Important Groups First: Encouraging Disentanglement in Variational Autoencoders

Gil Halevi

Acknowledgements

I would like to thank Ms. Kahini Wadhawan for assisting me in completing this work.

Contents

List of Figures and Tables

Abstract

Machine learning models represent data as vectors for a variety of tasks. Creating vector representations that are *disentangled*—where every dimension in the vector represents one human-interpretable aspect of the data—has been gaining interest recently. Disentangled representations allow these models to be more easily understood and diagnosed. We experiment with a modification to the Variational Autoencoder (VAE), the state-of-the-art model for disentanglement, to encourage it to store information more important in representing an image in certain dimensions. We hope to mitigate two common issues with VAEs: that they are inconsistent in performing disentanglement, and that disentanglement often comes at the cost of quality of recovering images from their representations. Our method proves competitive to existing methods, with similar results in terms of disentanglement consistency and the tradeoff between image quality and disentanglement.

1 Introduction

Representation learning is a crucial task in machine learning, allowing machines to efficiently utilize normally very complex data, such as pictures, by compressing them down into vectors (lists of numbers, where each number is referred to as a dimension). Recently, much research has come out focusing on the learning of *disentangled* representations, particularly for images. A disentangled vector representation is one in which every dimension corresponds to a specific attribute of the image, usually measured as one of the image's generative factors.

Several papers have theorized or explored the benefits of disentanglement. One advan-

tage of disentangled representations is that they are more likely to be human interpretable, as each dimension corresponds to a factor that humans can understand (Chen et al. 2016). Interpretability is important due to the important real-world uses of machine learning, such as disease detection (Rajpurkar et al. 2017), signature verification (Sekhar et al. 2019) and object detection in autonomous vehicles (Lu et al. 2017). There can be serious consequences for the failure of machine learning in these applications, and interpretable models allow for accountability due to easier diagnosis of the model's mistakes. Disentangled representations have also been shown to be less susceptible to adversarial attacks (Willetts et al. 2019). These are attacks where other adversarial machine learning models are trained to fool a primary model with image perturbations that imperceptible to humans (Szegedy et al. 2013). For example, the primary model might be one that classifies images, and the adversarial model would be able to produce misclassified images that look nearly identical, to humans, to real images.

Additionally, disentangled representations may lead to better performance on other machine learning tasks. In particular, it has been theorized that disentangled representations will reduce the number of samples necessary for optimization on other tasks (Bengio et al. 2013), though this has been disputed by evidence found by Locatello et al. (2018). van Steenkiste et al. (2019) find that disentangled representations help in abstract visual reasoning tasks.

2 Review of Literature

Autoencoders are a type of machine learning model first created by Ballard (1987). These models learn to take a form of data, turn it into a vector representation (called a latent vector), and are able to reconstruct the form of data using the vector representation. In this paper, the form of data that will be discussed is an image. Autoencoders work using two networks trained simultaneously: the encoder, which encodes images into latent vectors; and the decoder, which decodes these latent vectors back into the image.

Kingma and Welling (2013) introduce the variational autoencoder(VAE). In this autoencoder, instead of each image being encoded as a single latent vector, each image is encoded as two vectors: one of means and one of standard deviations. To decode this representation, a single vector is generated, with each of its dimensions being picked randomly from a Normal distribution with the mean and standard deviation of the corresponding dimensions. This vector is then decoded by the decoder to reconstruct the image. The loss function of the autoencoder then becomes two-fold: punishing the model for creating inaccurate representations, known as reconstruction loss, and punishing the model for having its distributions vary from the standard mean 0 and standard deviation 1 distribution, known as KL Divergence. The latter loss is implemented to ensure that the autoencoder does not create extremely varied means, and standard deviations that are close to zero, which would effectively turn the VAE into a standard autoencoder.

Higgins et al. (2017) introduce the β-VAE. This is simply a VAE where the KL divergence is upweighted by term β, thereby increasing the restraint on the representations. Increasing β was shown to increase the model's reconstruction loss by limiting the amount of information that can be transmitted through each vector. However, it was also shown to

help the model achieve better disentanglement.

Factor-VAE (Kim and Mnih 2018), β-TCVAE (Chen et al. 2018) and DIP-VAE (Kumar et al. 2017) are three additional variations introduced to the VAE. These methods add onto the β-VAE objective by adding an additional term to the loss function to penalize the total correlation(TC) between the dimensions in the representations. TC is a measure of how much information is shared between different variables, so if each dimension corresponds to an independent factor as desired, this correlation should be minimized. However, TC is impossible to calculate directly. Factor-VAE estimates it using the density ratio trick (Nguyen et al. 2010; Sugiyama et al. 2012), β-TCVAE uses a tractable biased Monte-Carlo estimate, and DIP-VAE matches the moments of a distribution and a factorized prior.

