Coupled VAE: Improved Accuracy and Robustness of a Variational Autoencoder

Shichen Cao   SHICA@BU.EDU
Department of Mechanical Engineering
Jingjing Li   JLI0203@BU.EDU
Department of Mathematics & Statistics
Kenric Nelson   KENRIC.NELSON@GMAIL.COM
Department of Electrical & Computer Engineering
Mark Kon   MKON@BU.EDU
Department of Mathematics & Statistics
Boston University
Boston, MA 02215, USA

Abstract

We present a coupled Variational Auto-Encoder (VAE) method that improves the accuracy and robustness of the probabilistic inferences on represented data. The new method models the dependency between input feature vectors (images) and weighs the outliers with a higher penalty by generalizing the original loss function to the coupled entropy function, using the principles of nonlinear statistical coupling. We evaluate the performance of the coupled VAE model using the MNIST dataset. Compared with the traditional VAE algorithm, the output images generated by the coupled VAE method are clearer and less blurry. The visualization of the input images embedded in 2D latent variable space provides a deeper insight into the structure of a new model with coupled loss function: the latent variable has a smaller deviation, and a more compact latent space generates the output values. We analyze the histogram of the likelihoods of the input images using the generalized mean, which measures the model's accuracy as a function of the relative risk. The neutral accuracy, which is the geometric mean and is consistent with a measure of the Shannon cross-entropy, is improved. The robust accuracy, measured by the -2/3 generalized mean, is also improved.

1. Introduction

A challenge for machine learning is the development of methodologies that assure the accuracy and robustness of inferences given limited training samples. The variational autoencoder contributes to this goal by learning a statistical model of the data which is optimized with a cost function based on the cross-entropy of the inference and the divergence from a simple model such as the normal distribution. In this paper, we show that accuracy and robustness can be improved by utilizing a generalization of the cross-entropy and divergence. This generalization is referred to as the coupled entropy because it models a long-range

correlation between the states of distributions, thereby providing a method to modify the cost of outliers. Whereas the entropy measures the average uncertainty of distribution with equal weighting of each state, the coupled-entropy adds/subtracts additional weight to the tails of the distribution for positive/negative coupling, respectively. The use of positive coupling for the cross-entropy and divergence costs of the variational autoencoder enables the learning of a robust inference model.

Our study builds from the work of Kingma and Welling (2014) on variational autoencoders and Tran et al. (2017) on deep probabilistic programming. Variational autoencoders use unsupervised learning method to train encoder and decoder neural networks. Between the encoder and decoder, the parameters of a multidimensional distribution are learned to form a compressed latent representation of the training data (Bowman et al., 2015). It is an effective method for generating complex datasets such as images and speech. VAE can also be used in forecasting from static images as well as in facial expression editing. Zalger (2017) implemented the application of VAE for aircraft turbomachinery design and Xu et al. (2018) used VAEs to achieve unsupervised anomaly detection for seasonal KPIs (key performance indicators) in web applications. Autoencoders can use a variety of latent variable models, but restricting the models can enhance performance. Sparse autoencoders add a penalty for the number of active hidden layer nodes used in the model. Variational autoencoders further restrict the model to a probability distribution $q_\phi(\mathbf{z}|\mathbf{x})$ specified by a set of encoder parameters ϕ learned via variational inference. As stated by Blei et al. (2016), the goal of variational inference is to approximate a conditional density of latent variables given observed variables. The decoder learns a set of parameters θ for a generative distribution $q_\theta(\mathbf{x}'|\mathbf{z})$, where $\mathbf{x}$ is the input training data; $\mathbf{z}$ is the latent variable; $\mathbf{x}'$ is the output generated data. The loss function is determined by a variational bound on the likelihood, which consists of two terms, the expected log-likelihood of the generated data (cross-entropy) and the divergence between the learned model and a prior distribution of model, which will be defined in the next section.

In this study, we draw upon the principles of Nonlinear Statistical Coupling (NSC) (Nelson and Umarov, 2010; Nelson et al., 2017) to analyze and improve the accuracy and robustness of a variational autoencoder (Nelson, 2020). NSC is an interpretation of non-extensive statistical mechanics (Tsallis, 2009), which focuses on the role of nonlinear coupling κ in generalizing entropy and its related functions. The approach defines a family of heavy-tail (positive coupling) and compact-support (negative coupling) distributions which maximize the generalized entropy function. Using the MNIST dataset of handwritten numerals, we show that the measure of robustness and accuracy based on generalized information theory is improved by incorporating the coupled entropy into the loss function of the variational autoencoder.

The next two sections provide a description of the variational autoencoder and the MNIST dataset used for our evaluation. Section 4 introduces the generalized metrics which are use to measure the robustness and accuracy. In section 5, the improved autoencoder is evaluated using the MNIST handwritten numeral test set. Section 6 discusses the results using a simplified 2-dimensional latent variable. Section 7 contains the conclusions.
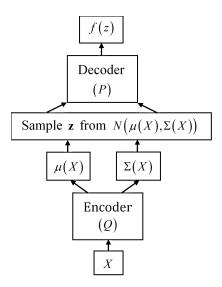
2. The Variational Autoencoder

Figure 1: The variational autoencoder consists of an encoder, a probability model and a decoder.

A variational autoencoder consists of an encoder, a decoder, and a loss function. The encoder is a neural network that converts high-dimensional information from the input data into a low-dimensional hidden, latent representation $\mathbf{z}$. Information lost in the compression, which is a necessary selection of good models for the representation. While in general autoencoders can learn a variety of representations, VAEs especially learn the parameters of a probability distribution. The model used here learns the means and standard deviations θ of a multivariate Gaussian distribution and stores this information in a two-layer space. Figure 1 represents the basic structure of an autoencoder.

