DISENTANGLING ACTION SEQUENCES: DISCOVERING CORRELATED IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentanglement is a highly desirable property of representation due to its similarity with human's understanding and reasoning. This improves interpretability, enables the performance of down-stream tasks, and enables controllable generative models. However, this domain is challenged by the abstract notion and incomplete theories to support unsupervised disentanglement learning. We demonstrate the data itself, such as the orientation of images, plays a crucial role in disentanglement and instead of the factors, and the disentangled representations align the latent variables with the action sequences. We further introduce the concept of disentangling action sequences which facilitates the description of the behaviours of the existing disentangling approaches. An analogy for this process is to discover the commonality between the things and categorizing them.

Furthermore, we analyze the inductive biases on the data and find that the latent information thresholds are correlated with the significance of the actions. For the supervised and unsupervised settings, we respectively introduce two methods to measure the thresholds. We further propose a novel framework, fractional variational autoencoder (FVAE), to disentangle the action sequences with different significance step-by-step. Experimental results on dSprites and 3D Chairs show that FVAE improves the stability of disentanglement.

1 INTRODUCTION

The basis of artificial intelligence is to understand and reason about the world based on a limited set of observations. Unsupervised disentanglement learning is highly desirable due to its similarity with the way we as human thinking. For instance, we can infer the movement of a running ball based on a single glance this is because the human brain is capable of disentangling the positions from a set of images. It has been suggested that a disentangled representation is helpful for a large variety of downstream tasks (Schölkopf et al., 2012; Peters et al., 2017). According to Kim & Mnih (2018), a disentangled representation promotes interpretable semantic information. That brings substantial advancement, including but not limited to reducing the performance gap between humans and AI approaches (Lake et al., 2017; Higgins et al., 2018). Other instances of disentangled representation include semantic image understanding and generation (Lample et al., 2017; Zhu et al., 2018; Elgammal et al., 2017), zero-shot learning (Zhu et al., 2019), and reinforcement learning (Higgins et al., 2017b). Despite the advantageous of the disentangling representation approaches, there are still two issues to be addressed including the abstract notion and the weak explanations.

Notion the conception of disentangling factors of variation is first proposed in 2013. It is claimed in Bengio et al. (2013) that for observations the considered factors should be explanatory and independent of each other. The explanatory factors are however hard to formalize and measure. An alternative way is to disentangle the ground-truth factors (Ridgeway, 2016; Do & Tran, 2020). However, if we consider the uniqueness of the ground-truth factors, a question arise here is how to discover it from multiple equivalent representations? As a proverb "one cannot make bricks without straw", Locatello et al. (2019) prove the impossibility of disentangling factors without the help of inductive biases in the unsupervised setting.

Explanation There are mainly two types of explanations for unsupervised disentanglement: information bottleneck, and independence assumption. The ground-truth factors affect the data independently, therefore, the disentangled representations must follow the same structure. The

approaches, holding the independence assumption, encourage independence between the latent variables (Schmidhuber, 1992; Tian Qi Chen & Duvenaud, 2018; Kim & Mnih, 2018; Kumar et al., 2018; Lopez et al., 2018). However, the real-world problems have no strict constraint on the independence assumption, and the factors may be correlative. The other explanation incorporates information theory into disentanglement. Burgess et al.; Higgins et al.; Insu Jeon et al.; Saxe et al. suggest that a limit on the capacity of the latent information channel promotes disentanglement by enforcing the model to acquire the most significant latent representation. They further hypothesize that the information bottleneck enforces the model to find the significant improvement.

In this paper, we first demonstrate that instead of the ground-truth factors the disentangling approaches learn actions of translating based on the orientation of the images. We then propose the concept of disentangling actions which discover the commonalities between the images and categorizes them into sequences. We treat disentangling action sequences as a necessary step toward disentangling factors, which can capture the internal relationships between the data, and make it possible to analyze the inductive biases from the data perspective. Furthermore, the results on a toy example show that the significance of actions is positively correlated with the threshold of latent information. Then, we promote that conclusion to complex problems. Our contributions are summarized in the following:

- We show that the significance of action is related to the capacity of learned latent information, resulting in the different thresholds of factors.
- We propose a novel framework, fractional variational autoencoder (FVAE) to extracts explanatory action sequences step-by-step, and at each step, it learns specific actions by blocking others' information.

We organize the rest of this paper as follows. Sec.2 describes the development of unsupervised disentanglement learning and the proposed methods based on VAEs. In Sec.3, through an example, we show that the disentangled representations are relative to the data itself and further introduce a novel concept, i.e., disentangling action sequences. Then, we investigate the inductive biases on the data and find that the significant action has a high threshold of latent information. In Sec.4, we propose a step-by-step disentangling framework, namely fractional VAE (FVAE), to disentangle action sequences. Finally, we conclude the paper and discuss the future work in Sec.5

2 UNSUPERVISED DISENTANGLEMENT LEARNING

We first introduce the abstract concepts and the basic definitions, followed by the explanations based on information theory and other related works. This article focuses on the explanation of information theory and the proposed models based on VAEs.

