Video Generation using Multimodal VAE

Anubhav Mittal

EE491A

Supervised by

Ravindra Yadav, PhD student, Department of Electrical Engineering
Prof. Vinay P. Namboodiri, Department of Computer Science and Engineering
Prof. K. Vasudevan, Department of Electrical Engineering

November 20, 2019

Introduction

- We first give a brief overview of generative models : likelihood based (VAEs) and GANs.
- We move onto the describe the existing literature on audio and video generation using these methods.
- We then introduce a new method for video generation, which we base on the multimodal VAE model defined in Wu and Goodman, 2018. As this did not achieve respectable results, we introduce changes to the model by replacing the MSE error with a learned similarity metric (Larsen et al., 2016)
- Finally, we discuss a class of flow based generative models (Dinh et al., 2014, 2016; Kingma and Dhariwal, 2018) which has gained popularity in recent times. We also suggest how we can use these approaches in video generation.

Generative modelling

- Generative models capture the joint probability $p(X, Y)$, or just $p(X)$ if there are no labels. A generative model includes the distribution of the data itself, and tells you how likely a given example is.
- Discriminative models capture the conditional probability $p(Y|X)$. A discriminative model ignores the question of whether a given instance is likely, and just tells you how likely a label is to apply to the instance.

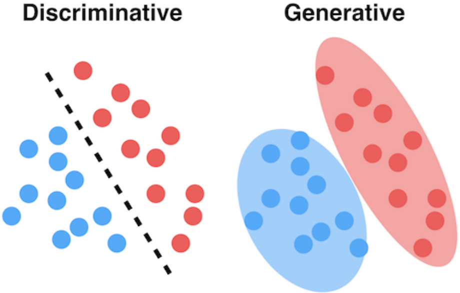


Figure: Discriminative vs Generative models

Generative Adversarial Networks (GANs)

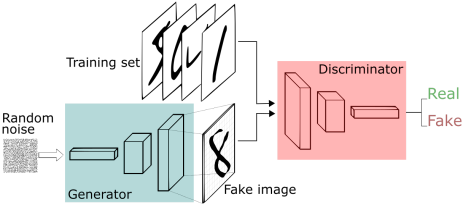


Figure: Generative Adversarial Network (Image courtesy: towardsdatascience.com)

- A GAN consists of two networks: the generator network $G_{\theta_g}(z)$ maps latents z to data space while the discriminator network assigns probability $y = D_{\theta_d}(x) \in [0, 1]$ that x is an actual training sample and probability $1 - y$ that x is generated by our model through $x = G_{\theta_g}(z)$ with $zp(z)$

Generative Adversarial Networks (GANs)

- The GAN objective is to find the binary classifier that gives the best possible discrimination between true and generated data and simultaneously encouraging Gen to fit the true data distribution.
- We thus aim to maximize/minimize the binary cross entropy:

$$L_{GAN} = log(D_{\theta_d}(x)) + log(1 - D_{\theta_d}(G_{\theta_g}(z)))$$

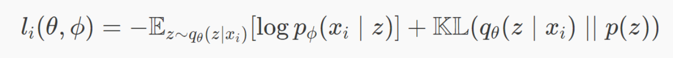
with respect to D / G with x being a training sample and $z \sim p(z)$.

Variational Autoencoders (VAEs)

- A VAE consists of two networks that encode a data sample x to a latent representation z and decode the latent representation back to data space, respectively:

$$z \sim Enc(x) = q_\theta(z|x), x' \sim Dec(z) = p_\phi(x|z)$$

- The VAE regularizes the encoder by imposing a prior over the latent distribution $p(z)$. Typically $z \sim \mathcal{N}(0, I)$ is chosen. The VAE loss is the sum of the negative expected log likelihood (the reconstruction error) and KL divergence between the obtained posterior and its prior:



$$l_i(\theta, \phi) = -\mathbb{E}_{z \sim q_\theta(z|x_i)}[\log p_\phi(x_i \mid z)] + \mathbb{KL}(q_\theta(z \mid x_i) \mid\mid p(z))$$

