The tradeoff for β-TCVAE

Deep Learning 2019-2020

Hidde Folkertsma (s2759799) Ivo de Jong (s3174034)
René Flohil (s2548925) Travis Hammond (s2880024)

May 2020

Abstract

β-TCVAEs have been suggested as an improvement on β-VAEs by prioritizing the Total Correlation in the VAEs latent space as a loss components. This paper explores how training a β-TCVAE may make a tradeoff on the other terms in the loss function compared to a β-VAE. The networks were trained on the Stanford Dogs dataset, and showed no clear tradeoffs between β-VAEs and β-TCVAEs, though both showed a tradeoff on reconstruction loss. Since the output of the models was not adequate, this paper does not provide evidence against a tradeoff.

1 Introduction

Variational Autoencoders (VAEs) aim to compress data into a latent space, allowing reconstruction back to the input domain [5]. It is an unsupervised learning algorithm that attempts to extract meaningful features from data by minimizing both the reconstruction loss and the Kullback-Leibler divergence between the latent variables and the unit normal distribution, in an attempt to obtain “meaningful” features. This approach has been a subject of study to improve the representational quality of the latent space.

1.1 The variational autoencoder

Like traditional autoencoders, VAEs encode vectors $\mathbf{x} \in \mathcal{X}$ into a latent representation $\mathbf{z}$, and then decode it with (hopefully) minimal reconstruction loss. More formally, for our model to be representative, every $\mathbf{x}$ should have one or more latent representations $\mathbf{z}$ that decodes into $\hat{\mathbf{x}}$ similar to $\mathbf{x}$, using a function $f(\mathbf{z};\theta)$. f is a deterministic function with random $\mathbf{z}$ and fixed θ, and is therefore a random variable in $\mathcal{X}$. Therefore we can represent it as a Gaussian distribution $p(\mathbf{x}|\mathbf{z};\theta)$. We want to optimize θ such that we can sample $\mathbf{z}$ and the resulting $f(\mathbf{z};\theta)$ will be similar to the $\mathbf{x}$ in our dataset. VAEs avoid describing the latent variables; they simply assume that all $\mathbf{z}$ can be sampled from $\mathcal{N}(0, I)$ and rely on a sufficiently powerful f to transform the uniformly distributed variables. Mathematically the goal is to maximize the following equation:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z})dz \tag{1}$$

We could approximate $p(\mathbf{x})$ by taking many vectors $\mathbf{z}_i$ and averaging $p(\mathbf{x}|\mathbf{z}_i)$, but this would require a very large amount of samples. Moreover, for most of these sampled $\mathbf{z}$, $p(\mathbf{x}|\mathbf{z})$ will be close to zero. Therefore we would like to sample only $\mathbf{z}$ that are likely to produce $\mathbf{x}$, and compute $p(\mathbf{x})$ from them. In order to achieve this we introduce a new (usually Gaussian for continuous data) distribution $q(\mathbf{z}|\mathbf{x})$ which takes a vector $\mathbf{x}$ and produces $\mathbf{z}$ vectors that are likely to produce $\mathbf{x}$. The idea is that the space of likely $\mathbf{z}$ under Q is smaller than the space under $p(\mathbf{z})$, so we can compute $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})]$ more easily.

Using $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})]$, we can infer $p(\mathbf{x})$ as follows. First, we construct a Kullback-Leibler divergence between $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z}|\mathbf{x})$ (note that $p(\mathbf{z}|\mathbf{x})$ is intractable: it describes the vectors $\mathbf{z}$ that are likely to produce $\mathbf{x}$).

$$\mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log q(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})] \tag{2}$$

Rewriting using Bayes' rule:

$$\mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log q(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{x}|\mathbf{z}) - \log p(\mathbf{z})] + \log p(\mathbf{x}) \tag{3}$$

$p(\mathbf{x})$ is not in the expectation because it does not depend on $\mathbf{z}$. We can rewrite Equation 3 as follows:

$$\log p(\mathbf{x}) - \mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})] = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{4}$$

Equation 4 is the basis of the VAE. The left hand side of the equation is what we want to maximize, but it is infeasible to do so. However, we can use gradient descent to maximize the right hand side, effectively circumventing the issue. The right hand side is the loss function of the VAE. $q(\mathbf{z}|\mathbf{x})$ will approximate the intractable $p(\mathbf{z}|\mathbf{x})$, meaning we optimize $p(\mathbf{x})$ directly because the KL term on the left hand side of Equation 4 will approach zero.

