NashAE: Disentangling Representations through Adversarial Covariance Minimization

Eric Yeats[1], Frank Liu[2], David Womble[2], and Hai Li[1]

[1] Duke University, Durham NC 27708
{eric.yeats, hai.li}@duke.edu
[2] Oak Ridge National Laboratory, Oak Ridge TN 37830
{liufy, womblede}@ornl.gov

Abstract. We present a self-supervised method to disentangle factors of variation in high-dimensional data that does not rely on prior knowledge of the underlying variation profile (e.g., no assumptions on the number or distribution of the individual latent variables to be extracted). In this method which we call NashAE, high-dimensional feature disentanglement is accomplished in the low-dimensional latent space of a standard autoencoder (AE) by promoting the discrepancy between each encoding element and information of the element recovered from all other encoding elements. Disentanglement is promoted efficiently by framing this as a minmax game between the AE and an ensemble of regression networks which each provide an estimate of an element conditioned on an observation of all other elements. We quantitatively compare our approach with leading disentanglement methods using existing disentanglement metrics. Furthermore, we show that NashAE has increased reliability and increased capacity to capture salient data characteristics in the learned latent representation.

Keywords: representation learning, autoencoder, adversarial, minmax game

1 Introduction

Deep neural networks (DNNs) have proven to be extremely high-performing in the realms of computer vision [7], natural language processing [26], autonomous control [16], and deep generative models [5,13], among others. The huge successes of DNNs have made them almost ubiquitous as an engineering tool, and it is very common for them to appear in many new applications. However, as we rush to deploy DNNs in the real world, we have also exposed many of their shortcomings.

One such shortcoming is that DNNs are extremely sensitive to minute perturbations in their inputs [6,25] or weights [8], causing otherwise high-performing models to suddenly be consistently incorrect. Additionally, DNNs trained on image classification tasks are observed to predict labels confidently even when the image shares no relationship with their in-distribution label space [9,17]. Furthermore, DNNs are known to perpetuate biases in their training data through

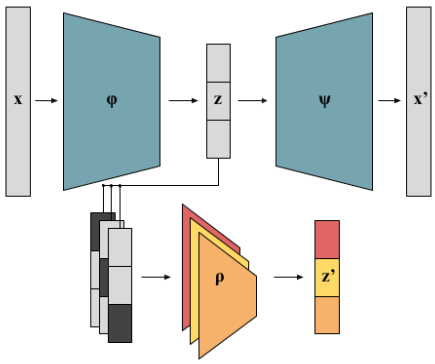


Fig. 1: Depiction of the proposed disentanglement method, NashAE, for a latent space dimensionality of $m = 3$. An autoencoder (AE), composed of the encoder ϕ and decoder ψ, compresses a high dimensional input $\mathbf{x} \in \mathbb{R}^n$ to a lower dimensional latent vector $\mathbf{z} \in \mathbb{R}^m$, and decompresses $\mathbf{z}$ to approximate $\mathbf{x}$ as $\mathbf{x}'$. An ensemble of m independently trained regression networks ρ takes m duplicates of $\mathbf{z}$ which each have an element removed, and each independent regression network tries to predict the value of its missing element using knowledge of the other elements. Disentanglement is achieved through equilibrium in an adversarial game in which ϕ minimizes the element-wise covariance between the true latent vector $\mathbf{z}$ and concatenated predictions vector $\mathbf{z}'$

their predictions, exacerbating salient issues such as racial and social inequity, and gender inequality [1]. While this small subset of examples may appear to be unrelated, they are all linked by a pervasive issue of DNNs: their lack of interpretability [23,24]. The fact that DNNs are treated as black boxes, where engineers lack a clear explanation or reasoning for why or how DNNs make decisions, makes the root cause of DNNs' shortcomings difficult to diagnose.

A promising remedy to this overarching issue is to clarify learning representations through feature disentanglement: the process of learning unique data representations that are each only sensitive to independent factors of the underlying data distribution. It follows that a disentangled representation is an inherently interpretable representation, where each disentangled unit has a consistent, unique, and independent interpretation of the data across its domain.

Several works have pioneered the field of feature disentanglement, moving from supervised [14] to unsupervised approaches [2,4,10,11], which we focus on in this work. Chen et al. [4] present InfoGAN, an extension of the GAN framework [5] that enables better control of generated images using special noise variables. Higgins et al. [10] introduce β-VAE, a generalization of the VAE framework [13] that allows the VAE to extract more statistically independent, disentangled representations.

