SAVGAN: SELF-ATTENTION BASED GENERATION OF TUMOUR ON CHIP VIDEOS

Sandeep Manandhar[1], Irina Veith[2,3], Maria Carla Parrini[2,3], and Auguste Genovesio[1]

[1]IBENS, ENS, [2]Institut-Curie, [3]INSERM,PSL Univ, 75005, Paris

1. Introduction

Generation of videomicroscopy sequences will become increasingly important in order to train and evaluate dynamic image analysis methods. The latter are crucial to the study of biological dynamic processes such as tumour-immune cell interactions. However, video generation is not an easy task because of

- **the huge amount of data each sample contains**
- **the complexity to account for all the trajectories and interaction of objects present within them**

To this end, we propose a self-attention based generative model for videomicroscopy sequences that aims to take into account for the full range of interactions within a spatio-temporal volume of 32 frames. To reduce the computational burden of such a strategy, we consider the Nyström approximation of the attention matrix. This approach leads to significant improvements in reproducing the structures and the proper motion of videomicroscopy sequences as assessed by a range of existing and proposed quantitative metrics.

2. Saptio-Temporal Self-Attention

Self-Attention is a technique to consider relationship between distant (here spatio-temporally) points in a feature map. However, the complexity of computing a full self-attention map as shown in eq. (1) is $\mathcal{O}(n^2)$. This operation becomes expensive when considering long duration videos.

$$\mathbf{S} = \sigma\left(\frac{\mathbf{A}}{\sqrt{C}})\right). \tag{1}$$

Therefore we rely on Nyström approximation [3] of the eq.(1) which reduces the complexity to $\mathcal{O}(n)$ when a small subset of landmarks $\tilde{\mathbf{x}}$ are chosen. Landmarks are randomly sampled in each training iteration.

$$\tilde{\mathbf{S}} = \sigma\left(\frac{\mathbf{f}(\mathbf{x})^T\mathbf{g}(\tilde{\mathbf{x}})}{\sqrt{C}}\right)\sigma\left(\frac{\mathbf{f}(\tilde{\mathbf{x}})^T\mathbf{g}(\tilde{\mathbf{x}})}{\sqrt{C}}\right)^{+}\sigma\left(\frac{\mathbf{f}(\tilde{\mathbf{x}})^T\mathbf{g}(\mathbf{x})}{\sqrt{C}}\right). \tag{2}$$

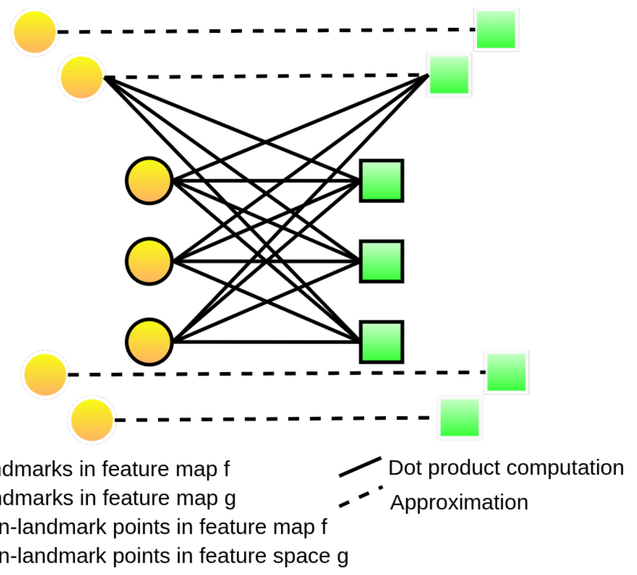

Figure 1. Nyström Approximation scheme. Dot products of landmarks and non-landmark points are not depicted for the sake of neatness.

Equation 2 breaks the big $\mathbf{A}$ into three smaller matrices which are subjected to softmax function separately.

3. Network Architecture

The schematic below shows the generator and the discriminators of SAVGAN. The output and the channel sizes are labelled above and under the corresponding layers, respectively.

