Machine Learning (2023) 112:2297–2315
https://doi.org/10.1007/s10994-022-06299-1

Autoencoding slow representations for semi-supervised data-efficient regression

Oliver Struckmeier[1] · **Kshitij Tiwari**[2] · **Ville Kyrki**[1]

Received: 11 November 2021 / Revised: 7 October 2022 / Accepted: 27 December 2022 /
Published online: 25 January 2023

Abstract
The slowness principle is a concept inspired by the visual cortex of the brain. It postulates that the underlying generative factors of a quickly varying sensory signal change on a different, slower time scale. By applying this principle to state-of-the-art unsupervised representation learning methods one can learn a latent embedding to perform supervised downstream regression tasks more data efficient. In this paper, we compare different approaches to *unsupervised slow representation learning* such as L_p norm based slowness regularization and the SlowVAE, and propose a new term based on Brownian motion used in our method, the S-VAE. We empirically evaluate these slowness regularization terms with respect to their downstream task performance and data efficiency in state estimation and behavioral cloning tasks. We find that slow representations show great performance improvements in settings where only sparse labeled training data is available. Furthermore, we present a theoretical and empirical comparison of the discussed slowness regularization terms. Finally, we discuss how the Fréchet Inception Distance (FID), commonly used to determine the generative capabilities of GANs, can predict the performance of trained models in supervised downstream tasks.

Keywords Unsupervised representation learning · Slowness principle · Data-efficient downstream tasks

Editors: Bo Han, Tongliang Liu, Quanming Yao, Mingming Gong, Gang Niu, Ivor W. Tsang, Masashi Sugiyama.

✉ Oliver Struckmeier
oliver.struckmeier@aalto.fi

Kshitij Tiwari
kshitij.tiwari@oulu.fi

Ville Kyrki
ville.kyrki@aalto.fi

[1] Intelligent Robotics, Aalto University, Maarintie 8, 02150 Espoo, Finland

[2] Perception Engineering Group, University of Oulu, Erkki Koiso-Kanttilan Katu 3, 90014 Oulu, Finland

1 Introduction

Learning representations that capture general high-level information from abundant unlabeled sensory data remains a challenge for unsupervised representation learning. Research in neuroscience suggests that a major difference between state-of-the-art deep learning architectures and the human brain is that cells in the brain do not react to single stimuli, but instead extract invariant features from sequences of fast-changing sensory input signals (Bengio & Bergstra, 2009). Evidence found in the hierarchical organization of simple and complex vision cells shows that time-invariance is the principle after which the cortex extracts the underlying generative factors of these sequences and that these factors usually change slower than the observed signal (Wiskott & Sejnowski, 2002; Berkes & Wiskott, 2005; Bengio & Bergstra, 2009).

Computational neuroscientists have named this paradigm the *slowness principle* wherein individual measurements of a signal may vary quickly, but the underlying generative features vary slowly. For example, individual pixel values in a video change rapidly during short periods of time, but the scene itself usually changes in a slower time scale. This principle has found application in *Slow Feature Analysis* (SFA) as proposed in Wiskott and Sejnowski (2002). SFA has shown promising results in computational neuroscience but little research has explored the possible applications of the underlying slowness principle to state-of-the-art unsupervised representation learning methods used in deep learning. Previous research has employed temporal contrastive $L1$ and $L2$ losses in end-to-end tasks (Mobahi et al., 2009; Sermanet et al., 2018; Zou et al., 2012) such as classification and view-point invariant robot imitation learning. Only recently an unsupervised representation learning method, the SlowVAE (Klindt et al., 2020), has been leveraging the observed statistics of natural transitions in observation space to extend the VAE objective with a sparse temporal prior. The SlowVAE slowness prior has been evaluated on a variety of disentanglement metrics, but not with respect to downstream task data efficiency.

In this paper, we put existing methods for slow representation learning using Variational Autoencoders (VAEs) (Kingma & Welling, 2013) into a shared context and compare them from a theoretical and empirical point of view. We show that different priors used to enforce slowness can be included as a general slowness regularization term to the evidence lower bound (ELBO) of the VAE objective. Additionally, we propose a new slowness regularization term based on a Brownian motion prior for latent space evolution which is used in our method, the *S-VAE*. We empirically compare the β-VAE, the SlowVAE, S-VAE and $L1$/$L2$ slowness-based VAEs with respect to their performance and data efficiency on downstream regression tasks such as odometry estimation and behavioral cloning.

Furthermore, we investigate quantitative measures for the quality of latent representations and find that the Fréchet Inception Distance proposed in Heusel et al. (2017) correlates with the downstream task performance. Being able to predict the downstream task performance without the need for ground truth labels, greatly accelerates the hyperparameter search during the unsupervised pre-training of VAE models and helps identify good models without the need to perform the downstream task.

2 Related work

2.1 Slowness principle

The slowness principle is based on the assumption that the true generative factors of a signal vary on slower time scales than raw sensory signals. Research in computational neuroscience suggests that cell structures in the visual cortex have emerged based on the underlying principle of extracting slowly varying features from the environment (Berkes & Wiskott, 2005). Leveraging this principle, we can extract higher-level invariant scene information which usually changes slower than for example the individual pixel values of a video.

The most well-known application of the slowness principle is the slow feature analysis method (SFA) introduced in Wiskott and Sejnowski (2002). SFA is an unsupervised learning algorithm designed to extract linearly decorrelated features by expanding and transforming the input signal such that it can be optimized for finding the most slowly varying features from an input signal (Wiskott & Sejnowski, 2002). Extending the SFA method to nonlinear features has shown that the learned features share many characteristics with those of complex cells in the V1 cortex (Berkes & Wiskott, 2005). Further applications of the slowness principle include transformation invariant object detection (Franzius et al., 2011), pre-training of neural networks for improved performance on the MNIST dataset (Bengio & Bergstra, 2009) and the self-organization of grid cells, structures in the rodent brain used for navigation (Franzius et al., 2007a, b).

2.2 Contrastive learning and the slowness principle

The objective of contrastive learning is to learn to embed data using a metric score to express (dis-)similarity of data points. Contrastive learning has been successfully applied in reinforcement learning (Laskin et al., 2020) and recently for object classification in SimCLR (Chen et al., 2020a, b). These methods use a contrastive loss on augmented versions of the same observation, effectively learning transformation invariant features from images, and show that these representations benefit reinforcement learning and image classification tasks.

When using time as the contrastive metric we talk about *time-contrastive* learning. Time-contrastive learning has been applied successfully to learning view-point-invariant representations for learning from demonstration with a robot (Sermanet et al., 2018). Similar to our work, Mobahi et al. used the coherence in video material to train a Convolutional Neural Network (CNN) for a variety of specific tasks (Mobahi et al., 2009). While training two CNNs in parallel with shared parameters, in alternating fashion a labeled pair of images was used to perform a gradient update minimizing training loss followed by selecting two unlabeled images from a large video dataset to minimize a time-contrastive loss based on the $L1$ norm of the representations at each layer. The experiments showed that supervised tasks can benefit from the additional pseudo-supervisory signal and that features invariant to pose, illumination or clutter can be learned. Compared to the above methods where a specific task is learned end-to-end with temporal similarity as an additional supervisory signal, we use the slowness principle as a bias on model and data to learn task-agnostic representations that facilitate data efficient downstream task learning.