While these highly successful methods are considered to be unsupervised, they still have a considerable amount of prior knowledge built into their operation. InfoGAN requires prior knowledge of the number and form of disentangled factors to extract, and β-VAE encounters bottleneck capacity issues and inconsistent results with seemingly innocuous changes to hyperparameters, requiring finetuning with some supervision [2,10].

We propose a new method, NashAE, to promote a sparse and disentangled latent space in the standard AE that does not make assumptions on the number or distribution of underlying data generating factors. The core intuition behind the approach is to reduce the redundant information between latent encoding elements, regardless of their distribution. To accomplish this, this work presents a new technique to reduce the information between encoded continuous and/or discrete random variables using just access to samples drawn from the unknown underlying distributions. We empirically demonstrate that the method can reliably extract high-quality disentangled representations according to the metric presented in [10], and that the method has a higher latent feature capacity with respect to salient data characteristics.

The paper makes the following contributions:

- We develop a method to quantify the relationship between random variables with unknown distribution (arbitrary continuous or discrete/categorical), and show that it can be used to promote statistical independence among latent variables in an AE.
- We provide qualitative and quantitative evidence that NashAE reliably extracts a set of disentangled continuous and/or discrete factors of variation in a variety of scenarios, and we demonstrate the method's improved latent feature capacity with regard to salient data characteristics.
- We release the Beamsynthesis disentanglement dataset, a collection of time-series data based on particle physics studies and their associated data generating factor ground truth.

The code for all experiments and the Beamsynthesis dataset can be found at: `https://github.com/ericyeats/nashae-beamsynthesis`.

2 Related Work

Autoencoders. Much of this work derives from autoencoders (AE), which consist of an encoder function followed by a decoder function. The encoder transforms high-dimensional input $\mathbf{x} \sim X$ into a low-dimensional latent representation $\mathbf{z}$, and the decoder transforms $\mathbf{z}$ to a reconstruction of the high-dimensional input $\mathbf{x}'$. AE have numerous applications in the form of unsupervised feature learning, denoising, and feature disentanglement [11,20,22]. Variational autoencoders (VAEs) [13] take AEs further by using them to parameterize probability distributions for X. VAEs are trained by maximizing a lower bound of the likelihood of X, a process which involves conforming the encoded latent space $\mathbf{z} \sim Z$ with a prior distribution P. Adversarial AEs [21], like VAEs, match encoded

distributions to a prior distribution, but do so through an adversarial procedure inspired by Generative Adversarial Networks (GANs) [5].

Unsupervised Disentanglement Methods. One of the most successful approaches to feature disentanglement is β-VAE [10], which builds on the VAE framework. β-VAE adjusts the VAE training framework by modulating the relative strength of the $D_{\mathrm{KL}}(Z||P)$ term with hyperparameter β, effectively limiting the capacity of the VAE and encouraging disentanglement as β becomes larger. Higgins et al. [10] note a positive correlation between the size of the VAE latent dimension and the optimal β hyperparameter to do so, requiring some hyperparameter search and limited supervision. Another important contribution of Higgins et al. [10] is a metric for quantifying disentanglement which depends on the accuracy of a linear classifier in determining which data generating factor is held constant over a pair of data batches.

Multiple works have augmented β-VAE with loss functions that isolate the Total Correlation (TC) component of $D_{\mathrm{KL}}(Z||P)$, further boosting quantitative disentanglement performance in certain scenarios [3,12]. Another VAE-based work proposed by Kumar et al. [15] directly minimizes the covariance of the encoded representation. However, simple covariance of the latent elements fails to capture more complex, nonlinear relationships between the elements. Our work employs regression neural networks to capture complex dependencies.

Chen et al. [4] present InfoGAN, which builds on the GAN framework [5]. InfoGAN augments the base GAN training procedure with a special set of independent noise inputs. A tractable lower bound on MI is maximized between the special noise inputs and output of the generator, leading to the special noise inputs resembling data generating factors. While the method claims to be unsupervised, choosing its special noise inputs requires prior knowledge of the number and nature (e.g. distribution) of factors to extract.

Limitations of Unsupervised Disentanglement. Locatello et al. [19] demonstrate that unsupervised disentanglement learning is fundamentally impossible without incorporating inductive biases on both models and data [19]. However, they assert that given the right inductive biases, the prospect of unsupervised disentanglement learning is not so bleak. We incorporate several inductive biases in our method to achieve unsupervised disentanglement. First, our approach assumes that disentangled learning representations are characterised by being statistically independent. Second, we posit that breaking up the latent factorization problem into multiple parts by individual masking and adversarial covariance minimization helps boost disentanglement reliability. In terms of models and data, we employ the network architectures and data preparation suggested by previous works in unsupervised disentanglement. Under such conditions, NashAE has demonstrated superior reliability in retrieving disentangled representations.

