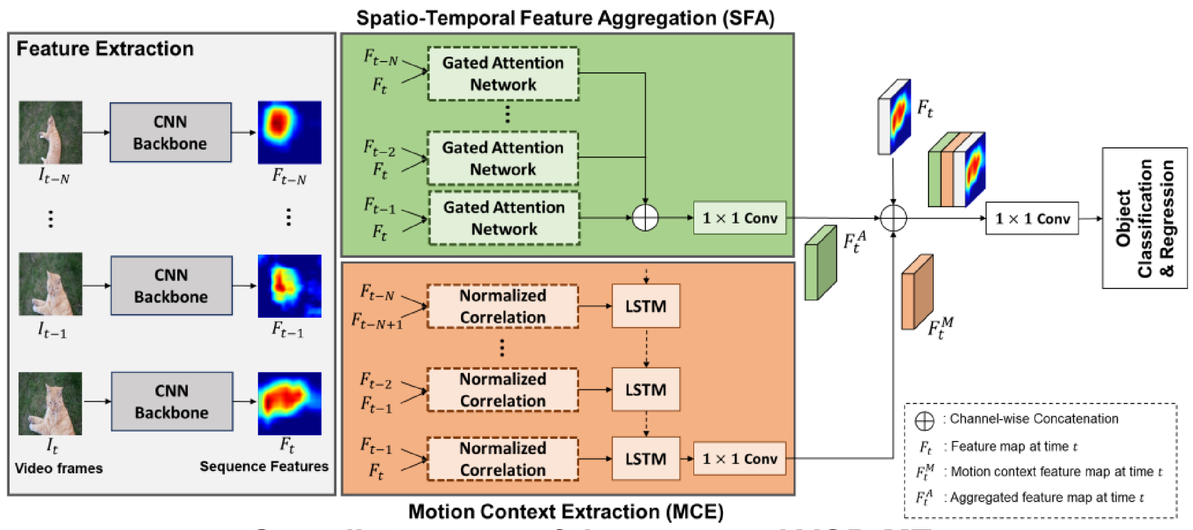
Video object detection using object's motion context and sptio-temporal feature aggregation

Jaekyum Kim[1]*, Junho Koh[1]*, Byeongwon Lee[2], Seungji Yang[2] and Jun Won Choi[1]
[1]Signal Processing & Artificial-intelligence Laboratory Hanyang University, Seoul Korea
[2]T3K Vision AI Product Center of Excellence, SK Telecom, Seongnam Korea
*** Indicates the equal contribution**

Summary

- We propose a novel one-stage video object detection method called VOD-MT, which exploits both temporal redundancy and motion context obtained from the image sequence.
- For exploiting temporal redundancy, spatio-temporal feature aggregation block generate the spatio-temporal feature maps by combining the spatial feature maps with the attention weights determined by the gated attention block.
- The proposed method extracts the motion cue by applying LSTM to the correlation maps between the adjacent frames.
- The experiments on the ImageNet VID benchmark [1] demonstrate that the proposed method achieves the significant gain over the baseline object detector and outperforms the existing one-stage video object detectors.

Proposed Video Object Detector



< Overall structure of the proposed VOD-MT >

Overall Structure

- The proposed VOD-MT builds on two main blocks: 1) spatio-temporal feature aggregation (SFA) block and 2) motion context extraction (MCE) block.
- The feature maps extracted from the backbone CNN network are fed into the gated attention network and the outputs are aggregated to produce the spatio-temporal feature maps F_t^A and the motion context extraction block computes the normalized correlation map between each pair of feature maps and finally extracts the motion feature F_t^M using LSTM.
- Finally, the feature map F_t for the present frame, the aggregated spatio-temporal feature map F_t^A, and the motion feature F_t^M are concatenated and used for final box regression and classification.
- The proposed method can be applied to every one-stage detectors and we have developed the VOD-MT based on the SSD [2] and RetinaNet [3].

Gated Attention Network

- The role of the gated attention network is to scale the pair of the feature maps F_t and $F_{t-\tau}$ with the attention weights according to the relevance to the current frame feature maps.
- The feature maps F_t and $F_{t-\tau}$ are concatenated and the 3×3 convolutional layers and sigmoid function are applied to the concatenated feature maps to obtain two attention maps.
- These attention maps contain the weights and multiplied to each feature map F_t and $F_{t-\tau}$.
- The weighted feature maps are concatenated and passed through 1×1 convolutional layer to produce the final spatio-temporal feature map.