2.1 THE CONCEPT

Disentanglement learning is fascinating and challenging because of its intrinsic similarity to human intelligence. As depicted in the seminal paper by Bengio et al., humans can understand and reason from a complex observation to the explanatory factors. A common modelling assumption of disentanglement learning is that the observed data is generated by a set of ground-truth factors. Usually, the data has a high number of dimensions, hence it is hard to understand, whereas the factors have a low number of dimensions thus simpler and easier to be understood. The task of disentanglement learning is to uncover the ground-truth factors, such factors are invisible to the training process in an unsupervised setting. The invisibility of factors makes it hard to define and measure disentanglement (Do & Tran, 2020).

Furthermore, it is shown in Locatello et al. (2019) that it is impossible to unsupervisedly disentangle the underlying factors for the arbitrary generative models without inductive biases. In particular, they suggest that the inductive biases on the models and the data should be exploited. However, they do not provide a formal definition of the inductive bias and such a definition is still unavailable.

2.2 INFORMATION BOTTLENECK

Most of the dominant disentangling approaches are the variants of variational autoencoder (VAE). The variational autoencoder (VAE) is a popular generative model, assuming that the latent variables obey

a specific prior (normal distribution in practice). The key idea of VAE is maximizing the likelihood objective by the following approximation:

$$\mathcal{L}(\theta,\phi;x,z)=\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)]-D_{\mathrm{KL}}(q_\phi(z|x)||p(z)) \tag{1}$$

which is know as the evidence lower bound (ELBO); where the conditional probability $P(x|z)$,$Q(z|x)$ are parameterized with deep neural networks.

Higgins et al. find that the KL term of VAEs encourages disentanglement and introduce a hyperparameter β in front of the KL term. They propose the β-VAE maximizing the following expression:

$$\mathcal{L}(\theta,\phi;x,z)=\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)]-\beta D_{\mathrm{KL}}(q_\phi(z|x)||p(z)). \tag{2}$$

β controls the pressure for the posterior $Q_\phi(z|x)$ to match the factorized unit Gaussian prior $p(z)$. Higher values of β lead to lower implicit capacity of the latent information and ambiguous reconstructions. Burgess et al. propose the Annealed-VAE that progressively increases the information capacity of the latent code while training:

$$\mathcal{L}(\theta,\phi;x,z,C)=\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(x|z)]-\gamma|D_{\mathrm{KL}}(q_\phi(z|x)||p(z))-C| \tag{3}$$

where γ is a large enough constant to constrain the latent information, C is a value gradually increased from zero to a large number to produce high reconstruction quality. As the total information bottleneck gradually increasing, they hypothesize that the model will allocate the capacity of the most improvement of the reconstruction log-likelihood to the encoding axes of corresponding factor. However, they did not exploit why each factor makes the different contributions to the reconstruction log-likelihood.

2.3 Other related work

The other dominant direction initiates from the prior of factors. They assume that the ground-truth factors are independent of each other, and a series of methods enforce the latent variables have the same structure as the factors. FactorVAE (Kim & Mnih, 2018) applies a discriminator to approximately calculate the total correlation (TC, Watanabe (1960)); β-TCVAE (Tian Qi Chen & Duvenaud, 2018) promotes the TC penelty by decomposing the KL term; DIP-VAE (Kumar et al., 2018) identifies the covariance matrix of $q(z)$. However, the real-world problems have no strict constrain in the prior of factors. For instance, the length of hair and the gender are two independent factors to describe a person, but the real situation is that women are more likely to have long hair.

3 Disentangling action sequences

For machines and humans, disentangling the underlying factors is a challenging task. For instance, there are more than 1,400 breeds of dogs in the world, and it seems impossible for an ordinary person to distinguish all of them just by looking at their pictures. The challenge in disentangling the underlying factors is mainly due to the complexity of establishing relationships without supervision, where the corresponding models should contain some level of prior knowledge or inductive biases. However, it is possible to determine the differences without having extensive knowledge. For example, one may mistakenly identify the breed of a dog by looking at a picture but is almost impossible to misrecognize a dog as a cat. Therefore, in practice, discovering the differences or similarities is often a much easier task than that of uncovering the underlying factors, and it does not also need much prior knowledge. One may conclude that discovering the commonalities between the things is an important step toward disentanglement.

3.1 Disentangled representations

Although the disentangling approaches can learn informative, independent, and meaningful representations on dSprites (Matthey et al., 2017), these models disentangle the ground-truth factors by accident. In other words, there is no reason for the latent variables to have any structure since there always exists a decoder that can map the latent variable into the desired output. To determine the representations that the current approaches learn, here we design a dataset family to reveal the common aspects of these representations.

