Learning Disentangled Representations from Sequential Data

Aniket Das, Yatin Dandi

***Abstract*—We study the problem of learning disentangled representations through Variational Autoencoders, with special attention to sequential data such as animated sprites. In particular, we present, to the best of our knowledge, the only public implementation of Disentangled Sequential Autoencoder [1], a novel variational autoencoder architecture that enforces disentangled representations for the time invariant and time variable content in the data, just by careful design of the probabilistic graphical model, without any modifications to the ELBO objective. We also study and implement possible modifications to the loss function to improve sample quality. Implementation and ongoing experiment results are available in https://github.com/yatindandi/Disentangled-Sequential-Autoencoder**

I. Introduction

THE problem of learning disentangled representations in deep generative models is significant from various perspectives. It captures the human ability to represent visual content through independent factors of variation. From an information theoretic perspective, the latent code of a disentangled representation captures the most salient ways in which observations can differ from one another as the components of the disentangled representation vector are ideally statistically independent. From a practical viewpoint, because disentangled representations compress the content into a compact and interpretable form, they become semantically meaningful inputs for a variety of supervised learning tasks. Most of the recent work on learning disentangled representations through VAEs, such as the β-VAE [2] and the β-TCVAE [3] are based on modifying the ELBO objective by giving more weight to the term that enforces disentanglement

In the context of sequential data, the problem of disentangling the time invariant features shared across the sequence from the time variable features is one that arises naturally. A popular example is the task of learning disentangled representations for the content and dynamics of videos. Denton and Birodkar 2017. [4] approach this by training two separate networks, each with their own distinct losses for disentangling content and dynamics and achieve state of the art results in predicting future frames conditioned on the previously observed frames. Among these models, we focus on the Disentangled Sequential Autoencoder [1] because it enforces the disentanglement of content and dynamics simply by inegnious construction of the probabilistic graphical model of a Variational Autoencoder. A latent code f is explicitly used to represent content, and a sequence of latent codes $z_1, z_2, ..., z_T$ associated to each frame represents the dynamics, such as pose and position. In contrast to [4], the model focuses on learning the entire data distribution as opposed to predicting the next frame. To the best of our knowledge, our implementation of is the only publicly available implementation.

II. Probabilistic Model

The model trains using standard Variational Inference by maximizing the Evidence Lower Bound on the log likelihood of the training data. Using Bayes' Theorem,

$$P_\theta(x_{1:T}) = \frac{P_\theta(x_{1:T}|z_{1:T}, f)P_\theta(z_{1:T}, f)}{P_\theta(z_{1:T}, f|x_{1:T})}$$

We sample z_t and f from two independent Gaussians $\mathcal{N}(0, 1)$ and hence the prior has the form:

$$\begin{aligned} P_\theta(z_{1:T}, f) &= P_\theta(f)P_\theta(z_{1:T}) \\ &= P_\theta(f)\prod_{t=1}^{t} P_\theta(z_t) \end{aligned}$$

The generative distribution $P_\theta(x_{1:T}|z_{1:T}, f)$ is a factored Gaussian over the pixel values and is modelled by a deconvolutional neural network architecture. For the variational posterior $q_\phi(z_{1:T}, f|x_{1:T})$, we experiment with two types of factorization structures:

A. Full q Inference Model

It allows $z_{1:T}$ to depend on both f and $x_{1:T}$. Conditioning on f allows time invariant content to influence the dynamics. The distribution is then defined as:

$$q_\phi(z_{1:T}, f|x_{1:T}) = q_\phi(f|x_{1:T})q_\phi(z_{1:T}|f, x_{1:T})$$

We use the following model for $q_\phi(z_{1:T}|f, x_{1:T})$

$$q_\phi(z_{1:T}|f, x_{1:T}) = q_\phi(f|x_{1:T})\prod_{t=1}^{T} q_\phi(z_t|, z_{<t}, x_t, f)$$

The ELBO is then derived as follows:

$$\begin{aligned} logP_\theta(x_{1:T}) &= \mathbb{E}_{z_{1:T}, f\sim q_\phi(z_{1:T}, f|x_{1:T})}[logP_\theta(x_{1:T})] \\ &= \mathbb{E}_{z_{1:T}, f}[log\frac{P_\theta(x_{1:T}|z_{1:T}, f)P_\theta(f)\prod_{t=1}^{t} P_\theta(z_t)}{P_\theta(z_{1:T}, f|x_{1:T})}] \\ &= \mathbb{E}_{z_{1:T}, f}[log\frac{P_\theta(x_{1:T}|z_{1:T}, f)P_\theta(f)\prod_{t=1}^{t} P_\theta(z_t)\, q_\phi(z_{1:T}, f|x_{1:T})}{P_\theta(z_{1:T}, f|x_{1:T})\qquad q_\phi(z_{1:T}, f|x_{1:T})}] \\ &= \mathbb{E}_{z_{1:T}, f}[logP_\theta(x_{1:T}|z_{1:T}, f)] - \mathbb{E}_{z_{1:T}, f}[log\frac{q_\phi(f|x_{1:T})}{P_\theta(f)}] + \\ &- \sum_{t=1}^{T}\mathbb{E}_{z_{1:T}, f}[log\frac{q_\phi(z_t|z_{<t}, x_t, f)}{P_\theta(z_t)} + \mathbb{E}_{z_{1:T}, f}[log\frac{q_\phi(f, z_{1:T}|x_{1:T})}{P_\theta(f, z_{1:T}|x_{1:T})}] \\ &\geq \mathbb{E}_{z_{1:T}, f}[logP_\theta(x_{1:T}|z_{1:T}, f)] - D_{KL}(q_\phi(f|x_{1:T})||P_\theta(f)) \end{aligned}$$

$$-\sum_{t=1}^{T} D_{KL}(q_\phi(z_t|z_{<t}, x_t, f)||P_\theta(z_t))$$

$q_\phi(f|x_{1:T})$ and $q_\phi(z_t|z_{<t}, x_t, f)$ are chosen to be factored Gaussians. $q_\phi(f|x_{1:T})$ is modelled by a bidirectional LSTM over the sequence $x_{1:T}$ followed by an affine layer. $q_\phi(z_t|z_{<t}, x_t, f)$ is modelled by a bidirectional LSTM conditioned over both $x_{1:T}$ and f by feeding the concatenated vector $[x_t, f]$ to it at every timestep, followed by an RNN stacked over the bi-LSTM

