β-Multivariational Autoencoder for Entangled Representation Learning in Video Frames

Fatemeh Nourilenjan Nokabadi[1], Setareh Rezaee Oshternian[2]

[1]LVSN-REPARTI, Université Laval, Québec, Canada

[2]University Medical Center Groningen, University of Groningen, The Netherlands

Introduction

- A new motion modelling for objects in video sequences is proposed, where the fundamental parameters are dependent on each other with a covariance matrix.
- β-Multivariational Autoencoder (βMVAE) is developed to learn an MGD prior from video frames for use as part of a single object-tracking in the form of a decision-making process.
- By using U-Net instead AE neural network, both posterior estimation and segmentation of the network have been improved.

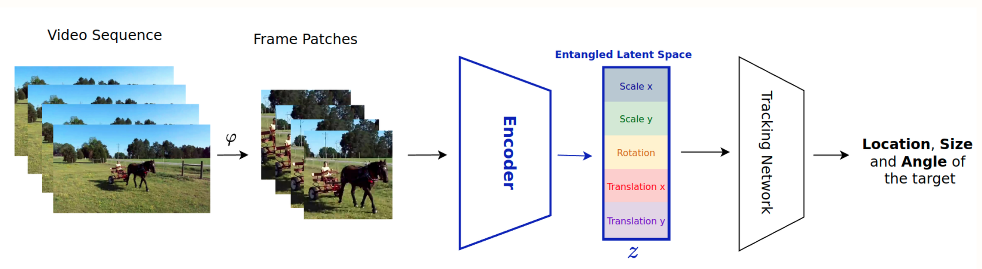

Figure 1: Overview of the entangled representation application in the future tracking system.

Preliminaries

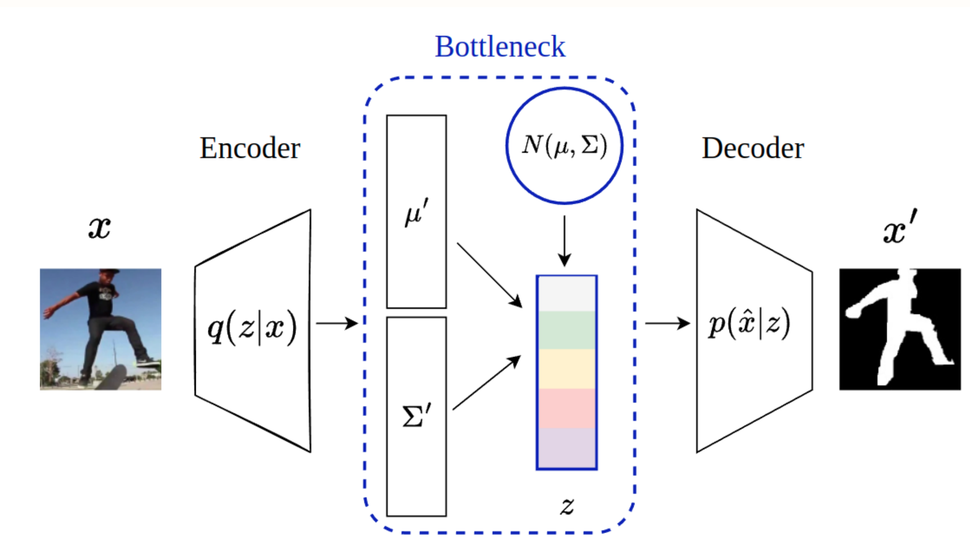

Figure 2: Overview of our proposed method.

$$\mathcal{L}_{\beta\text{VAE}} = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \beta D_{KL}(q_\phi(z|x)\|p(z)) \quad (1)$$

$$\mathcal{L}_{\beta\text{MVAE}} = \mathcal{L}_{cons}(\hat{x}, gt) + \beta\mathcal{L}_{KL} \quad (2)$$

$$\mathcal{L}_{cons}(\hat{x}, gt) = \mathcal{L}_{ce}(\hat{x}, gt) + \mathcal{L}_{\mathcal{J}}(\hat{x}, gt) \quad (3)$$