Video generation using GANs

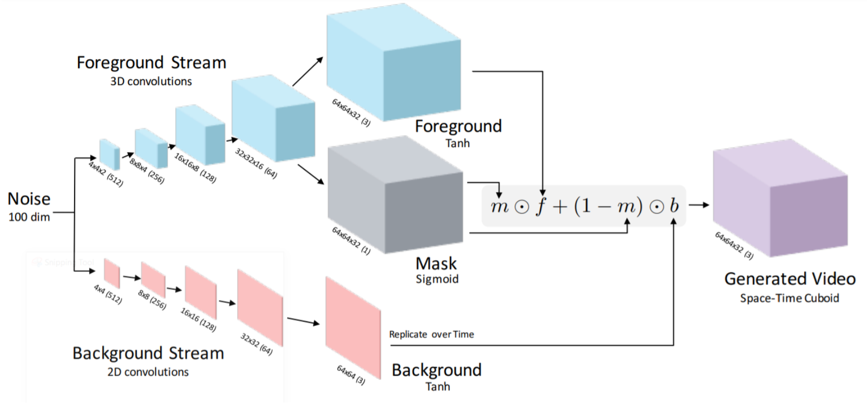


Figure: Video generator network. Image courtesy: Vondrick et al., 2016

Video generation using VAEs

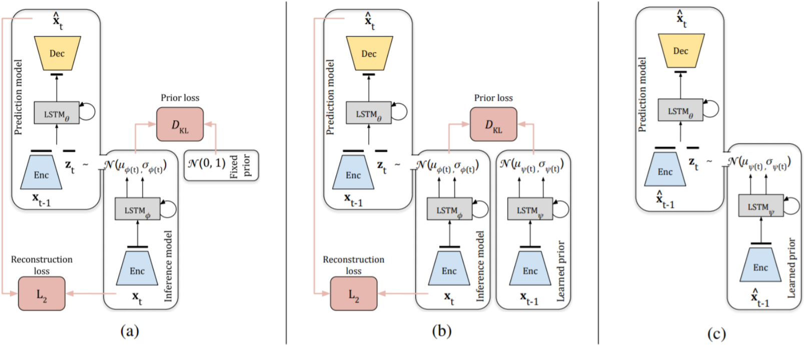


Figure: The video generation model by Denton and Fergus. (a) Training with a fixed prior (SVG-FP); (b) Training with learned prior (SVG-LP); (c) Generation with the learned prior model. The red boxes show the loss functions used during training. Image courtesy: Denton and Fergus, 2018

Multimodal generative model

- In the multimodal setting we assume the N modalities, $x_1, ..., x_N$, are conditionally independent given the common latent variable, z, then the ELBO becomes:

$$\mathrm{ELBO}(X) \triangleq \mathbb{E}_{q_\phi(z|X)}[\sum_{x_i \in X} \lambda_i \log p_\theta(x_i|z)] - \beta\, \mathrm{KL}[q_\phi(z|X), p(z)].$$

- They approximate the joint posterior as a product of experts of individual posteriors computed by the encoder network of VAE and the prior for latent variables. As PoE does not uniquely specify its component gaussians, to train the individual sub-networks, we need to also train for each of the 2^N modalities. To prevent the computational intractability, they propose the loss function for an iteration to be the sum of the joint ELBO, individual ELBOs and k ELBO terms using k randomly chosen subsets, X_k:

$$\mathrm{ELBO}(x_1, ..., x_N) + \sum_{i=1}^{N} \mathrm{ELBO}(x_i) + \sum_{j=1}^{k} \mathrm{ELBO}(X_j)$$