In contrast, the GP-VAE proposed by Fortuin et al. (2020) is a model for learning temporal dynamics for problems such as reconstructing missing input features, especially in a medical context. The core idea of the GP-VAE is to learn a latent embedding of high dimensional sequential data and to model latent dynamics using a Gaussian Process prior. The authors claim this prior facilitates representations that are smoother and allow the reconstruction of missing features in the input space. Compared to the GP-VAE, the methods presented in this paper address the problem of learning good representations for downstream tasks, instead of learning temporal dynamics of the signal. Another noteworthy method for learning temporal dynamics is the temporal difference variational autoencoder (TD-VAE) (Gregor et al., 2019) which learns representations that encode an uncertain belief state from which multiple possible future scenarios can be rolled out.

2.3 Disentanglement

Another concept related to unsupervised representation learning, especially when talking about the β-VAE, is disentanglement. Although there is no clear definition of disentanglement yet, most works (Bengio et al., 2013; Locatello et al., 2019a; Klindt et al., 2020) agree on the common notion that disentangled representations should approximate the ground truth generative factors of the observed data while the dimensions should be largely independent of each other. Ideally, each disentangled factor represents one ground truth factor that led to the generation of the observation data. Disentanglement in the context of the β-VAE has been discussed more in-depth by Burgess et al. (2018). In the β-VAE pressure on the latent bottleneck of the autoencoder limits how much information can be transmitted per sample while at the same time trying to maximize the data log likelihood. This is done by enforcing a unit Gaussian prior on the latent distributions which results in the embedding of data points on a set of representational axes where nearby points on the axes are also close in data space. This regularization results in these axes being the main contributors to improvements in the data log-likelihood and therefore often coincide with the ground truth generative factors.

Some of the claimed benefits of disentangled representations are better downstream task data efficiency and interpretability (Schölkopf et al., 2012; Bengio et al., 2013; Peters et al., 2017). However, in their research (Locatello et al., 2019a, b) show that various disentanglement methods are not able to generate disentangled representations without implicit biases on model and data and that more disentanglement does not necessarily lead to better downstream task data-efficiency. In a later work by Locatello et al. (2020) the aforementioned challenges were addressed and the authors showed that with weak supervision it is possible to learn fair and generalizable representations.

2.4 Fréchet inception distance

The Fréchet Inception Distance (FID), as introduced in Heusel et al. (2017), is a measure for the generative capabilities of deep generative models. It measures how similar the images generated by GANs are to images from the real data distribution. The FID is an improvement of the Inception Score (IS) introduced by Salimans et al. (2016) which only evaluates the distribution of the generated images and does not compare it to the true data distribution which has been shown to fail when comparing models (Barratt & Sharma, 2018). The FID is computed by comparing the activation distributions of an Inception-v3 neural network pre-trained on the ImageNet dataset for the generated and true data using the Wasserstein-2 distance between the

real data distributions (μ, C), a Gaussian normal distribution with mean μ and covariance C, and $(\mu_{\text{pred}}, C_{\text{pred}})$ describing the generated data distribution.

3 Autoencoding slow representations

Variational Autoencoders (VAEs) (Kingma & Welling, 2013) are a popular tool for dimensionality reduction and representation learning. Let q_ϕ be the variational approximate posterior distribution obtained from a VAE's encoder network with parameters ϕ and $\boldsymbol{z}$ be a latent vector such that $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{o})$ where $\boldsymbol{o}$ is an observation. The decoder network denoted by $p_\theta(\boldsymbol{o} \mid \boldsymbol{z})$ is parameterized by θ. Since it is computationally not tractable to directly maximize the log probability of the data, a lower bound $\mathcal{L}$ is used for optimization:

$$\max_{\phi,\theta} \log p_\theta(\boldsymbol{o}) \geq \mathcal{L} = \mathcal{L}_{\text{rec}} - \mathcal{L}_{\text{b}} \tag{1}$$

where $\mathcal{L}_{\text{rec}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{o})}[\log p_\theta(\boldsymbol{o} \mid \boldsymbol{z})]$ is the reconstruction quality, measured by comparing the true observation and the decoded observations. $\mathcal{L}_{\text{b}} = D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{o}) \| p(\boldsymbol{z}))$ is imposing a unit Gaussian prior $p(\boldsymbol{z})$ on the representations in the bottleneck of the VAE. To make the sampling process differentiable (and thus trainable using gradient based optimization), the variational distribution is usually reparameterized as a Gaussian $q_\phi = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$.

Higgins et al. (2017) proposed the β-VAE, which adds a parameter β to scale the weight of the pressure on the bottleneck to allow a trade-off between disentanglement of the latent factors and reconstruction quality.

Unlike in the β-VAE, where training data is assumed to be *i.i.d*, slow representation learning methods assume sequential data. The core idea of slow representation is to use this property as a weak supervision signal to extract better high-level representations. The slowness constraint is incorporated in the VAE optimization target as

$$\begin{aligned} \mathcal{L}_{\text{rec}} \text{ subject to } \mathcal{L}_{\text{b}} &< \epsilon_1, \\ \mathcal{L}_{\text{slow}} &< \epsilon_2 \end{aligned} \tag{2}$$

Rewriting Eq. (2) under the Karush–Kuhn–Tucker conditions (KuhnandA & Tucker, 1951; Karush , 1939) results in

$$\mathcal{F} = \mathcal{L}_{\text{rec}} - \beta(\mathcal{L}_{\text{b}} - \epsilon_1) - \gamma(\mathcal{L}_{\text{slow}} - \epsilon_2) \tag{3}$$

With the simplification of $\beta, \epsilon_1, \gamma, \epsilon_2 \geq 0$, we can derive the lower bound

$$\mathcal{F} \geq \mathcal{L} = \mathcal{L}_{\text{rec}} - \beta\mathcal{L}_{\text{b}} - \gamma\mathcal{L}_{\text{slow}}. \tag{4}$$

with β being the β-VAE parameter and γ being a parameter to control the weight of the general slowness regularization term $\mathcal{L}_{\text{slow}}$. In the following, we present various slowness regularization terms used in existing works and a new slowness regularization term used in our method, the S-VAE.