We create a toy dataset family— each dataset contains 40x40 images which are generated from an original image of a 11x5 rectangle, by translating on a 64x64 canvas. Each image on the dataset

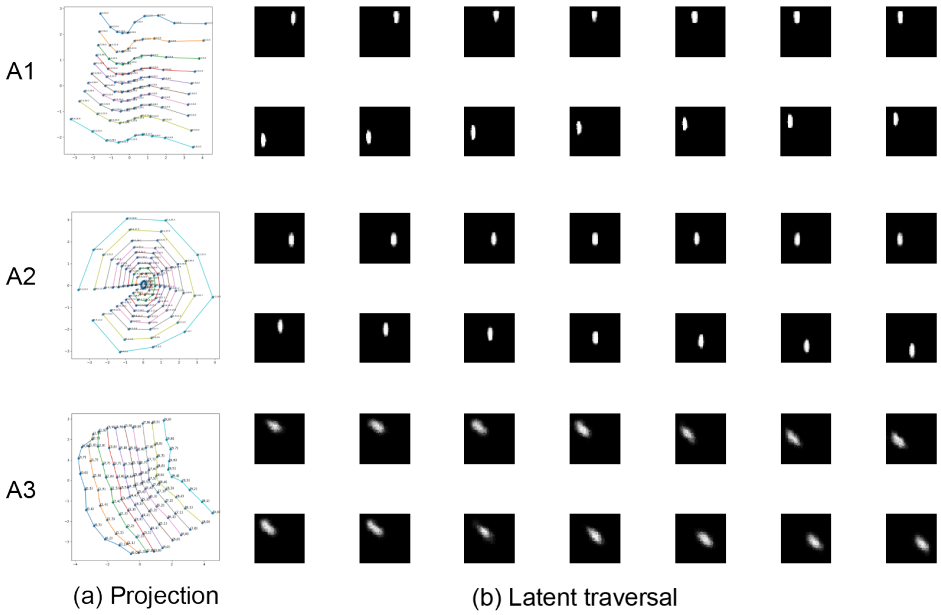


Figure 1: Leaned presentations on A1, A2, and A3. (a) The factors are projected into the latent space. (b) Each row denotes the latent traversal on the specific dataset.

has a unique label (position of X, position of Y) to describe the ground-truth factors. In this dataset family, there are two variables: the orientation of the rectangle and the way to determine the two factors. There are infinite solutions to determine these two factors; the polar coordinate system and the Cartesian coordinate system are the most common solution. We then create a baseline dataset, A1, with a horizontal rectangle in the Cartesian coordinate system and obtain its variants. A2 differs in the positions determined by the polar coordinate system, and A3 differs in the orientation (45 degrees) of the images. For the experiment settings, we choose the well-examined baseline model, β-VAE ($\beta = 50$), and the backbone network follows the settings in Locatello et al. (2019).

As it is shown in Fig. 1(a), we visualize the learned representation on the latent space. In the left column, each point represents an image with the annotation of the corresponding factors. We also link the points with the same controlling factor with coloured lines to show the ground-truth transformation. We argue that the factors are not the key to disentanglement since the learned representations are changed while the factors are unchanged (A1, A3), and the learned representations do not change while the factors are changed (A1, A2). From Fig. 1, one can see that the invariant in these three cases is that they learn a representation moving along the direction of the long side of the rectangle and the orthogonality direction. The disentangled representations consider the rectangle as the reference system.

3.2 Action Sequences

We assume that images of the observed data consist of a set of actions such as rotation, translation, and scaling. Given a subset of images from original data set, we define the **meaningful action** as a sequence, i.e. an ordered permutation of elements from the subset. To obtain such sequences, one can get an action by traversing the ground-truth factors:

$$\mathbb{S} = \{\boldsymbol{x}_i\} = T(\{\boldsymbol{c}_i\}), \tag{4}$$

where T is a transformation, $\boldsymbol{c}_i$ is the set of factors, $\boldsymbol{x}_i$ is an image from the dataset. Therefore, the disentanglement learning can be interpreted as a process of learning meaningful action sequences. Furthermore, these actions have associated parameters to illustrate the changing processes in detail, i.e., a parameterized action sequence represents an action. We define an action sequence as a subset of images from the dataset, and it reveals the relationships amongst the images. The formal definition of action sequences can be described as the latent traversal in Higgins et al. (2017a).