3 NashAE Methodology

Our approach starts with a purely deterministic encoder ϕ, which takes input observations $\mathbf{x} \sim X$ and creates a latent representation $\mathbf{z} = \phi(\mathbf{x})$. Where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^m$, and typically $n \gg m$. Furthermore, ϕ employs a sigmoid activation function σ at its output to produce $\mathbf{z}$ such that $\mathbf{z} = \sigma(\zeta)$ and $\mathbf{z} \in [0, 1]^m$, where ζ is the output of ϕ before it is passed through the sigmoid non-linearity. A deterministic decoder ψ maps the latent representation $\mathbf{z}$ back to the observation domain $\mathbf{x}' = \psi(\mathbf{z})$. To achieve disentanglement, the AE is trained with two complementary objectives: (1) reconstructing the observations, and (2) maximizing the discrepancy between each latent variable and predicted values of the variable using information of all other variables. The intuitions behind each are the following. First, reconstruction of the input observations $\mathbf{x}$ is standard of AEs and ensures that they learn features relevant to the distribution X. Second, promoting discrepancy between i-th latent element and its prediction (conditioned on all other $j \neq i$ elements) reduces the information between latent element i and all other elements $j \neq i$.

For the reconstruction objective, the goal is to minimize the mean squared error:

$$\mathcal{L}_R = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim X}\big|\big|\mathbf{x}' - \mathbf{x}\big|\big|_2^2. \tag{1}$$

Reconstructing the input observation $\mathbf{x}$ ensures that the features of the latent space are relevant to the underlying data distribution X. The following subsection describes the adversarial game loss objectives, which settle on an equilibrium and inspire the name of the proposed disentanglement method, NashAE.

Adversarial Covariance Minimization

In general, it is difficult to compute the information between latent variables when one only has access to samples of observations $\mathbf{z} \sim Z$. Since the underlying distribution Z is unknown, standard methods of computing the information directly are not possible. To overcome this challenge, we propose to reduce the information between latent variables indirectly using an ensemble of regression networks which attempt to capture the relationships between latent variables. The process is computationally efficient; it uses simple measures of linear statistical independence and an adversarial game.

Consider an ensemble of m independent regression networks ρ, where the output of the i-th network ρ_i corresponds to a missing i-th latent element. The objective of each ρ_i is to minimize the mean squared error:

$$\mathcal{L}_{\rho_i} = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim X}\big(\rho(\overline{\mathbf{z}_i})_i - \mathbf{z}_i\big)^2, \tag{2}$$

where $\mathbf{z}_i$ is the i-th true latent element, and $\overline{\mathbf{z}_i}$ is the latent *vector* with the i-th latent element masked with 0 (i.e., all elements of the latent vector are present except $\mathbf{z}_i$).

We call ρ the predictors, since they are each optimized to predict one missing value of $\mathbf{z}$ given knowledge of all other $\mathbf{z}$. If all their individual predictions are concatenated together, they form $\mathbf{z}'$ such that each $\mathbf{z}'_i = \rho(\overline{\mathbf{z}_i})_i$.

For the disentanglement objective, we want to choose encodings $\mathbf{z} \sim \phi(X)$ that make it difficult to recover information of one element from all others. This leads to a natural minmax formulation for the AE and predictors:

$$\min_{\phi,\psi} \max_{\rho} \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim X}\left[\left|\left|\mathbf{x}' - \mathbf{x}\right|\right|_2^2 - \left|\left|\mathbf{z}' - \mathbf{z}\right|\right|_2^2\right]. \tag{3}$$

In general, each predictor attempts to use information of $\overline{\mathbf{z}_i}$ to establish a one-to-one linear relationship between $\mathbf{z}'$ and $\mathbf{z}$. Hence we propose to use covariance between $\mathbf{z}'$ and $\mathbf{z}$ across a batch of examples to capture the degree to which they are related. In practice, we find that training the AE to minimize the summed covariance objective between each of the $\mathbf{z}'_i$ and $\mathbf{z}_i$ random variable pairs,

$$\mathcal{L}_A = \sum_{i=1}^{m} \mathrm{Cov}(\mathbf{z}'_i, \mathbf{z}_i), \tag{4}$$

is more stable than maximizing $\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim X}||\mathbf{z}' - \mathbf{z}||$ and leads to more reliable disentanglement outcomes. Hence, in all the following experiments we train the AE to minimize this summed covariance measure, $\mathcal{L}_A$. Furthermore, one can show that the fixed points of the minmax objective (3) are the same as those of training ϕ to minimize $\mathcal{L}_R + \mathcal{L}_A$ for disentangled representations (see supplementary material).

