
Early Visual Concept Learning with Unsupervised Deep Learning

Irina Higgins
Google DeepMind
irinah@google.com

Loic Matthey
Google DeepMind
lmatthey@google.com

Xavier Glorot
Google DeepMind
glorotx@google.com

Arka Pal
Google DeepMind
arkap@google.com

Benigno Uria
Google DeepMind
buria@google.com

Charles Blundell
Google DeepMind
cblundell@google.com

Shakir Mohamed
Google DeepMind
shakir@google.com

Alexander Lerchner
Google DeepMind
lerchner@google.com

Abstract

Automated discovery of early visual concepts from raw image data is a major open challenge in AI research. Addressing this problem, we propose an unsupervised approach for learning disentangled representations of the underlying factors of variation. We draw inspiration from neuroscience, and show how this can be achieved in an unsupervised generative model by applying the same learning pressures as have been suggested to act in the ventral visual stream in the brain. By enforcing redundancy reduction, encouraging statistical independence, and exposure to data with transform continuities analogous to those to which human infants are exposed, we obtain a variational autoencoder (VAE) framework capable of learning disentangled factors. Our approach makes few assumptions and works well across a wide variety of datasets. Furthermore, our solution has useful emergent properties, such as zero-shot inference and an intuitive understanding of “objectness”.

1 Introduction

State-of-the-art AI approaches still struggle with some scenarios where humans excel [21], such as knowledge transfer, where faster learning is achieved by reusing learnt representations for numerous tasks (Fig. 1A); or zero-shot inference, where reasoning about new data is enabled by recombining previously learnt factors (Fig. 1B). [21] suggest incorporating certain “start-up” abilities into deep models, such as intuitive understanding of physics, to help bootstrap learning in these scenarios. Elaborating on this idea, we believe that learning basic visual concepts, such as the “objectness” of things in the world, and the ability to reason about objects in terms of the generative factors that specify their properties, is an important step towards building machines that learn and think like people. We believe that this can be achieved by learning a disentangled posterior distribution of the generative factors of the observed sensory input by leveraging the wealth of unsupervised data [4, 21]. We wish to learn a representation where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors [4]. With a disentangled representation, knowledge about one factor could generalise to many configurations of other factors, thus capturing the “multiple explanatory factors” and “shared factors across tasks” priors suggested by [4]. Unsupervised disentangled factor learning from raw image data is a major open challenge in AI. Most previous attempts require a priori knowledge of the number and/or nature of the data

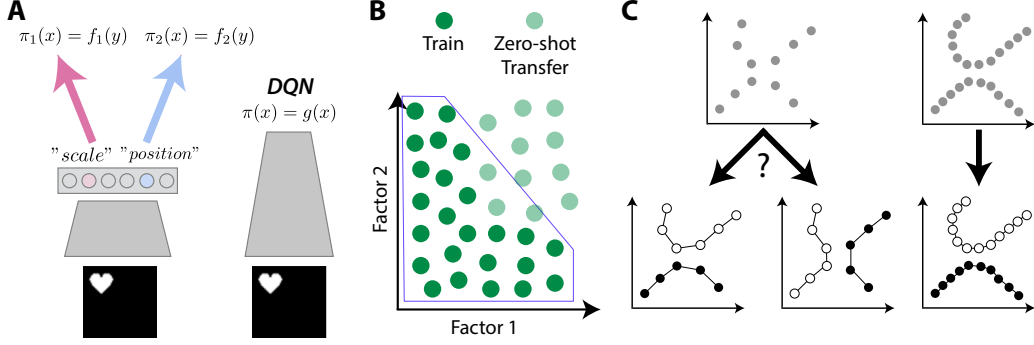


Figure 1: **A:** Disentangled representations of data generative factors allow for fast knowledge transfer between different reinforcement learning (RL) policies. State of the art RL models without such representations (e.g. DQN by [23]) require complete re-learning of low-level features for different tasks [21]. **B:** Models are unable to generalise to data outside of the convex hull of the training distribution (light blue line) unless they learn about the data generative factors and recombine them in novel ways. **C:** Sparse data points do not provide enough information for an unsupervised model to identify where the data manifold should lie. Data generated using factors densely sampled from continuous distributions makes manifold learning less ambiguous.

generative factors [16, 25, 35, 34, 13, 20, 8, 33, 17]. This is infeasible in the real world, where the newborn learner may have no a priori knowledge and little to no supervision for discovering the generative factors. So far any purely unsupervised approaches to disentangled factor learning have not scaled well [11, 30, 9, 10].

We propose a deep unsupervised generative approach for disentangled factor learning inspired by neuroscience [2, 3, 24, 15]. We apply similar learning constraints to the model as have been suggested to act in the ventral visual stream in the brain [28]: redundancy reduction, an emphasis on learning statistically independent factors, and exposure to data with transform continuities analogous to those human infants are exposed to [2, 3]. We show that the application of such pressures to a deep unsupervised generative model can be realised in the variational autoencoder (VAE) framework [19, 26]. Our main contributions are the following: 1) we show the importance of neuroscience inspired constraints (data continuity, redundancy reduction and statistical independence) for learning disentangled representations of continuous visual generative factors; 2) we devise a protocol to quantitatively compare the degree of disentanglement learnt by different models; and 3) we demonstrate how learning disentangled representations enables zero-shot inference and the emergence of basic visual concepts, such as “objectness”.