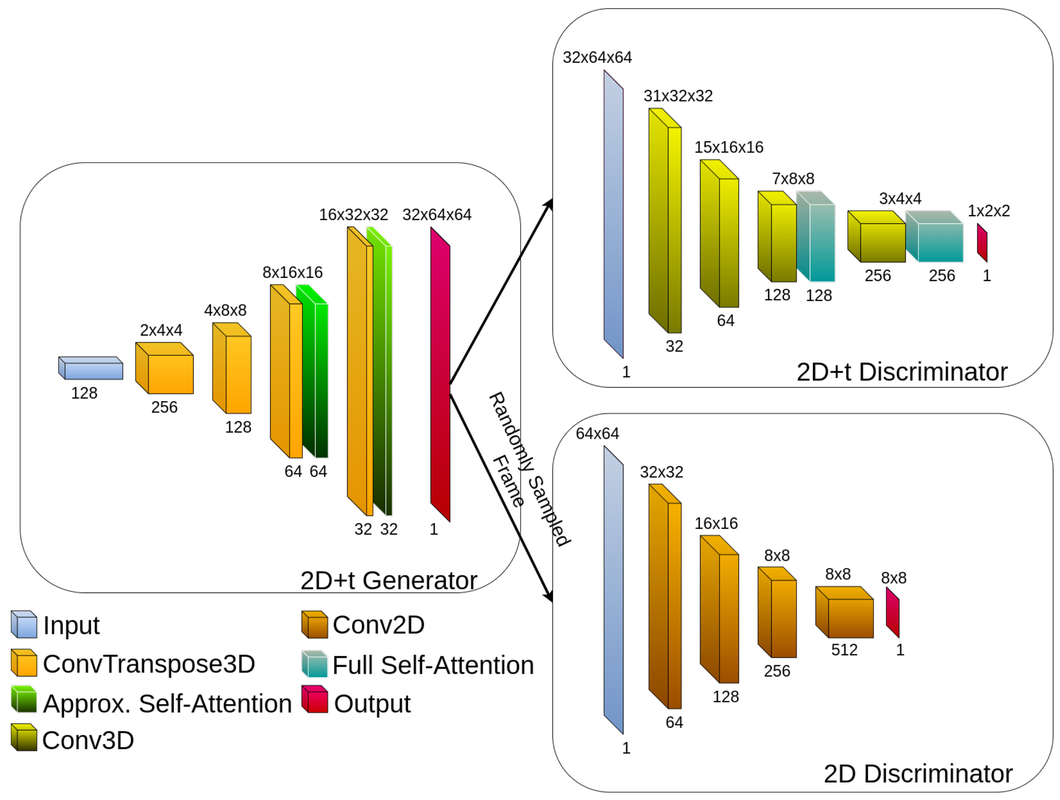

Figure 2. Network Architecture

Loss function

We employ Wasserstein loss with gradient penalty term to optimize the generator loss (L_G) and the discriminator losses (L_I and L_V). The losses are given as follows:

$$L_V = \mathop{\mathbb{E}}_{\tilde{p}\sim\mathbb{P}_g}[\mathcal{D}_V(\tilde{p})] - \mathop{\mathbb{E}}_{p\sim\mathbb{P}_r}[\mathcal{D}_V(p)] + \lambda \mathop{\mathbb{E}}_{\hat{p}\sim\mathbb{P}_{\hat{p}}}[(\|\nabla_{\hat{p}}\mathcal{D}_V(\hat{p})\|_2 - 1)^2], \tag{3}$$

$$L_I = \mathop{\mathbb{E}}_{\tilde{q}\sim\mathbb{Q}_g}[\mathcal{D}_I(\tilde{q})] - \mathop{\mathbb{E}}_{q\sim\mathbb{Q}_r}[\mathcal{D}_I(q)] + \lambda \mathop{\mathbb{E}}_{\hat{q}\sim\mathbb{Q}_{\hat{q}}}[(\|\nabla_{\hat{q}}\mathcal{D}_I(\hat{q})\|_2 - 1)^2], \tag{4}$$

$$L_G = \mathop{\mathbb{E}}_{\tilde{p}\sim\mathbb{P}_g}[\mathcal{D}_V(\tilde{p})] + \lambda_i \mathop{\mathbb{E}}_{\tilde{q}\sim\mathbb{Q}_g}[\mathcal{D}_I(\tilde{q})], \tag{5}$$

4. Results

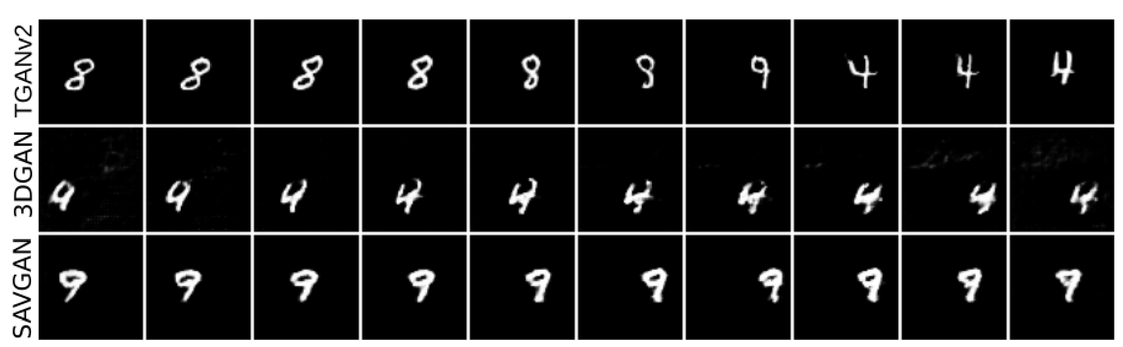

Figure 3. The generated moving digit for MMNIST dataset is consistent throughout the sequence with SAVGAN. The flipping of digit from frame to frame is more frequent in the other two methods.