1.1.1 Maximizing the objective and the reparametrization trick

In order to maximize the right hand side of Equation 4, we must maximize the first term and minimize the second term. Starting with the second term, recall that $p(\mathbf{z}) = \mathcal{N}(0, I)$ and that $q(\mathbf{z}|\mathbf{x})$ is a Gaussian. This means we only need to learn the parameters of q, for which we use a neural network. The first term is tricky; we can sample $\mathbf{z}$ vector(s) from $q(\mathbf{z}|\mathbf{x})$, and use that as an approximation for $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$, but this leads to a problem when backpropagating because that operation has no gradient. The solution (called the 'reparametrization trick' in [5]) is to move the sampling to an input layer where we first sample $\mathcal{N}(0, I)$ and then use the mean μ and covariance Σ of $q(\mathbf{z}|\mathbf{x})$ to compute $\mathbf{z}$. A more detailed description is given in [5].

1.2 β-VAE and β-TCVAE

Introduced in 2017, the β-VAE [9] is an improvement of the original VAE, which showed to qualitatively outperform the original VAE on CelebA [7], chairs [6], and faces [2]. It does so by introducing one additional hyperparameter β which balances the reconstruction accuracy against the disentanglement of the latent variables. Besides outperforming regular VAEs, it also outperforms other unsupervised techniques such as InfoGAN [8].

While identifying an explanation on why β-VAEs work, an adaptation β-TCVAE was developed [10]. In this adaptation the KL term in the loss function is decomposed into three terms, of which only the *Total Correlation* (TC) term is weighted with β. It was determined that increasing the weight of this factor is what made β-VAEs work in the first place, but since β-TCVAEs increase the weight of that factor exclusively, β-TCVAEs should perform even better.

1.3 Decomposition of the VAE loss function

The first loss component in Equation 4 is the reconstruction loss. This component describes the (log)-likelihood of an image $\mathbf{x}$ given a latent space encoding $\mathbf{z}$ as $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$. This term can be intuitively understood as the "accuracy" of the results: how similar the output images are to the input images. The second term is the KL divergence between the distribution of the latent space and the defined prior for the latent space can be considered as a kind of regularization. Putting these two together defines the loss function of the original VAE:

$$\mathcal{L}(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{5}$$

The β-VAE showed that the disentanglement would be qualitatively better if the KL divergence term gets an increased weight. To do this they introduce a hyperparameter $\beta > 1$ which gives the loss for the β-VAE:

$$\mathcal{L}(\beta, \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \beta\, \mathrm{KL}[q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})] \tag{6}$$

The development of the $\beta-$TCVAE decomposed this KL divergence into a sum of 3 terms. The first term is referred to as the index-code mutual information (MI). This is the mutual information between the input data and the latent data. It is defined as $\mathrm{KL}(q(\mathbf{z}|\mathbf{x})||q(\mathbf{z})p(\mathbf{x}))$. The second term is the Total Correlation (TC), which is found to be the driving term that made β-VAEs work well [10]. It measures the dependence between the latent variables, so a penalty ensures that the latent variables are statistically independent. It can be denoted as $\text{KL}(q(\mathbf{z})||\prod_i q(\mathbf{z}_i))$. The final term is the dimensionwise KL. This is the sum of the KL divergence for each individual latent variable. Adding a penalty for this prevents any latent variable to steer too far away from the prior normal distribution. This is mathematically expressed as $\sum_i \text{KL}(q(\mathbf{z}_i)||p(\mathbf{z_i}))$.

