

TEXT-TO-IMAGE GENERATION AND VISUAL QUESTION ANSWERING

Natalie Parde
parde@uic.edu

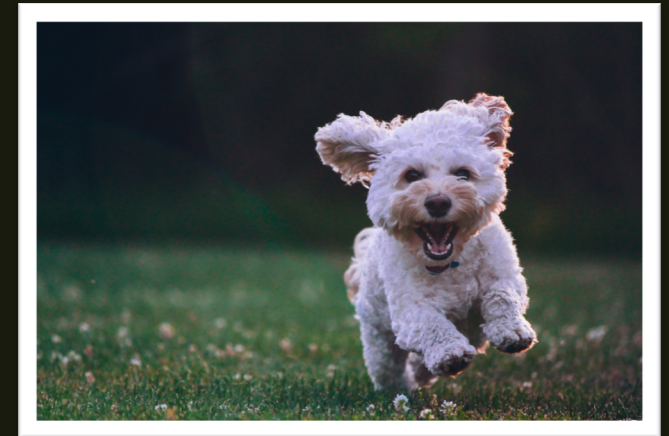
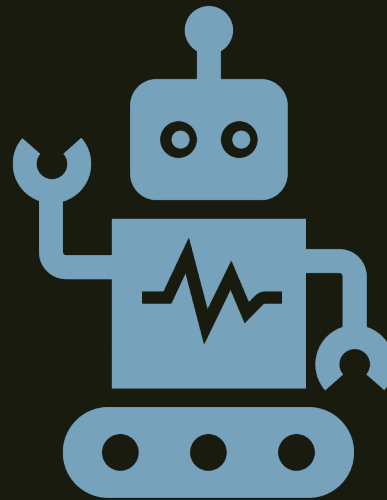
CS 594: Language and Vision
Spring 2019

What is text-to-image generation?




The task of automatically generating an image, given a text description.

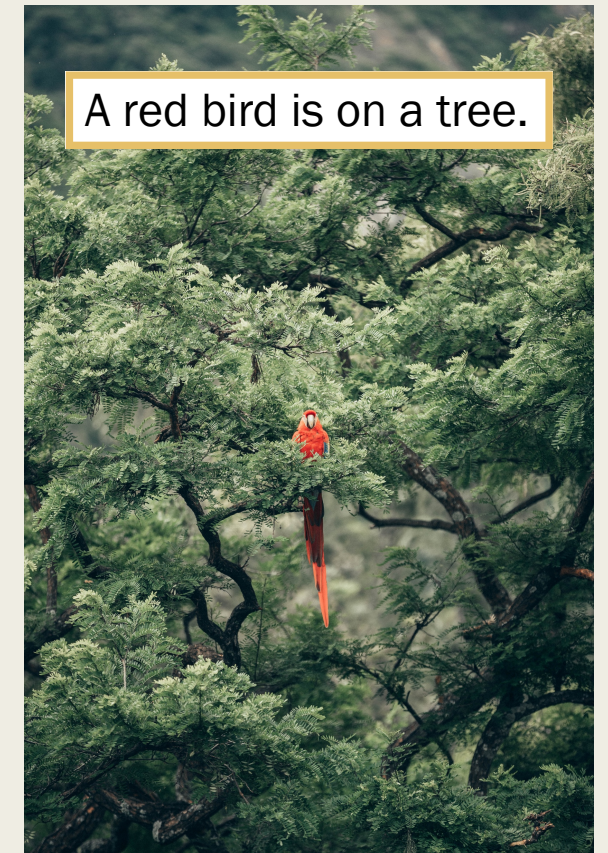
- Almost like reverse image captioning!