3.1 Lp-norm slowness

A straightforward way to describe $\mathcal{L}_{\text{slow}}$ is to compute the L_p-norm of the means $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ of two encoded latent distributions from two distinct yet sequential observations $\mathbf{o}_i, \mathbf{o}_j \in \mathbf{D} \mid j > i$ as

$$\mathcal{L}_{L_p}(\boldsymbol{o}_j, \boldsymbol{o}_i) = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^p \tag{5}$$

where $j > i$. Following Eq. (4), the full formulation of a L_p-slow VAE ELBO is therefore

$$\begin{aligned}
&\mathcal{L}(\phi, \theta, \beta, \gamma, \boldsymbol{o}_j, \boldsymbol{o}_i) \\
&\quad = \mathbb{E}_{q_\phi(z_j, z_i | \boldsymbol{o}_j, \boldsymbol{o}_i)}[\log p_\theta(\boldsymbol{o}_j, \boldsymbol{o}_i \mid z_j, z_i)] \\
&\qquad - \beta D_{KL}(q_\phi(z_i \mid \boldsymbol{o}_i) \| p(z_i)) \\
&\qquad - \gamma(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^p,
\end{aligned} \tag{6}$$

where ϕ and θ parameterize encoder and decoder. The hyperparameters β and γ are weights for the strength of the disentanglement and slowness regularization.

3.2 SlowVAE

The SlowVAE by Klindt et al. (2020), is based on a study of the statistics of natural transitions of object and object mask properties in two acknowledged video-object segmentation datasets. The authors base this prior on experimental evidence that the transitions of ground truth factors in such datasets can be approximated by generalized Laplace distributions, indicating that the temporal transitions are sparse. In other words, only few of the ground truth generative factors of a signal change in one transition. This assumption is expressed in a Laplacian prior on latent transitions, encouraging axis alignment. The SlowVAE slowness loss term is defined as

$$\mathcal{L}_{\text{SlowVAE}}(\boldsymbol{o}_{i+1}, \boldsymbol{o}_i) = \mathbb{E}_{q_\phi(z_i | o_i)}\left[D_{KL}(q_\phi(z_{i+1} \mid \boldsymbol{o}_{i+1})) \| p(z_{i+1} \mid z_i))\right]. \tag{7}$$

where $p(z_{i+1} \mid z_i)$ is the Laplacian prior on the transition.

Consequently, the SlowVAE ELBO is defined as

$$\begin{aligned}
\mathcal{L}(\phi, \theta, \beta, \gamma, \boldsymbol{o}_{i+1}, \boldsymbol{o}_i) =& \mathbb{E}_{q_\phi(z_{i+1}, z_i | \boldsymbol{o}_{i+1}, \boldsymbol{o}_i)}\left[\log p_\theta(\boldsymbol{o}_{i+1}, \boldsymbol{o}_i \mid z_{i+1}, z_i)\right] \\
&- \beta D_{KL}(q_\phi(\mathbf{z_i} \mid \mathbf{o_i}) \| p(\mathbf{z_i})) \\
&- \gamma \mathbb{E}_{q_\phi(z_i | o_i)}\left[D_{KL}(q_\phi(z_{i+1} \mid \boldsymbol{o}_{i+1})) \| p(\mathbf{z}_{i+1} \mid z_i))\right].
\end{aligned} \tag{8}$$

3.3 S-VAE

We propose the S-VAE as an alternative point of view on slow representation learning. The key idea is to directly incorporate the slowness principles such that "underlying generative factors change on a slower time scale" (Berkes & Wiskott, 2005). Thus the S-VAE enforces that observations close in time have similar latent representations and reduces the strength of this assumption with growing temporal separation. The temporal similarity in the S-VAE is expressed as an additive stochastic process with stationary uncertainty, which is known to approach Brownian motion in the limit. Considering two distinct yet sequential observations $\mathbf{o}_i, \mathbf{o}_j \in D \mid j > i$, the difference of the corresponding latent representations is given by the approximate difference distribution,

$$q_\theta(z_j - z_i \mid \mathbf{o}_j, \boldsymbol{o}_i) = \mathcal{N}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_i) \equiv q_\theta(\Delta z \mid \boldsymbol{o}_j, \boldsymbol{o}_i). \tag{9}$$

We impose a prior $p(\Delta z)$ on the approximate difference distribution in Eq. (9). For two increments z_j, z_i of a Brownian motion, the prior distribution is defined as

$$z_j - z_i = \sqrt{\Delta t} \cdot N \sim \mathcal{N}(0, \Delta t \lambda \boldsymbol{I}) \equiv p(\Delta z), \tag{10}$$

where $N \sim \mathcal{N}(0, \Delta t \lambda \boldsymbol{I})$ and $\Delta t = j - i$ and λ is a parameter corresponding to the variance of the prior distribution.

The S-VAE loss term is computed as the Kullback–Leibler (KL) divergence between the approximate difference distribution in Eq. (9) and the prior distribution in Eq. (10) as

$$\mathcal{L}_{\text{slow}}(\boldsymbol{o}_i, \boldsymbol{o}_j) = D_{KL}(q_\theta(\Delta z \mid \boldsymbol{o}_j, \boldsymbol{o}_i) \| p(\Delta z)). \tag{11}$$

resulting in the S-VAE ELBO

$$\begin{aligned} &\mathcal{L}(\theta, \phi, \beta, \gamma, \boldsymbol{o_j}, \boldsymbol{o_i}) \\ &\quad = \mathbb{E}_{q_\theta(z_j, z_i \mid \boldsymbol{o}_j, \boldsymbol{o}_i)}[\log p_\phi(\boldsymbol{o}_j, \boldsymbol{o}_i \mid z_j, z_i)] \\ &\qquad - \beta D_{KL}(q_\theta(z_i \mid \boldsymbol{o}_i) \| p(z_i)) \\ &\qquad - \gamma D_{KL}(q_\theta(\Delta z \mid \boldsymbol{o}_j, \boldsymbol{o}_i) \| p(\Delta z)). \end{aligned} \tag{12}$$

Following from Eq. (10), the Brownian motion prior $p(\Delta z)$ in the S-VAE ELBO explicitly takes the temporal separation of consecutive observations into account and relaxes the prior when temporal separation grows. A more detailed derivation can be found in Appendix A.

3.4 Discussion

Using the L_p-norm to enforce temporal similarity has been explored by Mobahi et al. (2009) where a $L1$ norm was used to enforce similarity between two halves of a siamese CNN architecture to leverage temporal similarity of observations during training. Other applications of a L_p-norm to enforce temporal similarity can be found in Zou et al. (2012) and (Sermanet et al., 2018), where a $L1$ norm and a triplet loss (based on the L2 norm) respectively were used to achieve viewpoint-invariance by enforcing similarity of representations from different angles according to temporal similarity. Cadieu and Olshausen (2012) claim that there are no significant differences between $L1$ and $L2$ norm for enforcing temporal similarity in latent space. The main differences between existing research and the representation learning methods presented in this work are that they do not operate in a variational setting and are learned end-to-end instead.