Burgess et al. (2018) analyze the cause of disentanglement in β-VAE. They show that β-VAE learns dimensions that contribute different amounts to the reconstruction loss. That is, the β-VAE begins training with dimensions that store no information about the image. It then learns a few dimensions that will help it reconstruct the image as accurately as possible, lowering the reconstruction loss the most. After learning these dimensions, it then learns the dimensions that will further reduce the reconstruction loss the most, repeating this process until the reconstruction loss is as optimized as possible. This process is similar to Principal Component Analysis, a method for extracting the most important components in a high-dimensional space of vectors. Indeed, research has described the direct relationship between the VAE's loss function and the objective of PCA (Lucas et al. 2018; Rolinek et al. 2019).

Despite being state-of-the-art, there are still large challenges in using VAEs for disen-

tanglement. In particular, Locatello et al. (2018) show that they obtain quite inconsistent disentanglement scores, with even the random seed (i.e. pure randomness) impacting the scores more than differences in regularization constants (i.e. β value) or the type of loss function (i.e. β-VAE vs. β-TC-VAE vs. DIP-VAE). It has also been noted (Higgins et al. 2017) that there is often a tradeoff between reconstruction quality of the image (measured by the reconstruction loss) and disentanglement scores.

3 Statement of Purpose

In this research, we evaluate a modification to the loss function of the VAE to more explicitly encourage this PCA-like behavior. This modification is made in hopes that it will lead to more consistent disentanglement, and that it will mitigate the tradeoff between reconstruction quality and disentanglement scores.

4 Methods

We base our model on the Variational Autoencoder, which uses encoder e with parameters θ to convert image x into latent vector z, and then decoder d with parameters ϕ to reconstruct the image as $\hat{x}$:

$$e_\theta(x) = \{\mu, \sigma\} \tag{1}$$

$$z \sim N(\mu, \sigma) \tag{2}$$

$$d_\phi(z) = \hat{x} \tag{3}$$

z is sampled from the latent dimension by first generating a random value from distribution N(0,1) for each dimension. Each of these values are then multiplied by the corresponding dimension in the σ vector and then added to the corresponding dimension in the μ vector.

The loss function is then two-fold. The KL loss measures the difference between the distribution that z is sampled from and the standard $N(0, 1)$ distribution, using the negative KL divergence. The reconstruction loss measures the similarity between $\hat{x}$ and x by computing the cross entropy of their respective pixels, treating $\hat{x}$ as a bernoulli distribution.

$$L_{rec} = logp_{\phi}(x|z) \tag{4}$$

$$L_{KL} = -D_{KL}(N(\mu, \sigma)||N(0, 1)) \tag{5}$$

$$L_{total} = L_{rec} + \beta L_{KL} \tag{6}$$

where β was introduced as part of the β-VAE to encourage disentanglement. For our new model, known as Grouped-VAE, we treat the latent vector as being made up of G different groups:

$$e_{\theta}(x) = z_1, z_2, \ldots, z_G \tag{7}$$

We expand the reconstruction loss to be made up of G terms. Each term $L_{rec,i}$ of the reconstruction loss measures how well the image can be reconstructed using group i and every group that came before it. We estimate this by concatenating these groups and then padding the concatenated vector with zeros to be the same size as z. This creates a vector

that can then be fed in to the decoder to create $\hat{x}_i$ to be compared to x.

$$d_\phi(z_1, z_2 \dots z_i) = \hat{x}_i \tag{8}$$

$$L_{rec,i} = p_\phi(x|z_1, z_2 \dots z_i) \tag{9}$$

For every reconstruction loss $L_{rec,i}$, we propagate the loss back to the full decoder, but only to the part of the encoder responsible for generating z_i. This ensures that each z_i is optimized solely for generating the image given only the groups before it. The KL divergence remains the same, taking into account the entire latent vector. The full loss is then:

$$L_{rec} = L_{rec,1} + L_{rec,2} + \dots + L_{rec,G} \tag{10}$$

$$L = L_{rec} + \beta L_{KL} \tag{11}$$

The idea behind Grouped-VAE is that we ensure, explicitly, that each group of dimensions reduces the reconstruction loss as much as possible given every group that came before it. This ensures that the model satisfies the PCA-like quality without relying on the model "stumbling" upon this solution while training, which we hypothesize may be where much of the variation in disentanglement score comes in. In addition, because we do not have to increase β, thus lowering the amount of information that can be stored in each dimension, we hope that the reconstruction loss will be decreased as well.

