Artificial Intelligence

Perception — Speech/Language Understanding and Vision

Outline

- Speech Recognition
- Language Processing
- Language Understanding
- Vision

Communication

- “Classical” view (pre-1953):
  language consists of sentences that are true/false
  (compare with logic)

- “Modern” view (post-1953):
  language is a form of action

- Why?
  it can be used to change the actions of the agents.

Communication on the Receiving End

- Perception — Agent perceives speech W in context C
- Analysis — Agent infers possible meanings P_1, … P_n
- Disambiguation — Agent infers intended meaning P_i
- Incorporation — Agent incorporates P_i

Speech Recognition

- Speech recognition is the process of recognizing and translating the spoken language to text.

- Speech recognition technologies benefit from knowledge and research in the computer science, linguistics and computer engineering fields.

- Speech recognition has a long history with several waves of major innovations.

    - speech as probabilistic inference — What is the most likely word sequence, given the speech signal that is noisy, variable, and ambiguous?

    - Hidden Markov Models — a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal.

    - Deep learning

Language Processing

- Stop-word removal
- Part-of-speech tagging
- Tokenization
- Parsing
- …
- Aim to shape language sentences into a set format and process the text in a literal sense.

Language Understanding

- Interpret the language
- Derive meaning
- Identify context
- Draw insights
- …
- Aim to extract the context and intent, or to understand what was meant from the text.

Real Language

- Real human languages provide many problems for natural language processing and understanding:
  - ambiguity
  - anaphora
  - indexicality
  - vagueness
  - discourse structure
  - metonymy
  - metaphor
  - noncompositionality

Vision

- Agent's visual perception of the world is an image or video (sequence of images):
  $S = g(W)$, where g = graphics

- Vision seeks to gain high level and meaningful information from images or videos.

- Can we do vision as inverse graphics?
  $W = g^{-1}(S)$

- Problem: massive ambiguity!

Slightly Better Approach

- Luckily, we don’t need to completely recover the exact world scene. Just extract information needed for:
    - navigation
    - manipulation
    - recognition/identification
- Bayesian inference of world configurations:
  $$P(W \mid S) = \frac{P(S \mid W) \times P(W)}{P(S)}$$
  where P(W) is the “prior knowledge of A world”
- Compare $P(S \mid W_1) \times P(W_1)$ and $P(S \mid W_2) \times P(W_2)$ to see which world is more likely given the image S.
- Problems: how do we get the prior knowledge of all possible worlds?

Image Processing and Pattern Recognition

- Image Processing
    - Input and output are both images
    - including operations such as image noise reduction, contrast enhancement, image sharpening, edge detection, etc.
- Segmentation and Recognition
    - Input are images, while output are attributes extracted from the images
    - Object Recognition
- High-level Processing/Recognition
    - “Making sense” of an ensemble of recognized objects
    - Image analysis
    - Computer vision

Learning

- Treat object recognition as a classification problem
- Solve it using learning techniques, especially deep learning algorithms
- Advantage:
  - computationally feasible
  - quite accurate when trained with good quality dataset
- Disadvantage:
  - Biased with only pre-set class labels
  - Disastrous when trained with skewed dataset

Summary

- Both speech/language understanding and vision are hard
  — noise, ambiguity, complexity

- Prior knowledge is essential to constrain the problems in both fields