REFERENCES

[1] O. Russakovsky et al, "Imagenet large scale visual recognition challenge," in *IJCA*, 2015.
[2] W. Liu et al, "Ssd: signle shot multibox detector," in *ECCV*, 2016.
[3] T. Y. Lin et al, "Focal loss for dense object detection," in *ICCV*, 2017.
[4] M. Liu et al, "Mobile video object detection with temporraly-aware feature maps," in *CVPR*, 2018.
[5] X. Chen et al, "Temporally identity-aware ssd with attentional lstm," in *IEEE transactions on cybernetics*, 2019.
[6] B. Zhao et al, "Deep spatial-temporal joint feature representation for video object detection," in *Sensors*, 2018.

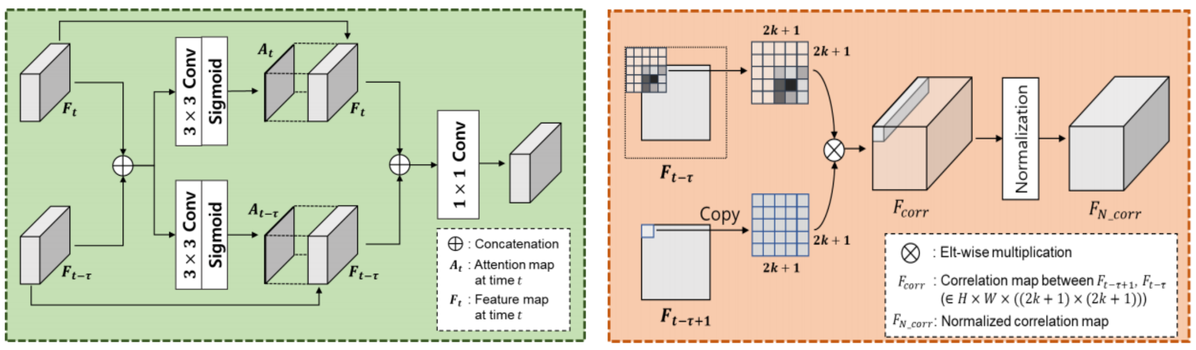
Motion Context Extraction

- The role of the Motion Context Extraction (MCE) block is to produce the motion context using normalized correlation block and LSTM network.
- The normalized correlation block computes the correlation map between $F_{t-\tau}$ and $F_{t-\tau+1}$ following equation.

$$F_{corr}(p,q,i,j) = F_{t-\tau+1}(p,q) \cdot F_{t-\tau}(p+i, q+j),$$

where $-k \le i,j \le k$ are the offsets, $F_{t-\tau+1}(p,q)$ denotes the feature vector at (p,q)th pixel of the feature map $F_{t-\tau+1}$ and $x \cdot y$ denotes the normalized inner product of two vectors x and y.

- The correlation maps for all pairs of adjacent frames are input to the convolutional LSTM network to find the representation of temporal information from the sequence of the correlation maps.
- Finally, the output of the LSTM is passed through 1×1 convolution layer to produce the motion context.



< Structure of the gated attention network and normalized correlation >

Experiment Results

Abation study on the ImageNet VID validation set

Method	Proposed VOD-MT based on SSD300				
	(a)	(b)	(c)	(d)	(e)
SFA block		✓		✓	✓
MCE block			✓	✓	✓
Post-processing (Seq-NMS)					✓
mAP (%)	66.7	69.04 (↑2.34)	69.00 (↑2.30)	71.03 (↑4.33)	73.20 (↑6.50)
run-time (ms)	23	32	38	55	-

Method	Proposed VOD-MT based on RetinaNet				
	(a)	(b)	(c)	(d)	(e)
SFA block		✓		✓	✓
MCE block			✓	✓	✓
Post-processing (Seq-NMS)					✓
mAP (%)	77.89	78.60 (↑0.71)	78.33 (↑0.44)	79.23 (↑1.34)	80.17 (↑2.28)
run-time (ms)	110	123	129	157	-

- Table presents mAP and run-time of the ablation study of the proposed method (a) to the method (e) and the proposed method achieved the gain of 4.33% and 1.34% over the SSD and RetinaNet, respectively.

Network	Backbone	Base Detector	mAP (%)
LSTM-SSD [4]	MobileNet	SSD	54.4
TSSD(-OTA) [5]	VGG-16	SSD	65.4
Method of [6]	VGG-16	SSD	69.5
Proposed VOD-MT	VGG-16	SSD	71.0

- We compare the proposed VOD-MT with state-of-the-art one-stage video object detectors and the proposed VOD-MT significantly outperforms the existing methods.

Conclusions

- In this paper, we proposed a new one-stage video object detector, which exploits the temporal redundancy of the adjacent frames and motion context to enhance the detection performance
- The experiments conducted on the ImageNet VID benchmark demonstrate that the proposed method achieves the significant performance gain over the state-of-the-art one-stage video object detectors.