2 Constraints to encourage disentangled factor learning

The infants’ ventral visual stream learns basic visual concepts through exposure to unsupervised data during the first few months of life [28, 5]. We hypothesise that a deep unsupervised model should be able to learn similar representations if exposed to similar data streams and put under the same learning constraints as the visual brain. In this section we elaborate on this hypothesis.

Continuously transformed data Up to around 3 months of age human babies are unable to focus beyond 8-10 inches [22]. Their visual cortices are learning from a large unsupervised dataset of objects transforming continuously against a blurred background [6]. Computational neuroscience simulations of the ventral visual pathway suggest that the response properties of neurons in the inferior temporal cortex arise through a Hebbian learning algorithm that relies on the fact that nearest neighbours of a particular object in pixel space are transforms of the same object [24]. This notion can be generalised within the manifold learning framework. As shown in Fig. 1C, sparse samples from data transformation manifolds provide little information for unsupervised models about the manifold shapes. This ambiguity may be resolved through either dense sampling of the manifolds or by adding supervised signals. The importance of vast quantities of unlabeled data for the success of unsupervised approaches in learning disentangled factor representations was pointed by [4]. Here we specify a particular aspect of the data we believe is important for such learning. We postulate that it is

important that the *observed data is generated using factors of variation that are densely sampled from their respective continuous distributions*. We leave the learning of discrete factors to future work.

Redundancy reduction and independence According to [2], one of the main functions of the sensory brain is redundancy reduction, where redundancy is defined as the difference between the maximum entropy that a channel can transmit, and the entropy of messages actually transmitted. Sensory redundancy reduction is facilitated through learning statistically independent components within the data [3]. We hypothesise that an unsupervised deep model encouraged to perform redundancy reduction and to learn statistically independent components from continuous data, as described in the section above, will learn basic visual concepts similar to those learnt by the ventral visual stream. Such constraints have been considered before [32, 29, 27], but no scalable unsupervised solution capable of disentangled factor learning based on these ideas yet exists.

We start by specifying an unsupervised deep generative model for learning latent factors $\mathbf{z} \in \mathbb{R}^m$ that, when combined in a non-linear way, generate the observed data \mathbf{x} . For a given observation, we describe the plausible posterior configurations of such generative latent factors \mathbf{z} by a probability distribution $q_\phi(\mathbf{z}|\mathbf{x})$. We aim to maximise the probability of the observed data \mathbf{x} on average over all possible samples from the latent factors \mathbf{z} . This corresponds to the optimisation problem in Eq. 1.

$$\max_{\phi, \theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] \quad (1)$$

In order to learn disentangled representations of the generative factors we introduce a constraint that encourages the distribution over latent factors \mathbf{z} to be close to a prior that embodies the neuroscience inspired pressures of redundancy reduction and independence prior. This results in a constrained optimisation problem shown in Eq. 2, where ϵ specifies the strength of the applied constraint.

$$\max_{\phi, \theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] \quad \text{subject to } D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon. \quad (2)$$

Writing Eq. 2 as a Lagrangian we obtain the familiar variational free energy objective function shown in Eq. 3 [19, 26], where $\beta \geq 0$ is the inverse temperature or regularisation coefficient.

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

If we set the disentangled prior to be an isotropic unit Gaussian ($p(\mathbf{z}) = \mathcal{N}(0, I)$), the variational bound in Eq. 3 matches well the desiderata proposed by [2, 3]. It adds redundancy reduction pressure by constraining the capacity of the latent information channel \mathbf{z} , while preserving enough information to enable reconstruction. The isotropic nature of the Gaussian prior puts implicit independence pressure on the learnt posterior. Varying β changes the degree of applied learning pressure during training, thus encouraging different learnt representations. When $\beta = 0$, we obtain the standard maximum likelihood learning. When $\beta = 1$, we recover the Bayes solution. We postulate that in order to learn disentangled representations of the continuous data generative factors it is important to *tune β to approximate the level of learning pressure present in the ventral visual stream*.

3 Experiments

3.1 Learning disentangled factors in a 2D dataset

We first demonstrate that a VAE can learn disentangled generative factors when exposed to a dataset with continuous transformations as defined in Sec. 2. We use a synthetic binary dataset of 737,280 2D shapes (heart, oval and square) generated from the Cartesian product of four factor values v_k defined in vector graphics: position X (32 values), position Y (32 values), scale (6 values) and rotation (40 values over the 2π range). To ensure smooth affine object transforms, each two subsequent values for each factor v_k were chosen to ensure minimal differences in pixel space given 64x64 pixel image resolution. We used randomly sampled batches of size 100 to train a fully connected VAE with $m = 10$ latent units and various β values until convergence (see Tbl. 1 in Appendix for details). After training, a VAE with $\beta = 0.01$ learnt a good (while not perfect) disentangled representation of the data generative factors, and its decoder learnt to act as a rendering engine (Fig. 2A). The most informative units z_i have the highest KL divergence from the unit Gaussian prior ($p(z) = \mathcal{N}(0, I)$), while the uninformative latents have KL divergence close to zero. Throughout the rest of the paper we illustrate disentangling performance of various models using latents with the highest KL divergence from the prior.