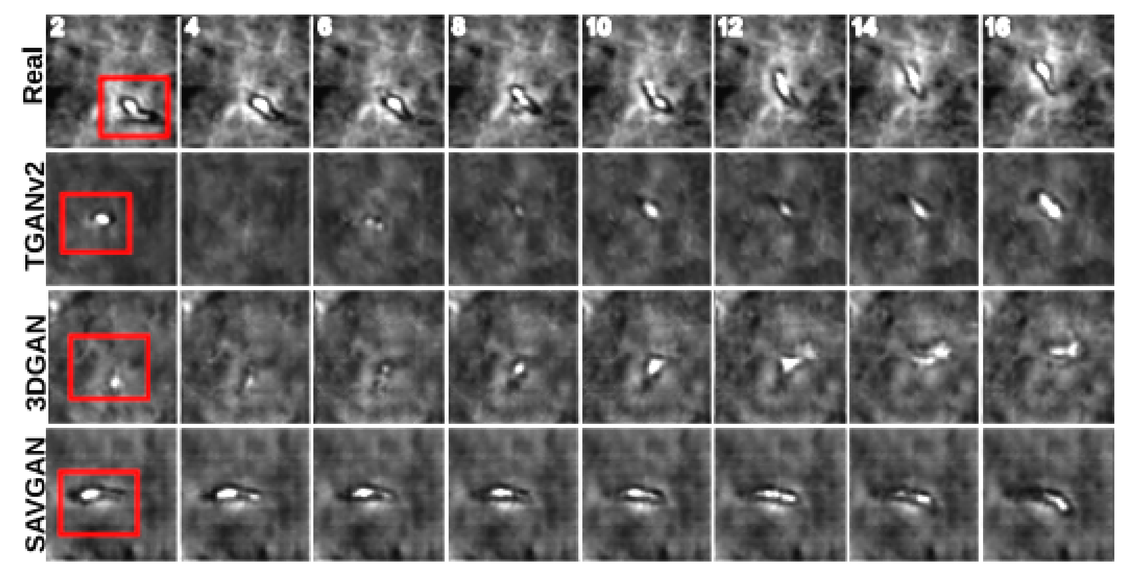

Figure 4. SAVGAN produces better sequence of dynamic immune cell for TOC dataset.

5. Comparison

We compare our results with TGANv2 [1] and as an ablation, our network without the use of self-attention layers (3DGAN). We use MMNIST dataset which is a synthetic video dataset depicting a single digit moving in a horizontal direction bouncing off the frame edges throughout a sequence. This dataset provides a way to quantitatively assess the performance of different networks in a clear manner. We then use the tumour-on-chip video dataset to generate realistic videos of immune cell activity. Two popular metrics [2]: FVD and AACD scores have been used for the comparison regarding this dataset.

Dataset	**TOC**			**MMNIST**		
Method	**TGANv2**	**3DGAN**	**SAVGAN**	**TGANv2**	**3DGAN**	**SAVGAN**
FVD	738.75	729.25	**389.3**	84.97	**80.4**	81.29
AACD	0.739	2.26	**0.477**	1.218	1.040	**0.469**
Flips/seq	-	-	-	3.42	3.1	**1.76**
Res18-cosTS	-	-	-	0.63	0.69	**0.76**

Table 1. Scores in bold face are better.

6. Remarks

In this work, we proposed SAVGAN, a self-attention based video generative adversarial network. The architecture is made up of a mix of convolutional and self-attention modules to better capture spatio-temporal interactions. Furthermore, in order to reduce the computational and memory load, we implemented a Nyström approximation of the full self-attention matrix. We believe SAVGAN provides a robust method to generate consistent videos and should benefit other tasks of interest such as domain transfer or segmentation in biological imaging analysis, and possibly in other fields.

References

[1] Saito et al. "Train Sparsely, Generate Densely: Memory-efficient Unsupervised Training of High-resolution Temporal GAN". In: *IJCV*. May 2020.

[2] Unterthiner et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges". In: *CoRR* abs/1812.01717 (2018).

[3] Xiong et al. "Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention". In: *AAAI*. 2021.