In contrast, the S-VAE and SlowVAE both use the variance of the predicted latent distributions, albeit in different ways depending on the applied prior. Figure 1 shows the conceptual differences between the S-VAE and SlowVAE for a pair of observations. The SlowVAE Laplace prior expresses the assumption that transitions in latent space are sparse, or in other words, are axis aligned with the ground truth generative axes. This bears similarity to the definition of disentanglement and can be understood as disentangling latent transitions. The S-VAE does not make such an assumption on the nature of the transitions. Instead, it is based on the assumption that observations close in time must also be similar in latent space. Increasing temporal distance in observation space is taken into account by the increasing uncertainty of the Brownian motion prior. This allows the S-VAE to benefit from the closed-form solution of the Brownian motion and to handle pairs of observations

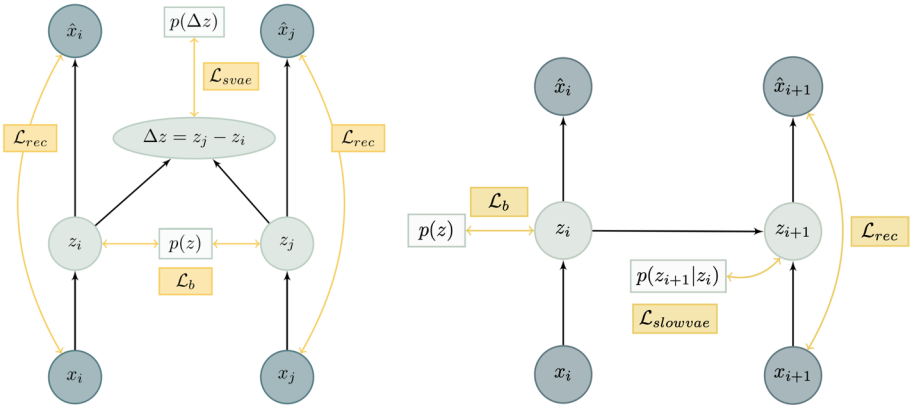


Fig. 1 Conceptual difference between the S-VAE (ours) and the SlowVAE. The S-VAE expresses slowness through a prior $p(\Delta z)$ on the similarity of z_j and z_i, which is relaxed when $\Delta t = j - i$ grows. The SlowVAE imposes a sparsity prior on z_{t+1}, assuming that latent transitions are sparse

from any point in the temporal sequence as opposed to the SlowVAE, which has been evaluated on consecutive observations.

In Klindt et al. (2020), the authors investigated the performance of the SlowVAE when varying the temporal distance between observations up to 1 second when training. Analysis of the natural transitions showed that with increasing temporal distance between frames, the estimated kurtosis parameter α of the fitted Laplace distribution increased, effectively moving closer to Gaussianity ($\alpha = 2$). This is in line with the central limit theorem stating that the sum of i.i.d. random variables approaches the normal distribution when the number of terms increases.

Based on these differences, we aim to study the following hypotheses experimentally. Existing literature suggests that when compared to the β-VAE, L_1 and L_2, slowness regularization is beneficial for various of downstream tasks. The S-VAE and SlowVAE are expected to perform better than the L_p-norm slowness regularization as they also consider the variance of the latent distributions. Furthermore, we hypothesize that the S-VAE outperforms the SlowVAE with a kurtosis of $\alpha = 1$ when the temporal separation between observations grows.

4 Empirical comparison

In this section, we compare the previously introduced slow representation learning methods to the baseline β-VAE with respect to their data efficiency when learning downstream regression tasks. We analyze the influence of the slowness hyperparameter γ on the latent representations by visualizing the latent spaces.

Although the use case is different, we compared the TD-VAE (Gregor et al., 2019) to the slow methods and the β-VAE and found that it is outperformed by all methods. We did not include those results as we were not confident in their thoroughness. The problem is that, to our knowledge, there is no official code repository for the TD-VAE, and could not

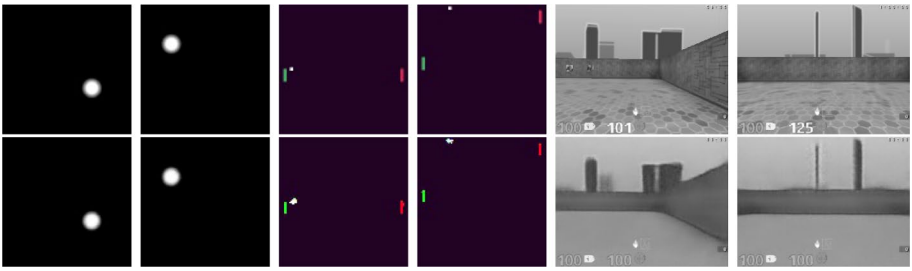


Fig. 2 Random images from the training data (top) and their reconstructions (bottom) obtained from the S-VAE

reproduce the original results on the more complex DeepMind Lab experiment with the given implementation instructions.

4.1 Experimental setting

Three experiments have been conducted in which the goal was to learn downstream tasks in a semi-supervised way from video data. The semi-supervised process consists of two steps.

First, a VAE model is trained on abundant unlabeled video data in an unsupervised way. The VAE models use the encoder component to encode two observations o_i and o_j from the input video sequence such that the slowness loss terms can be computed for training. Following the ablation study in Klindt et al. (2020), which indicates that there is a sweet spot for the temporal separation of consecutive observations, we also vary Δt during training to take into account a wider variety of temporal separation.

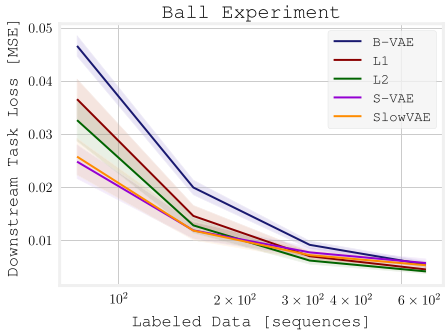
Second, a downstream task is learned using embedding from the pre-trained VAE model while keeping the encoder network weights frozen. The two encoded latent representations z_i and z_j are concatenated and used to learn a downstream task end-to-end. The downstream tasks involve temporal tasks like velocity and odometry estimation or behavioral cloning.

Figure 2 shows random observations from the training dataset of all three experiment domains and their reconstructions generated using the S-VAE. The domains are a synthetic dataset of a ball bouncing in a $2D$ world, two reinforcement learning agents playing the game of Pong and a human randomly moving an agent in a $3D$ world in the DeepMind Lab environment (Beattie et al., 2016). The experiment setup and neural network architecture are described in more detail in Appendix C.

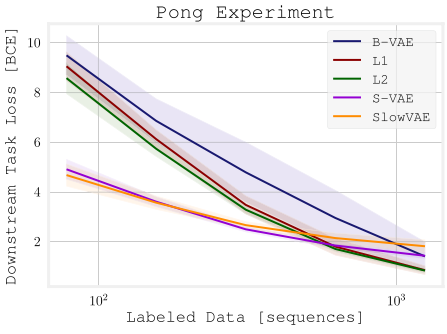
4.2 Evaluation of downstream task performance

After the VAE training step, the encoder network is frozen and the downstream tasks are learned. For each downstream task, multiple models with varying amounts of labeled data available (from sparse to abundant) are trained. The downstream task performance is measured by computing the loss on a previously unseen labeled test set.

