Vision and Language

Chenxi Liu
2018/11/27
Three Levels of Vision

High-level vision
- 3D model: hierarchically organized parts and relations

Mid-level vision
- 2 1/2 D-sketch: local surface depths and orientations
- Complex features
- Symmetries
- Parts
- Edge-grouping
- Region grouping
- Occlusion evidence

Low-level vision
- Primal sketch: local 2D tokens: edges, blobs, contours, etc.

Visual input
- Input image
Three Levels of Vision

- **Low-Level:**
  - Edge detection
  - ...

- **Mid-Level:**
  - Depth estimation
  - ...

- **High-Level:**
  - Image classification
  - Object detection
  - Semantic segmentation
  - ...
  - IS THERE MORE?
Vision and Language

- High-level vision is basically about semantics
- We use natural language to express semantics
- Using “person, bicycle, car, horse” to describe a scene is fundamentally limited
- In general, we will need phrases, sentences, paragraphs...
Tasks

Image Captioning

A boy feeding a giraffe

giraffe on right

Referring Expression

Image Retrieval

A boy feeding a giraffe

How many giraffes?

Visual Question Answering/Turing Test

Two
Neural Network for Vision

- Intuition:
  - Local regions are grouped together
  - The same operation can be applied across different locations
- Convolutional Neural Network (CNN):
Neural Network for Language

- **Intuition:**
  - There is a “state” that summarizes everything in history
  - The same operation can be applied across different time steps

- **Recurrent Neural Network (RNN):**

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Standard RNN

The repeating module in a standard RNN contains a single layer.

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Long-Short Term Memory (LSTM)

The repeating module in an LSTM contains four interacting layers.

http://colah.github.io/posts/2015-08-Understanding-LSTMs/
Encoder RNN vs Decoder RNN

https://towardsdatascience.com/sequence-to-sequence-tutorial-4fde3ee798d8
Tasks

Image Captioning
- A boy feeding a giraffe

Referring Expression
- giraffe on right

Image Retrieval
- A boy feeding a giraffe

Visual Question Answering/Turing Test
- How many giraffes?
- Two
Image Captioning

A boy feeding a giraffe
Neural Network Model Design

- **Input:**
  - Domain?

- **Output:**
  - Domain?
Neural Network Model Design

- **Input:**
  - Domain: Vision
  - Model?

- **Output:**
  - Domain: Language
  - Model?
Neural Network Model Design

- **Input:**
  - Domain: Vision
  - Model: CNN
  - Need spatial?

- **Output:**
  - Domain: Language
  - Model: RNN/LSTM
  - Encoder/Decoder?
Neural Network Model Design

- **Input:**
  - Domain: Vision
  - Model: CNN
  - Need spatial: Probably no

- **Output:**
  - Domain: Language
  - Model: RNN/LSTM
  - Encoder/Decoder: Decoder

Demo!

- [https://www.captionbot.ai/](https://www.captionbot.ai/), powered by Microsoft
Referring Expression

giraffe on right
Neural Network Model Design

- **Input:**
  - Domain?

- **Output:**
  - Domain?
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model?

- **Output:**
  - Domain: Vision
  - Model?
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model: CNN & RNN/LSTM
  - Encoder/Decoder?

- **Output:**
  - Domain: Vision
  - Model: CNN
  - Need spatial?
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model: CNN & RNN/LSTM
  - Encoder/Decoder: Encoder

- **Output:**
  - Domain: Vision
  - Model: CNN
  - Need spatial: Yes

Demo!

- [http://vision2.cs.unc.edu/refer/comprehension](http://vision2.cs.unc.edu/refer/comprehension), powered by UNC
Visual Question Answering

How many giraffes?

Two
Neural Network Model Design

- **Input:**
  - Domain?

- **Output:**
  - Domain?
Neural Network Model Design

- Input:
  - Domain: Vision & Language
  - Model?

- Output:
  - Domain: Language
  - Model?
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model: CNN & RNN/LSTM
  - Need spatial?
  - Encoder/Decoder?

- **Output:**
  - Domain: Language
  - Model: MLP or RNN/LSTM
  - (If RNN/LSTM) Encoder/Decoder?
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model: CNN & RNN/LSTM
  - Need spatial: Probably no
  - Encoder/Decoder: Encoder

- **Output:**
  - Domain: Language
  - Model: MLP or RNN/LSTM
  - (If RNN/LSTM)
    - Encoder/Decoder: Decoder
Neural Network Model Design

- **Input:**
  - Domain: Vision & Language
  - Model: CNN & RNN/LSTM
  - Need spatial: Probably no
  - Encoder/Decoder: Encoder

- **Output:**
  - Domain: Language
  - Model: MLP or RNN/LSTM
  - (If RNN/LSTM) Encoder/Decoder: Decoder

---

“How many horses are in this image?”

STM coder
Demo!

- [http://vqa.cloudcv.org/](http://vqa.cloudcv.org/), powered by Georgia Tech
Other tasks?

- E.g., language as input, vision as output. What is a good name for this task?
Other tasks?

- E.g., language as input, vision as output. What is a good name for this task?
- Conditional Image Synthesis:

Take-home Messages

● When vision goes to high-level, it seems eventually inevitable to involve language
● In the deep learning era, CNN is usually used for the vision domain, and RNN/LSTM is usually used for the language domain
● Many fun tasks (image captioning, referring expression, visual question answering) with vision and/or language as input/output
Thank you!