Machine Learning for Visual Analytics and Compression

YAO WANG, PROFESSOR,
ELECTRICAL AND COMPUTER ENGINEERING AND BIOMEDICAL ENGINEERING
TANDON SCHOOL OF ENGINEERING
NEW YORK UNIVERSITY
HTTPS://WP.NYU.EDU/VIDEOLAB/
OCT. 10, 2018
Current and Recent Research Interests

- Joint optimization of video coding and delivery in networked video applications
  - 360 degree video streaming
- Medical image analysis and applications
- Video analytics
- Video coding and video adaptation
- Perceptual video quality modeling
Robust Vehicle Tracking at Urban Intersections

- **Challenges**
  - Severe occlusion in dense traffic
  - Vanishing point (non-bird eye view)
  - Shadows and illumination changes

- Developing a deep learning network that can **simultaneously detect and track** a video object
  - Detect bounding tubes that cover moving objects in short video segments
  - Extension of faster region-CNN, which detects bounding boxes in individual frames

- **Thanks:** *Chenge Li, Yilin Song*
State of Art: Region-CNN

- Extract features from an image (e.g. VGG)
- Generate object proposals (bounding boxes of vary size)
  - Each proposal specified by box position parameters and objectness score
- Refine proposals and classify each detected box

Extending Region-CNN For Moving Object Detection in Video

- Consider a video segment consisting of multiple frames
- Use 3D and 2D convolution for feature extraction (C3D and VGG)
- Generate object proposals (bounding tubes of various sizes and orientations)
- Refine proposals and classify each detected tube (car, van, bus, pedestrian, …)
Chenge Li, Gregory Dobler, Yilin Song, Xin Feng, Yao Wang, "TrackNet: Simultaneous Detection and Tracking of Multiple Objects",
• State-of-art learnt image coding: train different networks for different target rates.

• Our goal: train a single multi-layer network that can provide optimal performance at multiple bit rates or generate layered bit streams

• Each layer learns additional features to reconstruct the residue, optimized for best rate distortion performance

• Important for networked video applications with dynamically varying bandwidths

• Thanks: Chuanmin Jia and Zhaoyi Liu
Loss function considers both distortion and rate:

$$L(\theta_f, \theta_g, \theta_r) = D(x, g(f(x; \theta_f) + q); \theta_g) + \lambda R(f(x; \theta_f) + q; \theta_r)$$

- Train multiple networks using different $\lambda$, to reach different rate-distortion points
- Take the set of $\lambda$ that achieve the lower convex hull of resulting rate-distortion points
- Complete different networks for different target rates (°
Preliminary Results

Learning for compression and analytics simultaneously

• No need to decompress and then do visual analytics at receiver
• Camera site can do compression and visual analytics simultaneously
• We choose saliency detection as a generic analytics task in the hope that the learnt features are good for other analytics tasks
• The trained compression branch has the potential to yield better visual quality because the encoded features should contain visually important features.
• Jointly optimize for both rate-distortion and saliency prediction
• Thanks: Alp Aygar, Chuanmin Jia
Left: predicted saliency map, middle: ground truth, right: input color image. Training and testing Data from [http://salicon.net](http://salicon.net)
Acknowledgement

NYU WIRELESS Industrial Affiliates