Convolutional Gated Recurrent Networks for Video Segmentation

Mennatullah Siam*, Sepehr Valipour*, Martin Jagersand, Nilanjan Ray

UNIVERSITY OF ALBERTA

Introduction

- Motion used to be main cue for video segmentation.
- Current state of the art uses deep networks that do not consider motion / dynamics in video.
- The relatively few attempts that were made to incorporate temporal data into deep networks did not result in a consistent and significant improvement over single image segmentation.
- Using recurrent architectures is shown to be effective for data streams. E.g. text classification, speech synthesis and translation.
- Regular recurrent architectures are not practical for image processing.
  - They are fully connected
  - They do not preserve spatial connectivity.

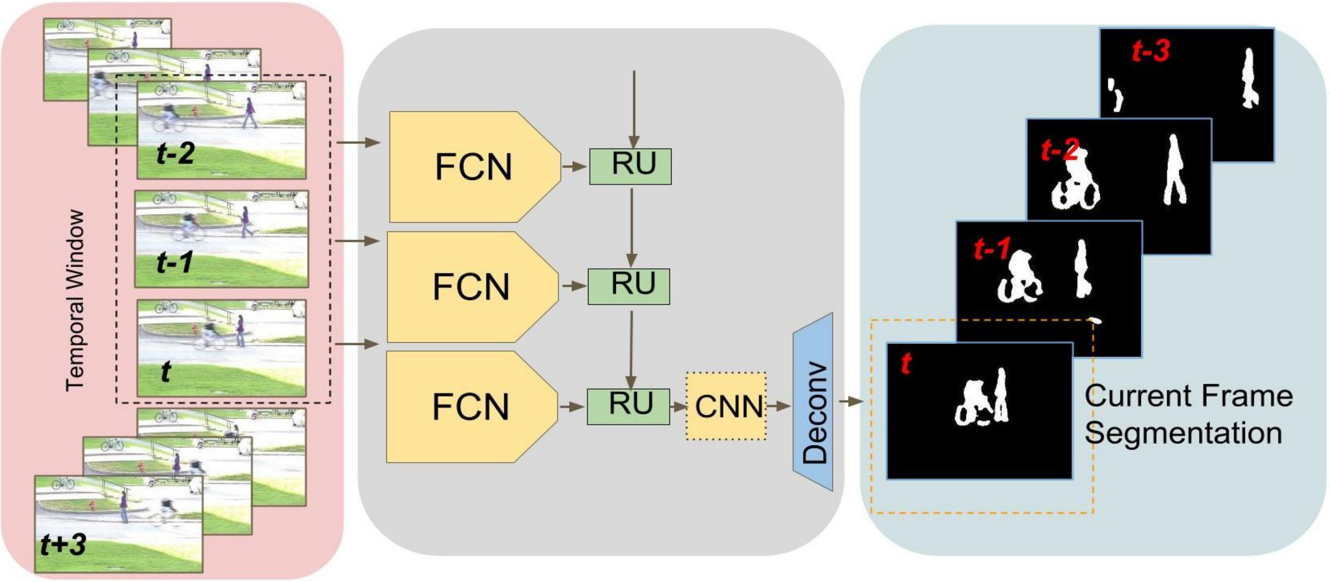
We propose a recurrent fully convolutional network that is able to process a video stream online and produce segmentation using both the current image information and the implicit observed dynamics of the sequence.

Overview

- We embed a fully convolutional network inside a convolutional gated recurrent unit. Our network takes in a sliding window of images and produces a segmentation corresponding to the last image in it.
- Each image is processed by the FCN network. Its output along with the hidden state are convolved with the the weights and produce gates weights and next hidden state.

Conv-GRU units:

