RESEARCH ACTIVITY OVERVIEW

Experimenting with Autoencoders to Encode World State Representations

M. Sc. Blaž Meden

Ferrol, Galicia, 16.9.2019

Introduction

- We explored the possibility of using autoencoders to represent compressed world states in a given scenario.
- First step was to evaluate autoencoder reconstruction capabilities. Two models were used (AE, VAE), with different latent sizes (16, 64), different attention factors for foreground and background (0.2, 0.33, 0.5, 0.67, 0.8) and also different object size factors (1, 2, 3).
- Second step included introduction of forward model (FM) in between encoder and decoder as a simple densely connected network, consisting of 5 dense layers in size of latent space. FM has two inputs, one is the latent space in time t, second is the action. Action is defined as a global angle of robot direction of movement, normalized in between -1 and 1 and repeated into tensor of the same size as latent space.

Overview

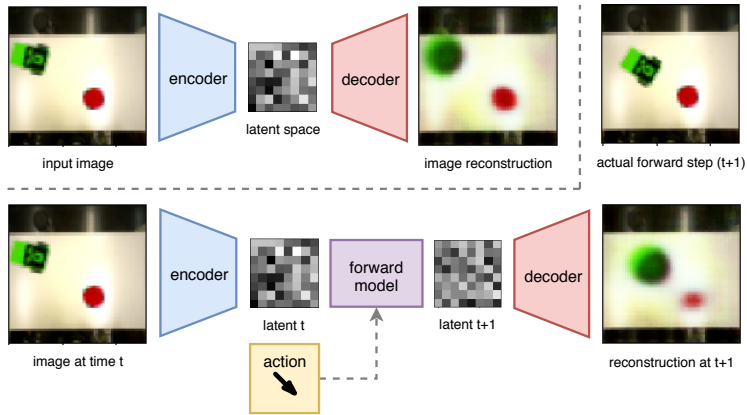
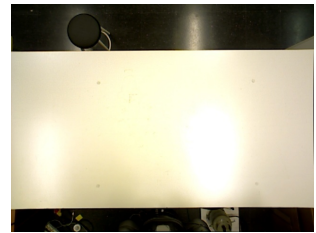


Figure 1: Pipeline of our experiments. Top marked section evaluated the reconstruction quality of images. Section at the bottom included separately trained forward model to predict latent representation, which can be used to reconstruct image in next time step.

Dataset

(a)

(b)



(c)

Figure 2: Building blocks of our generated dataset: a) a robot, b) an object of interest, and c) the background image.

Dataset

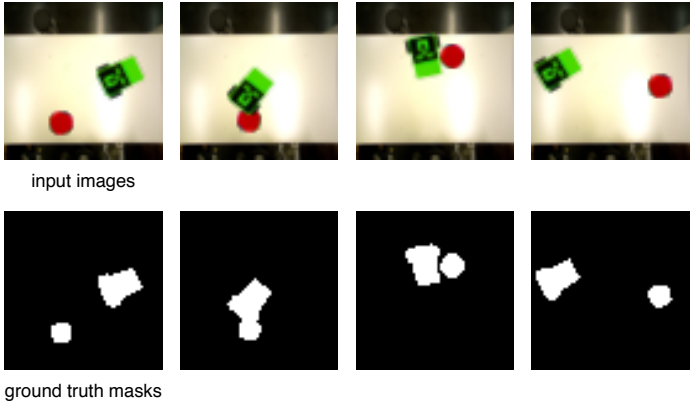
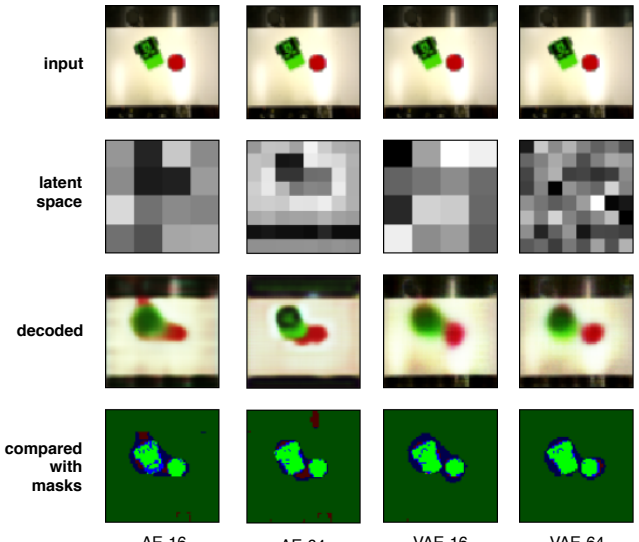


Figure 3: Generated image examples (top) and corresponding ground truth masks (bottom).