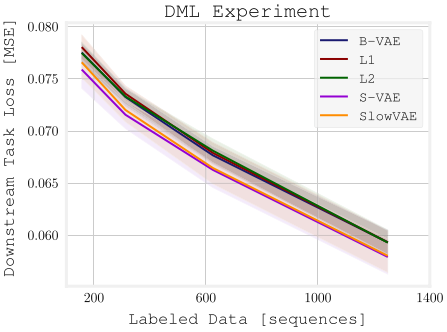
Figure 3 shows a plot of the average downstream task performance for each method (baseline β-VAE, L_1, L_2, S-VAE and SlowVAE) against the amount of available labeled data. Mean and standard deviation are computed over multiple runs with different seeds and hyperparameter configurations (γ and β).



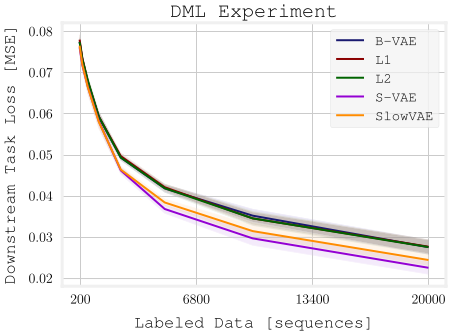
(a) Ball experiment.



(b) Pong experiment.



(c) DeepMind Lab experiment (labeled data sparse).

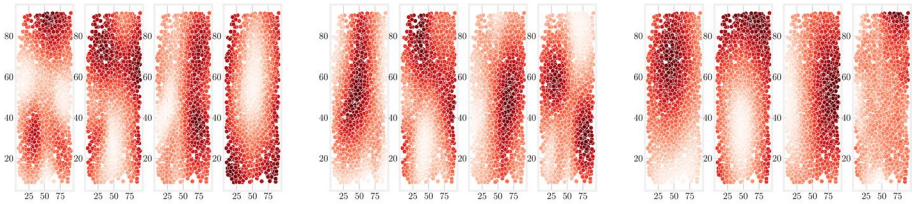


(d) DeepMind Lab experiment (labeled data abundant).

Fig. 3 Results of the downstream data-efficiency experiment. MSE/BCE Loss between true and predicted label in the downstream task vs. the amount of labeled data used to train the model

Figures 3a, c show the Ball and Pong experiment in the case where labeled data is sparse. In those cases, the S-VAE and SlowVAE outperform the L_1 and L_2 slowness regularization terms. The β-VAE without temporal regularization is outperformed by all methods in the Ball and Pong experiment. S-VAE and SlowVAE achieve the same performance as the baseline β-VAE with up to an order of magnitude fewer data and significantly better performance. In the DeepMind Lab experiment, the difference is less pronounced in the sparse data case, due to the complexity of estimating 6-DOF odometry from a 3D world with less than 1500 labeled examples. However, Fig. 3d shows that the S-VAE outperforms the other methods when more labeled data is available. Furthermore, in the DeepMind Lab experiment, the L_1 and L_2 methods yield no significant improvement over the β-VAE. We theorize that in this more complex task, taking into account the covariances in the slowness regularization term is important to learn good representations.

To summarize, slow methods generally yield better downstream task performance. Taking into account the variance of the latent distributions when applying slowness (S-VAE and SlowVAE) improves performance further.



(a) β-VAE with $\beta = 2.0$ and $\gamma = 0.0$.

(b) S-VAE with $\beta = 2.0$ and $\gamma = 0.05$.

(c) S-VAE with $\beta = 2.0$ and $\gamma = 5.0$.

Fig. 4 Visualization of the $4D$ latent space of the Ball experiment for 3 hyperparameter configurations. Each dimension is shown with an individual scatter plot in which x and y axis are the ground truth position of the ball in the environment. The color is the value of the latent representation at the given position. Slowness regularization increases from $\gamma = 0$ (β-VAE) in **a** to $\gamma = 5.0$ in **c**

4.3 Slow latent spaces

Next, we investigate how the slowness hyperparameter γ influences the latent representations by visualizing the latent spaces in the Ball experiment. Figure 4 shows a scatter plot where the x- and y-axis describe the ground truth position of the ball in the arena and the color represents the latent value. Figure 4a shows four plots, one for each latent dimension of a β-VAE. Figure 4b, c show the same for variations of the S-VAE with the same parameter β, but increasingly higher slowness regularization γ. We can see that increasing the strength of the slowness regularization increases the continuity of the latent space. For the β-VAE we observe multiple discontinuities that separate regions with similar latent values, whereas the S-VAE model with increasing slowness regularization exhibits visibly smooth representations.

5 Predicting downstream task performance

While a qualitative analysis showed that increasing slowness regularization makes the latent space smoother, this analysis is not generally suited to measure the quality of a learned embedding space or to predict downstream task performance. In the Ball experiment, we can visualize the latent space using the ground truth position of the ball, which is difficult for high-dimensional latent spaces or complex downstream tasks with higher-dimensional generative factors or observations.

In this section, we want to investigate further how one can predict downstream task performance without the need for labeled data or human intervention. Let us consider the predictive capabilities of each of the three components of the general formulation in Eq. (4): disentanglement, slowness and reconstruction performance.

The extent to which *disentanglement* can predict downstream task performances has been investigated by Locatello et al. in Locatello et al. (2019a). In their work, the authors questioned the benefits disentangled representations have for learning downstream tasks and criticized that currently, all disentanglement metrics require ground truth labels. Furthermore, disentanglement measures are supervised methods relying on abundant labeled data and are usually tailored for specific tasks. Thus we do not consider existing disentanglement metrics to predict downstream task performance.

The second option is to measure if a latent space exhibits the *slowness* properties and correlate this measure with downstream task performance. To this end, we experimented with label-agnostic metrics that measure the slowness by how smooth the latent space is. We measured the length of a trajectory in latent space defined by N encoded latent representations $z_1, \dots, z_N$ and compared it to the euclidean distance between z_1 and z_N. According to the slowness principle and the qualitative analysis in Fig. 4 we hypothesize that higher slowness regularization with a qualitatively smoother latent space has fewer jumps and therefore, the trajectory in latent space is shorter and more similar to the euclidean distance. We observed that, as expected, higher slowness regularization leads to less fragmented and shorter latent trajectories. However, this metric does not correlate with downstream task performance.

As the remaining option, we show that the *generative capabilities* of a model allow us to predict downstream task performance. The core idea is that a model capable of decoding "realistic" images from random samples drawn from the latent distributions encodes more useful information in the latent space. To measure this, we use the FID, which is commonly used to evaluate the generative capabilities of Generative Adversarial Networks (GANs). In practice, we used the *python-fid* package with standard parameters. The code can be found on GitHub. In this implementation, the FID is computed using the 2048 dimensional features extracted from the *pool3* layer of the pre-trained Inception Net.

Figure 5 shows a scatter plot of the FID for all combinations of γ and β plotted against downstream task performance.