2.1 VAE loss function

The decoder, which forms a complementary process to that of the encoder, decompresses and reconstructs the information from the low-dimensional hidden representation back to the parameters of the output data probability distribution. The output also includes the weights and biases ϕ. The distribution of $\mathbf{x}$ is either a Bernoulli or Gaussian. The decoder reads the data from the latent representation $\mathbf{z}$ and outputs specific distribution parameters to generate a new reconstruction $\mathbf{x}'$. The objective is to minimize the loss of information in the reconstruction, which is measured by the log-likelihood $log p_\phi (\mathbf{x}|\mathbf{z})$ of input data given the model and decoder parameters. The loss function of the variational autoencoder is set to map the loss onto some real numbers intuitively representing the loss of information during the encoding and decoding processes. The training process is to minimize the loss functions. For dataset $\mathbf{X} = \left\{ \mathbf{x}^{(i)} \right\}_{i=1}^{N}$ consisting of N independent and identically distributed samples, the loss function for the ith data point or image $\mathbf{x}^{(i)}$ in the original VAE algorithm (Kingma and Welling, 2014) is

$$L\left(\mathbf{x}^{(i)}\right) = -D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}^{(i)}\right) \parallel p\left(\mathbf{z}\right)\right) + \mathbb{E}_{q\left(\mathbf{z}|\mathbf{x}^{(i)}\right)}\left[\log p\left(\mathbf{x}^{(i)}|\mathbf{z}\right)\right]. \tag{1}$$

The first right-hand side is the negative Kullback-Leibler divergence between the variational approximation q and the intractable posterior p, and the second right-hand side is called the expected reconstruction error, which is referred to as the cross-entropy. The prior distribution of each dimension of $\mathbf{z}$ follows a standard Gaussian distribution. The posterior distribution of $\mathbf{z}$ given $\mathbf{x}^{(i)}$ ($\mathbf{z}|\mathbf{x}^{(i)}$) follows a Gaussian distribution with mean vector $\mu^{(i)}$ and covariance matrix $diag(\sigma_1^2, \cdots, \sigma_d^2)^{(i)}$. Meanwhile, all dimensions of $\mathbf{z}$ and $\mathbf{z}|\mathbf{x}^{(i)}$ are mutually independent. Let J be the dimensionality of $\mathbf{z}$; then the Kullback-Leibler divergence

simplifies to

$$
\begin{aligned}
& - D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}^{(i)}\right)||p\left(\mathbf{z}\right)\right) \\
& = \int_{-\infty}^{\infty} q\left(\mathbf{z}|\mathbf{x}^{(i)}\right)\left(\log p\left(\mathbf{z}\right) - \log q\left(\mathbf{z}|\mathbf{x}^{(i)}\right)\right)dz \\
& = \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\left((\sigma_j)^2\right) - (\mu_j)^2 - (\sigma_j)^2\right)
\end{aligned}
\tag{2}
$$

The expected reconstruction error (cross-entropy) $E_{q(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p\left(\mathbf{x}^{(i)}|\mathbf{z}\right)\right]$ can be estimated by sampling, that is,

$$
\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p\left(\mathbf{x}^{(i)}|\mathbf{z}\right)\right] = \frac{1}{L}\sum_{l=1}^{L}\left(\log p\left(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}\right)\right)
\tag{3}
$$

And if data $\mathbf{x}$ given $\mathbf{z}$ follows a multivariate Bernoulli with dimension D,

$$
\log p\left(\mathbf{x}|\mathbf{z}\right) = \sum_{i=1}^{D}(x_i \log y_i + (1 - x_i)\log(1 - y_i))
\tag{4}
$$

Therefore, the regular loss function can be calculated by,

$$
L\left(\mathbf{x}^{(i)}\right) = -D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}^{(i)}\right) \| p\left(\mathbf{z}\right)\right) + \frac{1}{L}\sum_{l=1}^{L}\left(\log p\left(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}\right)\right)
\tag{5}
$$

For this research, the loss function is modified to improve the robustness of the variational autoencoder, which will be discussed in section 4.

2.2 Comparison with other generative machine learning methods

The paradigm generative adversarial networks form a recent advance in generative machine learning methods. The basic idea of GANs was published in a 2010 blog post by Niemitalo (2010). The name 'GAN' was introduced by Goodfellow et al. (2014). Compared with variational autoencoders, generative adversarial networks are used for optimizing generative tasks specifically. Though GANs can set models with a true latent space, as is the case with BiGAN and ALI (Donahue et al., 2017; Dumoulin et al., 2017), which are designed to improve the performance of GANs, GANs cannot generate a reasonable result when the data is high-dimensional. By contrast, as a probabilistic model, the specific goal of a variational autoencoder is to marginalize out non-informative variables during the training process. The ability to set complex priors enables prior expert knowledge to be incorporated. Due to the characteristic of latency in generative machine learning methods, combining the latent representation with many existing models is now an improved method for sequence modeling. Bayesian networks form another generative model. Judea Pearl proposed the Bayesian network paradigm in 1985. Bayesian networks have a strong ability to capture the symbolic figures of input information (Pearl, 1985) and combine the objective probabilities with subjective estimates for both qualitative and quantitate modeling. The whole concept

of Bayesian networks is built on Bayes' theorem. Due to the non-restriction between distribution families and variables, as well as the properties of neural networks, Deep Bayesian networks are now used to compute the complex data. Furthermore, another effective way to solve the posteriority of the distribution derived from neural networks is to train and predict by variational inference techniques (Goodfellow et al., 2016). Compared to the original Bayesian network, the basic building blocks of deep networks provides multiple loss functions to make multi-target prediction, transfer learning, and various outputs depending on different situations. The improvement of the deeper architectures, VAE, specifically, keeps growing.

Other generative models are now commonly combined with a variational autoencoder to improve performance. Ebbers et al. (2017) developed a VAE with a Hidden Markov Model (HMM) as the latent model for discovering acoustic units. Dilokthanakul et al. (2016) studied the use of Gaussian mixture models as the prior distribution of the VAE to perform unsupervised clustering through deep generative models. He showed a heuristic algorithm called "minimum information constraint," and it is capable of improving the unsupervised clustering performance with his model. Srivastava and Sutton (2017) presented the effective autoencoding variational Bayes based inference method for latent Dirichlet allocation (LDA). This model solves the problems caused by autoencoding variational Bayes by the Dirichlet prior and by component collapsing. Also, this model matches traditional methods inaccuracy with much better inference time.