The difference between the β-VAE and the β-TCVAE is simple: β-VAE increases the weight for all 3 of these components, whereas β-TCVAE only increases the weight of the TC component. Using this we can define a single loss function where different constraints on hyperparameters α, β and γ determine the type of VAE

$$\mathcal{L}(\alpha,\beta,\gamma,\mathbf{x},\mathbf{z}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]}_{\text{(reconstruction)}} - \alpha \underbrace{\text{KL}[q(\mathbf{z}|\mathbf{x})||q(\mathbf{z})p(\mathbf{x})]}_{\text{(i) Index-Code MI}} - \beta \underbrace{\text{KL}[q(\mathbf{z})||\prod_i q(\mathbf{z}_i)]}_{\text{(ii) Total Correlation}} - \gamma \underbrace{\sum_i \text{KL}[q(\mathbf{z}_i)||p(\mathbf{z_i})]}_{\text{(iii) Dimensionwise KL}} \quad (7)$$

Here it is required for β-VAE that $\alpha = \beta = \gamma > 1$, whereas for β-TCVAE $\alpha = \gamma = 1$ and $\beta > 1$. Note that setting $\beta = 1$ would result in a regular VAE.

1.4 Aim of this research

The presented research will aim to investigate whether the β-TCVAEs have a worse performance compared to β-VAEs with respect to the MI, the dimensionwise KL and the reconstruction accuracy, in order to get a better performance on the Total Correlation.

It is hypothesized that the loss components for MI, dimensionwise KL and reconstruction will be higher for β-TCVAEs compared to β-VAEs, but that the loss for TC will go down. We make this prediction because the β-TCVAE adds additional weight to the TC term for the loss, which means that the relative weight over the other components will be lower. Therefore, the training will place less focus on these terms, and their loss may therefore be higher.

The hypothesis will be tested by applying a β-VAE and a β-TCVAE to the Stanford Dogs dataset [4] where the losses will be investigated over training time. In addition, we aim to reproduce the results shown in [10]. To make a valid conclusion the VAEs should be applied as a valid procedure. To achieve this the images are cropped to fit the annotated dogs and a deep convolutional architecture will be used to sufficiently encode and decode the images.

2 Methods

Both types of VAE will be tested using the same deep convolutional architecture over 3 values of $\beta \in \{1, 3, 6\}$. Previous research found an appropriate range of β to be $[1, 10]$ [10], but considering computational limitations only limited samples from this range are drawn. With $\beta = 1$ both models are the same, and are also the same as the classical VAE [10], but with $\beta > 1$ the β-TCVAE adds weight to the TC term, whereas the β-VAE adds weight to the full KL term. Both VAEs are trained using Adam [5] for stochastic gradient-based optimization with batches of 64 images. All experiments are run for 50 epochs.

2.1 Data

The Stanford dogs dataset consists of 12000 training images and 8580 testing images of dogs such as in Figure 1. As Figure 1 shows these images contain a lot of irrelevant details, so to overcome that the images were cropped using provided bounding box annotations. Considering the different poses and 120 different breeds this still leaves a large amount of variation in each cropped image.

In order to apply the images to the VAEs each image was stretched/squished into 64x64 RGB pixels, which results in an input shape of 3x64x64 for the VAEs. The 8-bit colour values were divided by 255 to normalize the images.