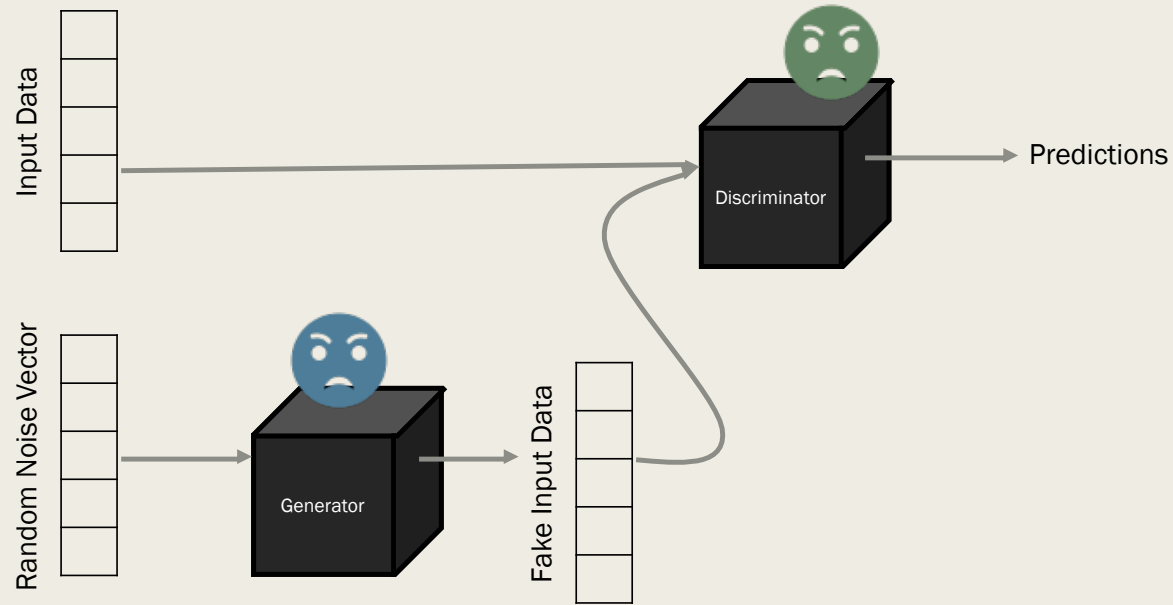
A small, fluffy white dog.



Highly Multimodal Process

- To generate high-quality images, models must:
 - *Identify concepts and their relationships in text descriptions*
 - *Associate the concepts with visual characteristics*
 - Red = 
 - Bird = 
 - Tree = 
 - *Associate the relationships with spatial references*





Virtually all recent work in text-to-image generation has made use of generative adversarial networks (GANs).

Refresher: Generative Adversarial Networks

- Comprised of two neural networks that act as adversaries of one another
- Generative model rather than discriminative
 - *Generative: Learn the probability distributions of features associated with classes*
 - *Discriminative: Learn the boundary between classes*