Proposed Method

Object Motion Modeling

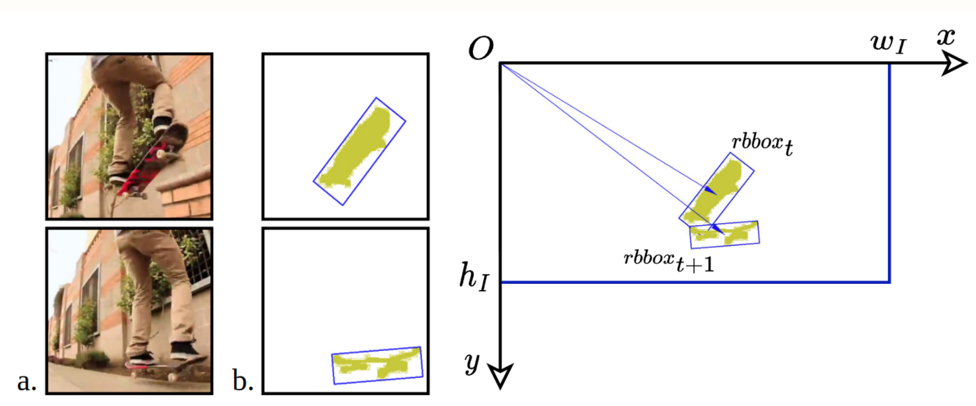

Figure 3: Motion modelling in two successive frames.

$$[x', y', 1]^T = G_t \times [x, y, 1]^T \quad (4)$$

where G_t is:

$$G_t = \begin{bmatrix} \Delta s_x \times \cos\theta & -\Delta s_y \times \sin\theta & \Delta x \\ \Delta s_x \times \sin\theta & \Delta s_y \times \cos\theta & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

Bottleneck structure and distribution

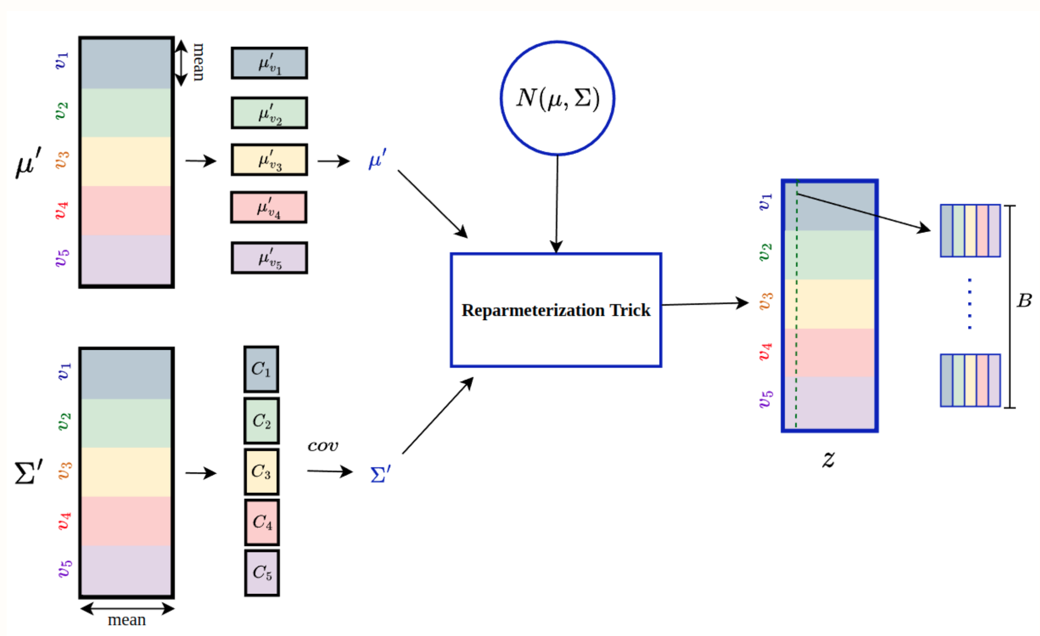

Figure 4: The posterior parameters, μ' and Σ', computed by two FC layers following the encoder's output.

Reparameterization trick

Using the diagonal elements of Σ', the variable's variances σ_i's are computed in this step. Then, we obtain the coefficients of the linear transformation to map the prior samples p to the encoder's posterior q using the lower and upper bounds of each distribution. The coefficients are calculated as $a_i = \frac{\sigma_i'}{\sigma_i}$ and $b_i = \mu_i' - \mu_i \times \frac{\sigma_i'}{\sigma_i}$ for the i^{th} part of the bottleneck, therefore:

$$z_i = a_i\epsilon_i + b_i \quad (6)$$

where z_i is forming the i-th part of our latent set of samples z.