Figure 1: Sample Rhodesian Ridgeback from Stanford Dogs

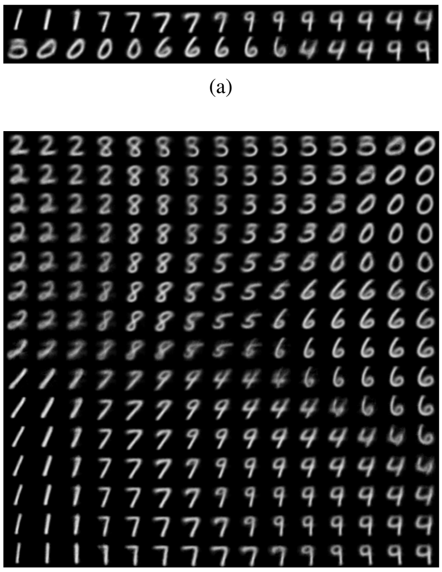
2.2 Architecture

In order to create an architecture that is able to process the complex image space of the Stanford Dogs dataset we looked at research that has been done previously. In 2018 researchers were able to get an accuracy of 8.18% on a validation set of 1015 images using a combination of convolutional layers, rectified linear unit (ReLU) activations, and max pooling layers followed by dense and dropout layers [11]. Other architectures use a combination of either convolutional layers with differing kernel sizes or dense layers. The disentanglement library, a VAE library created by the team at Google Brain Zurich [12], proposes such systems for disentanglement challenges.

Variations on these architectures were tested on the MNIST dataset [3] using a β-VAE with $\beta = 1$ and a latent dimension size of 2 checking for proper reconstruction and disentanglement.

Unfortunately, the architecture described in [11] was not fruitful. The pooling layers in the encoder and subsequent upsampling layers in the symmetric decoder were hindering proper reconstruction with the used image size. A combination of convolutional layers and dense layers produced better results. Tests on architectures of up to 5 dense layers only achieved in producing blurry circles with no disentanglement for the Stanford Dogs dataset and produced no disentanglement for MNIST. This prompted us to test an architecture using only convolutional layers which decreased the reconstruction loss. Additionally, increasing the image size produced better reconstruction results.

The final architecture was determined to be as followed: the encoder consists of solely three convolutional 2D layers with kernel size 3x3, strides of 2 and filters doubling the image dimension every consecutive layer before being flattened. The only dense layers present are the two that split the flattened convolved image into the mean and log variance. The decoder should be symmetric to the encoder and thus consists of 1 dense layer shaping the latent dimensions into the flattened convolved image, and three transposed convolutional layers with kernel size 4x4, strides of 2 and filters being halved each step. All layers used ReLU activations when applicable and all convolutional layers used 'same' padding. The result of a 50 epoch run on MNIST is shown in Figure 2.



(a)

(b)

Figure 2: Reconstruction result of the proposed architecture for MNIST. Each row in (a) is the result of altering the value of one of the latent dimensions. The grid in (b) is the result of combining varied values of the two latent dimensions.

2.3 Evaluation

From the experiment described above we gather how the loss components move over time for each value of β. From this the epoch with the lowest validation loss will be extracted, as that is the most representative of a final result from training a model. In this epoch we can investigate the values of the loss components to see whether the β-TCVAE does indeed make a trade-off in order to improve on the TC term.

3 Results

Table 1 shows the loss and their components for the experiment as described above. For each situation consisting of the VAE type and the β value the epoch with the lowest validation loss is presented.

It is crucial to be aware that the total loss is based on the scaled individual components. Therefore this is determined by the experiment setting, regardless of the behaviour of the training. Specifically, for β-TCVAE the total loss becomes incredibly low. This is because it is in part determined by the TC term (about -70) multiplied by β. The total loss therefore does not give any indication of performance compared to other settings of β.

An interesting finding is that the reconstruction loss goes up with β. This might be caused by the gradient descent ”focussing” more on the scaled term, therefore failing to minimize the reconstruction loss. This effect is visible in both VAEs, but is more prominent in the β-VAE. This increase in reconstruction loss is a clear balance against the KL loss, which goes down correspondingly as it, or its components are considered with more weight.

