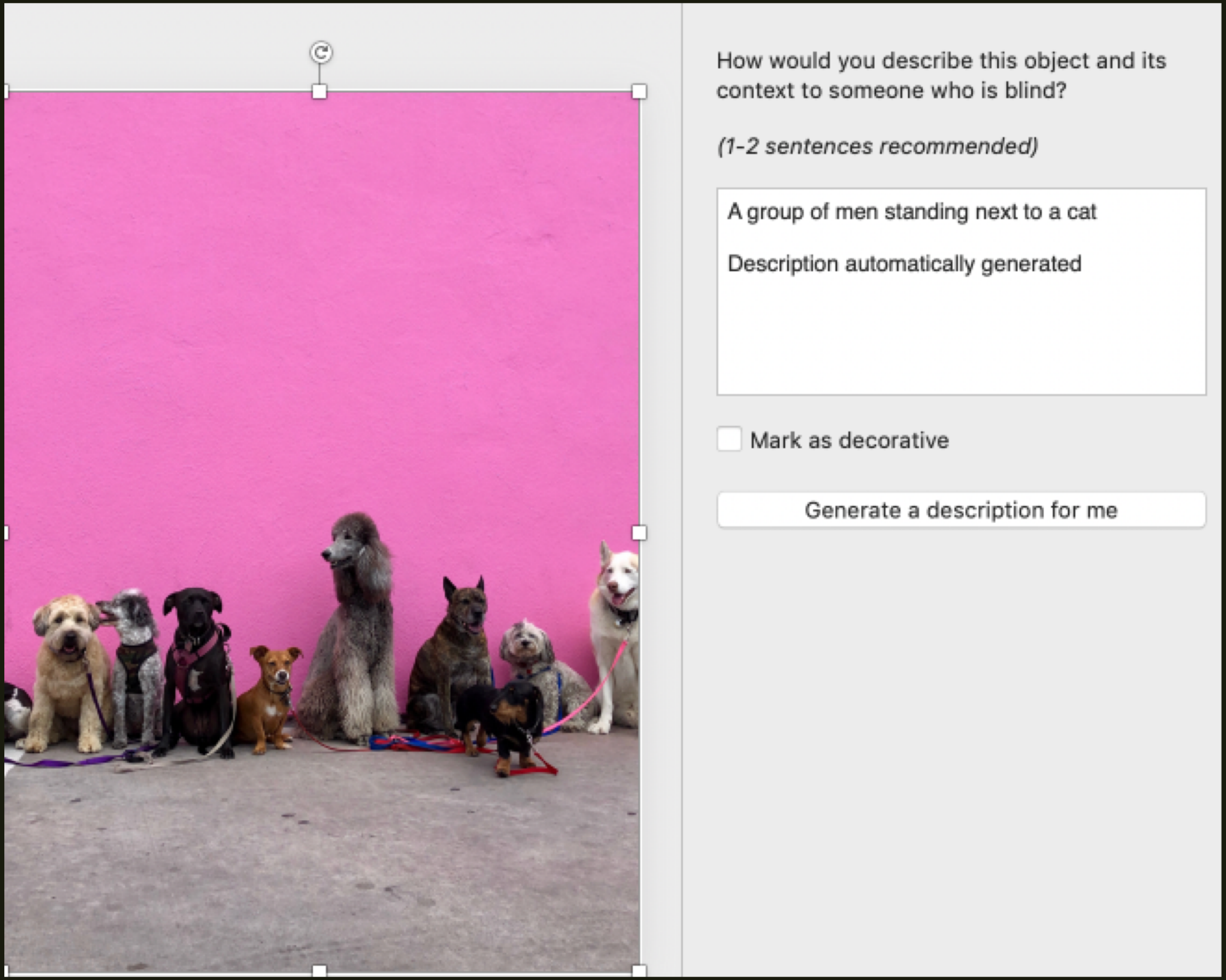
CS 594: Language and Vision
Spring 2019

What is automated image captioning?

Automatically generating descriptions of image content.

What is image-text alignment?

Automatically aligning existing descriptions, keyphrases, or tags with images.



Challenging problem of substantial interest to many groups!

Two tasks involved in generating good image captions

Image Understanding

- What concepts are illustrated in the image?

Natural Language Generation

- How can these concepts be described in an appropriate, grammatically correct way?

Neither of these is sufficient without the other!

Group of dogs in front are stands pink wall.

A group of men are standing next to a cat.

A group of dogs are sitting in front of a pink wall.