3. Use of MNIST database for evaluation

The MNIST, handwritten digit database is a large database of handwritten digits consisting of a training set of 60,000 images and a test set of 10,000 images widely used for evaluating machine learning and pattern recognition methods. The digits have been size-normalized and centered in a fixed-size image. Each image in the database contains 28 by 28 grey-scale pixels. Pixel values vary from 0 to 255. Zero means the pixel is white, or background, while 255 means the pixel is black, or foreground (LeCun et al., 1998).

For this research, we used the MNIST database as the input. Specifically, input x is a batch of the 28 by 28-pixel photo of a handwritten number. The encoder encodes the data, which is 784-dimensional for each image in a batch into the latent layer space z. For our experiment, the dimension of space z can be chosen from 2 to 20. Taking the latent layers z as the input, the probability distribution of each pixel is computed using a Bernoulli or Gaussian distribution by the decoder. The decoder outputs corresponding 784 parameters and decodes the remodeled value to generate the images at the last step. We used specific numbers of images from the training set as the batch size and fixed epochs for the most modeling process. Additionally, in the learned MNIST manifold, visualizations of learned data and reproduced results can be plotted in the research.

4. Accounting For Risk with Coupled-entropy

Machine learning algorithms, including the VAE, have achieved efficient learning and inference for many image processing applications. Nevertheless, assuring accurate forecasts of the uncertainty is still a challenge. Problems such as outliers and overfitting impact the robustness of scientific prediction and engineering systems. This paper concentrates on assessing and improving the robustness of the VAE algorithm.

4.1 Assessing probabilistic forecasts with the generalized mean

First, proper metrics are needed to evaluate the accuracy and robustness of machine learning algorithms, such as VAE. The arithmetic mean and the standard deviations are widely used to measure central tendency and fluctuation, respectively, of a random variable. Nevertheless, a random variable formed by the ratio of two independent random variables has a central tendency determined by the geometric mean, as described by McAlister (1879). Thus, probabilities which are formed as ratios need the geometric mean to measure the central tendency.
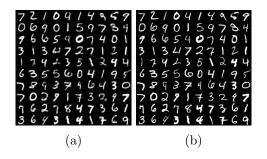

(a)   (b)

Figure 2: Example set of (a) MNIST input images and (b) VAE generated output images.

A Risk Profile, which is the spectrum of the generalized means of probabilities, was introduced to evaluate the central tendency and fluctuations of probabilistic inferences [5]. The generalized mean $(\frac{1}{N} \sum_{i=1}^{N} p_i^r)^{\frac{1}{N}}$ is a translation of generalized information-theoretic metrics back to the probability domain, and is derived in the next section. It's use as a metric for evaluating and training inference algorithms is related to the Wasserstein distance (Frogner et al., 2015), which incorporates the generalized mean. The accuracy of the likelihoods is measured with robust, neutral, and decisive risk bias using the $r = -\frac{2}{3}$, $r = 0$ (geometric), and $r = 1$ (arithmetic) means, respectively. For simplicity, we refer to these three metrics as the robustness, accuracy, and decisiveness. The label "accuracy" is used for the neutral accuracy, since "neutralness" is not appropriate and "neutral" does not express that this metric is the central tendency of the accuracy. Summarizing:

$$Decisiveness \ (\text{Arithmetic Mean}) : \frac{1}{N} \sum_{i=1}^{N} p_i \tag{6}$$

$$Accuracy \ (\text{Geometric Mean}) : \prod_{i=1}^{N} p_i^{\frac{1}{N}} \tag{7}$$

$$Robustness \ (-2/3 \, \text{Mean}) : \left(\frac{1}{N} \sum_{i=1}^{N} p_i^{-\frac{2}{3}} \right)^{-\frac{3}{2}} \tag{8}$$
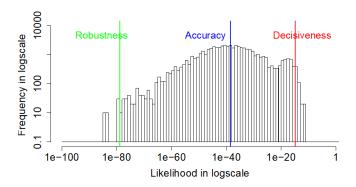
6

Figure 3: The likelihood for the input images under the VAE model. The extremely small value of -2/3 mean metric indicates the poor robustness of the VAE model, which can be improved.

And similar to the standard deviation, the arithmetic mean and -2/3 mean play roles as measures of the fluctuation. We will use these metrics to assess the probability inferences and a related generalization of the loss functions will be used to improve the robustness of the training. Figure 3 shows an example of input images from the MNIST dataset and the generated output images produced by the VAE. Despite the blur in some output images, the VAE succeeds in generating very similar images with the input. However, the histogram in Figure 4 describing the likelihood for the input data x under the trained VAE shows that the probabilities of ground truth range over a large scale. The arithmetic mean or Decisiveness is 10^{-15} . The geometric mean or Accuracy, is 10^{-37} . The -2/3 mean or Robustness is 10^{-77} . The neutral accuracy is near the mode of the histogram. The minimal value of -2/3 mean metric indicates the poor robustness of the VAE model, which can be improved.