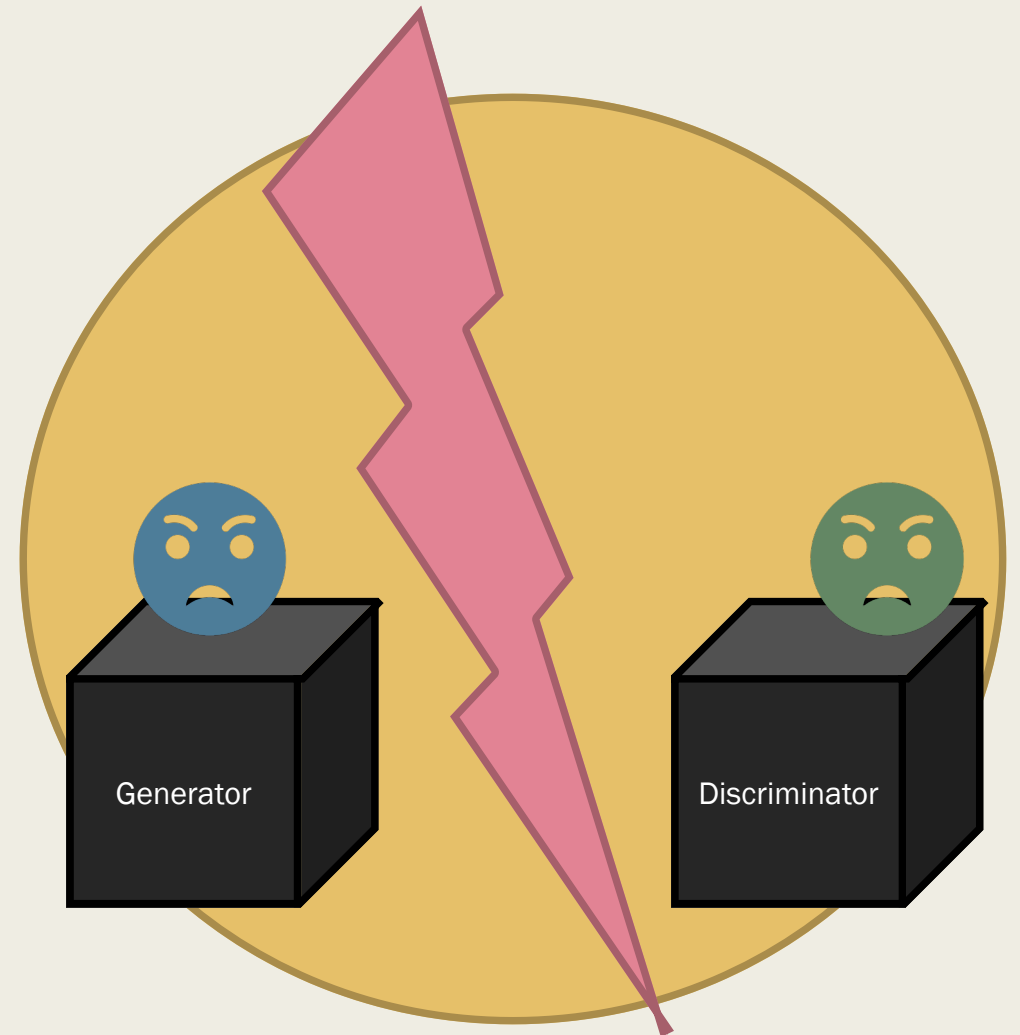


What is the label, given what we know?

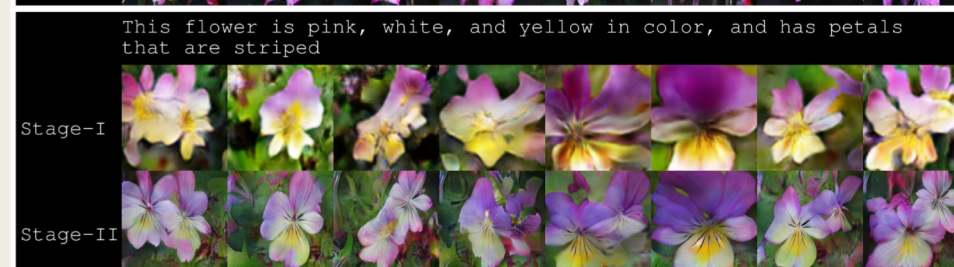
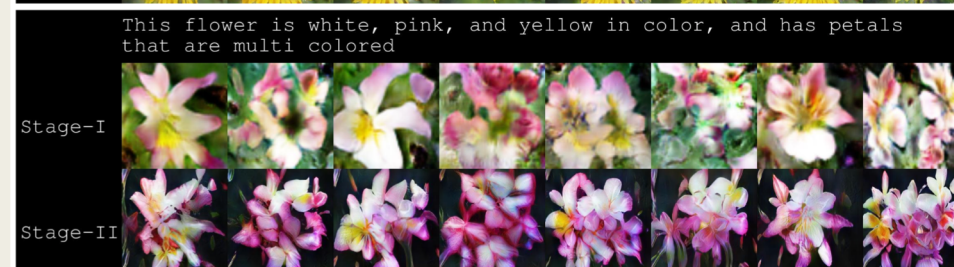
How do we know that this is the label?

Refresher: Generative Adversarial Networks

- Generator: Goal is to generate fake instances that are passable enough that the discriminator doesn't detect them
- Discriminator: Goal is to discriminate between real instances and generated fake instances



What does the text-to-image generation process look like?



How is text-to-image generation performance typically measured?

- Inception Score

- Seeks to measure *image quality and diversity*

- *Basic idea:*

- Classify generated images using the Inception network

- Consider quality to correlate with predictability (low entropy)

- Consider diversity to correlate with *unpredictability* (high entropy)

- Combine the criteria as follows:

- $IS(image) = \exp(\mathbb{E}_{image}[KL(p(label|image) \parallel p(label))])$

Text-to-Image Generation Datasets

- Typical image captioning datasets
 - COCO: <http://cocodataset.org/#home>
- Caltech-UCSD Birds: <http://www.vision.caltech.edu/visipedia/CUB-200.html>
- Oxford 102 Flowers:
<http://www.robots.ox.ac.uk/~vgg/data/flowers/102/>

What is this dog sitting inside of?



A teepee.

What holiday is this dog ready for?

St. Patrick's Day.



Blue.

What color bandanna is the dog on the left wearing?

Where is this dog?



At the playground.

What is visual question answering (VQA)?