5 Experimentation

We test this model on the dsprites dataset (Higgins et al. 2017), containing black-and-white images of 2D shapes generated using five factors: shape (ellipse, heart or square), rotation, size, x-position and y-position. These are all treated as categorical variables, with separate discrete values of each factor. We use a latent vector of size 6. For grouped-VAE, we split the vector into two groups, both of size 3, and we set β to 1. Following (Locatello et al. 2018), we train our encoder and decoder as feed-forward networks. The encoder uses ReLU activation functions, with two hidden states, both of size 1200. The decoder has three hidden states of size 1200 and uses Tanh activation functions. Our learning rate is 0.0001, and our batch size is 64. We run each configuration 10 times with a random seed each time to be able to assess consistently.

As a disentanglement metric, we use the Mutual Information Gap(MIG) (Chen et al. 2018), which was shown to correlate with most other disentanglement metrics (Locatello et al. 2018). MIG measures the normalized difference between the mutual information(MI) of the two dimensions that share the most MI with a factor of variation. This is then averaged over all factors of variation:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(f_k)}\left(I(d_{j_k},f_k)-\max_{j\neq j_k}I(d_j,f_k)\right) \tag{12}$$

where H is the entropy function, I is the function for mutual information, f_k is a factor of the image, d_j is a dimension of the latent vector and $j_k = \arg max_j I(d_j, f_k)$.

To quantify other qualities of the latent representation, we introduce a measure, the Mutual Information Concentration (MIC), of a factor in a subset S of the latent dimen-

sions. This is defined as the sum of the MIs of a factor with each dimension in the set, subtracted by the sum of the MIs of a factor with each dimension outside the set. This is then normalized:

$$\text{MIC}(f, S) = \frac{1}{H(f)} \left(\sum_{n \in S} I(d_n, f) - \sum_{n \notin S} I(d_n, f) \right) \tag{13}$$

We estimate mutual information by generating representations for the 10,000 images in our testing set, and then bin each dimension into 20 bins. We then use scikit-learn (Pedregosa et al. 2011) to calculate mutual information.

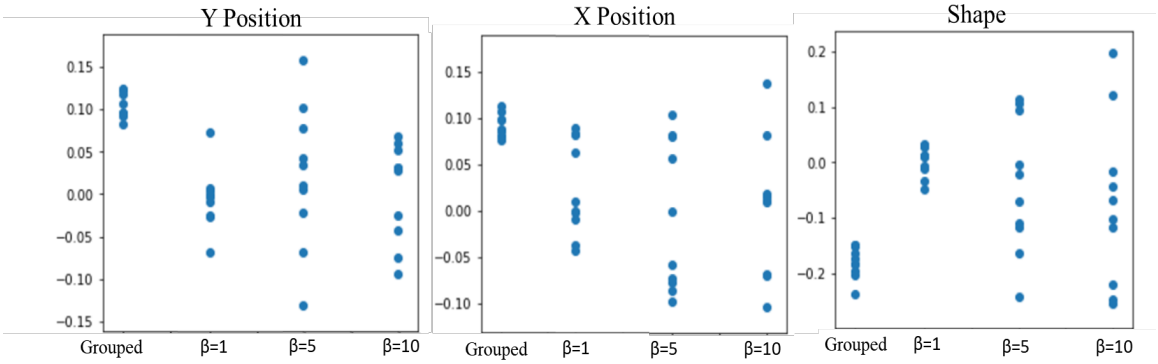

Figure 1: The MIC of three of the factors with the first three dimensions of the latent vector. MIC is displayed on the y axis, with each dot representing one trial. The first column shows the scores of Grouped-VAE, while the next columns display the scores of β-VAE with three β different values.

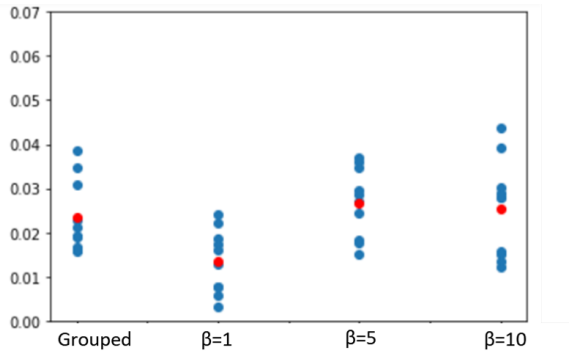

Figure 2: The MIG of the models. MIG is displayed on the y axis, with each blue dot representing one trial. The red dots represent the mean. The first column shows the scores of Grouped-VAE, while the next columns display the scores of β-VAE with three different β values.