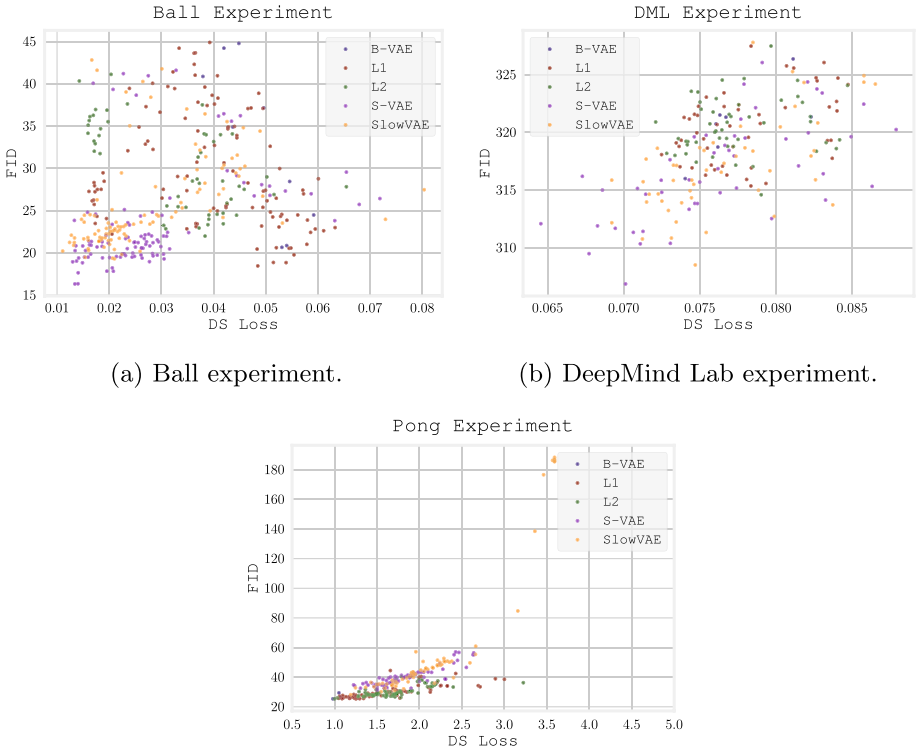
We observe correlation between low FID and low downstream task loss for the S-VAE and SlowVAE in all three experiments.[1] In the pong experiment, one can clearly identify outliers by their high FID. Upon further inspection, in those hyperparameter configurations, the SlowVAE failed to reconstruct the observations and thus led to weak downstream task performance. These models usually had high values for γ and β, indicating that for those models, the applied regularization was too strong. In the Ball experiment, the L_1 and L_2 methods did not exhibit correlation. We hypothesize that the hyperparameter search in this experiment could have been even broader, exploring stronger regularization to find better models. This is further supported by the fact that the L_1 and L_2 models in the Ball experiment could reconstruct the ball with minimal reconstruction errors.

In conclusion, the generative capabilities of models can be used to predict the downstream task performance of VAE models. Using the FID as an indicator for downstream task performance allows more targeted exploration of hyperparameter ranges. We think that the FID as a tool to express the experiment-specific and general properties of generative models when learning downstream tasks should be further explored.

6 Empirical comparison of slowness regularization

As discussed in Sect. 3.4, the priors in the SlowVAE and S-VAE interpret slowness differently. The SlowVAE looks at slowness from a transition perspective, assuming that transitions are sparse. While transitions with small Δt have been shown to be sparse, results in Klindt et al. (2020) indicate that the kurtosis of the Laplacian fit to the transitions increases

[1] We exclude the β-VAE in this discussion since there are not enough parameter configurations that allow conclusions about correlation.



(a) Ball experiment.

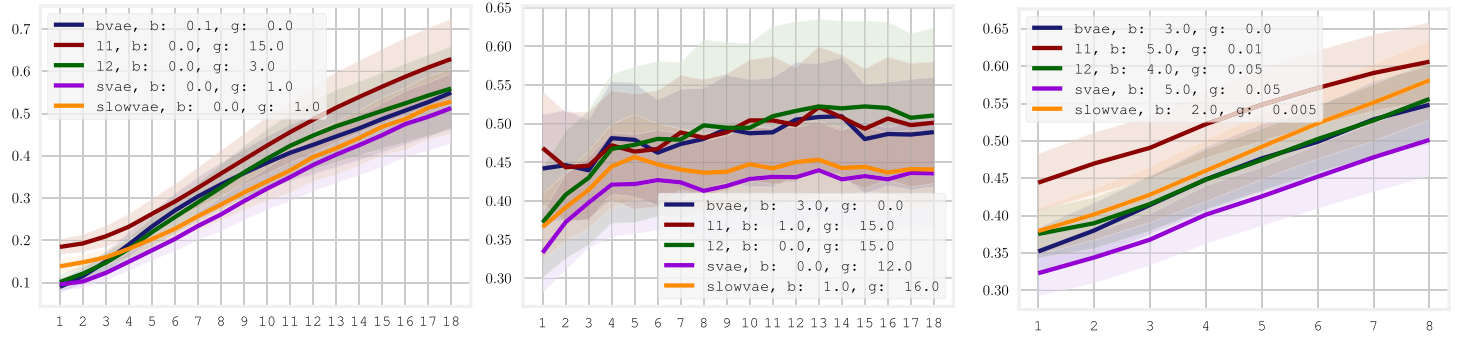
(b) DeepMind Lab experiment.

(c) Pong experiment (labeled data abundant and outliers).

Fig. 5 Scatter plot visualization of FID plotted against downstream task performance. Each point represents one model colored by method and with different hyperparameters γ and β

with temporal separation. The S-VAE, on the other hand, does not put a prior on the transition but instead enforces similarity based on temporal distance. Furthermore, the S-VAE explicitly incorporates Δt in the training process through the Brownian motion prior [see Eq. (10)]. Therefore, using the Brownian motion slowness prior to train an embedding should yield better performance in downstream tasks where observations are further apart in time.

To investigate this hypothesis, we used the best-performing models for each method presented in Fig. 3 and trained a latent dynamics model to predict future latent representations in a sequence. The first two observations of a sequence are encoded and a random observation Δt steps ahead in the sequence is drawn for the dynamics model to predict. The performance of the dynamics model is expressed as the mean squared error between the predicted and true future latent representation. Figure 6 shows the average latent dynamics error for predictions of varying Δt. In all environments, the S-VAE generally has a lower error than the SlowVAE and the other methods. Interestingly the β-VAE is performing on par with the L_2 slowness regularization. Both methods methods perform better or as good as the SlowVAE for short-term predictions in the Ball and DeepMind Lab latent dynamics experiment. This is in line with the findings by Klindt et al. (2020), showing that the SlowVAE has a sweet spot for the temporal



(a) Ball experiment. (b) Pong experiment. (c) DeepMind Lab experiment.

Fig. 6 Latent dynamics error vs. sequence length for the best performing model of each method. The latent dynamics error is computed as the mean squared error between true and predicted latent vector. The latent vectors are normalized to account for the different scales of latent representations across the methods

separation, roughly when $\Delta t > 0.4$ seconds. The Brownian motion prior in the S-VAE yields better performance across all values for Δt explored in this experiment.

