IIT Jodhpur

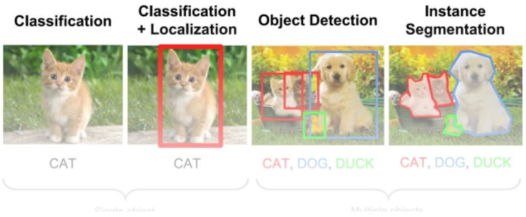
Biological Vision and Applications
Module 06-02: Neural Network based attention models

Hiranmay Ghosh

Classification-localization-segmentation

Role of attention

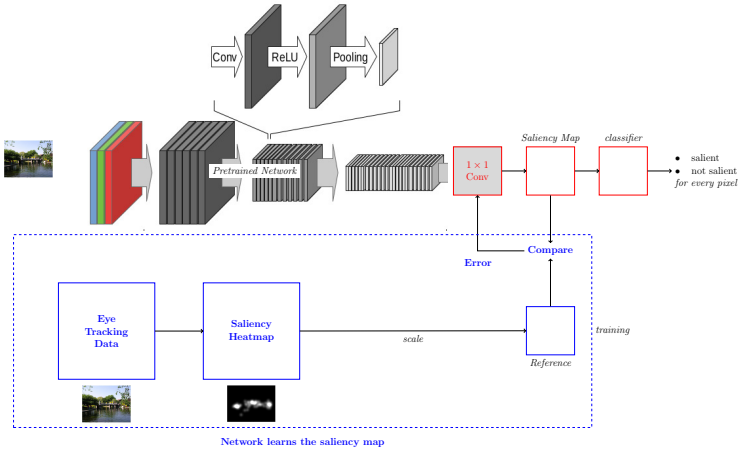
Progressive development of CNN based image processing



Attention and object recognition goes hand-in-hand

- The objects determine focus of attention
- Object recognition / segmentation takes place where there is attention

Basic Architecture



Does it implement bottom-up attention or top-down attention?

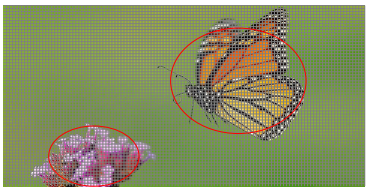
Attention and object detection

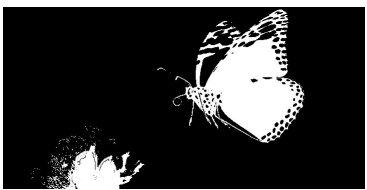
- Use CNN pre-trained for object detection
  - Not enough training data for saliency
  - Objects lead to saliency
- In neural network based architectures
  - Attention and object detection complement each other
    - Find salient locations (where objects are likely to be there)
    - Detect objects at those locations

Soft attention vs. hard attention

- Soft attention
  - Graded saliency values
  - Fixation traverses from location with highest saliency to lowest
- Hard attention
  - Binary saliency values
  - Fixation at the region with saliency $= 1$
  - One or very few "salient" objects in a scene
- NN based attention models generally use hard attention

How to train?





Objects and saliency

Saliency-cut algorithm

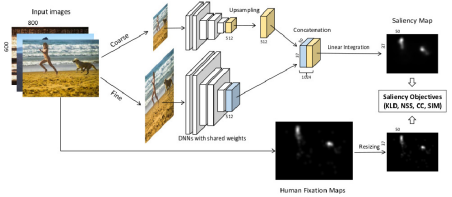
- Saliency of the nearby pixels should be similar
- The image can be divided into 'superpixels'
  - Areas of near uniform color/texture
- Adjust saliency values to encourage locations in nearby superpixels to have homogeneous saliency
  - Minimize:
    $O(S) = \sum_i (s_i^{new} - s_i)^2 + \sum_{i,j} w_{ij} (s_i^{new} - s_j^{new})^2$
  - Weights w_{ij} decreases with physical distance
  - Optimal weights are learned

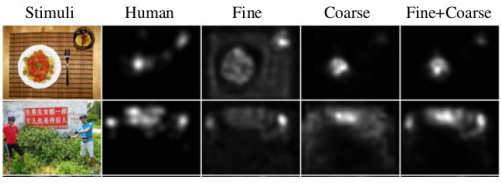
Something like graph-cut algorithm

Graph-cut algorithm (slide deck)