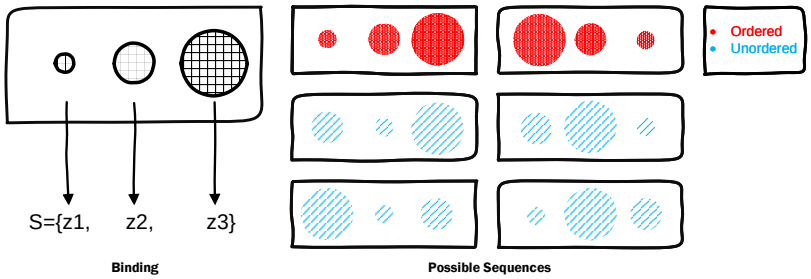
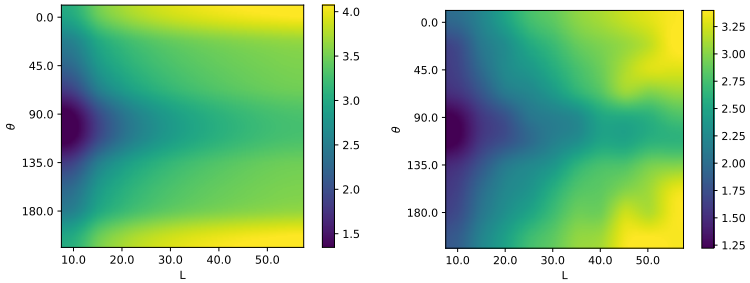


Figure 2: An example of action sequences. There are three images with differing sizes on the dataset (left box), and the AE learns a representation with one dimension of latent space (the double-arrow lines). So the total number of possible sequences is six. Two of them are meaningful, and the others are somewhat random.



(a) Entropy (b) KL divergence

Figure 3: Comparison of entropy and KL divergence on A4.

For an autoencoder (AE) architecture, the neural networks are flexible enough to approximate any actions. For clarity, we define an action as a sequence of images controlled by the ground-truth factors, and an action sequence is the approximation of this action generated by a neural network. Due to the strong expression of neural networks, the sequences may be organized in any given order. As shown in 2, the neural network can form any possible sequences, and among them, the ordered sequences follow human's intuition—the circle becomes larger or smaller. Using VAEs, minimizing the objective increase the overlap between the posterior distributions across the dataset Burgess et al. (2018), and it then leads to disentanglement and emerges human intuition.

3.3 Significance of Sequences

Although Locatello et al. (2019) addresses the importance of inductive biases, few studies investigate the biases for disentanglement on the data. As we have shown in Sec. 3.1, the orientation of the rectangle can affect the direction of disentangled representation. We assume that there are much more inductive biases on the data which have not been yet exploited. It is suggested in Burgess et al. (2018) that the latent components have different contributions to the objective function. Here, we investigate the reasons for different contributions and show that the capacity of latent information is correlated to the significance of the sequences. To measure the significance of actions, we use the information among this action:

$$H(\mathbb{S}) = -\int_{\mathbb{S}} p(\boldsymbol{x}) \log(p(\boldsymbol{x})) d\boldsymbol{x}, \tag{5}$$

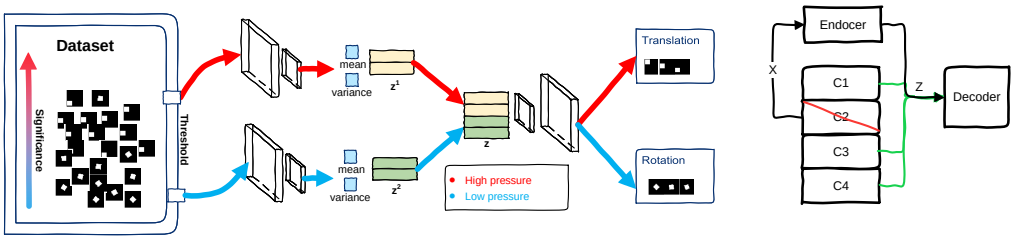
where $\mathbb{S}$ is the action controlled by the factor, .

We design a translation family (A4) with two controllable parameters, θ, L to determine the significance of the sequence; θ is the orientation of the rectangle; L is the maximal distance of translating. All datasets have only one factor which is the position of X-axis, and also contain one sequence. If the value of L is small enough, the images in the sequence are almost the same, this sequence is referred

to as insignificant. Both θ, L have different effects on the significance of the sequence. As shown in Fig. 3, the trend of KL divergence is consistent with that of entropy. Please note that the maximum for both are reached when $\theta = 90$ and L is at its maximum. These imply that the significance is closely related to information among the sequence.

Therefore, a higher significance of the sequence results in more information on the latent variables. In contrast, by the gradual increase of the pressure on the KL term, the latent information decreases until it reaches zero. One can infer that there exists a threshold of latent information. Experimental results show a positive correlation between the significance of data and the threshold of latent information.

