Disentangling via dynamics

Benjamin & Andrew

Abstract

1 Introduction

A fundamental goal in machine learning is to build algorithms which can learn to predict future states, given a set of training data. In order to make such a problem tractable, certain inductive biases are adopted which constrain the possible solutions to a smaller set. In the case of visual object recognition and classification, one commonly adopted inductive bias is that the appearance of a given object can be determined by a small set of independent generative factors, such as the object's identity, size, rotation, lighting, and color. A learned representation having these properties is said to be *disentangled* [1, 2], that is, object properties with different semantic meanings are not distributed across latent variables.

Recent work has attempted to develop variational autoencoders (VAE) [3] that can learn disentangled representations, either by regularizing reconstruction error [4], targeting a particular point on a rate-distortion curve [5], or level of overlap of latent factors [6]. These approaches, however, are quite general and do not necessarily lead to truly disentangled representations [5, 6].

We hypothesize that a further inductive bias is necessary to produce truly disentangled representations. Objects are not static; they exist in a continuous world. The putative generative factors that determine an object's appearance are likely to remain stable (in the case of object identity) or vary smoothly and slowly over time (in the case of position, rotation, etc.) [7]. We propose that a representation which factorizes static properties and their dynamics will lead to better disentanglement.

TODO: motivate methods with related work

- briefly discuss Archer paper [8]
  - even more briefly discuss Johnson paper [9]
    - They have LDS SVAE, which uses structured VAE to model a latent LDS graphical model. SLDS VAE uses latent switching LDS to represent "continuous latent states that evolve according to a discrete library of linear dynamics"
- Related work: [10, 11, 12, 13, 14, 15, 16]
- Krishnan2015 "Deep Kalman Filters"
  - Use MLP or RNNs rather than PGMs
  - Very similar. ELBO + Kalman filter
  - Uses NNs for nonlinear dynamics
- Krishnan et al 2017 "Structured Inference Networks for Nonlinear State Space Models" further generalizes this. Still using RNNs.
- Fraccaro et al 2017 "A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning"
  - Separate latent representations: object rep from recognition network, and a latent state describing its dynamics

Preprint. Work in progress.

  - "Kalman VAE"
  - "At each time step t, a variational auto-encoder compresses high-dimensional visual stimuli x_t into latent encodings a_t. The temporal dynamics in the learned a_t-manifold are modelled with a linear Gaussian state space model that is adapted to handle complex dynamics (despite the linear relations among its states z_t). The parameters of the state space model are adapted at each time step, and non-linearly depend on past a_t's via a recurrent neural network"
  - They model a bouncing ball "video". I think we can do better?
- Fraccaro et al 2016 "Sequential Neural Models with Stochastic Layers"
  - Has separate layers of deterministic and stochastic latent variables
- Gregor 2015 "DRAW" uses RNN to refine image reconstructions sequentially

TODO: briefly summarize our contributions and tasks

2 Methods

Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T)$ the observations and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_t, \ldots, \mathbf{z}_T)$ the corresponding latent embeddings, where $\mathbf{x}_t \in \mathbb{R}^m$ and $\mathbf{z}_t \in \mathbb{R}^n$.

Following [8] we structure the inverse covariance Σ^{-1} of the approximate posterior $q(z|x) = \mathcal{N}(\mu_\phi(x), \Sigma_\phi(x))$ to have a blog tri-diagonal structure.

$$\Sigma_\phi(x)^{-1} = \begin{bmatrix} D_0 & B_0^T & & \\ B_0 & D_1 & B_1^T & \\ & \ddots & \ddots & B_{T-1}^T \\ & & B_{T-1} & D_T \end{bmatrix}$$

This particular structure embodies that z_t conditionally only depends on z_{t-1} described by the precision matrix B_0.

The matrices B correspond to the partial correlations of latent variables between adjacent time-points. We here propose to impose linear dynamics for transitions in latent space

$$\mathbf{z}_t = A_t \mathbf{z}_{t-1}$$

where $A_t = L\Lambda_t L^{-1}$ is the dynamics matrix with eigenvector basis L and eigenvalues Λ_t at time-point t. In particular, for factorized latent representations we set $L = I$.

Allowing different eigenvalues for each time-point t admits complex dynamics. We will later impose regularizing constraints on the variation of the eigenvalues across time-points.