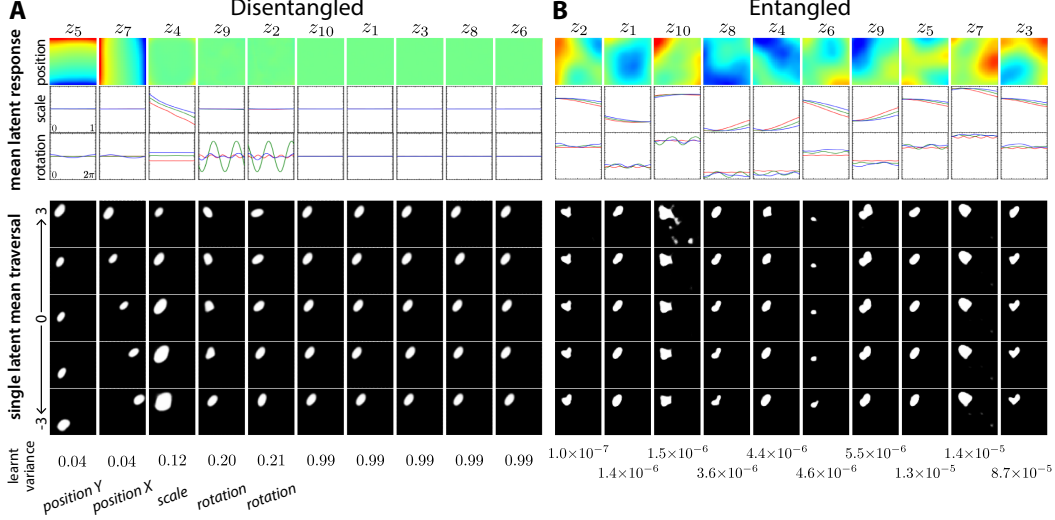


Figure 2: **A:** Disentangled representation learnt with $\beta = 0.01$. Each column represents a latent z_i , ordered according to the learnt Gaussian variance (last row). Row 1 (position) shows the mean activation (red represents high values) of each latent z_i as a function of all 32x32 locations averaged across objects, rotations and scales. Row 2 (scale) shows the mean activation of each unit z_i as a function of scale (averaged across rotations and positions). Row 3 (rotation) shows the mean activation of each unit z_i as a function of rotation (averaged across scales and positions). *Square* is red, *oval* is green and *heart* is blue. Rows 4-8 (second group) show reconstructions resulting from the traversal of each latent z_i over three standard deviations around the unit Gaussian prior mean while keeping the remaining 9/10 latent units fixed to the values obtained by running inference on an image from the dataset. After learning, five latents learnt to represent the generative factors of the data, while the others converged to the uninformative unit Gaussian prior. **B:** Similar analysis for an entangled representation learnt with $\beta = 0$.

Fig. 2A demonstrates the selectivity of each latent z_i to the continuous data generating factors: $z_i^\mu = f(v_k) \forall v_k \in \{v_{\text{position}X}, v_{\text{position}Y}, v_{\text{scale}}, v_{\text{rotation}}\}$ (top three rows), where z_i^μ stands for the learnt Gaussian mean of latent unit z_i . The effect of traversing each latent z_i on the resulting reconstructions is shown in the bottom five rows of Fig. 2A. It can be seen that latents z_7 and z_5 learnt to encode X and Y coordinates of the objects respectively; unit z_4 learnt to encode scale; and units z_2 and z_9 learnt to encode rotation. The frequency of oscillations in each rotational latent corresponds to the rotational symmetry of the corresponding object (2π for heart, π for oval and $\pi/2$ for square). Furthermore, the two rotational latents seem to encode \cos and \sin rotational coordinates, while the positional latents align with the Cartesian axes. While such alignment with human intuition is not guaranteed, empirically we found it very common. Fig. 2B demonstrates that a model with inappropriate learning pressures ($\beta = 0$) does not learn about the generative factors in the data and instead learns a dense entangled latent representation.

3.2 Quantifying disentangling

We have devised a metric to quantitatively approximate the degree of disentanglement within the learnt latent representations. The metric uses a linear classifier to predict which factor caused the transition between two frames in the dataset, where the frames are identical apart from a random change in a single generative factor. We use a low VC dimension classifier that has no capacity to do the disentangling itself to ensure that good classification performance can be achieved only if the generative factors are already disentangled in the latent space \mathbf{z} . The classifier has to learn a mapping $G(\mathbf{z}_{diff}) : \mathcal{R}^m \rightarrow \mathcal{R}^k$, where m is the dimensionality of the latent space \mathbf{z} , k is the number of factors in the dataset (in our case four: scale, rotation, position X and position Y) and $\mathbf{z}_{diff} = \frac{|\mathbf{z}_{start}^\mu - \mathbf{z}_{end}^\mu|}{\max(|\mathbf{z}_{start}^\mu - \mathbf{z}_{end}^\mu|)}$ is the change in the latent space corresponding to a change in a single generative factor in pixel space (see Alg. 1 in Appendix for details). Classification performance is reported for 5,000 test samples in Fig. 3. VAE that learnt a disentangled representation of the data



Figure 3: **A:** Factor change classification accuracy for the original 2D shapes dataset (heart, oval and square). Ground truth uses data generating vectors v . PCA decomposition keeps the first ten components (explaining 60.8% of variance). Untrained refers to a VAE with random weights. Disentangled is a VAE with $\beta = 0.01$. Entangled uses $\beta = 0$. **B:** “Zero-shot Understanding” refers to a VAE that did not see particular combinations of the generative factors during training (see Sec. 3.4), but had to reason about them during factor change classification. A projection of the hypercube formed by the data generative factors is visualised on the right. Only the yellow subset was used for training. The held out factor combinations are shown in grey and were used to evaluate the factor change classification accuracy.

generating factors (model in Fig. 2A) achieved similar classification score to the one obtained using ground truth data generation vectors v_{diff} . Both scores are significantly higher than several varied baselines: an untrained VAE with the same architecture, a VAE that learnt an entangled representation (model in Fig. 2B), the top ten PCA components of the data (PCA_{diff}) or using the raw pixels (x_{diff}), see Fig. 3A.