4 PLAIN IMPLEMENTATION ACCORDING TO THRESHOLDS



(a) FVAE (b) Mixed label

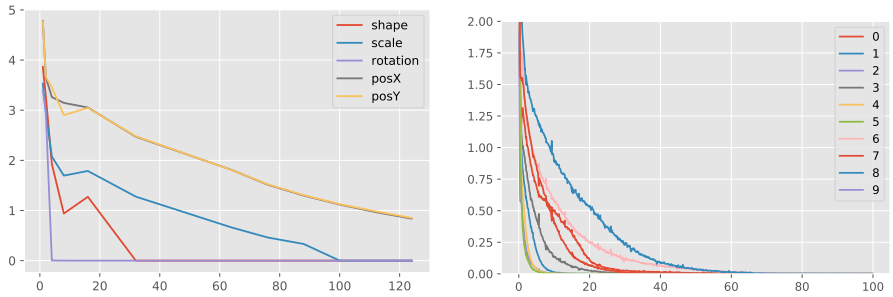
Figure 4: (a) The architecture of FVAE. Though the samples distribute in the dataset randomly, they have intrinsic significance. Under a high pressure (β), the significant actions can pass information alone the red line to itself along, and the information of insignificant actions is blocked. (b) The decoder receive the label information except for the target action.

We have discussed the situation with only one action and there are various thresholds for different significant actions. Particularly, if β is large enough, information of the insignificant actions will be blocked, leading to the disentangling process decays to a single factor discovery problem. From the modelling perspective, the learning process is similar to a single action learning problem. However, the difficulty of disentanglement is that different kinds of ground-truth actions are mixed, and a single and a fixed parameter β is unable to separate them. Therefore, a plain idea is to set different thresholds on the learning phases, and then in each phase, we enforce the model to learn specific actions by blocking information of the secondarily significant actions. We propose a fractional variational autoencoder (FVAE) which disentangles the action sequences step-by-step. The architecture of FVAE is shown in Fig. 4(a). The encoder consists of several groups of sub-encoders, and the inputs of the decoder are the concatenated codes of all sub-encoders. Besides, to prevent re-entangling the learned actions, we set different learning rates for the sub-encoders, which reduces the learning rate for the reset of N-1 groups and prevent the model from allocating the targeted action into the leaned codes. The training process of FVAE is similar to the common operation in chemistry for separating mixtures—distillation. To separate a mixture of liquid, we repeat the process of heating the liquid with different boiling points, and in each step, the heating temperature is different to ensure that only one component is being collected.

Discussion Although AnnealedVAE follows the same principles as the FVAE, it differs in the interpretation of the effects of beta, and it does not explicitly prevent mixing the factors. Moreover, the performance of AnnealedVAE depends on the choice of hyperparameter in practice Locatello et al. (2019). "A large enough value" is hard to determine, and the disentangled representation is re-entangled for an extremely large C. To address this issue, here we introduce two methods to determine the thresholds in each phase for the labelled and unlabelled tasks.

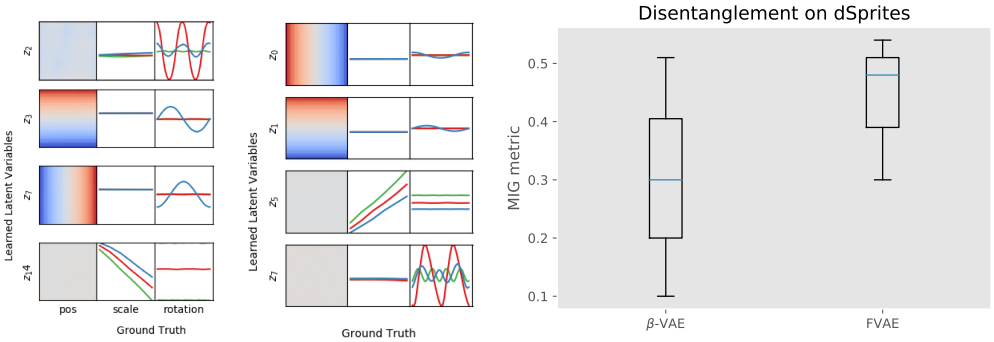
4.1 LABELLED TASK

For the labelled setting, we focus on one type of actions and clip the rest of them at first. However, the samples of one action are usually insufficient. For example, there are only three types of shapes on dSprites. Besides, the label information may be corrupted and only some parts of the dataset are labelled.



(a) (b)

Figure 5: β vs KL divergence on dSprites (left) and 3D Chairs (right).



