Recurrent Neural Networks for object detection

1st Bin Qasim Ahmad
Technical University of Munich
Department of Informatics
Munich, Germany
ahmad.qasim@tum.de

2nd Pettirsch Arnd
Technical University of Munich)
Department of Informatics
Munich, Germany
a.pettirsch@outlook.de

***Abstract*—ToDo**

***Index Terms*—TBD.**

I. Introduction

A. Image and Video Object Detection in general

- Image object detection history.
  - Bayesian methods before deep learning
  - ImageNet challenge and VID [15]
  - Deep Learning and AlexNet [16]
- Single stage and 2-stage image object detectors.
  - A two-stage pipeline firstly generates region proposals, which are then classified and refined. [17]
  - A single-stage method is often more efficient but less accurate. Directly regress on bounding boxes and classes. [18], [19]
- Why is video object detection harder?
  - Large size
  - Motion blur
  - Quality of the dataset
  - Partial occlusion
  - Unconventional Poses

B. Recurrent Neural Networks in general

ToDo

II. Feature-based Video Object Detection

A. Definition

ToDo

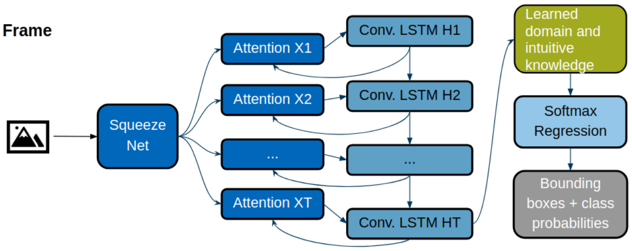
B. Recurrent Multi-fram Single Shot Detector for Video Object Detection

ToDo

C. Mobile Video Object Detection with Temporally Aware Feature Maps

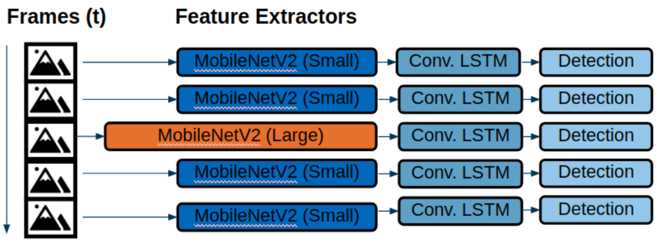
ToDo

D. Feature Selective Small Object Detection via Knowledge-based recurrent attentive neural networks



- Compute feature maps using a modified SqueezeNet architecture.
- Propagate the features through a Recurrent Attentive Neural Network, comprised of:
  - Attention Mechanism to detect key areas within the feature maps.
  - Convolutional LSTM for temporal feature propagation.
- Reverse gaussian feature maps are combined with the maps obtained from Conv. LSTM.
  - These feature maps are based on learnable mean and covariance terms.
  - This prior knowledge is derived from the assumption that traffic signs are always located at the bias of the center.

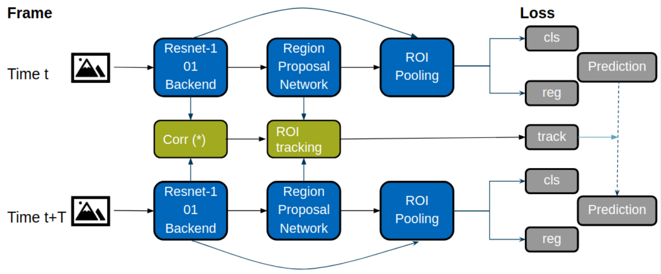
E. Looking fast and slow: memory-guided mobile video object detection



- Run multiple feature extractors sequentially or concurrently to obtain feature maps.
  - The idea is to use small and large feature extractors to optimize performance.
- Aggregate and refine these feature maps using convolutional LSTM based memory network.

    - To improve speed of LSTM network, add skip connections and LSTM state groups.
- Apply SSD-style detection on refined features to obtain classification and bounding boxes.
- Use a reinforcement learning based policy for selection of which feature extractor to run.
- Large and small frame extractors can run in parallel using asynchronous mode.

F. Detect to Track and track to detect



- Compute Convolutional feature maps using a Resnet-101 architecture.
- Use a RPN (region proposal network) to find candidate regions in the frame.
- ROI Pooling layer, to classify boxes and refine their coordinates (regression).
- Find correlation features between two frames' feature maps and do ROI tracking.
- Due to memory constraints, use tracklets, which are class-based optimal paths in video.

III. Box-Level-based Video Object Detection

A. Definition

ToDo

B. Object Detection from Video Tubelets with Convolutional Neural Networks

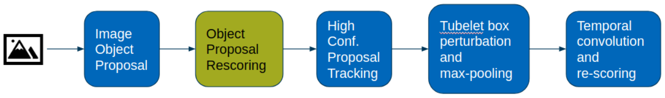
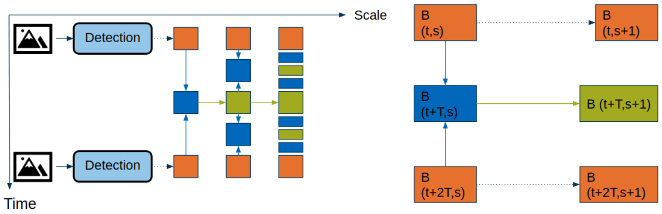


Image Object Proposal
Object Proposal Rescoring
High Conf. Proposal Tracking
Tubelet box perturbation and max-pooling
Temporal convolution and re-scoring

- Use selective search algorithm to generate around 2000 object proposals on each frame.
- Use GoogleNet for feature extraction and then 30 SVMs for 30 VID classes to generate object proposal scores for each object proposal.
- Track high confidence targets bi-directionally.
- Two kinds of Perturbations:
    - The first method is to generate new boxes around each tubelet box on each frame by randomly perturbing the boundaries of the tubelet box.
    - The second perturbation method is to replace each tubelet box with original object detections that have overlaps with the tubelet box beyond a threshold.
- Train a class-specific TCN using the tubelet features as input. The inputs are time series including detection scores, tracking scores and anchor offsets. The output values are probabilities whether each tubelet box contains objects of the class

C. Optimizing Video Object Detection via Scale-Time Lattice



- Apply object detection on keyframes extracted adaptively.
    - The extraction policy is based on number of objects and amount of movement in frames.
    - If higher number/movement of objects in frames then higher extraction rate.
- Propagation and refinement unit, propagates the frames temporally and refines spatially.
- For temporal propagation, use a small network such as resnet-18 to extract box features and a regressor to predict object movement from t to t + T.
- For spatial refinement, use a regressor to refine the bounding boxes over increasing scale.

D. Context Matters: Refining Object Detection in Video with Recurrent Neural Networks

ToDo

E. Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking

ToDo

IV. Flow-based Object Detection

A. Definition

Todo

B. Deep Feature Flow for Video Recognition

Todo

V. Comparison of different approaches

A. General

ToDo

B. Conclusion Perfomance

Todo

C. Conclusion Prediction Quality

Todo

VI. Outro

A. Conclusion

Todo

B. Further work

Todo

ACKNOWLEDGMENT

Todo

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first . . ."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove the template text from your paper may result in your paper not being published.