Dynamic Learning Term Project
Fall 2010

Visual Storytelling: Graphical Models for Dynamic Bayesian Learning from Video Stream Data

September 13, 2010

Ha-Young Jang
E-mail: hyjang@bi.snu.ac.kr

School of Computer Science and Engineering
Seoul National University

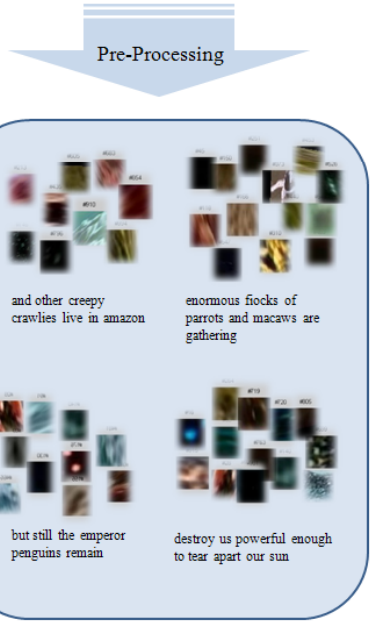
Tasks for the Project

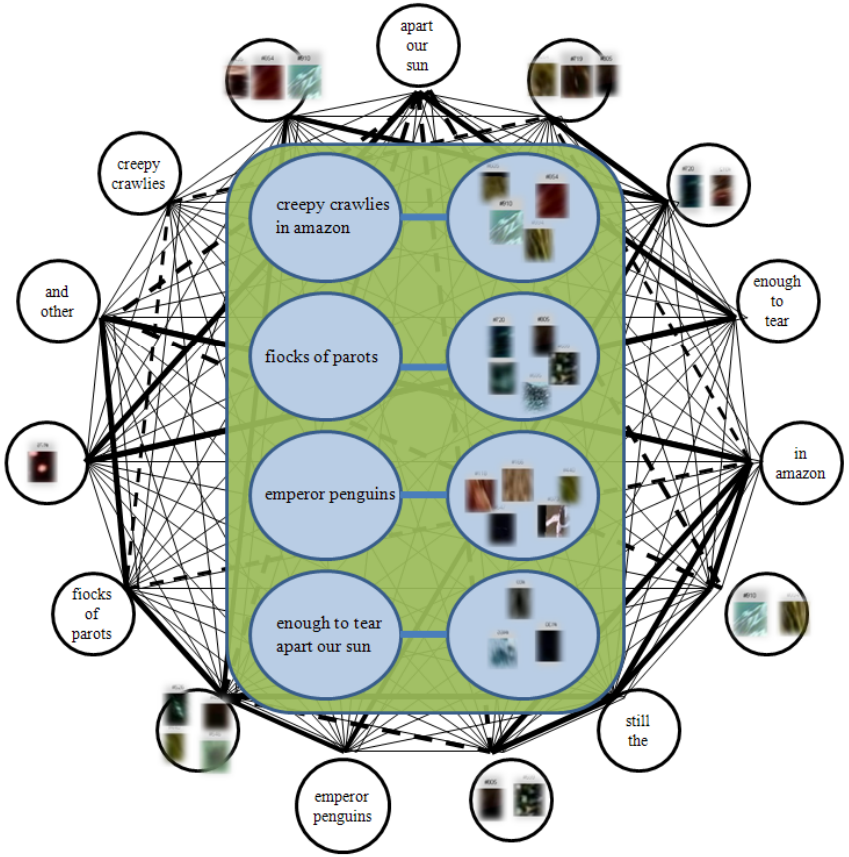
- Given
  - A sequence of *T* image-text pairs of Y(t) = (V(t), L(t)), t=1,…, *T*
  - V(t): a vector of visual words, L(t): a vector of linguistic words
  - E.g., from a 20-minute episode of *Friends*
- Construct
  - A dynamic system that learns to estimate the (mental memory) states to generate the future image-text sequences from a historical context of size *H*:

$$y_{t+1} = f(y_{t-H:t}, u_t; x_t)$$

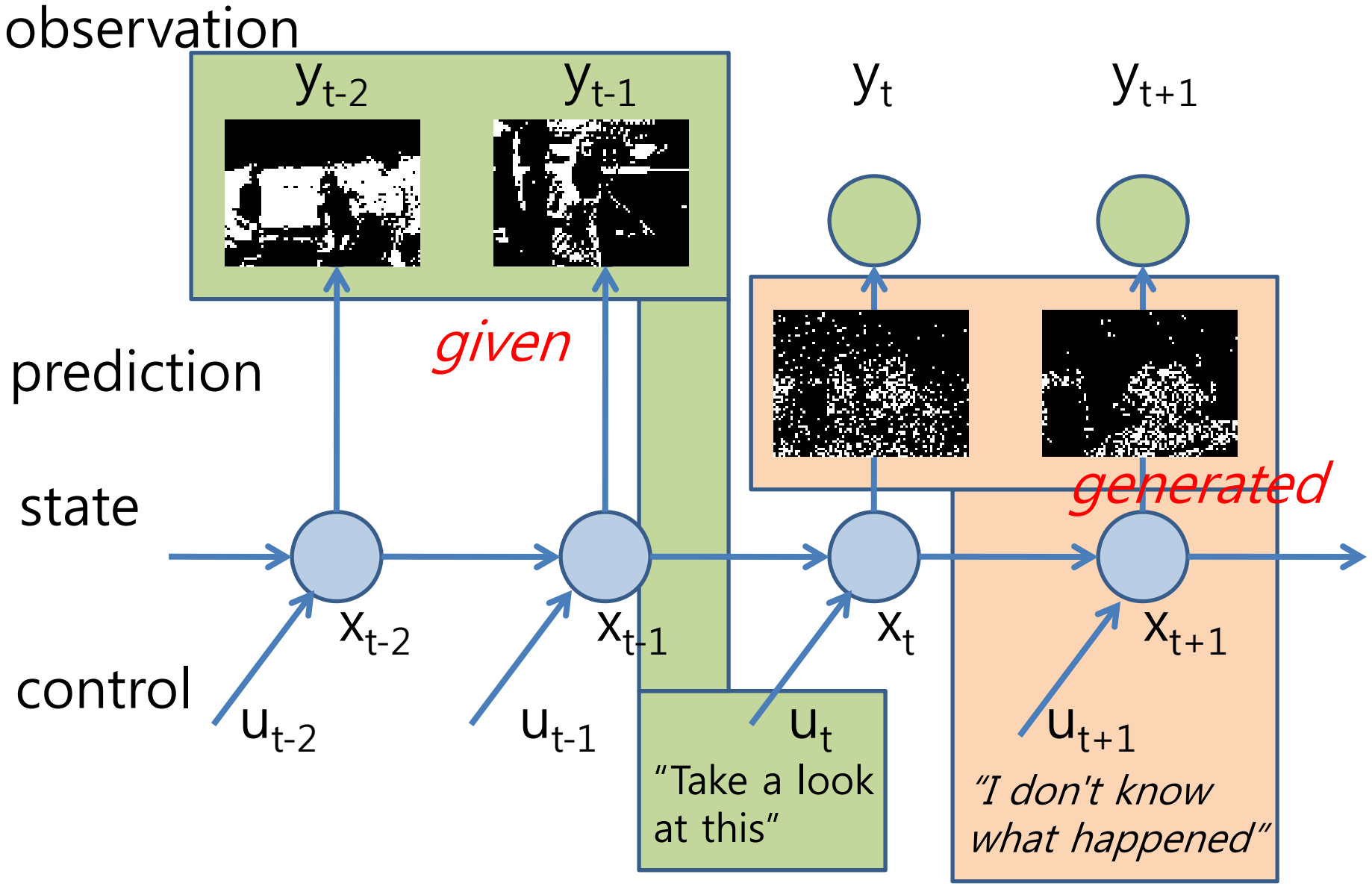
- To demonstrate
  - Visual storytelling: Given a sequence of pairs of (V(h), L(h)), h = t, t-1, t-2, …, t-H, generate (V(r),L(r)) for r = t+1, t+2, t+3, …, t+R
  - L2V translation (mental imagery): Given a series of texts, generate a series of images.
  - V2L translation (scene description): Given a series of images, generate a series of texts.

Dynamic Learning Memory





Project: Visual Storytelling



Data Sets

- 294 pairs of image and sentence from an episode of 'Friends'
- For each sentence, there are two screenshots
  - One is 200 by 150 color bitmap image
  - Another is 100 by 75 b/w bitmap image

- Defining the visual words by preprocessing the images is up to you!
- More sets of data (more episodes and different kinds of TV dramas) will be available on request.

Submission Guideline

- Team project
  - Number of teammates: 3
- Schedule
  - First report: Oct. 13, 24:00
    - Project title and abstract (conference paper format)
    - Motivation and questions asked
    - Experimental design
    - Methods
    - Results and discussion
  - Project presentations: Nov. 15-24
  - Final report: Dec. 6, 24:00
- Report format: English in conference paper style (AAAI style)

Tasks

- Main task: Visual storytelling: Given a sequence of pairs of (V(h), L(h)), h = t, t-1, t-2, …, t-H, generate (V(r),L(r)) for r = t+1, t+2, t+3, …, t+R
- L2V translation (mental imagery): Given a series of texts, generate a series of images.
- V2L translation (scene description): Given a series of images, generate a series of texts.
- Plus any additional tasks or variants of the above you want to address!
- In particular, you may use more data sets (video episodes) which will be available on request.
- Using more datasets will allow you to ask other questions, such as what stream of scenes (and dialogues) is coming from which kinds (title or genre) of drama?
- Asking more questions will be a definite plus in grading of the project reports.

Criteria for Project Grading

- Does the project address the essential dynamic learning problems on visual storytelling and their variants?
- How well the experiments are designed? That is, how broad and deep are the questions asked by the experiments?
- How advanced is the algorithm used to solve the problems?
- What innovations or new ideas were put to the existing methods to solve the problems?
- How well does the implemented system work?
- How well is the implemented system described?
- How good the experimental results are described and discussed?
- What are the lessons learned from the project?