(a) β-VAE (b) FVAE (c) Disentanglement

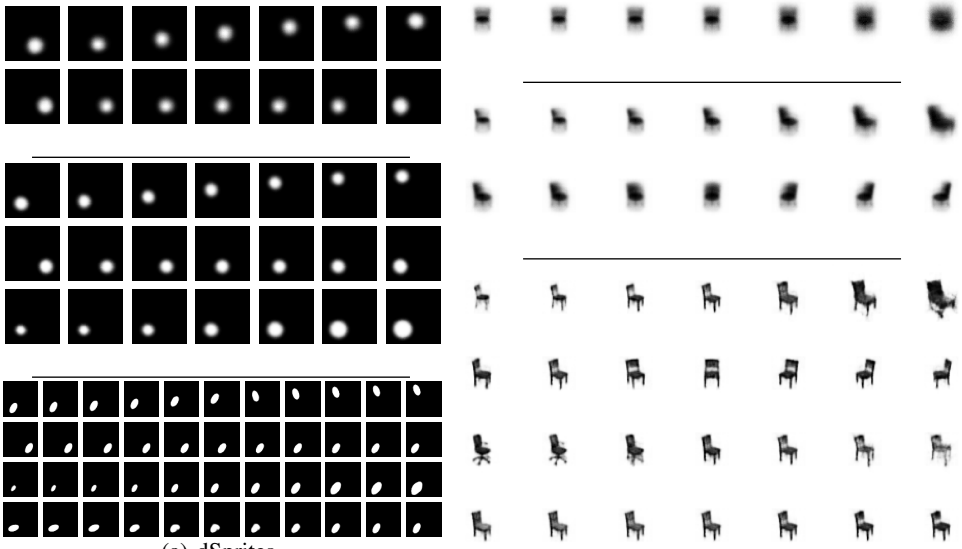
Figure 6: Comparison between β-VAE and FVAE. (a) and (b) show the relationship between $\boldsymbol{z}$ and factors on dSprites (only active units are shown). (c) The disentanglement scores of β-VAE and FVAE (MIG, Tian Qi Chen & Duvenaud (2018)).

To address these issues, we introduce the architecture shown in Fig. 4(b), in which the label information excepted for the target actions are directly provided to the decoder. We evaluate FVAE on dSprites dataset (involving five actions: translating, rotating, scaling, and shaping). We first measure the threshold of each action, and the result is shown in Fig. 5(a). One can see that the thresholds of translating and scaling are higher than the others. This suggests that these actions are significant and easy to be disentangled. This is in line with the results in Burgess et al. (2018); Higgins et al. (2017a).

According to these thresholds, we then arrange three stages for dSprites. At each stage, we set up a bigger number of β than the threshold of the secondary action. The pressure on the KL term also prevents the insignificant actions from being disentangled and ensures that the model only learns from the information of the target action. The training settings of each stage can be found in the Appendix A.2.1. As shown in Fig. 7(a), the translation factor is disentangled at first easily, while it is hard to distinguish the shape, orientation, and scale. Gradually, scaling and orientation also emerge in order. Nevertheless, it should be noted that the shaping is still hard to be separated. This could be attributed to the lack of commonalities between these three shapes on dSprites and motion compensation for a smooth transition. In other words, in terms of shape, the lack of intermediate states between different shapes is an inevitable hurdle for its disentanglement. Fig. 6 shows more substantial difference between the β-VAE and the FVAE. β-VAE has an unstable performance compared to FVAE, and position information entangles with orientation on some dimension.

4.2 Unlabelled Task

For the unlabelled setting, we introduce the annealing test to detect the potential components. In the beginning, a very large value for β is set to ensure that no action is learned. Then, we gradually



(a) dSprites (b) 3D Chairs

Figure 7: **FVAE disentangles action sequences step-by-step.** Latent traversals at each stage. The left is the results on dSprites. The right is the results on 3D Chairs

decrease β to disentangle the significant actions. There exists a critical point in which the latent information starts increasing and that point approximates the threshold of the corresponding action.

3D Chairs is an unlabelled dataset containing 1394 3D models from the Internet. Fig. 5(b) shows the result of the annealing test on 3D Chairs. One can recognize three points where the latent information suddenly increases: 60, 20, 4. Therefore, we arrange a three-stage training process for 3D Chairs (more details in the Appendix). As shown in Fig. 7(b), one can see the change of azimuth in the first stage. In the second stage, one can see the change of size and in the third stage, one can see the change of the leg style, backrest, and material used in the making of the chair.

5 Conclusion

We demonstrated an example of the effects of images' orientation on the disentangled representations. We have further investigated the inductive biases on the data by introducing the concept of disentangling action sequences, and we regarded that as discovering the commonality between the things, which is essential for disentanglement. The experimental results revealed that the actions with higher significance have a larger value of thresholds of latent information. We further proposed the fractional variational autoencoder (FVAE) to disentangle the action sequences with different significance step-by-step. We then evaluated the performance of FVAE on dSprites and 3D Chairs. The results suggested robust disentanglement where re-entangling is prevented.

This paper proposed a novel tool to study the inductive biases by action sequences. However, other properties of inductive biases on the data remain to be exploited. The current work focuses on an alternative explanation for disentanglement from the perspective of information theory. In the future, the influence of independence on disentanglement requires further investigation.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 35(8):1798–1828, 2013.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. (Nips), 2018.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. 2020.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating art by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

I. Higgins, Nicolas Sonnerat, Loic Matthey, A. Pal, C. Burgess, Matko Bosnjak, M. Shanahan, M. Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning hierarchical compositional visual concepts. *arXiv: Machine Learning*, 2018.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–13, 2017a.

Irina Higgins, Arka Pal, Andrei A. Rusu, Loïc Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: improving zero-shot transfer in reinforcement learning. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1480–1490. PMLR, 2017b.