The metrics derive from a translation of a generalized entropy function back to the probability domain. Use of the geometric mean for accuracy derives from the Boltzmann-Gibbs-Shannon entropy, which measures the average uncertainty of a system and is equal to the arithmetic average of the negative logarithm of the probability distribution,

$$H\left(\mathbf{P}\right) \equiv -\sum_{i=1}^{N} p_i \ln p_i = -\ln\left(\prod_{i=1}^{N} p_i^{p_i}\right) \tag{9}$$

Translating the entropy back to the probability domain via the inverse of the negative logarithm, which is the exponential of the negative, results in the weighted geometric mean of the probabilities

$$P_{avg} \equiv \exp\left(-H\left(\mathbf{P}\right)\right) = \exp\left(\ln\left(\prod_{i=1}^{N} p_i^{p_i}\right)\right) = \prod_{i=1}^{N} p_i^{p_i} \tag{10}$$

The role of this function in defining the central tendency of the y-axis of a density is illustrated with the Gaussian distribution. Utilizing the continuous definition of entropy for a density $f\left(x\right)$ for a random variable x, the neutral accuracy or central tendency of the

7

density is

$$f_{avg} \equiv \exp\left(-H\left(f\left(x\right)\right)\right) = \exp\left(\int_X f\left(x\right)\ln f\left(x\right)dx\right) \tag{11}$$

For the Gaussian, the average density is equal to the density at the mean plus the standard deviation $f\left(\mu \pm \sigma\right)$.

The use of the geometric mean as a metric for the neutral accuracy in the previous section is related to the cross-entropy between the reported probability of the algorithm and the probability distribution of the test set. The cross-entropy between a "quoted" probability distribution $\mathbf{q}$ and the distribution of the test set $\mathbf{p}$ is

$$H\left(\mathbf{p}, \mathbf{q}\right) \equiv -\sum_i p_i \ln q_i \tag{12}$$

In evaluating an algorithm, the actual distribution is defined by the test samples, which for equally-probable independent samples each have a probability of $p_i = \frac{1}{N}$. Translated to the probability domain, the cross-entropy becomes the geometric mean of the reported probabilities (7), thus, showing that use of the geometric mean of the probabilities as a measure of accuracy for reported probabilities is equivalent to the use of cross-entropy as metric of forecasting performance.

Likewise, the use of the generalized mean as a metric from robustness and decisiveness derives from a generalization of the cross-entropy. While there are a variety of proposed generalizations to information theory, in the Renyi and Tsallis entropies were both shown to translate to generalized mean upon transformation to the probability domain (Nelson et al., 2017) . Here we show that the derivation of this transformation uses the coupled Entropy, which derives from the Tsallis entropy, but utilizes a modified normalization. The nonlinear statistical coupling (or simply the coupling), has been shown to a) quantify the relative variance of a superstatistics model in which the variance of exponential distribution fluctuates according to a gamma distribution, and b) be equal to the inverse of the degree of freedom of the Student's t distribution. The coupling is related to the risk bias by the expression $r = \frac{-2\kappa}{1+\kappa}$, where the numeral 2 is associated with the power 2 of the Student's t distribution, and the ratio $r = \frac{-2\kappa}{1+\kappa}$ is associated with a duality between the positive and negative domains of the coupling. The coupled Entropy uses a generalization of the logarithmic function

$$\ln_\kappa\left(x\right) \equiv \frac{1}{\kappa}\left(x^\kappa - 1\right), \ \ x > 0 \tag{13}$$

which provides a continuous set of functions with power . The coupled entropy aggregates the probabilities of a distribution using the generalized mean and translates this to the entropy domain using the generalized logarithm. Using the equiprobable for the sample probabilities, the coupled cross-entropy "score" for the forecasted probabilities $\mathbf{p}$ for the events $\mathbf{e}$

$$S_\kappa\left(\mathbf{e}, \mathbf{p}\right) \equiv \frac{-2}{1+\kappa}\ln_{\left(\frac{-2\kappa}{1+\kappa}\right)}\left(\left(\frac{1}{N}\sum_{i=1}^N p_i^{\frac{-2\kappa}{1+\kappa}}\right)^{\frac{-1-\kappa}{2\kappa}}\right) \equiv \frac{1}{\kappa}\left(\left(\frac{1}{N}\sum_{i=1}^N p_i^{\frac{-2\kappa}{1+\kappa}}\right) - 1\right) \tag{14}$$

8

In order to improve performance against the robust metric, the training of the variational autoencoder needs to incorporate this generalized metric. To do so we derive a coupled loss function in the next subsection.

4.2 Coupled loss function

The cross-entropy, which measures the uncertainty of distribution relative to another distribution, underlies both the metrics described above and the loss function used for training. The cross-entropy is the sum of two components, the underlying uncertainty in the distribution $\mathbf{p}$ measured by the entropy and difference between the distributions measured by the Kullback-Leibler (KL) divergence. The Kullback-Leibler divergence is defined as

$$D_{KL}\left(\mathbf{p}||\mathbf{q}\right) \equiv -\sum_i p_i \ln\left(\frac{q_i}{p_i}\right) \tag{15}$$

In the VAE algorithm, the loss function consists of the KL-divergence between the posterior distribution $q\left(\mathbf{z}|\mathbf{x}^{(i)}\right)$ and a prior $p\left(\mathbf{z}\right)$ and the cross-entropy between the reported probabilities and the training sample distribution.

$$L\left(\mathbf{x}^{(i)}\right) = -D_{KL}\left(q\left(\mathbf{z}|\mathbf{x}^{(i)}\right) \parallel p\left(\mathbf{z}\right)\right) + \frac{1}{L}\sum_{l=1}^{L}\left(\log p\left(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}\right)\right) \tag{16}$$

In this paper, the loss function is modified by coupled generalizations of the KL-divergence and cross-entropy to improve the robustness of the VAE model. Under the assumption that states in the system are no longer independent, a generalized entropy in which the average uncertainty is measured when there is "nonlinear coupling" between the states (Nelson et al., 2017). The generalized mean, $\left(\sum p_i^{1-\frac{2\kappa}{1+\kappa}}\right)^{-\frac{1+\kappa}{2\kappa}}$, modeling long-range correlation between the states, aggregates the states. When the coupling $\kappa \to 0$, the generalized mean is asymptotically equal to the geometric mean. The mathematical form of coupled entropy function with power $\alpha = 2$ and coupling κ is defined as in (Nelson et al., 2017),

$$S_\kappa\left(\mathbf{p}\right) \equiv \frac{1}{2}\ln_\kappa\left(\left(\sum p_i^{1+\frac{2\kappa}{1+\kappa}}\right)^{\frac{-1}{\kappa}}\right) \equiv \frac{1}{\kappa}\left(\left(\sum p_i^{\frac{1+3\kappa}{1+\kappa}}\right)^{-1} - 1\right) \tag{17}$$

where $\ln_\kappa\left(x\right)$ is the generalization of the logarithm function, known as the coupled logarithm function.

