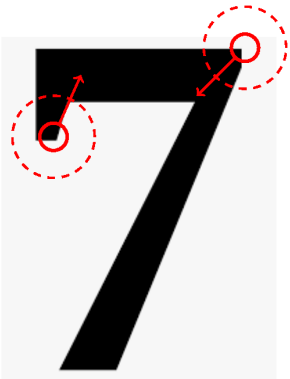
Biological Vision and Applications
Module 05-09: Recurrent attention models

Hiranmay Ghosh

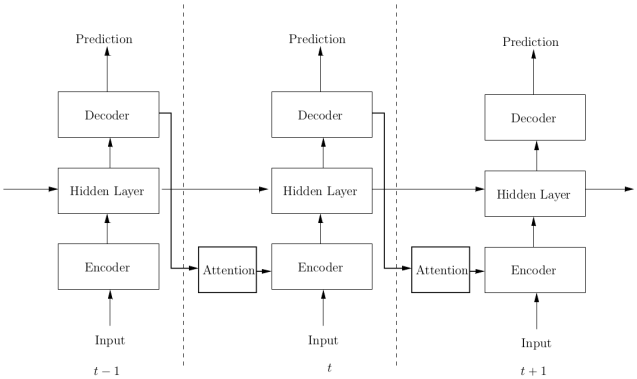
Saliency is dynamically constructed



- We look at a small part of a scene at a time
- Where we look at next depends on what we see
  - ... plus, the task at hand
- Saliency map of a scene is not computed in one go
  - Constructed dynamically over time
    - As and when needed ... Just in time
  - Saliency map for the whole image is never built
- Peripheral vision guides the direction of eye movement

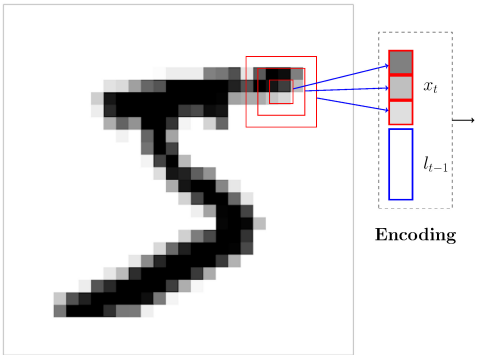
EdPuzzle assignment

Attention-based RNN Architecture



- RNN and the "Attention" module are trained together

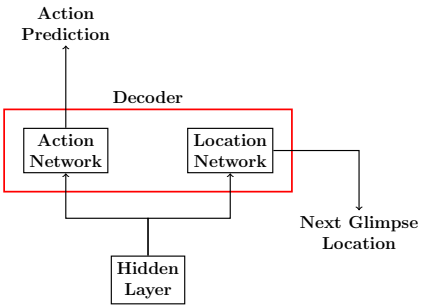
Implementation example



- Encoding
  - Glimpse: Encoded representation of visual field
  - Glimpse Network:
    - Image data + Location (x_t, l_{t-1})
    - Encoded to some internal representation with an NN
- Where do you look at the first glimpse? $l_0 = ?$

Mnih, et al. Recurrent models of visual attention (2014)

Decoder



- Each of Action and Location Networks is an NN
- "Action" can be different in different contexts:
  - Predicting the object
    - number, in our example
  - Driving a car
  - ...

Training

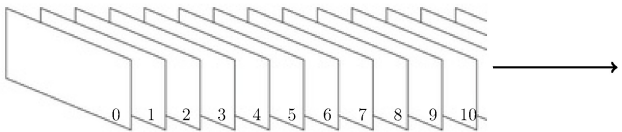
Reinforcement Learning

- Model for "hard attention"
- Training for optimal saccades
  - Training based on back-propagation does not work
  - Reinforcement learning used
  - Reward after each time-step
  - Biological system might follow similar "reward" based learning mechanism
- In the case of object recognition
  - Reward $r_t = 1$ if the object is classified correctly at time step t
  - $r_t = 0$ otherwise
- Positive reward is sparse
- System tries to maximize $\sum_t r_t$ over time

Discussions

- Attention and object recognition complements each other
- Example of “life-long learning”
- Network trained on a few patterns performs well for other patterns with little training
  - Example of transfer learning
- Robust against distractors (noisy patches on the image)

Recurrent Attention for Video

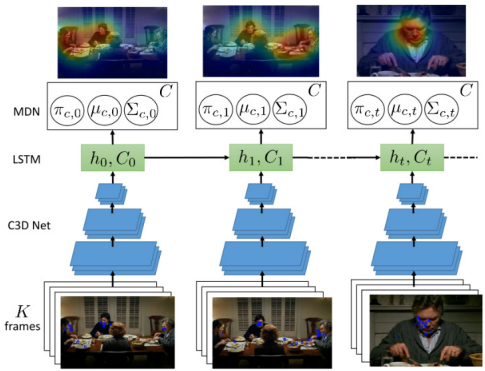


Why processing video frame by frame does not work ?

- Motion information is lost
- Saliency map for each frame depends on the earlier frames
- Too much data to be processed
  - There are lots of redundancies in video data (over successive frames)

Recurrent Attention Model for Video

Recurrent Mixture Density Network



- Soft attention model is used
- Prediction in the form of a GMM over space
  - There can be multiple salient objects

Bazzani & Larochelle. Recurrent mixture density network ... (2017)

https://www.youtube.com/watch?v=aXOwc17nx_s

Discussions

- Wasteful processing
  - Same frame processed multiple times
  - Alternate approach uses two layers of LSTM
    - Lower layer: short-term temporal variations (motion features)
    - Upper layer: long-term history learns to predict saliency
- Camera motion vs. object motion
  - Object motion matters
    - FG–BG separation
    - Assign weights to FG

Quiz

Quiz 05-09

End of Module 05-09