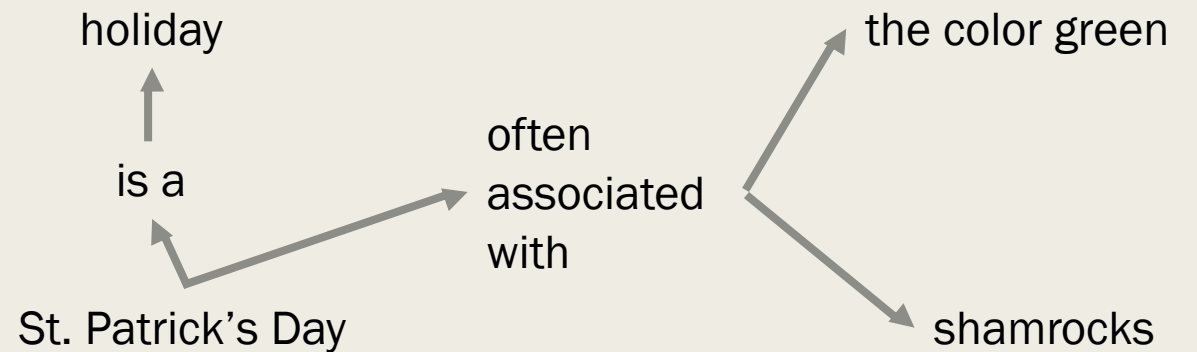
The task of automatically generating natural-language responses to questions about images.

Just like with text-to-image generation, this is a highly multimodal process!

- To generate good responses, systems must be able to:
 - *Build meaningful representations of images*
 - What concepts are represented, and how are they related to one another?
 - *Create commonsense inferences based on these representations*
 - If a dog is on a swing, does that mean the dog is at the playground?
 - What if the dog is on a swing *and* in front of a jungle gym?
 - *Interpret natural-language questions*
 - What information should the answer contain?
 - *Generate natural-language responses*
 - How should the answer be constructed?

VQA systems typically need some way to represent and reason over knowledge.

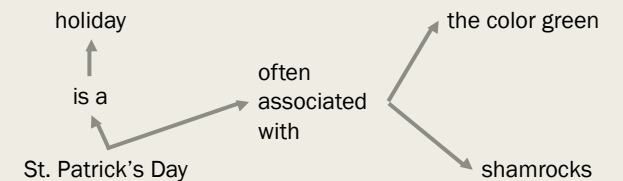
- **Knowledge Representation and Reasoning:** An entire separate subfield of artificial intelligence (!) concerned with storing facts and their relationships/dependencies in a formal language.
 - *Closely related to and often a crucial subcomponent of NLP systems.*





Unlike traditional QA systems, VQA systems need to perform most of their reasoning over concepts extracted from images!

What holiday is this dog ready for?





What color bandanna is the dog on the left wearing?

Many recent VQA systems decide which portions of an image to focus on using **attention**.

- Break an image into sub-regions
- Pay attention to only specific sub-regions of the image when deciding how to respond to an input question
 - *Select these sub-regions based on their probabilities of being relevant to the question*

How is visual question answering typically measured?

- Accuracy

- *Percentage of answers that were correct*

- Accuracy (VQA Challenge)

- $Accuracy(answer) = \min\left(\frac{\# \text{ humans who provided this answer}}{3}, 1\right)$
- *Maintains consistency with human accuracies*

How to address open-ended questions?

- Wu-Palmer Similarity (WUPS) sometimes used
 - *Semantic distance between words based on WordNet hierarchy*
- Problem(s) with this metric?
 - *Imperfect coverage*
 - *Incorrect answers often occur in very similar contexts to correct answers*



What color bandanna is the dog on the left wearing?

blue
gray
purple
red
yellow
pink
orange
green

Visual Question Answering Datasets

VQA Dataset:

<https://visualqa.org/download.html>

Visual7W:

<http://web.stanford.edu/~yukez/visual7w/>

Visual Genome:

<https://visualgenome.org/>

Additional VQA Resources

- VQA Challenge
 - Current: <https://visualqa.org/challenge.html>
 - 2018: https://visualqa.org/workshop_2018.html
 - 2017: https://visualqa.org/workshop_2017.html
 - 2016: https://visualqa.org/workshop_2016.html
- Presentation analyzing last year's VQA challenge results: <https://youtu.be/6SFJd3x-NI8>
- *Words, Pictures, and Common Sense*, presented by Devi Parikh: <https://youtu.be/me82jND3jLI>

Wrapping up....

- Overview of text-to-image generation
- GANs refresher
- Evaluating text-to-image generation
- Datasets for text-to-image generation
- Overview of visual question answering
- Knowledge representation for VQA
- Attention for VQA
- Evaluating VQA
- VQA datasets and resources