B. Factorized q Inference Model

The probabilistic model for the factorized version is:

$$q_\phi(z_{1:T}, f|x_{1:T}) = q_\phi(f|x_{1:T}) \prod_{t=1}^{T} q_\phi(z_t|x_t, z_{<t})$$

The factorization differs from [1], which conditions the dynamics of a frame z_t only on the pixel content of the frame, x_t and hence uses

$$q_\phi(z_{1:T}, f|x_{1:T}) = q_\phi(f|x_{1:T}) \prod_{t=1}^{T} q_\phi(z_t|x_t)$$

We find that conditioning on the dynamics of the previous frames better represents the variation of the dynamics over time and achieves better sample quality without sacrificing disentanglement.

The ELBO is derived in a similar fashion to obtain:

$$\mathbb{E}_{z_{1:T},f}[logP_\theta(x_{1:T}|z_{1:T}, f)] - D_{KL}(q_\phi(f|x_{1:T})||P_\theta(f))$$
$$-\sum_{t=1}^{T} D_{KL}(q_\phi(z_t|x_t, z_{<t})||P_\theta(z_t))$$

The model for $q_\phi(f|x_{1:T})$ and $q_\phi(z_t|z_{<t}, x_t)$ is similar to the one used in the previous section, except $q_\phi(z_t|z_{<t}, x_t)$ is not conditioned over f and hence only x_t is fed in at each timestep.

III. Experiments

We train on the Universal LPC Sprites dataset used in [1] and perform the task of reconstructing held out test data, style transfer by exchanging the dynamics of two sprites while retaining their content, and report the cosine similarities of f and z. The deconvolutional neural network that models $P_\theta(x_{1:T}|z_{1:T}, f)$ uses transposed convolutions of kernel size $4*4$ with 256 channels and batch normalization. For the inference model, an intermediate encoding vector having dimension 1024 is generated for each frame by a CNN whose architecture is symmetric to the one used for the generative model. Both the LSTM architectures used for generating f and z_t respectively as well as the RNN used for z_t have hidden layer sizes of 256. We use an l^2 loss for the log likelihood term and train the model for 50 epochs using the Adam optimizer with a learning rate of 0.001. We experiment with the dimensions of f and z_t and find that too high a dimension of f relative to z_t gives good results on the reconstruction task but fails to achieve proper disentanglement as observed in the style transfer task. We hypothesise that the dimensions of f and z_t depend on the relative complexities of the content and dynamics and find the dimensions of 64 and 32 for f and z_t respectively to work best for the LPC dataset. The model is able to reconstruct the test data quite well. For the style transfer task, we randomly choose 7 pairs of sprites such that the two members of a pair differ both in appearance and motion features. We generate their f and $z_{1:T}$ encodings and upon exchanging the latter, we observe that the sprites retain their original appearance but their style of motion has been interchanged. We also pick 12 pairs of sprites, generate their encodings and report the cosine similarities of f and $z_{1:T}$. We observe that sprites having similar motion patterns have a high cosine similarity of $z_{1:T}$, and those having dissimilar motion patterns have a low cosine similarity, irrespective of their appearance. Similar observations are made for the cosine similarity of f between sprites having similar appearance, irrespective of their style of motion.

IV. Ongoing Work

The key advantage of this architecture is that disentangling of content and motion is achieved solely by the probabilistic graphical model. Hence, the focus of ongoing work is primarily on modifying the loss function in order to achieve better sample quality, allow applications to more complex video datasets such as the KTH Actions Video Dataset. To this end, we are experimenting with the use of hierarchical prior distributions as used in Hsu. et. al. 2017 [5] and learned prior distributions as used in [6] for learning better disentangled representation of content and dynamics. We are also investigating ways to use the graphical model of [1] and applying it to approaches such as Adversarial Variational Bayes [7] to allow usage of much more expressive inference models that can better approximate the posterior distribution over the latent variables.

References

[1] Y. Li, S. Mandt. Disentangled Sequential Autoencoder. *International Conference on Machine Learning*, 2018

[2] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework *International Conference on Learning Representations*, Apr. 2017

[3] T. Chen, X. Li, R. Grosse, D. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders *International Conference on Learning Representations*, 2018

[4] E. Denton, V. Birodkar. Unsupervised Learning of Disentangled Representations from Video *Advances in Neural Information Processing Systems*. 2017

[5] W. Hsu, Y. Zhang, J. Glass Unsupervised Learning of Disentangled and-Interpretable Representations from Sequential Data *Advances in Neural Information Processing Systems*. 2017

[6] E. Denton, R. Fergus. Stochastic Video Generation with a Learned Prior *International Conference on Learning Representations*, 2018

[7] L. Mescheder, S. Nowozin, A. Geiger. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks *International Conference on Machine Learning* 2017