3.3 Factors affecting learning

In this section we investigate the sensitivity of disentangled factor learning in the VAE framework to the learning constraints of data continuity, redundancy reduction and independence.

Data continuity We hypothesised that data continuity is important for guiding unsupervised models towards learning the correct data manifolds (Sec. 2). To test this idea we measured how the degree of learnt disentangling changes with reduced continuity in the 2D shapes dataset. We trained a VAE with $\beta = 0.01$ (Fig. 2A) on subsamples of the original 2D shapes dataset, where we progressively decreased the generative factor sampling density. Reduction in data continuity negatively correlates with the average pixel wise (Hamming) distance between two consecutive transforms of each object (normalised by the average number of pixels occupied by each of the two adjacent transforms of an object to account for object scale). Fig. 4A demonstrates that as the continuity in the data reduces, the degree of disentanglement in the learnt representations also drops. This effect holds after additional hyperparameter tuning and can not solely be explained by the decrease in dataset size, since the same VAE can learn disentangled representations from a data subset that preserves data continuity but is approximately 55% of the original size (see Sec. 3.4).

Optimizing learning constraints We hypothesised that constrained optimisation is important for enabling deep unsupervised models to learn disentangled representations of the data generative factors (Sec. 2). In the VAE framework this corresponds to tuning the β coefficient. We find that larger latent z layer sizes m require higher constraint pressures (higher β values) (Fig. 4B). In some circumstances $\beta > 1$ is necessary. Furthermore, the relationship of β for a given m is characterised by an inverted U curve. When β is too low or too high the model learns an entangled latent representation due to either too much or too little capacity in the latent z bottleneck. We also note that VAE reconstruction quality is a poor indicator of learnt disentanglement. Good disentangled representations often lead to blurry reconstructions due to restricted capacity of the latent information channel z , while entangled representations often result in the sharpest reconstructions. Since VAE model selection is often performed based on reconstruction quality, this may be one of the reasons why the ability of VAEs to disentangle data generative factors has been overlooked before. Another reason may be the lack of transform continuity in many popular datasets (i. e. Multi-PIE [14]).

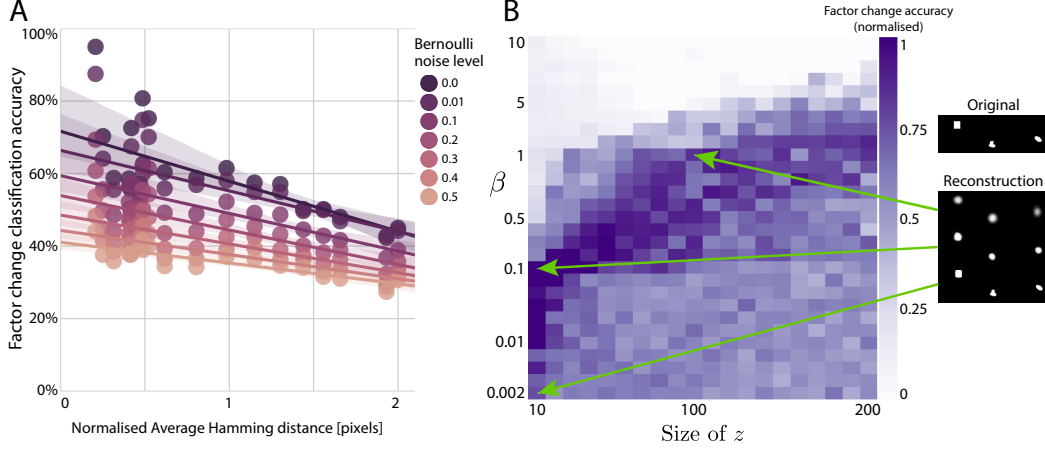


Figure 4: **A:** Negative correlation between data transform continuity and the degree of disentangling achieved by VAEs. Abscissa is the average normalized Hamming distance between each of the two consecutive transforms of each object. Ordinate is factor change classification accuracy from Sec. 3.2. Disentangling performance is robust to Bernoulli noise added to the data at test time, as shown by slowly degrading classification accuracy up to 10% noise level, considering that the 2D objects occupy on average between 2-7% of the image depending on scale. Fluctuations in classification accuracy for similar Hamming distances are due the different nature of subsampled generative factors (i.e. symmetries are present in rotation but are lacking in position). **B:** Positive correlation is present between the size of z and the optimal values of β for disentangled factor learning for a fixed VAE architecture. Note that β values are not uniformly sampled. The average of 25 disentanglement scores is shown. Good reconstructions are associated with entangled representations (lower disentanglement scores). Disentangled representations (high disentanglement scores) often result in blurry reconstructions.