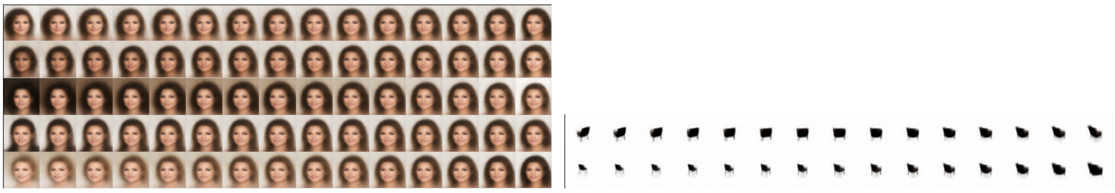
When looking at the decomposed KL loss it shows that the MI loss really is the only term that is strongly affected by changing β. This is particularly interesting when considering β-TCVAEs, where the weight is put on the TC term, but the increased performance occurs in the MI term. This may be explained by the fact that they are closely related, as TC is a generalization of MI [1].

β-VAE	Epoch	Total loss	Reconstruction loss	KL loss	(i) MI loss (10^{-3})	(ii) TC loss	(iii) DW-KL loss
$\alpha = \beta = \gamma = 1$	17	138.506	115.235	23.259	−0.104	−60.993	84.264
$\alpha = \beta = \gamma = 3$	35	170.578	135.616	11.628	−25.150	−71.965	83.644
$\alpha = \beta = \gamma = 6$	48	196.710	160.148	6.049	−406.400	−76.890	−83.390
β-TCVAE							
$\alpha = \gamma = 1, \beta = 1$	25	139.024	114.980	24.057	−0.039	−60.176	84.220
$\alpha = \gamma = 1, \beta = 3$	41	3.317	133.136	12.514	−16.280	−71.175	83.721
$\alpha = \gamma = 1, \beta = 6$	44	−218.615	154.167	7.275	−202.000	−76.010	83.481

Table 1: Validation loss and their components for each best epoch for the β-VAE and β-TCVAE across β values on the Stanford Dogs dataset. Be aware that the total loss is the weighted loss as described by the parameters, whereas the other columns are the unweighted values.

3.1 CelebA and chairs reconstruction

In an effort to reproduce the results shown in [10], we test our implementation on the same datasets, CelebA [7] and the 3D chairs dataset [6]. Results are shown in Figures 3a and 3b. The results shown are the best qualitative results, both in terms of reconstruction quality and in terms of the quality and amount of meaningful latent variables.



(a) orientation, skin color, background color, hair length, baldness

(b) azimuth, size

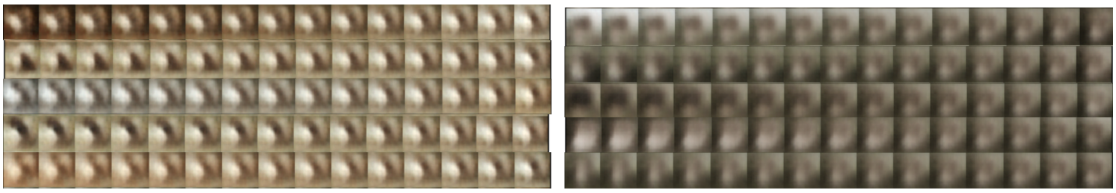
Figure 3: Reconstruction of the CelebA and chairs dataset. Left: CelebA, β-TCVAE, $\beta = 6$. Right: chairs, β-VAE, $\beta = 3$.

We successfully find latent variables semantically similar to the ones shown in the paper (baldness for CelebA and azimuth and size for the chairs), as well as some other interesting latent variables for CelebA. This confirms both the reproducibility of the paper and that our model works for simpler datasets.

3.2 Stanford Dogs reconstruction

Using the trained VAEs we are able to reconstruct the dog images. Figure 4 shows the reconstruction performed by the β-TCVAE with $\beta = 1$ and $\beta = 6$ for the VAE with the lowest reconstruction loss and total loss respectively.

The reconstruction is not clear enough to distinguish a dog in the image. There are no clear meaningful disentangled latent variables, unlike MNIST (where we see transformations from one number to another) or CelebA and chairs (where we see features such as orientation shift as the value of the latent dimension is altered). The disentanglement rather seems to be in more general image features such as



(a) $\beta = 1$

(b) $\beta = 6$

Figure 4: Reconstruction of the Stanford Dogs dataset. Both images are reconstructed by the β-TCVAE.

the shape, brightness and hue of the image. Due to the complex nature of the dataset these may be the only features that the VAEs were able to disentangle.