	Grouped	β=1	β=5	β=10
Average MIG	0.023	0.014	0.027	0.025
Average Reconstruction Loss	49.1	34.8	41.3	50.9

Table 1: The average MIG and reconstruction loss per image over all iterations. "Grouped" refers to Grouped-VAE, while the β values refer to β-VAE

6 Results

As evidence that our model produces the correct functionality, we display Figure 1. This shows the MIC of select factors in the first three dimensions of the image, which correspond to the first group in the case of the Grouped-VAE. Grouped-VAE stores information about the factors most important for reconstruction, the x and y position, in the first three dimensions of the image. Information about shape, a factor less important for reconstructing the image, is stored more in the last three dimensions. β-VAE shows no consistent bias for which dimensions the different factors are placed in, as expected.

Figure 2 and Table 1 then display the MIG and reconstruction loss of Grouped-VAE in comparison to β-VAE. As seen in Figure 2, the variability of the MIG scores in Grouped-VAE is not meaningfully different from that of β-VAE. Therefore, Grouped-VAE seems to have similar consistency in disentangling as the other β-VAE models. As seen in Table 1, its disentanglement scores are also similar to the other β-VAE models with similar reconstruction loss, so there is no clear benefit in reducing the trade-off between reconstruction loss and disentanglement

7 Conclusion

In this paper, we experiment with a new configuration of VAE model explicitly encouraged to have PCA-like behavior, where dimensions are learned to lower reconstruction loss as much as possible given previous dimensions. We find that it accomplishes this with reconstruction loss and disentanglement scores competitive with existing models. It does not, however, provide clear advantages in terms of reconstruction consistency or in mitigating the tradeoff between reconstruction loss and disentanglement.

Disentanglement may provide the key in many tasks in machine learning moving forward. With increased usage of machine learning in society, disentanglement may provide a flexible way of creating models that are human interpretable, due to the ease of understanding disentangled representations. Disentanglement may also be key in tasks relating to transfer learning, the use of the same information for learning different tasks. Disentangled representations, if they could be achieved, would provide robust ways of storing information that could then be used for a variety of distinct tasks.

With the prevalence of obstacles toward disentanglement, future research should continue to focus on new methods to make unsupervised disentanglement more consistent, and with less of a tradeoff between reconstruction loss and disentanglement. Future research could also focus on more concrete benefits of disentanglement, benefits that could perhaps be focused on in specialized models.

References

Ballard, D. H. (1987). Modular learning in neural networks. In *AAAI*, pp. 279–284.

Bengio, Y., A. Courville, and P. Vincent (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence 35*(8), 1798–1828.

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

Chen, T. Q., X. Li, R. B. Grosse, and D. K. Duvenaud (2018). Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620.

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180.

Higgins, I., L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR 2*(5), 6.

Kim, H. and A. Mnih (2018). Disentangling by factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P. and M. Welling (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumar, A., P. Sattigeri, and A. Balakrishnan (2017). Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

Locatello, F., S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem (2018). Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*.

Lu, J., H. Sibai, E. Fabry, and D. Forsyth (2017). No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*.

Lucas, J., G. Tucker, R. Grosse, and M. Norouzi (2018). Understanding posterior collapse in generative latent variable models. *OpenReview*.

Nguyen, X., M. J. Wainwright, and M. I. Jordan (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory 56*(11), 5847–5861.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Rajpurkar, P., J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Rolinek, M., D. Zietlow, and G. Martius (2019). Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415.

Sekhar, C., P. Mukherjee, D. S. Guru, and V. Pulabaigari (2019). Osvnet: Convolutional

siamese network for writer independent online signature verification. *arXiv preprint arXiv:1904.00240*.

Sugiyama, M., T. Suzuki, and T. Kanamori (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

van Steenkiste, S., F. Locatello, J. Schmidhuber, and O. Bachem (2019). Are disentangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*.

Willetts, M., A. Camuto, S. Roberts, and C. Holmes (2019). Disentangling improves vaes' robustness to adversarial attacks. *arXiv preprint arXiv:1906.00230*.