Latent Space Comparison



Reconstruction Results

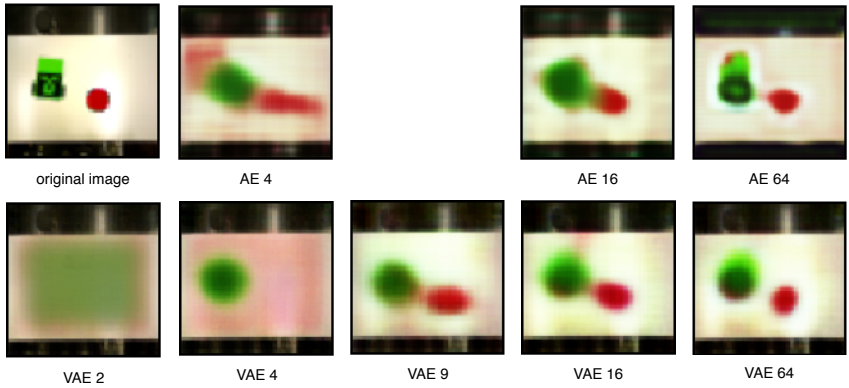


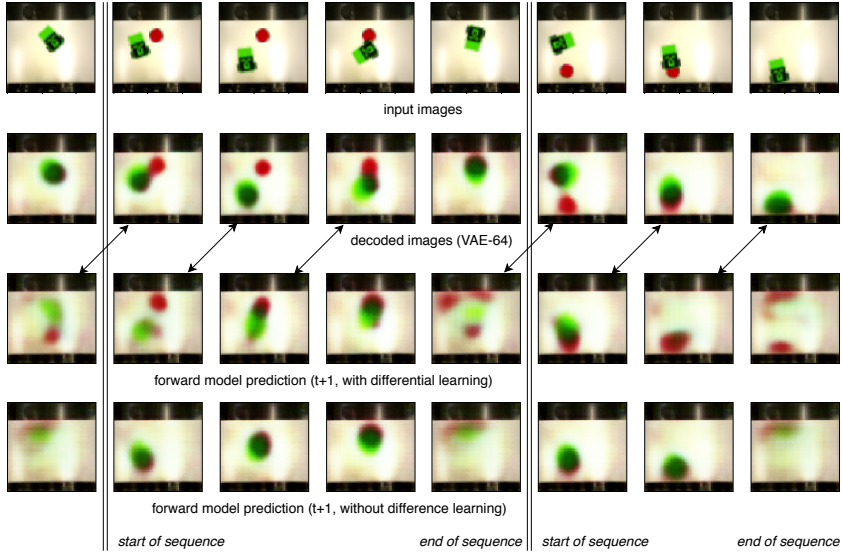
Figure 5: Qualitative reconstruction results from first part of the experiments. In left top corner an original reference image is shown. Top: image reconstructions with different sizes of latent spaces using regular AE (namely 4, 16 and 64; sizes 2 and 9 are omitted due to the topological constraints). Bottom: reconstructions using VAE (latent size 2, 4, 9, 16 and 64).

Reconstruction Results

Model	Latent size	Back att.	Obj. att.	Obj. size	IoU (avg)	IoU (stdev)
VAE	64	0.2	0.8	3	0.65520	0.07533
VAE	64	0.33	0.67	3	0.61377	0.07079
AE	16	0.2	0.8	3	0.57823	0.07041
VAE	64	0.2	0.8	2	0.57033	0.11629
AE	64	0.2	0.8	1	0.56386	0.15444
VAE	64	0.33	0.67	2	0.52029	0.11642
VAE	16	0.2	0.8	3	0.50127	0.07992
AE	64	0.2	0.8	3	0.50019	0.05898
VAE	64	0.02	0.98	1	0.47663	0.06301
VAE	16	0.33	0.67	3	0.46870	0.08014

Table 1: Top 10 rated AE / VAE models in terms of average Intersection over Union (IoU) during tests. Best results are achieved with larger latent space sizes.

Forward Model Results



Forward Model Results

Model	Diff. learning	Latent size	Obj. size	IoU (avg)	IoU (stdev)
VAE + FM	✓	64	3	0.35866	0.21140
VAE + FM	✗	64	3	0.35050	0.20697
VAE + FM	✓	16	3	0.19372	0.12674
VAE + FM	✗	16	3	0.18675	0.14988
AE + FM	✓	64	3	0.16381	0.12321
AE + FM	✗	64	3	0.14849	0.12366
AE + FM	✓	16	3	0.12763	0.10202
AE + FM	✗	16	3	0.05069	0.05329

Table 2: Results of second experiment in terms of average Intersection over Union (IoU). Various configurations were tested. VAE and AE were used to produce latent representations of the images and FM represents the forward model that was used to predict state in the next timestep. We tested latent sizes 64 and 16 with both autoencoders.

To conclude...

- We tested different parameters with AE and VAE. Larger latent spaces yielded best results. Having larger object sizes also helped quite a lot. VAE-64 performed best overall.
- VAE uses entire latent space to encode non-interpretable information, while AE performs downsample and upsample, using only the spatial area around moving robot (interpretable latent space).
- Forward model is able to predict the movement of the robot to some degree.
- Reconstruction step and forward model both introduce some uncertainty and errors to final reconstruction.
- Forward model cannot predict unknown robot state after the sequence is finished, since the next robot initialization is random. This probably impacts the quantitative results the most, although there are failure cases, where the next action and next image are within the same sequence.