Multimodal generative model

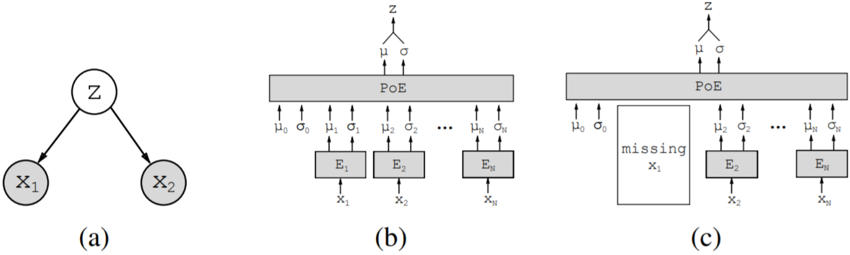


Figure: The model by Wu and Goodman. : (a) Graphical model of the MVAE. Gray circles represent observed variables. (b) MVAE architecture with N modalities. E_i represents the i-th inference network; μ_i and σ_i represent the i-th variational parameters; μ_0 and σ_0 represent the prior parameters. The product-of-experts (PoE) combines all variational parameters in a principled and efficient manner. (c) If a modality is missing during training, we drop the respective inference network. Thus, the parameters of $E_1, ..., E_N$ are shared across different combinations of missing inputs.. Image courtesy: Wu and Goodman, 2018

Method-I

- We use the multimodal model described above for video generation. The model is shown in Fig 7. Specifically, there are 2 modalities - the previous video frame x_{t-1} and the current video frame x_t. During training, these are passed into encoders followed by a LSTM, which gives us μ_a, σ_a and μ_b, σ_b respectively. Using the product of experts approach defined above, we compute loss according to the total ELBO:

  $$ELBO(x_a, x_b) + ELBO(x_a) + ELBO(x_b)$$

  where x_a, x_b correspond to x_t, x_{t-1} respectively. At test time, following an initial conditioning, the model should accept x_{t-1} as input, and produce both x_{t-1} and x_t as output.
- We trained the model on the moving MNIST dataset. The model parameters were similar those used in the Wu paper. However, the results were poor.

Method-I

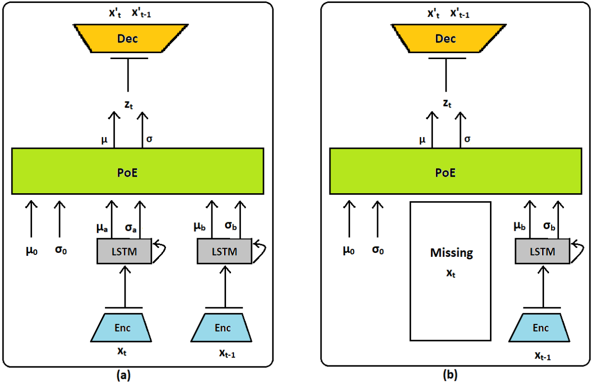


Figure: The first proposed model. (a) During training, both x_t and x_{t-1} are available. The model learns to generate both x_t' and x_{t-1}' and the dependency between the two. (b) During test time, only x_{t-1} is present. As the model has learned the dependency between the current and the next frame, it should be able to generate the next frame.

Method-I : Images

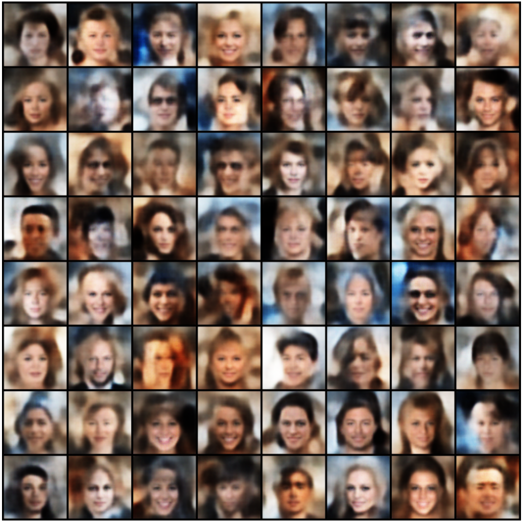


Figure: Images constructed from random noise in the first proposed model.