$$\ln_\kappa\left(x\right) \equiv \frac{1}{\kappa}\left(x^\kappa - 1\right), \; x > 0 \tag{18}$$

Therefore, the modified loss function contains two terms: negative coupled divergence and coupled cross-entropy. Coupled divergence is the generalization of KL divergence in equation

(15), which is defined as

$$D_\kappa \left(p\left(\mathbf{z}\right) || q\left(\mathbf{z}\right) \right)$$

$$\equiv \prod_{i=1}^{D1} \int_{-\infty}^{\infty} \frac{p(z_i)^{1+\frac{2\kappa}{1+\kappa}}}{\int_{-\infty}^{\infty} p(z_i)^{1+\frac{2\kappa}{1+\kappa}} dz_i} \frac{1}{2} \left(\ln_\kappa \left(q(z_i)^{-\frac{2}{1+\kappa}} \right) - \ln_\kappa \left(p(z_i)^{-\frac{2}{1+\kappa}} \right) \right) dz_i$$

$$= \prod_{i=1}^{D1} \frac{1}{\kappa} \int_{-\infty}^{\infty} \frac{\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z_i-\mu_i)^2}{2\sigma^2}} \right)^{1+\frac{2\kappa}{1+\kappa}}}{\int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z_i-\mu_i)^2}{2\sigma^2}} \right)^{1+\frac{2\kappa}{1+\kappa}} dz_i} \cdot \frac{1}{2} \left(\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \right)^{-\frac{2\kappa}{1+\kappa}} - \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z_i-\mu_i)^2}{2\sigma^2}} \right)^{-\frac{2\kappa}{1+\kappa}} \right) dz_i$$

$$(19)$$

where $D1$ is the dimensionality of $\mathbf{z}$. Coupled cross-entropy is the generalization of cross-entropy term in equation (12), which is defined as,

$$H_\kappa \left(\mathbf{x} \right) \equiv \sum_{i=1}^{D2} \left(x_i \frac{1}{2} \ln_\kappa \left((y_i)^{\frac{2}{1+\kappa}} \right) - (1 - x_i) \frac{1}{2} \ln_\kappa \left((1 - y_i)^{\frac{2}{1+\kappa}} \right) \right) \tag{20}$$

where $D2$ is the dimensionality of $\mathbf{z}$. The new loss function is the coupled loss function, which is written by

$$L\left(\mathbf{x}^{(i)} \right) = -D_\kappa \left(p\left(\mathbf{z}\right) \| q\left(\mathbf{z}\right) \right) + \frac{1}{L} \sum_{l=1}^{L} H_\kappa^{(l)} \tag{21}$$

Reasons that the coupled loss function can be used to improve the robustness of algorithm include: 1) Higher Uncertainty. The coupled entropy weights low probabilities with a higher cost, forcing the model to increase the probability learned for outliers in the training set. This ensures that outliers in the test set will be not be over-confident. 2) Penalty for Outliers. By modeling the correlation between samples, we are discounting the amount of available information. This forces the trained model to have more certainty and thereby be robust against outliers.

5. Results Using the MNIST Handwritten Numerals

We trained and tested the coupled VAE model using the MNIST dataset. The algorithm and experiments are developed with Python and the TensorFlow library. We set the dimensions of latent variables $\mathbf{z}$ to be 20, the batch size to be 5,000, and the number of epochs to be 100. Our Python code can be accessed on GitHub at https://github.com/Sission/Coupled-VAE-Improved-Robustness-and-Accuracy-of-a-Variational-Autoencoder.

The input images and output images for different values of coupling κ are shown in Figure 4. $\kappa = 0$ represents the original VAE model. Compared with the original algorithm, output images generated by the modified coupled VAE model show small improvements in detail and clarity. For instance, the fifth digit in the first row of the input images is

"4", but the output image in the original VAE is more like "9" rather than "4" while the coupled VAE method generates "4" correctly. For the seventh digit "4" in the first row, the generated image in the coupled VAE has an improved clarity than the regular VAE.

(a) Input Image (b) $\kappa = 0$ (c) $\kappa = 0.025$ (d) $\kappa = 0.05$ (e) $\kappa = 0.1$
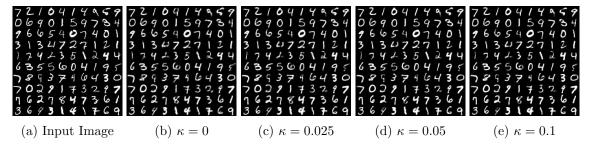
Figure 4: a) The MNIST input images and b) the output images generated by original VAE. c-e) The output images generated by modified coupled VAE model show small improvements in detail and clarity. For instance, the fifth digit in the first row of the input images is "4", but the output image in the original VAE is more like "9" rather than "4" while the coupled VAE method generates "4" correctly.
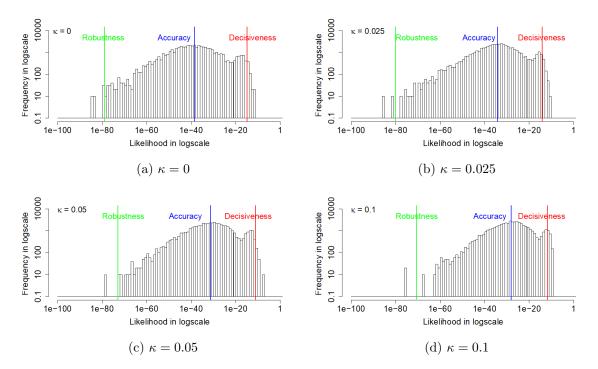

(a) $\kappa = 0$ (b) $\kappa = 0.025$

(c) $\kappa = 0.05$ (d) $\kappa = 0.1$

Figure 5: The histograms of likelihood for the input images with various κ values. The red, blue, and green lines represent the arithmetic mean (decisiveness), geometric mean (central tendency), and -2/3 mean (robustness), respectively. The minimal value of the robustness metric indicates that the original VAE suffers from poor robustness. As κ gets large, the geometric mean and the -2/3 mean metrics start to increase while the arithmetic mean metric almost keeps same.