3.4 Investigating qualities of learnt representations

In this section we show some of the desirable properties that arise from learning disentangled as opposed to entangled latent representations.

Learning statistically independent factors Computational neuroscience results suggest that the nature of representations learnt through Hebbian learning in the ventral visual stream in the brain relies on the statistics of the data. Statistically independent parts of the retinal inputs are allocated separate representations, while statistically dependent parts are grouped into a single representation [15]. We test whether the same holds for VAEs trained for disentangled factor learning. We use a dataset developed for psychophysical experiments to measure generative factor learning in humans [7] (unpublished). The dataset consists of a single “amoeba” object with four arms of varying length (Fig. 5A). The arms are pairwise coupled and the length of each arm within each pair is determined by a nonlinear factor (either quadratic or sigmoidal, see Fig. 5B). For example, growth in the values of the quadratic factor correspond to linear growth of arm three and quadratic growth of arm four. This means that during training the VAE sees the full range of lengths of each single arm, but it never sees certain combinations of lengths of pairs of arms (i.e. long arm three and a short arm four). We investigated whether a fully connected VAE (see Tbl 1 for architecture details) would learn representations of the two generative factors (sigmoidal and quadratic), or whether it would learn four separate representations, one for each arm (the latter would be expected if the VAE did not learn the statistical regularities in the data). We found the former to be true (Fig. 5B): the VAE learnt to allocate two latents (z_1 and z_2) to represent the sigmoidal and quadratic factors respectively, z_3 acted as a switch to split the quadratic factor space into two halves, while the remaining latents (z_4 - z_{10}) learnt the uninformative unit Gaussian prior ($p(z) = \mathcal{N}(0, I)$).

Generalisation to new latent factor combinations A model that understands the factorial structure of the data should be able to generalize its knowledge beyond the training distribution by recombining previously learnt factor values, thus performing zero-shot inference (Fig. 1B). We tested

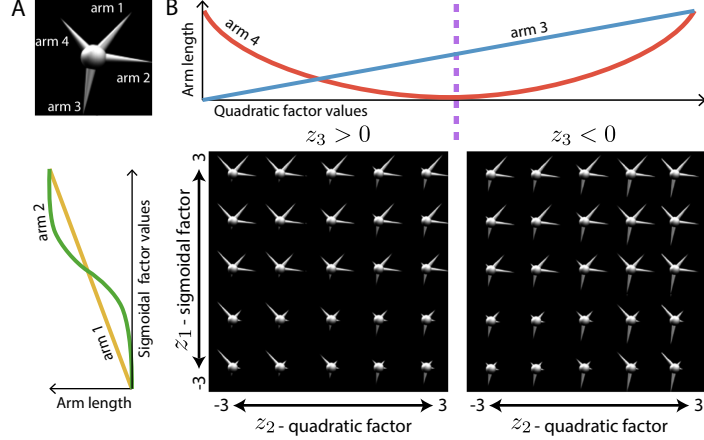


Figure 5: **A:** Amoeba object with four arms of varying length. **B:** Two non-linear generative factors determine the lengths of the pairwise grouped arms. Traversal over three standard deviations around the unit Gaussian prior mean for two latent units (z_1 and z_2) that learnt disentangled representations of the two generative factors. z_1 learnt the sigmoidal factor. z_2 learnt the quadratic factor. z_3 learnt to be a switch that determines which half of the quadratic factor is traversed by z_2 .

such properties of VAEs by training the architecture described in Sec. 3.1 on a subset of the full 2D shapes dataset. This subset preserved the original data continuity by traversing each individual generative factor fully, but some combinations of factors were never seen during training (i.e. the subset still contained all six scales across the three object identities, but there were no small squares present in any rotation or position). By dropping certain combinations of generative factors (Fig. 3B) we reduced the dataset size to approximately 55% of the original size. We then calculated the disentangling metric (Sec. 3.2) for a model with $\beta = 0.01$ or $\beta = 0$ trained on this subset. The disentangling metric was calculated using factor combinations that were excluded from the training subset. We found that the VAE with $\beta = 0.01$ learnt a disentangled representation and was able to reason well about the test data significantly outside of its training distribution (Fig. 3B). The model with $\beta = 0$ learnt an entangled representation and had worse generalization to test data outside of the convex hull of its training data distribution.

Learning basic visual concepts We argue that through learning disentangled representations of the data generative factors, VAEs may acquire basic conceptual understanding of the visual world, such as the “objectness” of things. Then, when presented with novel objects, the VAEs may still be able to reason about the properties of these objects, such as size or position, without necessarily knowing the identity of the new objects. A reinforcement learning framework built on top of such a VAE will then be able to preserve its policy performance without re-learning, hence moving towards the desiderata described in [21]. To test this we presented models that learnt disentangled ($\beta = 0.01$) or entangled ($\beta = 0$) representations on the original dataset of 2D objects (heart, oval and square) with new 2D objects (mushroom, rectangle and triangle) generated using the same four factors of variation (scale, rotation, position X and position Y). In order to visualise what exactly the VAE understands about the new 2D objects, we spliced together an encoder trained on the original dataset ($Enc_{orig} = p(z_{orig}|x_{orig})$) with a decoder trained on the new 2D shapes dataset ($Dec_{new} = p(x_{new}|z_{new})$) (Fig. 6A). We used a low VC dimension linear regressor to learn an alignment mapping $G : z_{orig} \rightarrow z_{new}$ using 50% of the new dataset. We then generated reconstructions $\hat{x}_{new} = Dec_{new}(G(Enc_{orig}(x_{new})))$ of the held out test data. Fig. 6B shows that the VAE that learnt a disentangled representation can reason well about location, scale, and rotation of the new objects despite the fact that its encoder has never seen the new objects. This is in contrast to the poor reconstructions produced by a VAE with an entangled representation.