Method-I : images

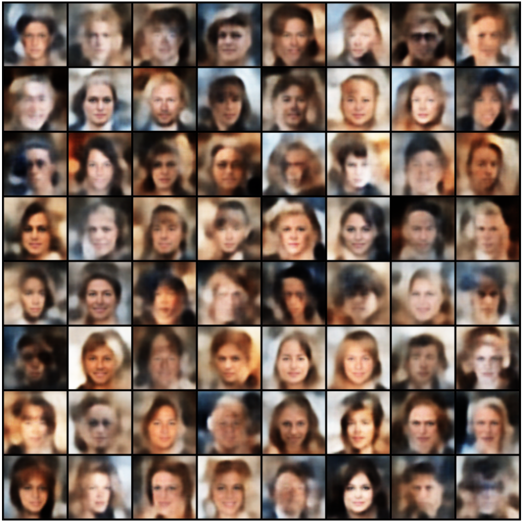


Figure: Images constructed with (attribute = male) in the first proposed model.

Learned similarity metric

- Element-wise reconstruction errors are not adequate for images and other signals with invariances.
- So in the work by Larsen et al., 2016, they propose replacing the VAE reconstruction (expected log likelihood) error term with a reconstruction error expressed in the GAN discriminator. To achieve this, let $Dis_l(x)$ denote the hidden representation of the l^{th} layer of the discriminator. They introduce a Gaussian observation model for $Dis_l(x)$ with mean $Dis_l(x')$ and identity covariance:

  $$p(Dis_l(x)|z) = \mathcal{N}(Dis_l(x)|Dis_l(x'), I)$$

  where x' is a sample from the decoder of VAE with x as input. Hence, the reconstruction loss of VAE is replaced by:

  $$\mathcal{L}_{llike}^{Dis_l} = -E_{q(z|x)}[logp(Dis_l(x)|z)]$$

Learned similarity metric

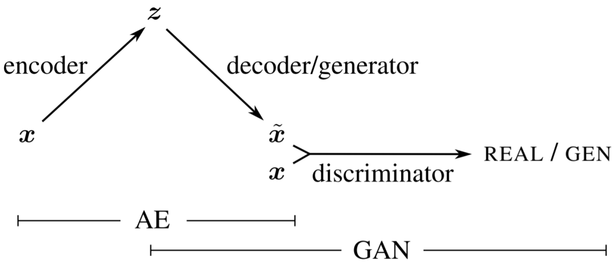


Figure: Overview of the network in Larsen et al. 2016. Image courtesy: their original paper.

- The total loss is given by the triple criterion:

$$\mathcal{L} = \mathcal{L}_{prior} + \mathcal{L}_{llike}^{Dis_l} + \mathcal{L}_{GAN}$$

Method-II

- We add a GAN discriminator at the output of the decoder/generator.
- The new ELBO becomes :

$$ELBO(X) = E_{q_\phi(z|X)}[\sum_{x_i \in X} \log p(Dis_l(x_i)|z)] + \sum_{x_i \in X}[E_{x_i \sim p_{data}} \log D_i(x) + E_{q_\phi(z|X)} \log(1 - D_i(Dec(z)_i))] - \beta KL[q_\phi(z|X), p(z)]$$

The first term is the modified reconstruction error, the second and third term are the GAN loss and the fourth term is the KL divergence between the posterior and its prior.

- For our case (two modalities), again, the total objective is:

$$ELBO(x_a, x_b) + ELBO(x_a) + ELBO(x_b)$$

Method-II : images

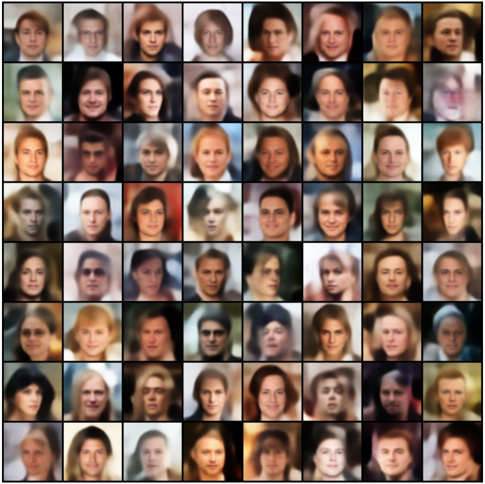


Figure: Images constructed with (attribute = male) in the second proposed model. Note that this model was not be trained till the losses were minimised, so there is still room for improvement.