(a) $\kappa = 0$

(b) $\kappa = 0.025$

(c) $\kappa = 0.05$
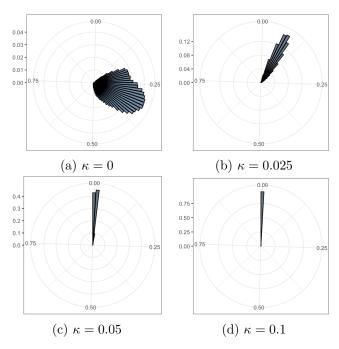
(d) $\kappa = 0.1$

Figure 6: The rose plots of the various standard deviation values in 20 dimensions. The range and average values of these standard deviations reduce as coupling increasing.

Figure 5 shows the likelihood histograms for 5000 input images with coupling values of $\kappa = 0, 0.025, 0.05, 0.1$. The red, blue, and green lines represent the arithmetic mean (decisiveness), geometric mean (central tendency), and -2/3 mean (robustness), respectively. When $\kappa = 0$, the minimal value of the robustness metric indicates that the original VAE suffers from poor robustness. As κ gets large, the geometric mean and the -2/3 mean metrics start to increase while the arithmetic mean metric almost keeps the same. However, when the coupling κ becomes large, the coupled loss function can quickly become infinity. For instance, when $\kappa = 0.2$, the loss function goes infinity at 53th epoch; when $\kappa = 0.5$, the loss function goes infinity at 8th epoch. In this case, the optimization of coupling values should be further investigated. The specific relationship between coupling κ and probabilities for input images is shown in Table 1. The increased robustness metric shows that the modified loss does improve the robustness of the original model.

Coupling κ	Arithmetic mean metric	Geometric mean metric	$-\frac{2}{3}$ mean metric
0	1.31×10^{-15}	2.41×10^{-39}	1.40×10^{-79}
0.025	6.61×10^{-15}	5.89×10^{-35}	9.91×10^{-81}
0.05	7.18×10^{-12}	5.80×10^{-32}	1.31×10^{-73}
0.1	1.34×10^{-12}	7.09×10^{-29}	2.57×10^{-71}

Table 1: The relationship between coupling κ with the probabilities for input data

Furthermore, compared with the original VAE model, the geometric mean, which measures the accuracy of the input image likelihood, is larger for the coupled algorithm. The improvement of this metric means that the input images(truth) are assigned to higher likelihoods in average by the coupled VAE model. Therefore, the modifications in section 3 also enhance the model's capability of capturing true and significant information.

The variance σ of latent variables $\mathbf{z}$ is shown in rose plots in Figure 6. The angular location of a bar represents the value of σ, clockwise from 0 to 1. The radius of the bar measures the frequency of different σ values from 0 to 100. As the coupling κ increases, the range and the average value of these standard deviations decrease. To be specific, when $\kappa = 0$, σ of all dimensions in all 5000 batches range from 0.09 to 0.72; when $\kappa = 0.025$, σ

ranges from 0.02 to 0.3; when $\kappa = 0.05$, σ ranges from 0.001 to 0.09; when $\kappa = 0.1$, σ ranges from 0.00007 to 0.06. These results may be the reason why images generated by modified coupled model have better clarity than those generated by regular VAE. Since the reduced standard deviation means less fluctuation of values of $\mathbf{z}$, the values of $\mathbf{x}'$ generated by those stable $\mathbf{z}$ in the decoder are more certain and concentrated. Thus, the output images, which are determined by values of $\mathbf{x}'$, are generated with higher clarity.

(a) $\kappa = 0$

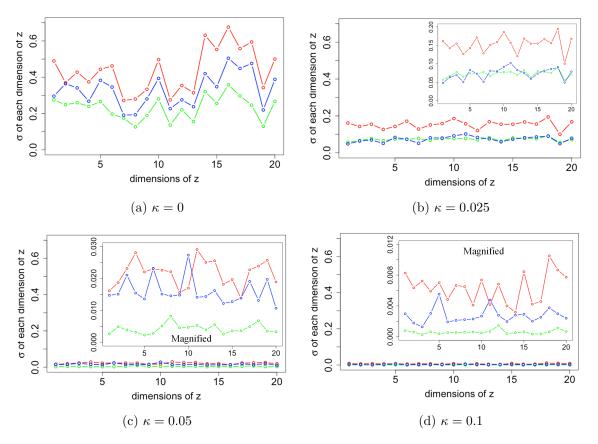(b) $\kappa = 0.025$

(c) $\kappa = 0.05$

(d) $\kappa = 0.1$

Figure 7: The standard deviation of latent variable samples near the three generalized mean metrics. The red, blue and green lines represent samples near the decisiveness, accuracy and robustness, respectively. As κ increases, values of σ are less fluctuant and decrease toward 0. Magnified plots are shown to visualize the results further. The general trend is for sigma to be more significant for samples near decisiveness, intermediate near the accuracy and smaller for samples near robustness. An exception is $\kappa = 0.025$, where sigma overlaps for near the robustness and accuracy.