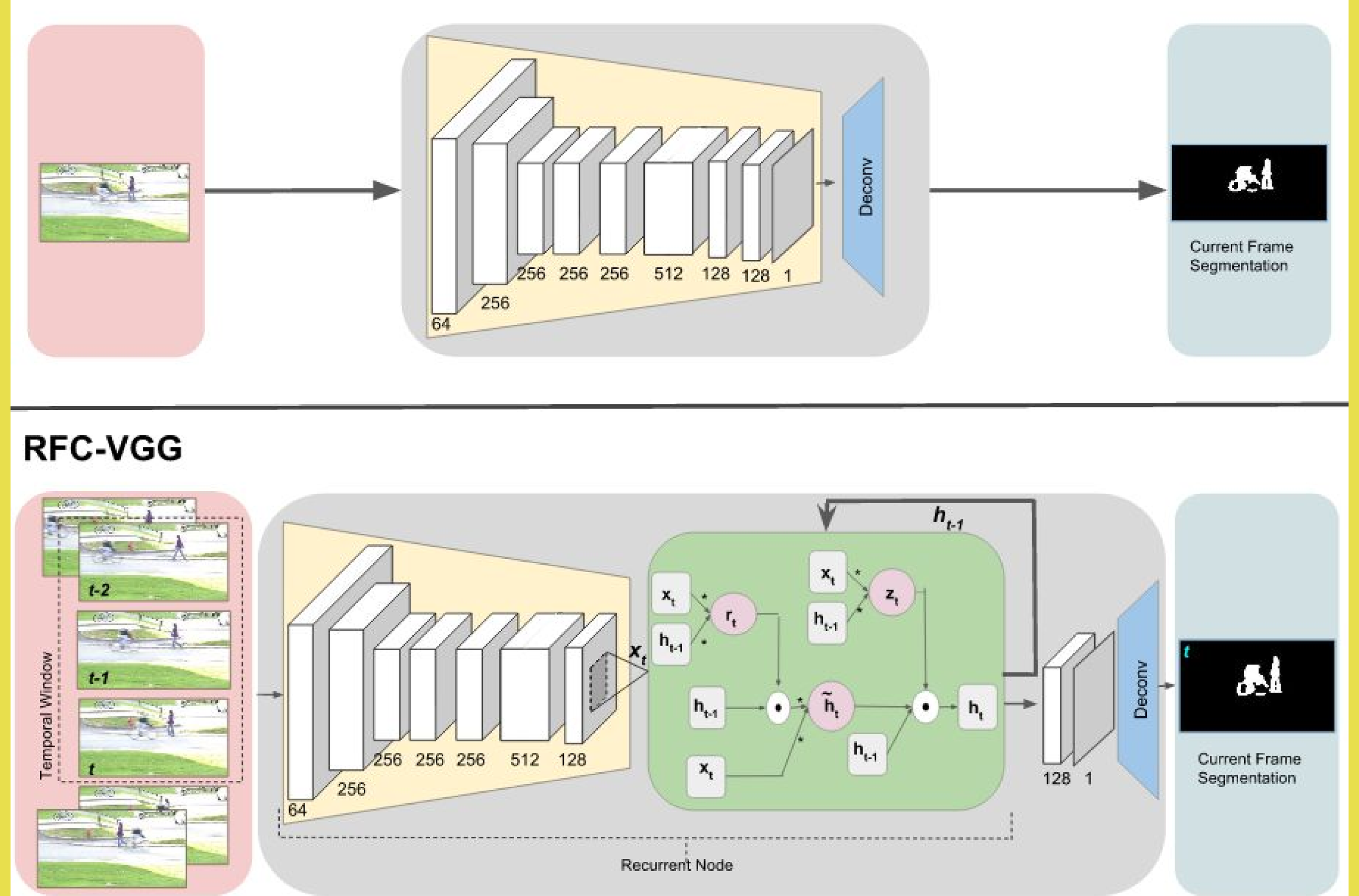
$$z_t = \sigma(W_{hz} * h_{t-1} + W_{xz} * x_t + b_z)$$
$$r_t = \sigma(W_{hr} * h_{t-1} + W_{xr} * x_t + b_r)$$
$$\hat{h}_t = \Phi(W_h * (r_t \odot h_{t-1}) + W_x * x_t + b)$$
$$h_t = (1 - z_t) \odot h_{t-1} + z \odot \hat{h}_t$$



Overview of Convolutional Gated Reccurrent FCN Method for Video Segmentation

Architecture

- Our original architecture RFC-VGG incorporating convolutional GRU.
- Another variant with Dilated Convolution is used in RFC-Dilated.