3.5 Other datasets

Additionally, we trained convolutional VAE architectures on a variety of datasets (including a 3D first person view maze navigation environment that shares many properties with the real world) and

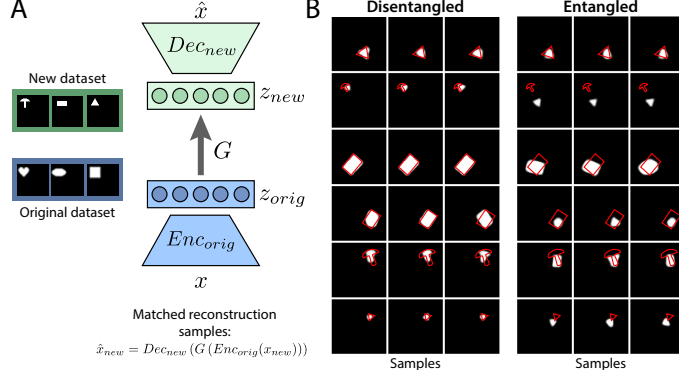


Figure 6: **A**: Model architecture used to visualise whether VAEs trained on the original dataset of 2D objects (*heart, oval and square*) can reason about new object identities (*mushroom, rectangle and triangle*). We splice an encoder trained on the original dataset (Enc_{orig}) with a decoder trained on the new dataset (Dec_{new}) using a linear regressor G , which learns to align the latent spaces z_{orig} and z_{new} . **B**: Samples from $G(z_{orig})$ when running inference through Enc_{orig} using novel 2D objects. Each row corresponds to a different ground truth image x_{new} (red outline). Disentangled VAE reasons well about the location, position and rotation of the novel objects, while slightly confusing object identities; the average normalized Hamming distance between original and reconstructed images over the whole new dataset is 0.42. Entangled VAE struggles to reason about the new objects. Its average normalized Hamming distance is 0.93.

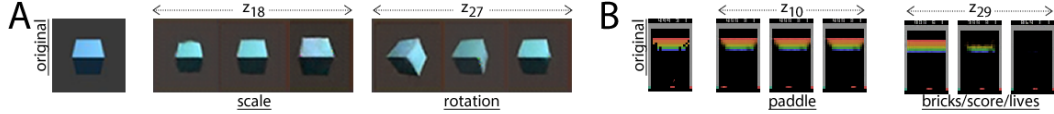


Figure 7: Best seen in animation at <http://tinyurl.com/jgbyzke>. Examples of disentangled factors learnt for different datasets. We run inference on an original image from each dataset, clamp all latent units to the values obtained, then traverse units z_i one at a time. Reconstructions shown are generated by traversing z_i 's with the lowest learnt prior variance for each dataset. **A**: synthetic dataset of 3D shapes with non-affine transformations. **B**: Atari game Breakout.

found them to robustly learn disentangled generative factor representations (see Tbl. 1 for details of VAE architectures and datasets). Some examples of learnt disentangled factors are shown in Fig. 7, however these are best seen in animations at <http://tinyurl.com/jgbyzke>. The examples of learnt factors include non-affine rotation of 3D shapes ($m = 10$, $\beta = 0.01$); the movement of the paddle or changing the score in the Atari game Breakout ($m = 30$, $\beta = 0.01$); the forward/rotational movement in a 3D first person view maze navigation environment ($m = 32$, $\beta = 1$), and the rotation of chairs in a dataset of 3D chairs [1] ($m = 10$, $\beta = 1$). Equivalent architectures that lacked the learning pressures necessary for disentangled factor learning ($\beta = 0$) could not disentangle the latent factors (results in <http://tinyurl.com/jgbyzke>).

4 Conclusion

In this paper we have shown that deep unsupervised generative models are capable of learning disentangled representations of the visual data generative factors if put under similar learning constraints as those present in the ventral visual pathway in the brain: 1) the observed data is generated by underlying factors that are densely sampled from their respective continuous distributions; and 2) the model is encouraged to perform redundancy reduction and to pay attention to statistical independencies in the observed data. The application of such pressures to an unsupervised generative model leads to the familiar VAE formulation [19, 26] with a temperature coefficient β that regulates the strength of such pressures and, as a consequence, the qualitative nature of the representations learnt by the model. Our approach does not depend on any a priori knowledge about the number or the nature of data generative factors, it is robust with respect to different VAE architectures, optimisation

parameters, datasets and noise. We have shown that learning disentangled representations leads to useful emergent properties. The ability of trained VAEs to reason about new unseen objects suggests that they have learnt from raw pixels and in a completely unsupervised manner basic visual concepts, such as the “objectness” property of the world. This is an important ability for the development of artificial intelligence that understands the world the same way humans do [21]. Furthermore, we have demonstrated the ability of VAEs trained for disentangled factor learning to generalise beyond the training data distribution in zero-shot inference scenarios. These are just the first demonstrations of how learning better representations in an unsupervised manner allows models to perform better on challenging machine learning tasks. We believe that using our approach as an unsupervised pre-training stage for supervised or reinforcement learning will produce significant improvements for scenarios such as transfer or fast learning.