We choose samples in which the likelihoods of input images are close to the three metrics and plotted the standard deviation σ of each dimension of the latent variable $\mathbf{z}$ of these samples in Figure 7. The red, blue and green lines represent samples near the decisiveness, accuracy and robustness, respectively. It shows that when $\kappa = 0$, the standard deviations of $\mathbf{z}$ range from 0.1 to 0.7. However, as κ increases, values of σ are less fluctuant and decrease toward 0. Magnified plots are shown to visualize the results further. The general

trend for σ is to be more significant for samples near decisiveness, intermediate near the accuracy and smaller for samples near robustness. An exception is $\kappa = 0.025$, where σ overlaps for samples near the robustness and accuracy. The histogram likelihood plots with a two-dimensional latent variable is shown in Figure 8. The increased values of arithmetic mean metric and -2/3 mean metric show that the accuracy and robustness of the output MNIST images in VAE model have been improved, consistent with the result in the 20-D model.
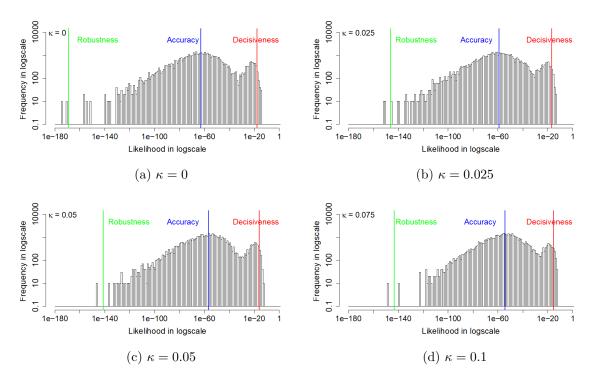

Figure 8: The histogram likelihood plots with a two-dimensional latent variable. Like the 20-D model, the increased values of arithmetic mean metric and -2/3 mean metric show that the accuracy and robustness of the VAE model have been improved.

6. Discussion

In order to understand the relationship between the increase in the coupling of the loss function and the decrease in the standard deviations of the Gaussian model, we examine a two-dimensional model which can be visualized.

Compared with the high-dimensional model, the probability likelihoods for the two-dimensional model are lower, indicating that the higher-dimensions does improve the model. Nevertheless, like the 20-dimensional model, the distribution of likelihood is compressed toward higher values as the coupling increases and, therefore, can be used to analyze the results further. Larger likelihood of input images and smaller standard deviations of latent variables are the two main changes as the coupling parameter for the modified loss function is increased. As a result, both the robustness and accuracy metrics increase. To be specific,

(a) $\kappa = 0$

(b) $\kappa = 0.025$

(c) $\kappa = 0.05$
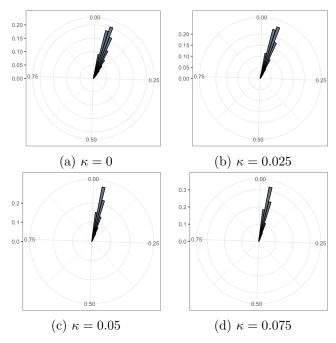
(d) $\kappa = 0.075$

Figure 9: The rose plots of the various standard deviation values in 2 dimensions. The range of standard deviation reduces as coupling increasing.

when κ increases from 0 to 0.075, the geometric mean metric increases from 1.20×10^{-63} to 4.67×10^{-55}, and the -2/3 mean metric increases from 5.03×10^{-170} to 5.17×10^{-144} while the arithmetic metric does not change very much. In this case, the input images will be assigned with higher probabilities by the coupled VAE method, which uses larger coupling values for the loss function.

The rose plots in Figure 9 show that both the range and the average of the standard deviations decrease when the coupling κ increases. The latent space plots shown in Figure 10 are the visualizations of images of the numerals from 0 to 9 Images are embedded in a 2D map where the axis is the values of the 2D latent variable. The same color represents images that belong to the same numeral, and they cluster together since they have higher similarity with each other. The distances between spots represent the similarities of images. The latent space plots show that the different clusters shrink together more tightly when coupling has a large value. The plots shown in Figure 11 are the visualization of the learned data manifold generated by the decoder network of the coupled VAE model. A grid of values from a two-dimensional Gaussian distribution is sampled. The distinct digits each exist in different regions of the latent space and smoothly transform from one digit to another. This smooth transformation can be quite useful when the interpolation between two observations is needed. Additionally, the distribution of distinct digits in the plot becomes more evenly, and the sharpness of the digits increases when κ increases.
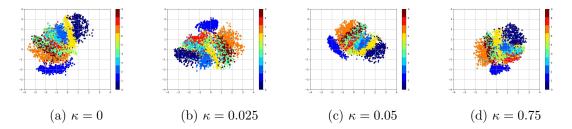

(a) $\kappa = 0$

(b) $\kappa = 0.025$

(c) $\kappa = 0.05$

(d) $\kappa = 0.75$

Figure 10: The plot of the latent space of VAE trained for 200 epochs on MNIST with various κ values. Different numerals cluster together more tightly as coupling κ increasing.

15

(a) $\kappa = 0$

(b) $\kappa = 0.025$

(c) $\kappa = 0.05$
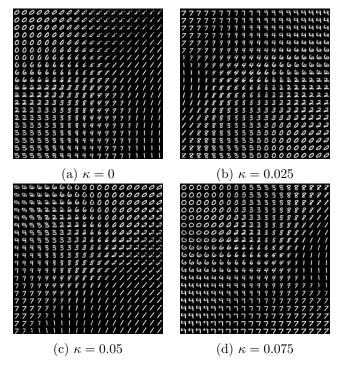
(d) $\kappa = 0.075$

Figure 11: The plot of visualization of learned data manifold for generative models with the axes to be the values of each dimension of latent variables. The distinct digits each exist in different regions of the latent space and smoothly transform from one digit to another.

The reasons that the likelihoods of input images increase and standard deviations of latent variable decrease are analyzed as follows.