Going from top to bottom in 4a we interpret the following features: amount of brown hue, size of the white blur, grayness, size of the black blur, amount of orange hue. In 4b we interpret the following features: white background, green hue/size of white blur, size of black ’head’, size of white blur (other shape), size of “dog’s body”.

4 Discussion

Considering the gathered losses it seems that the β-TCVAE does not make a ”trade-off” concerning the terms that are not TC. The results seem to show that only MI is getting a worse performance for β-TCVAE compared to β-VAE. However, it does still improve as β increases. Therefore it does not seem to be in a tradeoff against TC.

Conflictingly, the TC term does not get better for β-TCVAE compared to β-VAE. This seems to go against the design intention of the β-TCVAE.

While the losses give an clear suggestion that there is no tradeoff made, there is some questions to be put on the validity of the output of the VAEs. This might affect the validity of the loss behaviour that was found. The images shown in subsection 3.2 show that the best performing VAEs in either reconstruction or total loss were not able to clearly reconstruct the images of a dog. Instead disentanglement of features happened for general image features such as brightness, hue, location and size of black or white blurs. In contrast the same architecture works properly on the MNIST (shown in Figure 2), CelebA and chairs datasets. The difference between the two latter datasets and the Stanford Dogs dataset is that in the latter all pictures are taken from one perspective, reducing the variance in them greatly. The images are all taken from the front or from slight angles and the background is uniform. This might help in both reconstruction and disentanglement of features. In contrast, the images in the Stanford dogs dataset are all taken from different perspectives, the dogs are in different poses and locations and, most of all, the dog breeds vary widely. Whereas VAEs are able to encode a type of chair or the feature of gender in one latent dimension, the variety in the dogs dataset is too large to properly process.

To improve the performance of the VAEs several measures can be taken. First, vastly decrease the amount of dog breeds, as 120 classes is too much. Second, choose images containing dogs with similar orientations, these are more similar and thus easier to encode and reconstruct. Third, choose images with similar background. This shifts feature extraction to focus on the subject. Naturally, alternative architectures can be tried to achieve better results.

From our results there is no strong evidence of the β-TCVAE performing better than the β-VAE on the Stanford Dogs dataset. However, this lack of evidence may be attributable to the fact that reconstructing dogs from the Stanford Dogs dataset is not an appropriate task for the VAEs. Due to this critique, the present results should not be considered as discrediting β-TCVAEs, but it may serve as a reiteration that blindly applying the most recent deep learning technique is not necessarily going to give the best results on a specific dataset.

References

[1] S. Watanabe. “Information theoretical analysis of multivariate correlation”. In: *IBM Journal of research and development* 4.1 (1960), pp. 66–82.

[2] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. “A 3D face model for pose and illumination invariant face recognition”. In: *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee. 2009, pp. 296–301.

[3] Y. LeCun and C. Cortes. “MNIST handwritten digit database”. In: (2010). URL: `http://yann.lecun.com/exdb/mnist/`.

[4] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. “Novel Dataset for Fine-Grained Image Categorization”. In: *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO, June 2011.

[5] D. P. Kingma and M. Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[6] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. “Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 3762–3769.

[7] Z. Liu, P. Luo, X. Wang, and X. Tang. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. December 2015.

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

[9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. “beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.” In: *Iclr* 2.5 (2017), p. 6.

[10] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. “Isolating Sources of Disentanglement in Variational Autoencoders”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 2610–2620. URL: `http://papers.nips.cc/paper/7527-isolating-sources-of-disentanglement-in-variational-autoencoders.pdf`.

[11] C. Mayson. *Dog Breed Image Classification*. December 2018. URL: `https://medium.com/@claymason313/dog-breed-image-classification-1ef7dc1b1967`.

[12] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations”. In: *International Conference on Machine Learning*. 2019, pp. 4114–4124.