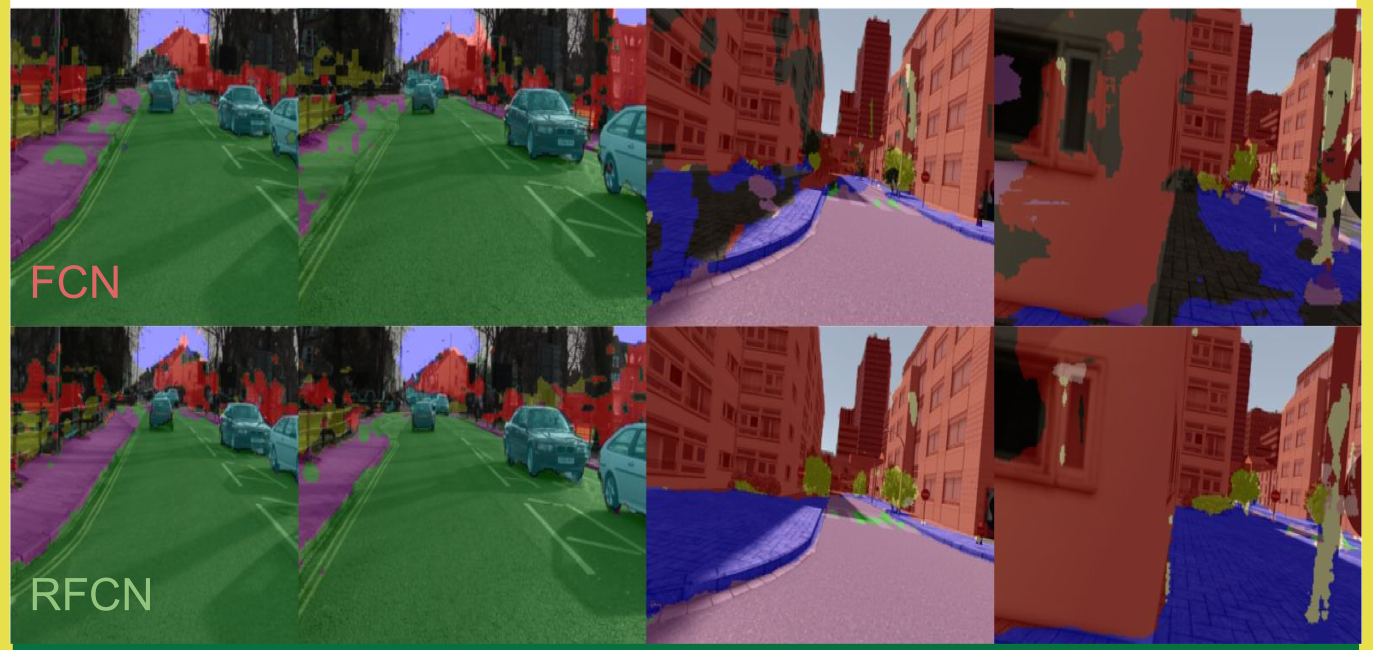


FC-VGG vs RFC-VGG Architecture for Segmentation.

Results

- Experiments on Moving MNIST, SegTrackV2 and DAVIS.

		Precision	Recall	F-measure	IoU
SegTrack V2	FC-VGG	0.7759	0.6810	0.7254	0.7646
	RFC-VGG	**0.8325**	**0.7280**	**0.7767**	**0.8012**
DAVIS	FC-VGG	0.6834	0.5454	0.6066	0.6836
	RFC-VGG	**0.7233**	**0.5586**	**0.6304**	**0.6984**



Qualitative comparison between FC-VGG(top) and RFC-VGG(bottom) on CamVid and Synthia.

Results cont'd

- Experiments Synthia, CityScapes, CamVid and AR-Drone collected sequences.

	Mean Class IoU	Per-Class IoU						
		Sky	Building	Road	Sidewalk	Vegetation	Car	Pedestrian
FC-Dilated	46.7	86.3	69.1	87.8	63.7	60.8	63.6	21.4
RFC-Dilated	**48.3**	**87.5**	**69.1**	**89.4**	**69.4**	**62.0**	**64.3**	**24.3**
Super Parsing[5]	42.0	-	-	-	-	-	-	-
Segnet[3]	46.4	-	-	-	-	-	-	-

	FCN	RFCN
Synthia	0.755	**0.812**
ARDrone	0.857	**0.871**

Qualitative comparison on AR-Drone.

Discussion

- RFCN give a consistent improvement over its baseline network. On average **5%**.
- Different type of recurrent units were tested
  - Conventional gated recurrent units can still improve the results over the baseline. Only practical for small images.
  - The Convolution Recurrent Units perform better.
  - Convolutional GRU is the winner.
- Different methods for training were tested.
  - ADADELTA is the best optimizer.
  - End-to-end training does better than stage by stage training.
- More convolutional layers added to baseline to verify the source of improvement.
  - These addition did not help or made the performance worse.
  - Therefore improvement is from using temporal data.

Conclusion

- In this paper, we presented a novel approach to incorporating temporal information for video segmentation.
- We tested the method on both synthesized and real data. We showed that by having a recurrent layer after either probability map or feature map can improve the performance.