Subproblems involved in image understanding

Object localization	*What objects are in the image?*
Attribute identification	*What are the key characteristics of these objects?*
Scene classification	*Where are these objects located?*
Entity relation	*How do these objects relate to one another and to the scene?*

Subproblems involved in natural language generation

Content selection

- *Which aspects of the image should be discussed?*

Content organization

- *What is the most effective way to discuss these elements?*

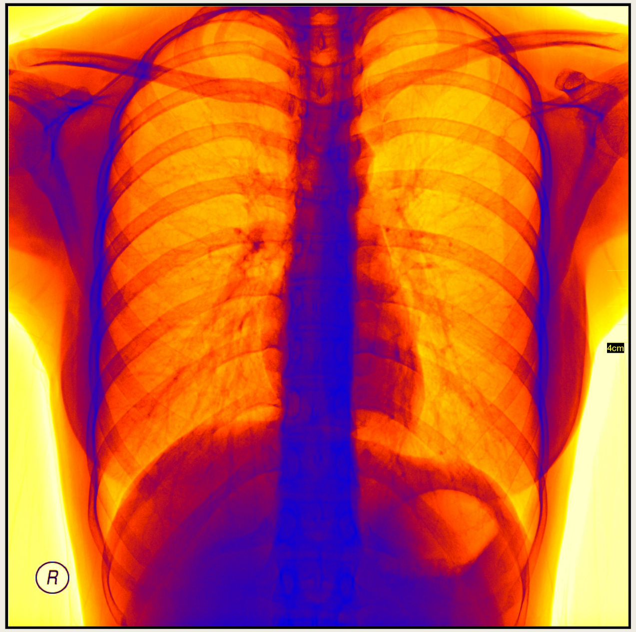
Surface realization

- *What words should be used to discuss these elements?*
- *Should any pronouns be used?*
- *What tense should be used?*
- *How can related information be aggregated?*

Additional Layers of Complexity

- Different audiences desire different types of descriptions
- Understanding some images requires contextual or common sense knowledge

A picture of a chest x-ray.



No apparent cardiopulmonary abnormalities.

Two general categories of image captioning models

Direct generation

- *First identify the key components, attributes, etc., and generate a description based on those components*

Retrieval-based

- *Find images similar to the test image, and generate a description based on the descriptions for those images*

Evaluating Image Captioning Approaches

Human evaluation metrics

- *Grammaticality*
- *Relevance*
- *Creativity*
- *“Humanness”*

Automatic evaluation metrics

- *BLEU*
- *ROUGE*
- *Translation Error Rate*
- *METEOR*
- *CIDEr*

Bilingual Evaluation Understudy (BLEU)

- Designed to assess machine translation
  - *Compares a generated text sample with one or more references*
- Best possible score: 1.0
- Worst possible score: 0.0
- Computed by finding the average percent of n-gram matches between the generated and reference samples

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

- Designed to assess machine translation and text summarization
- Based on BLEU
- Computes the percentage of n-grams from the reference text(s) that occur in the automatically-generated text
- Lots of variations:
  - *ROUGE-N: n-gram overlap*
  - *ROUGE-L: longest matching sequence of words*
  - *ROUGE-S: skip-gram overlap*

Translation Error Rate

- Designed to assess machine translation
- Based on edit distance
- Computes the number of changes needed to transform the automatically-generated text to (one of) the reference text(s), divided by the average number of words in the reference text
  - *Possible changes: insertion, deletion, substitution, shift*
- Lower score is better

Metric for Evaluation of Translation with Explicit ORdering (METEOR)

- Designed to assess machine translation
- Computes alignment score between generated text and reference text(s) based on exact, stem, synonym, and paraphrase matches
- Best score: 1.0
- Worst score: 0.0

Consensus-based Image Description Evaluation (CIDEr)

- Designed to assess image description
- Computes a score based on how closely the generated text matches *most* of the reference texts
  - *Incorporates TF-IDF scores (n-grams that occur across most image descriptions are weighted lower than those that occur across fewer image descriptions)*
- Higher score is better (can be > 1)

These metrics still fail to capture some important qualities!

How to measure and promote diversity/originality of image captions?

How to measure contextual descriptions for images occurring in a temporal sequence?

Resources

Datasets

- *COCO: http://cocodataset.org*
- *Conceptual Captions: https://ai.googleblog.com/2018/09/conceptual-captions-new-dataset-and.html*

Lectures

- *Automated Image Captioning with ConvNets and Recurrent Nets, by Andrej Karpathy: https://youtu.be/xKt21ucdBY0*
- *How we teach computers to understand pictures, by Fei Fei Li: https://youtu.be/40riCqvRoMs*

Wrapping up….

- Overview of automated image captioning
- Overview of image-text alignment
- Image captioning subtasks
  - *Image understanding*
  - *Natural language generation*
- Types of image captioning models
- Metrics for evaluating automatic image captioning
- Resources