Insu Jeon, Wonkwang Lee, and Gunhee Kim. Ib-gan: Disentangled representation learning with information bottleneck gan, 2019.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *35th International Conference on Machine Learning, ICML 2018*, 6:4153–4171, 2018.

Abhishek Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *ArXiv*, abs/1711.00848, 2018.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, and Marc Aurelio Ranzato. Fader networks:manipulating images by sliding attributes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *NIPS 30*, pp. 5967–5976. Curran Associates, Inc, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucie, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *36th International Conference on Machine Learning, ICML 2019*, 2019-June: 7247–7283, 2019.

Romain Lopez, Jeffrey Regier, N. Yosef, and Michael I. Jordan. Information constraints on auto-encoding variational bayes. In *NeurIPS*, 2018.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

J. Peters, D. Janzing, and B. Schölkopf. Elements of causal inference: Foundations and learning algorithms. 2017.

K. Ridgeway. A survey of inductive biases for factorial representation-learning. *ArXiv*, abs/1612.05299, 2016.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. 2018.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992. ISSN 08997667. doi: 10.1162/neco.1992.4.6.863.

B. Schölkopf, D. Janzing, J. Peters, Eleni Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *ICML*, 2012.

Roger B. Grosse Tian Qi Chen, Xuechen Li and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS 2018, 3-8 December Montreal, Canada*, pp. 2615–2625, 2018.

Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, 4(1):66–82, 1960.

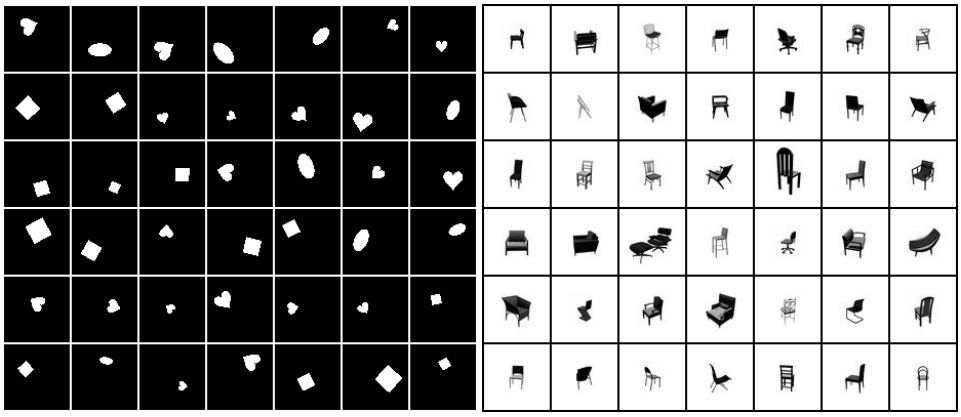
Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1004–1013, 2018.

Yizhe Zhu, Jianwen Xie, Bingchen Liu, and Ahmed Elgammal. Learning feature-to-feature translator by alternating back-propagation for zero-shot learning. *arXiv preprint arXiv:1904.10056*, 2019.

A Appendix

A.1 Datasets

A.2 dSprites and 3D Chairs



(a) (b)

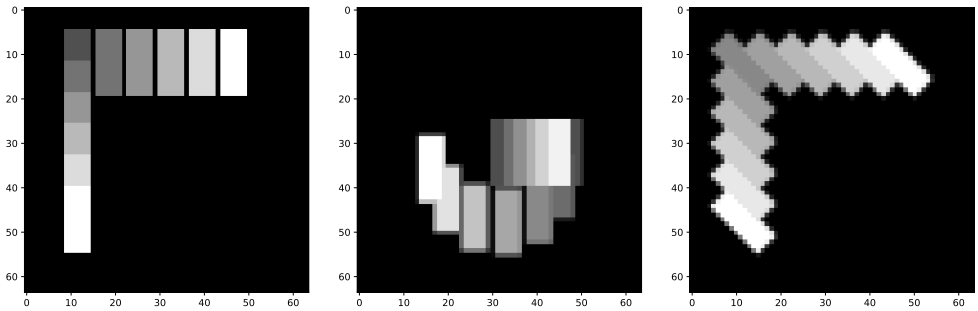
Figure 8: Real samples from the training dataset.

A.2.1 Designed Datasets

The measured thresholds are shown in Tab. 3 and 2. Therefore, at each stage, we set the β a higher value to block insignificant information, and the hyperparameters of each stage are shown in Tab. 1.

Phase	1	2	3
E_1	**5e-4**	5e-5	5e-5
E_2	0	**5e-4**	5e-5
E_3	0	0	**5e-4**
β_1	100	40	4
β_2	60	20	2

Table 1: Training settings on dSprites and 3D Chairs. E_i denotes the learning rate of i-th group of sub-encoder. β denotes the regularisation coefficient before the KL divergence, β_1 for dSprites, β_2 for 3D Chairs



(a) A1 (b) A2 (c) A3

Figure 9: Samples on A1, A2, A3. We visualize these samples by afterimage, and the image starts with a dark colour, then, it goes brighter as traversing the factor.