References

- [1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.
- [2] H. B. Barlow. *Sensory Communication*, chapter Possible principles underlying the transformation of sensory messages, page 217. M.I.T. Press, Cambridge MA, 1961.
- [3] H. B. Barlow, T. P. Krushal, and G. J. Mitchison. Finding minimal entropy codes. *Neural Computation*, 1:412–423, 1989.
- [4] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. In *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2013.
- [5] J. G. Bremner, A. M. Slater, and S. P. Johnson. Perception of object persistence: The origins of object permanence in infancy. *Child Development Perspectives*, 2015.
- [6] T. R. Candy, J. Wang, and S. Ravikumar. Retinal image quality and postnatal visual experience during infancy. *Optom Vis Sci*, 86(6):556–571, 2009.
- [7] M. Chadwick, A. Banino, and D. Kumaran. Amoeba dataset. (Unpublished).
- [8] B. Cheung, J. A. Levezey, A. K. Bansal, and B. A. Olshausen. Discovering hidden factors of variation in deep networks. In *Proceedings of the International Conference on Learning Representations, Workshop Track*, 2015.
- [9] T. Cohen and M. Welling. Learning the irreducible representations of commutative lie groups. *arXiv*, 2014.
- [10] T. Cohen and M. Welling. Transformation properties of learned visual representations. In *ICLR*, 2015.
- [11] G. Desjardins, A. Courville, and Y. Bengio. Disentangling factors of variation via generative entangling. *arXiv*, 2012.
- [12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.
- [13] R. Goroshin, M. Mathieu, and Y. LeCun. Learning to linearize under uncertainty. *NIPS*, 2015.
- [14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5), 2010.
- [15] I. Higgins and S. M. Stringer. The role of independent motion in object segmentation in the ventral visual stream: Learning to recognise the separate parts of the body. *Vision Research*, 51:553–562, 2011.
- [16] G. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. *International Conference on Artificial Neural Networks*, 2011.
- [17] T. Karaletsos, S. Belongie, and G. Rätsch. Bayesian representation learning with oracle constraints. *ICLR*, 2016.
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.
- [19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014.
- [20] T. Kulkarni, W. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. *NIPS*, 2015.
- [21] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *arXiv*, 2016.
- [22] S. J. Leat, N. K. Yadav, and E. L. Irving. Development of visual acuity and contrast sensitivity in children. *Journal of Optometry*, 2009.
- [23] V. Mnih, K. Kavukcuoglu, D. S. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou,

- H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 2015.
- [24] G. Perry, E. Rolls, and S. M. Stringer. Continuous transformation learning of translation invariant representations. *Experimental Brain Research*, 2010.
- [25] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. *ICML*, 2014.
- [26] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv*, 2014.
- [27] O. Rippel and R. P. Adams. High-dimensional probability estimation with deep density models. *arXiv*, 2013.
- [28] T. Schenk and R. D. McIntosh. Do we have independent visual streams for perception and action? *Cognitive Neuroscience*, 2010.
- [29] J. Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–869, 1992.
- [30] Y. Tang, R. Salakhutdinov, and G. Hinton. Tensor analyzers. In *Proceedings of the 30th International Conference on Machine Learning, 2013, Atlanta, USA*, 2013.
- [31] T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop. Technical report, COURSERA: Neural Networks for Machine Learning, 2012. Technical report, 2012.
- [32] M. A. O. Vasilescu and D. Terzopoulos. Multilinear independent components analysis. 2005, page 547–553, CVPR.
- [33] W. F. Whitney, M. Chang, T. Kulkarni, and J. B. Tenenbaum. Understanding visual concepts with continuation learning. *arXiv*, 2016.
- [34] J. Yang, S. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. *NIPS*, 2015.
- [35] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *Advances in Neural Information Processing Systems 27*. 2014.

A Appendix

A summary of all VAE architectures used in this paper can be seen in Tbl 1. Next we provide various auxiliary details for the different datasets.

Dataset	Optimiser	VAE architecture	
2D shapes	adagrad [12]	encoder	fc 4096-1200-1200-10 (ReLU)
		decoder	fc 10-1200-1200-1200-4096 (tanh)
3D shapes	adam [18]	encoder	conv 32x6x6 (2-1)-64x6x6 (2-1)-512-32 (tanh)
		decoder	deconv 32-512-32x4x4 (2-1)-64x4x4 (2-1)-128x4x4 (2-1)
Amoeba	adagrad [12]	encoder	fc 16384-400-205-10 (ReLU)
		decoder	fc 10-400-8392-16384 (ReLU)
Atari (breakout)	adagrad [12]	encoder	conv 3x48x80-64x6x6 (2)-32x6x6 (2)-32x5x5 (2)-30 (tanh)
		decoder	deconv 30-3840-SU(2)-64x5x5-SU(2)-64x5x5-SU(2)-3x5x5-3x48x80 (tanh)
Atari (other)	adam [18]	encoder	conv 32x6x6 (2)-64x6x6 (2-1)-64x6x6 (2-1)-512-various (ReLU)
		decoder	deconv reverse of encoder (ReLU)
3D chairs [1]	rmsprop [31]	encoder	conv 32x6x6 (2)-64x6x6 (2)-256-10 (ReLU)
		decoder	deconv reverse of encoder (ReLU)
3D game	rmsprop [31]	encoder	conv 3x64x64-32x4x4 (2)-32x5x5 (2)-64x5x5 (2)-64x4x4 (ReLU)
		decoder	deconv reverse of encoder (ReLU)

Table 1: Various VAE architectures and optimisers were used for different experiments to show robustness of our approach. For convolutional architectures the numbers in parenthesis indicate: (stride-padding). SU stands for spatial upsampling.