Table 1: Lower Bound (LB) and Upper Bound (UB) of the latent distributions.

variables	$\mathcal{N}(\mu,\sigma)$	$LB(\mu-\sigma)$	$UB(\mu+\sigma)$
v_1	$\mathcal{N}(1.06, 0.52)$	0.55	1.59
v_2	$\mathcal{N}(1.06, 0.54)$	0.53	1.61
v_3	$\mathcal{N}(0, 0.43)$	−0.43	0.43
v_4	$\mathcal{N}(0.07, 0.34)$	−0.27	0.41
v_5	$\mathcal{N}(0.08, 0.74)$	−0.66	0.82

Result and discussions

Posterior and log-likelihood evaluation

$$\mathcal{D}_{Mah} = (\mu - \mu') \times (\frac{\Sigma + \Sigma'}{2})^{-1} \times (\mu - \mu') \quad (7)$$

Method	Data	$\mathcal{D}_{Mah}$	NLL	MSE
βMVAE	training	0.10	4.21	0.71
	validation	0.05	5.32	0.64
	test set	0.06	-	0.65
βMVUnet	training	$1.05 \times e^{-6}$	1.98	0.30
	validation	$1.22 \times e^{-6}$	2.71	0.30
	test set	$0.47 \times e^{-6}$	-	0.24

Video object segmentation

Figure 5: Binary masks generated for several frames of DAVIS16 set. a) input frame, b) annotation, c) βMVAE result, d)βMVUnet output.

Saliency detection

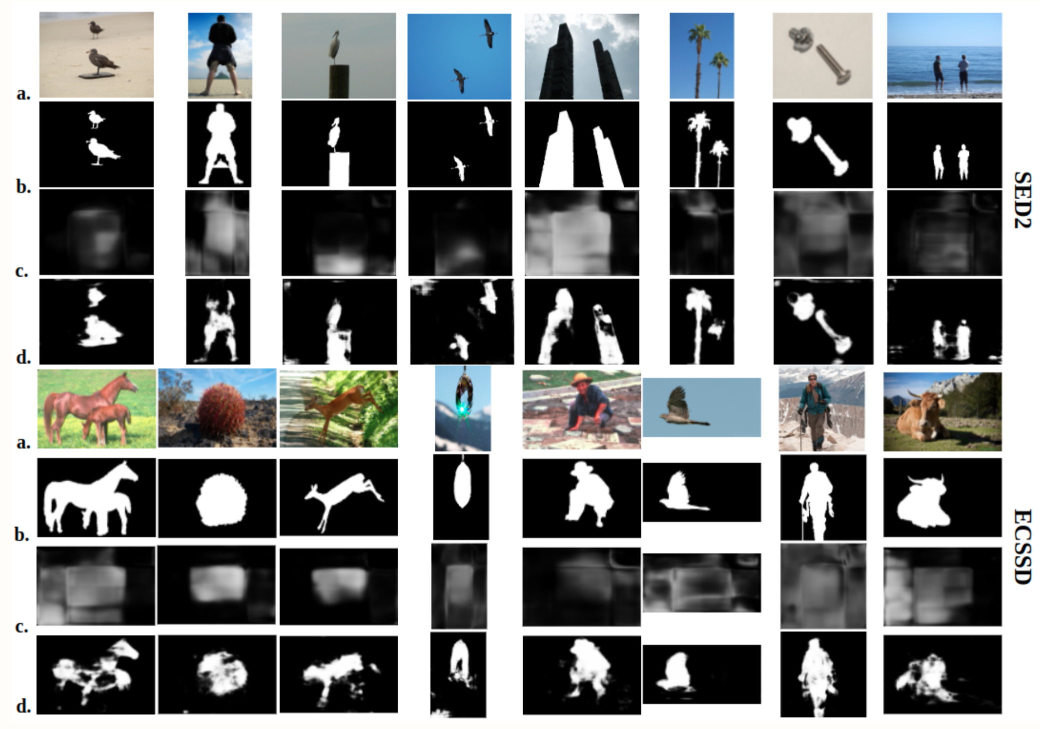

Figure 6: The visualized probabilistic maps as saliency maps for SED2 and ECSSD datasets. a)Input Image, b)Annotation, c)βMVAE map, d)β MVUnet map.

Summary and conclusions

1. We formulate a novel dynamic to model the single object's motion across video frames.
2. The βMVAE is developed to learn a multivariate Gaussian distribution with a full covariance matrix from raw pixels in addition to the object mask of the frame patches.
3. A novel trick is introduced for the bottleneck reparameterization to map a set of the prior samples to the posterior parameters to add the randomness in the proposed structure.
4. The bottleneck is directly trained by computing KullbackLeibler (KL) divergence between the prior and the estimated posterior instead of learning the expectation of the lower bound.
5. The outcomes of posterior estimation and segmentation mask creation are enhanced by the U-Net architecture.

Article Info

WiML Symposium @ ICML 2024, July 2024, Vienna, Austria