7 Conclusion

In this paper, we discuss the application of the slowness principle as an extension of the state-of-the-art β-VAE. We compare existing methods of slowness regularization such as $L1$ and $L2$ loss and the SlowVAE, a variation of the β-VAE imposing a Laplacian prior on the latent transitions. We also propose a new slowness regularization term based on a Brownian motion prior. We find that slow methods outperform the baseline β-VAE with respect to downstream task data efficiency. Furthermore, the results indicate that the S-VAE and SlowVAE perform similarly but better than the β-VAE and L_p-norm-based slowness regularization terms with respect to their data efficiency in downstream tasks. When learning a latent dynamics model to predict latent representations multiple steps ahead in time, the S-VAE exhibits superior performance due to its ability to adapt its Brownian motion prior to the temporal separation of observations. Lastly, we find that the Fréchet Inception Distance is a helpful measure to predict downstream task performance.

Appendix A Derivation of the S-VAE slowness regularization term

In the following, we will derive the S-VAE slowness regularization term following the derivation of the β-VAE (Higgins et al., 2017) and extend it to sequential observations.

In a sequential setting, we will maximize the marginal log-likelihood of two observations $\boldsymbol{o}_j$ and $\boldsymbol{o}_i$ over the latent factors as

$$\begin{aligned} &\max_{\phi,\theta} \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{D}}[\mathbb{E}_{q_\phi(z_j,z_i|\boldsymbol{o}_j,\boldsymbol{o}_i)}[\log p_\theta(\boldsymbol{o}_j,\boldsymbol{o}_i \mid z_j,z_i)]] \text{ subject to} \\ &D_{\mathrm{KL}}(q_\phi(z_j \mid \boldsymbol{o}_j) \mid\mid p(z)) < \epsilon \\ &D_{\mathrm{KL}}(q_\phi(z_i \mid \boldsymbol{o}_i) \mid\mid p(z)) < \epsilon \\ &D_{\mathrm{KL}}(q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \mid\mid p(\Delta z)) < \zeta. \end{aligned} \tag{A1}$$

As already discussed in Eqs. (2) and (3) we apply the KKT conditions to obtain the Lagrangian

$$\begin{aligned} \mathcal{F}(\phi,\theta,\beta,\gamma,\boldsymbol{o}_j,\boldsymbol{o}_i,z_j,z_i) =& \mathbb{E}_{q_\phi(z_j,z_i|\boldsymbol{o}_j,\boldsymbol{o}_i)}[\log p_\theta(\boldsymbol{o}_j,\boldsymbol{o}_i)] \\ &- \beta(D_{\mathrm{KL}}(q_\phi(z_j \mid \boldsymbol{o}_j) \mid\mid p(z)) - \epsilon) \\ &- \beta(D_{\mathrm{KL}}(q_\phi(z_i \mid \boldsymbol{o}_i) \mid\mid p(z)) - \epsilon) \\ &- \gamma(D_{\mathrm{KL}}(q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \mid\mid p(\Delta z)) - \zeta). \end{aligned} \tag{A2}$$

The ELBO can be obtained under the condition that $\beta,\epsilon,\gamma,\zeta \geq 0$ as

$$\begin{aligned}\mathcal{F}(\phi,\theta,\beta,\gamma,\boldsymbol{o}_j,\boldsymbol{o}_i,z_j,z_i) \geq & \mathcal{L}(\phi,\theta,\beta,\gamma,\boldsymbol{o}_j,\boldsymbol{o}_i,z_j,z_i)\\ =& \mathbb{E}_{q_\theta(z_j,z_i|\boldsymbol{o}_j,\boldsymbol{o}_i)}[\log p_\theta(\boldsymbol{o}_j,\boldsymbol{o}_i \mid z_j,z_i)]\\ &- \beta D_{\mathrm{KL}}(q_\phi(z_j \mid \boldsymbol{o}_j) \,||\, p(z))\\ &- \beta D_{\mathrm{KL}}(q_\phi(z_i \mid \boldsymbol{o}_i) \,||\, p(z))\\ &- \gamma D_{\mathrm{KL}}(q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \,||\, p(\Delta z))\end{aligned} \tag{A3}$$

where the first three terms (expectation and the β terms) of the r.h.s. are the familiar terms from the β-VAE. The term $-\gamma D_{\mathrm{KL}}(q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \,||\, p(\Delta z))$ is the S-VAE slowness loss term imposing the prior $p(\Delta z) \sim \mathcal{N}(0, \Delta t\lambda \boldsymbol{I})$ on the difference of latent distributions $q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \sim \mathcal{N}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i, \boldsymbol{\sigma}_j + \boldsymbol{\sigma}_i)$. The parameter λ allows adjustment of the variance of the prior distribution.

In practice, this KL Divergence is summed over the scalars for all N latent dimensions as

$$\begin{aligned}-\mathcal{L}_{\mathrm{slow}} =& - D_{\mathrm{KL}}(q_\phi(\Delta z \mid \boldsymbol{o}_j,\boldsymbol{o}_i) \,||\, p(z))\\ =& \sum_{n=1}^{N} -\log\left(\frac{\sigma_j^{(n)} + \sigma_i^{(n)}}{\Delta t\lambda}\right) + \frac{(\sigma_j^{(n)} + \sigma_i^{(n)})^2 + (\mu_j^{(n)} - \mu_i^{(n)})^2}{2\Delta t^2\lambda^2} + \frac{1}{2}.\end{aligned} \tag{A4}$$

Appendix B Slow VAE methods from an information theoretic point of view

In this section, we briefly want to discuss the connection between the FID score as a predictive metric for downstream task performance of a model and lowness regularization. Let us look at the β-VAE from the Information-theoretic point of view. We can write the β-VAE objective as $\max[I(z;y) - \beta I(o;z)]$ (Burgess et al., 2018; Alemi et al., 2016) where $I(z; y)$ is the mutual information of the intermediate representation z and the reconstruction task. The subtracted term $\beta I(o;z)$, usually referred to as the latent bottleneck, effectively encourages the model to discard less relevant (to predict the label y) information present in the input o by limiting the capacity of the latent information channels. In the case of reconstructing images, such information would be properties of the input signal that are not relevant for reconstruction. In the previous analysis, we observed that, as expected, increasing the latent bottleneck pressure by varying the parameter β reduces the number of latent dimensions used. We theorize that, similar to how the β-VAE discards less relevant information, the slow methods discard information about the input signal that changes in a shorter time scale. Hence fewer latent channels are used and better higher-level features are extracted. This draws a parallel to the core idea of the slowness principle: Discourage the encoding of irrelevant quickly changing components of the input signal and encourage encoding of slowly changing features. Further work on this end is required to show mathematical and experimental results that support our theory.

Appendix C Experiment details

In this section, we present a more detailed description of the experiments conducted. The experiment framework used in this research shares the same architecture with the Slow-VAE (Klindt et al., 2020) and the β-VAE (Kingma & Ba, 2014). The encoder consists of 5 convolutional layers with ReLU activation functions followed by a fully connected head

that returns mean and log-variance. The decoder consists of a fully connected layer and 5 deconvolutional layers with ReLU activation. For more details, please refer to the SlowVAE GitHub repository. The stride and filter size has been adjusted to accommodate rectangular images in the DeepMind Lab experiment.