A.1 2D shapes dataset

We trained the fully connected architecture in Tbl. 1 with cross-entropy cost function using adagrad [12] with learning rate of 1e-2.

A.2 Factor change classification

In order to quantify the degree of disentanglement learnt by the models we generated factor change data according to the pseudocode shown in Algorithm 1. We used a linear classifier to learn the identity of the generative factor that produced z_{diff} . We used a fully connected neural network mapping between input of size z_{diff} to an output of size 4 corresponding to the 4 generative factors (position X, position Y, scale and rotation) with softmax output nonlinearity and cross-entropy cost function. The classifier was trained with adagrad [12] with learning rate of 1e-2 until convergence.

Algorithm 1 Data generation for factor change quantification

```

1: procedure SAMPLEBATCH
2:   for  $n = 1, batch\ size$  do
3:      $objId \leftarrow$  randomly sample object identity
4:      $changeFactor \leftarrow$  randomly sample factor identity
5:      $changeDir \leftarrow$  Randomly sample the direction of change (+/-)
6:     for  $factor \in \{scale, rotation, positionX, positionY\}$  do
7:        $groundTruth_{start}^{factor} \leftarrow$  randomly sample factor value
8:        $groundTruth_{end} \leftarrow groundTruth_{start}$ 
9:        $groundTruth_{end}^{changeFactor} \leftarrow$  randomly sample a new value in the direction of  $changeDir$ 
10:       $x_{start} \leftarrow$  pixel representation of  $groundTruth_{start}$ 
11:       $x_{end} \leftarrow$  pixel representation of  $groundTruth_{end}$ 
12:       $z_{start} \leftarrow Enc(x_{start})$ 
13:       $z_{end} \leftarrow Enc(x_{end})$ 
14:       $z_{diff}^n \leftarrow \frac{|z_{start}^\mu - z_{end}^\mu|}{\max(|z_{start}^\mu - z_{end}^\mu|)}$ 

```

A.3 Zero shot inference regression

In order to map z_{orig} to z_{new} , we used a fully connected linear neural network with smooth L1 loss trained with adagrad [12] with learning rate of 1e-2 until convergence.

A.4 Amoeba dataset

We trained the fully connected architecture in Tbl. 1 with binary cross-entropy criterion and adagrad [12] optimizer with learning rate of 1e-2.

A.5 3D shapes dataset

We trained a convolutional VAE (see Tbl. 1) with learning rate 0.0001 on a dataset of three 3D objects (cylinder, cube and pyramid) with three factors of variation (6 scales, 60 out of plane rotations and 26 colours). The 3D objects were rotating around the z-axis over 2π using 60 equidistant steps. The objects were generated in Blender and the 6 scales and 6x6 position translations were generated for each object in each rotational position using ImageMagick. The full dataset contained 38,880 frames of size 64x64. The decoder had Gaussian outputs.

A.6 Atari dataset

We trained a convolutional VAE (see Tbl. 1) with learning rate 0.0001 on frames from the Atari games Breakout (z size 30, $\beta = 0.01$), SeaQuest (z size 10, $\beta = 5$), Frostbite (z size 100, $\beta = 5$) and Enduro (z size 100, $\beta = 1.75$). The Atari dataset consisted of 1 million frames collected from a trained DQN agent [23]. The frames were pre-processed as described in [23]. The continuity of the dataset enabled the VAE to learn disentangled representations of the independent factors in the data (see video visualisations at <http://tinyurl.com/jgbyzke>). The decoder had Gaussian outputs. The model was trained using adam [18] optimizer with learning rate of 1e-4.

A.7 3D chairs dataset

For the 3D chairs dataset [1] we trained a convolutional VAE (see Tbl. 1) on 82 chair identities. The images were cropped and downsampled to 100x100 pixels. The decoder had Gaussian outputs. The model was trained using rmsprop [31] optimizer with learning rate of 1e-5.

A.8 3D first person view maze navigation game dataset

We also trained a convolutional VAE (see Tbl. 1) on frames from a first person view 3D first person maze navigation game environment. The game frames were made greyscale and downsampled to 84x84 pixels. The dataset contained 1 million frames. This environment shares many properties with the real world: it is continuous and the dynamics of visual scene changes are similar to those experienced in the real world. After training the VAE was able to learn disentangled representations of several factors of variation present in the 3D game world (see video visualisations at <http://tinyurl.com/jgbyzke>). For example, certain single latent units learnt to represent changes in light, forward/backward movement and rotational movement. The VAE also learnt to allocate single latent units to represent the change in score and the rotation of the little character head at the bottom of the screen. For this experiment z size was set to 32 and $\beta = 1$. The decoder had Gaussian outputs. The model was trained using rmsprop [31] optimizer with learning rate of 1e-4.