The approximate posterior for a given time point t is an isotropic Gaussian (following standard variational mean-field approximation), thereby reflecting the conditional independence assumption that $q(z_t|x_t) = \prod_i^n q(z_t^{(i)}|x_t)$. We therefore set D_t to I. [*unclear on that part*]

2.1 Unconstrained Eigenvalues

In the first case, eigenvalues of A_t can change with every time-point and the mean and eigenvalues are parametrized by a neuronal network:

$$\Lambda_t = \text{NN}_{\phi\Lambda}(\mathbf{x}_t, \mathbf{x}_{t-1})$$
$$\mu_t = \text{NN}_{\phi\mu}(\mathbf{x}_t)$$

2.2 Constraining dynamics

2.2.1 Slow dynamics

To encourage slow dynamical changes in latent space (and therefore learning of slow changing/predictable features) we can constrain the transitional dynamics to be slow. In the first step we therefore add a regularizing term to the loss:

$$\sum_{t=0}^{T-1} ||\text{diag}(\Lambda_t) - \mathbf{1}||_1$$

2.2.2 Different dynamics for latents

Instead of applying the same regularization to all eigenvalues, we can instead specify a prior over eigenvalues. In particular, let $\lambda_t^{(i)}$ be the eigenvalue corresponding to the i-th latent dimension for the transition from time-point t to $t+1$. For instance, defining the regularizing loss as

$$\sum_{0}^{T-1} \sum_{i}^{n} \gamma(i) |\lambda_t^{(i)} - 1|$$

where the regularizing hyperparameter γ allows for different strengths of the 'slow' dynamics regularizer for each latent variable (e.g., $\gamma(i) \propto i$). (*this probably imposes an implicit prior on lambda - there is probably a better way to get this prior than just tuning the regularization.*).

2.3 Future directions or additional parts

Depending on how well things go (and make sense), this can be the next steps:

2.3.1 Allowing for interactions in latent space

Some generative factors can only be expressed in more than one dimension (e.g. angle of rotation of an object). The diagonal construction of A might therefore be to constrained.

Hence, the idea would be to relax A to deviate from diagonal, hence $A_t = \text{NN}(\mathbf{x_t})$ thereby allowing for interactions in latent space, while still regularizing:

$$\sum_{0}^{T-1} ||A_t - I||_1$$

2.3.2 Enforcing low-dimensional subspaces

What we actually want (instead of having an A matrix that has small deviations from an identity matrix) are small subspaces (i.e., small Jordan blocks in the Jordan decomposition of A). Therefore, this step would try to actually enforce this constraint directly.

2.4 Assessment of disentanglement

2.4.1 Qualitative

- Are two generative objects (i.e., an object vs. the background) that change with different dynamics captured in different latent dimensions?
- Can we recover bimodal distributions of learned eigenvalues?

2.5 Quantitative

Using established metrics, we will compare the disentanglement of latent representations across the different model variants as well as with established models in the literature.

missing part: construcing the full covariance matrix from

3 Results

4 Discussion

References

[1] James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, 2007.

[2] Y. Bengio. *Learning Deep Architectures for AI*, volume 2. 2009.

[3] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *ICLR*, pages 1–14, 2014.

[4] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts With A Constrained Variational Framework. *ICLR*, pages 1–12, 2017.

[5] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. In *ICML*, 2017.

[6] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling Disentanglement in Variational Auto-Encoders. 2018.

[7] Laurenz Wiskott and Terrence Sejnowski. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, 14(4):715–770, 2002.

[8] Evan Archer, Il Memming Park, Lars Buesing, John Cunningham, and Liam Paninski. Black box variational inference for state space models. *ICLR*, pages 1–11, 2016.

[9] Matthew J. Johnson, David Duvenaud, Alexander B. Wiltschko, Sandeep R. Datta, and Ryan P. Adams. Composing graphical models with neural networks for structured representations and fast inference. *NIPS*, 2016.

[10] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A Recurrent Neural Network For Image Generation. 2015.

[11] Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep Kalman Filters. (2000):1–17, 2015.

[12] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to Control: A Locally Linear Latent Dynamics Model for Control from Raw Images. pages 1–9, 2015.

[13] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data. (ii):1–13, 2016.

[14] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential Neural Models with Stochastic Layers. (Nips), 2016.

[15] Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning. (section 5), 2017.

[16] Rahul G Krishnan, Uri Shalit, and David Sontag. Structured Inference Networks for Nonlinear State Space Models. *AISTATS*, 2017.