The hyperparameter λ for the S-VAE was selected such that when plotting the SlowVAE and S-VAE loss functions, the functions would be roughly in the same value range. We expect that training more S-VAE models to find better values for λ could further improve the performance of the S-VAE.

The downstream task model consists of 2 fully connected layers with 50 nodes each and ReLU activation functions, followed by one more fully connected layer with a sigmoid activation function.

C.1 Ball experiment

This task aims to predict the velocity of a white ball on a black background in a square-shaped environment from two consecutive frames. We generated sequences of 20 frames of a ball bouncing in a 100×100 pixels. For each sequence, the ball is placed in a random position and initialized with a random direction velocity vector. Upon reaching the border of the environment, the ball's velocity vector is flipped to mimic the principle "incident angle equals emergence angle". Overall, 10000 labeled sequences consisting of 2 datapoints each were generated for training and 500 sequences for testing.

During the *unsupervised representation learning step*, we use the full dataset without labels to train models, exploring hyperparameters β and γ. The hyperparameter ranges were the same as in the SlowVAE (1.0-16.0 for both β and γ). The training consisted of 75 epochs with a batch size of 32 and a learning rate $1e-04$.

In the *supervised downstream task* we use subsets of $(1, 1/2, 1/4, \ldots, 1/128)$ of the full labeled dataset with their labels to train the downstream task of predicting the ball velocity/position from two consecutive frames. The downstream tasks are trained by freezing the encoder networks trained in the previous step and feeding two latent representations into the downstream task model to predict the ball velocity. The downstream loss is computed as the mean squared error between true and predicted label. The downstream experiment was averaged over 10 seeds, effectively initializing the downstream neural network differently for each run.

C.2 Pong experiment

The goal of this experiment is to learn the policy for two agents playing the game of pong. The agents can take 3 actions: No action, move up and move down. The dataset consists of 10000 sequences of length 20. Sequences that contained reset events of the environment, for example, when a game was won/lost were discarded during dataset generation to obtain uninterrupted sequences. In this experiment, a batch size of 32 was used and the training was performed for 175 epochs. The downstream training of this experiment was conducted similarly to the Ball experiment, obtaining two consecutive latent representations, concatenating them and predicting action probabilities. Subsets of $(1, 1/2, 1/4, \ldots, 1/128)$ of the full labeled dataset with their labels were used to train the downstream task. The loss is computed as the Binary Cross Entropy loss between the predicted and one-hot encoded action probabilities. The downstream experiment was averaged over 5 seeds, effectively initializing the downstream neural network differently for each run.

C.3 DeepMind lab dataset

In this experiment, the goal is to learn the 6-DOF motion vector of an agent exploring a DeepMind Lab environment (Beattie et al., 2016). A dataset of 20, 000 sequences of size 20 was generated while recording a human moving around in the environment. Subsets of $(1, 1/2, 1/4, \ldots, 1/128)$ of the full labeled dataset with their labels were used to train the downstream task. The loss is computed as the MSE loss between the predicted and true $6D$ odometry vector. The downstream experiment was averaged over 5 seeds, effectively initializing the downstream neural network differently for each run.

Acknowledgements The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project. The authors wish to acknowledge CSC—IT Center for Science, Finland, for computational resources.

Author Contributions Oliver Struckmeier, Kshitij Tiwari and Ville Kyrki contributed to the conception and design of the work, analysis and interpretation of results, and writing and editing of the manuscript. Implementation, experimentation and data collection were performed by the first author, Oliver Struckmeier.

Funding Open Access funding provided by Aalto University. This project was partially funded by the Human Brain Project Second Specific Grant Agreement (SGA) 2, project number 680093.

Code availability The code to reproduce the results is available at https://github.com/Oleffa/slow-representation-learning.

Declarations

Conflict of interest The authors have no financial or non-financial interests to disclose.

Ethical An earlier version of this work has been presented at the IJCAI 2021 Workshop on Weakly Supervised Representation Learning but has not been published. No experiments involving humans or animals have been conducted.

References

Alemi, A. A., Fischer, I., Dillon, J. V., & Murphy, K. (2016). Deep variational information bottleneck. arXiv preprint arXiv:1612.00410.

Barratt, S., & Sharma, R. (2018). A note on the inception score. arXiv preprint arXiv:1801.01973.

Beattie, C., Leibo, J. Z., Teplyashin, D., Ward, T., Wainwright, M., Küttler, H., Lefrancq, A., Green, S., Valdés, V., Sadik, A., & Schrittwieser, J. (2016). Deepmind lab. arXiv preprint arXiv:1612.03801.

Bengio, Y., & Bergstra, J. S. (2009). Slow, decorrelated features for pretraining complex cell-like networks. In *Advances in neural information processing systems* (pp. 99– 107).

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798–1828.

Berkes, P., & Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision, 5*(6), 9.

Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599.

Cadieu, C. F., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation, 24*(4), 827–866.
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709.
Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029.
Fortuin, V., Baranchuk, D., Rätsch, G., & Mandt, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics* (pp. 1651–1661). PMLR.
Franzius, M., Sprekeler, H., & Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology, 3*(8), 166.
Franzius, M., Vollgraf, R., & Wiskott, L. (2007). From grids to places. *Journal of Computational Neuroscience, 22*(3), 297–299.
Franzius, M., Wilbert, N., & Wiskott, L. (2011). Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation, 23*(9), 2289–2323.
Gregor, K., Papamakarios, G., Besse, F., Buesing, L., & Weber, T. (2019). Temporal difference variational auto-encoder. In *International conference on learning representations*. https://openreview.net/forum?id=S1x4ghC9tQ
Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6629– 6640).
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*. https://openreview.net/forum?id=Sy2fzU9gl
Karush, W. (1939). *Minima of functions of several variables with inequalities as side constraints* (M.Sc. dissertation, Department of Mathematics, University of Chicago).
Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114.
Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., & Paiton, D. (2020). Towards nonlinear disentanglement in natural data with temporal sparse coding. arXiv preprint arXiv:2007.10930.
Kuhnanda, H., & Tucker, W. (1951). Nonlinear programming. In *Proc. 2 Nd Berkeley symp. on mathematical statistics and probability* (pp. 481– 492).
Laskin, M., Srinivas, A., & Abbeel, P. (2020). Curl: Contrastive unsupervised representations for reinforcement learning. In *Proceedings of the 37th annual international conference on machine learning (ICML)*.
Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. arXiv preprint arXiv:1905.13662.
Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning* (pp. 4114–4124).
Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., & Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International conference on machine learning* (pp. 6348–6359). PMLR.
Mobahi, H., Collobert, R., & Weston, J. (2009). Deep learning from temporal coherence in video. In *Proceedings of the 26th annual international conference on machine learning* (pp. 737–744).
Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. In *NIPS*.
Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. arXiv preprint arXiv:1206.6471.
Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., & Brain, G. (2018). Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)* (pp. 1134–1141). IEEE.
Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770.
Zou, W., Zhu, S., Yu, K., & Ng, A. (2012). Deep learning of invariant features via simulated fixations in video. *Advances in Neural Information Processing Systems, 25,* 3203–3211.