Factor	Threshold
Shape	32
Scale	100
Orientation	5
Position X	120+
Position Y	120+

Table 2: Thresholds of actions on dSprites.

Factor	Threshold
Chair size	60
Leg style	20
Swivel	2
Unknown	2

Table 3: Thresholds of actions on 3D Chairs.

A.3 SAMPLES

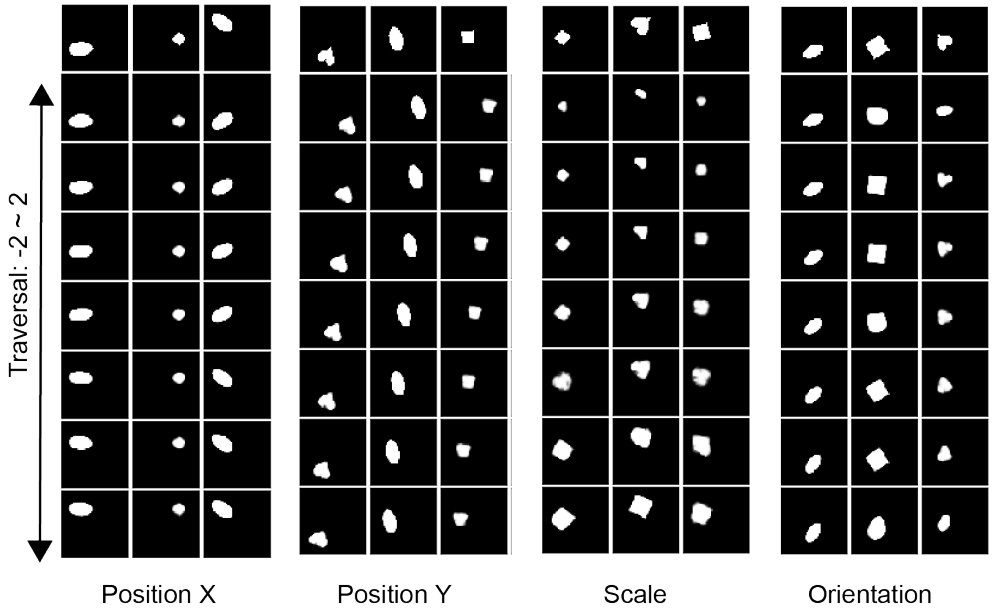


Figure 10: Latent traversal plots for FVAE on dSprites. The top row show the real samples from dSprites.

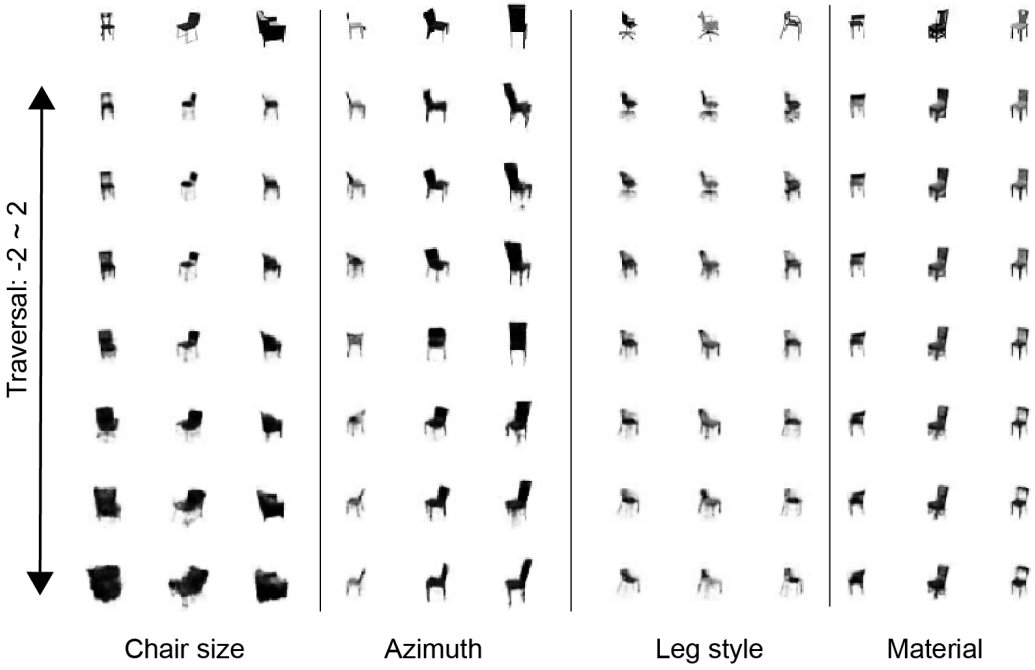


Figure 11: Latent traversal plots for FVAE on 3D Chairs. The top row show the real samples from 3D Chairs.