Multi-scale analysis

SALICON: Saliency in Context

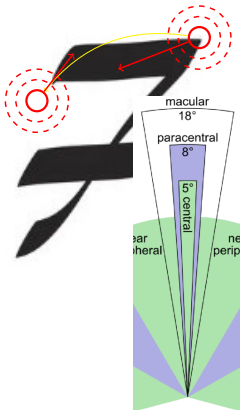




- Coarse level captures context; fine level captures local contrasts
  - Usually 2 or 3 levels of resolution is found to be sufficient

Recurrent Attention Models

Saliency is dynamically constructed

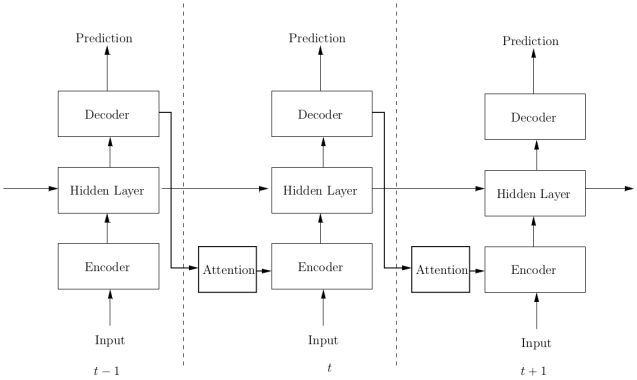


- We look at a small part of a scene at a time
- Where we look at next depends on what we see
  - Macular/peripheral vision guides the direction of eye movement
  - ... plus, the task at hand
- Saliency map of a scene is not computed in one go
  - Constructed dynamically over time
    - As and when needed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

Recall EdPuzzle Assignment – Visual Attention: Just in time representation

Attention-based RNN Architecture

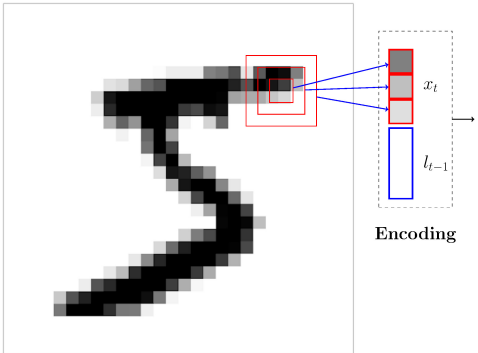
AB-RNN



- RNN and the "Attention" module are trained together

Implementation example

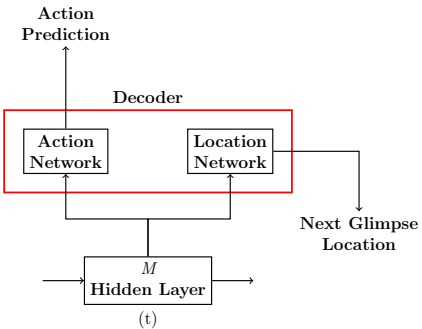
Attention-based object recognition



- 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

Decoder



- Each of Action and Location Networks is an NN
- "Action" can be different in different contexts:
  - Predicting the object
  - Locating a target
  - Navigating (car, pedestrian, drone, ...)
  - ...

Training

Reinforcement Learning

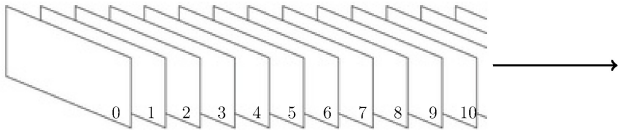
- Model for "hard attention"
- Training for optimal saccades
  - Training based on back-propagation does not work
  - Reinforcement learning used
  - Reward after each time-step
- In the case of object recognition
  - Reward $r_t = 1$ if the object is classified correctly at time step t
  - $r_t = 0$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

Reinforcement learning (tutorial slides)

Discussions

- Attention and object recognition goes hand-in-hand
  - Example of task-based attention
  - 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

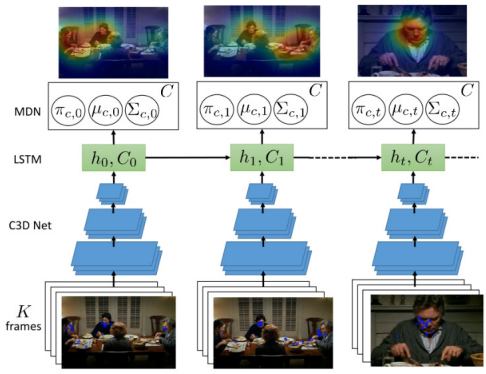


Why processing video frame by frame does not work ?

- 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 06-02

End of Module 06-02