1. Why does the latent variable have a smaller deviation in the coupled VAE model?

In the coupled VAE algorithm, the loss function is modified to coupled entropy function via the nonlinear statistical coupling. If we consider the states of the latent variable to be locations where an image will be "stored", then the "nonlinear coupling" models the dependency between these states. The coupled VAE method considers long-range correlation between the states. If we interpret the dependency between states to be "similarity", we can explain the tighter clustering with increased coupling as a result of modeling the dependency. That is because if different states, which are representing the images, have more similarities, they will be closer to each other. The shrinkage between numerals corresponds to the decreased variation of the latent variable, thus explaining the smaller standard deviations for the coupled VAE method.

2. Why do the probabilities of the input images increase in the coupled VAE method?

The probability of an input image for the decoder model can be calculated by

$$p\left(x|z\right) = \frac{p\left(x, z\right)}{p\left(z\right)} = \frac{p\left(z|x\right)p\left(x\right)}{p\left(z\right)} \tag{22}$$

where z given x follows a Gaussian distribution; $p(x)$ is the prior distribution of the input data x, which follows a Bernoulli distribution; $p(z)$ is the prior distribution of latent variable z, which follows a standard Gaussian distribution.

In our modified algorithm, $p(x)$ and $p(z)$ stay the same, while the density $p\left(z|x\right)$ changes. In the traditional VAE method, we assumed

$$p_1\left(z|x\right) = \frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma_1^2}} \tag{23}$$

while our coupled VAE method, assumed

$$p_2\left(z|x\right) = \frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma_2^2}} \tag{24}$$

16

where $\sigma_2 > \sigma_1$. So, the input data has a smaller range of probabilities, and the average density values increase. Furthermore, the standard deviation decreases as the coupling increases. In this case, the range of probabilities of input shrinks and the geometric mean of density values increases.

7. Conclusion

The coupled VAE method succeeds in increasing the probability likelihood of input images. We document the improvement by analyzing the histogram of the likelihoods for the input data using the arithmetic mean, geometric mean, and -2/3 mean, which represent decisiveness, accuracy, and robustness, respectively. Both the accuracy and the robustness are increased by increasing the coupling of the loss function. However, when coupling gets large, the modified loss function cannot converge. The modification of loss function changes the latent space in the model. The latent variable has smaller standard deviations as coupling κ increases. In this case, the learned images are compressed into a more compact 2D space, influencing the probabilities for the input data in the generative model. The clarity of the output images also shows small improvements with increases in the coupling for the loss function. For future work, we plan to assume the coupled Gaussian distribution to be the prior and posterior distribution of latent variables. This may be helpful to achieve a greater separation between the numerals into distinct clusters similar to what has been achieved with the t-Stochastic Neighborhood Embedding methods (Van Der Maaten and Hinton, 2008). If so, it may be possible to improve the decisiveness of the likelihoods in addition to further improvements in the accuracy and robustness.

References

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112:859–877, Jan 2016. URL http://arxiv.org/abs/1601.00670.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning (CoNLL), arXiv:1511.06349*, 2015. URL http://arxiv.org/abs/1511.06349.

Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders, Nov 2016. URL http://arxiv.org/abs/1611.02648.

Jeff Donahue, Trevor Darrell, and Philipp Krähenbühl. Adversarial feature learning. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2017.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2017.

Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. Hidden markov model variational autoencoder for acoustic unit discovery. *Interspeech*, 2017. URL http://dx.doi.org/10.21437/Interspeech.2017-1160.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2053–2061. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5679-learning-with-a-wasserstein-loss.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016. URL http://www.deeplearningbook.org/.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR), arXiv: 1312.6114v10*, 2014. URL https://arxiv.org/pdf/1312.6114.pdf.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits, 1998. URL http://yann.lecun.com/exdb/mnist/index.html.

Donald McAlister. Xiii. the law of the geometric mean. *Proceedings of the Royal Society*, 29(196-199):367–376, Dec 1879. ISSN 0370-1662.

Kenric P. Nelson. Reduced perplexity: A simplified perspective on assessing probabilistic forecasts. In Min Chen, Jon M. Dunn, Amos Golan, and Aman Ullah, editors, *Info-Metrics Volume*. Oxford University Press, 2020. URL http://arxiv.org/abs/1603.08830.

Kenric P. Nelson and Sabir Umarov. Nonlinear statistical coupling. *Physica A: Statistical Mechanics and its Applications*, 389(11):2157–2163, 2010. ISSN 03784371.

Kenric P. Nelson, Sabir R. Umarov, and Mark A. Kon. On the average uncertainty for systems with nonlinear coupling. *Physica A: Statistical Mechanics and its Applications*, 468:30–43, Feb 2017. ISSN 03784371.

Olli Niemitalo. A method for training artificial neural networks to generate missing data within a variable context, 2010. URL https://web.archive.org/web/20120312111546/http://yehar.com:80/blog/?p=167.

Judea Pearl. Bayesian netwcrks: A model cf self-activated memory for evidential reasoning. Technical report, University of California, 1985. URL http://ftp.cs.ucla.edu/pub/stat{_}ser/r43-1985.pdf.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *ICLR 2017*, Mar 2017. URL http://arxiv.org/abs/1703.01488.

Dustin Tran, Matthew D Hoffman, Rif A. Saurous, Eugene Brevdo, Kevin Murphy, and David M Blei. Deep probabilistic programming. In *Fifth International Conference on Learning Representations, arXiv:1701.03757*, 2017. URL https://arxiv.org/abs/1701.03757.

Constantino Tsallis. *Introduction to nonextensive statistical mechanics: Approaching a complex world*. Springer New York, 2009. ISBN 9780387853581.

Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using T-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Haowen Xu, Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, and Dan Pei. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web. arXiv:1802.03903*, pages 187–196, New York, New York, USA, 2018. ACM Press. ISBN 9781450356398. URL http://dl.acm.org/citation.cfm?doid=3178876.3185996.

Jonathan Zalger. Application of variational autoencoders for aircraft turbomachinery design. Technical report, Stanford Univ., 2017. URL http://cs229.stanford.edu/proj2017/final-reports/5231979.pdf.