In the adversarial loss $\mathcal{L}_A$, the optimization objective of the encoder ϕ is to adjust its latent representations to minimize the covariance between each $\mathbf{z}'_i$ and $\mathbf{z}_i$. Using minibatch stochastic gradient descent (SGD), the encoder ϕ can use gradient passed through the predictors ρ to learn exactly how to adjust its latent representations to minimize the adversarial loss. Assuming that ρ can learn faster than ϕ, each i-th covariance term will reach zero when $\mathbb{E}[\mathbf{z}_i|\mathbf{z}_j, \forall j \neq i] = \mathbb{E}[\mathbf{z}_i]$ everywhere.

In the following experiments, we weight the sum of $\mathcal{L}_R$ and $\mathcal{L}_A$ with the hyperparameter $\lambda \in [0, 1)$ in order to establish a normalized balance between the reconstruction and adversarial objectives:

$$\mathcal{L}_{R,A}(\lambda) = (1 - \lambda)\mathcal{L}_R + \lambda\mathcal{L}_A. \tag{5}$$

Intuitively, higher values of λ result lower covariance between elements of $\mathbf{z}$ and $\mathbf{z}'$, and eventually the equilibrium covariance settles to zero. In the special case where all data generating factors are independent, the AE can theoretically achieve $\mathcal{L}_{R,A} = 0$.

Proposed Disentanglement Metric: TAD

In the following section, we find that NashAE and β-VAE could achieve equally high scores using the β-VAE metric. However, the β-VAE metric fails to capture

a key aspect of a truly disentangled latent representation: change in one independent data generating factor should correspond to change in just one disentangled latent feature. This is not captured in the β-VAE disentanglement metric since the score can benefit from *spreading* the information of one data generating factor over multiple latent features. For example, duplicate latent representations of the same unique data generating factor can only increase the score of the β-VAE metric.

Furthermore, a disentanglement metric should quantify the degree to which its set of independent latent axes aligns with the independent data generating factor ground truth axes. In essence, a unique latent feature should be a confident predictor of a unique data generating factor, and all other latents should be orthogonal to the same data generating factor. Intuitively, the greater number of latent axes that align uniquely with the data generating factors and the more confident the latents are as predictors of the factors, the higher the metric score should be.

For these reasons, we design a disentanglement metric for datasets with binary attribute ground truth labels called Total AUROC Difference (TAD). For a large number l of examples which we collect a batch of latent representations z of the shape (l, m), we perform the following to calculate the TAD:

1. For each independent ground truth attribute, calculate the AUROC score of each latent variable in detecting the attribute.
2. For each independent ground truth attribute, find the maximum latent AUROC score $a_{1,i}$ and the next-largest latent AUROC score $a_{2,i}$, where i is the index of the independent ground truth attribute under consideration.
3. Take $\sum_i a_{1,i} - a_{2,i}$ as the TAD score, where i indexes over the independent ground truth attributes.

The TAD metric captures important aspects of a disentangled latent representation. First, each AUROC difference $a_{1,i} - a_{2,i}$ captures the degree to which a unique attribute is detected by a unique latent representation. Second, summing the AUROC difference scores for each independent ground truth attribute quantifies the degree to which the latent axes confidently replicate the ground truth axes. See the supplementary material for more details on how TAD is calculated and for a discussion relating it with other work.

4 Experiments

The following section contains a mix of qualitative and quantitative results for four unsupervised disentanglement algorithms: NashAE (this work), β-VAE [10], FactorVAE [12], and β-TCVAE [3]. The results are collected for disentanglement tasks on three datasets: Beamsynthesis, dSprites [10], and CelebA [18]. Please refer to the supplementary material for details on algorithm hyperparameters, network architectures, and data normalization for the different experiments.

Beamsynthesis is a simple dataset of 360 time-series data of current waveforms constructed from simulations of the LINAC (linear particle accelerator)

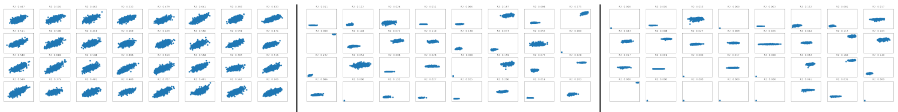
portion of high-energy particle accelerators. The dataset contains two ground truth data generating factors: a categorical random variable representing the *frequency* of the particle waveform which can take on one of the three values $(10, 15, 20)$ and a continuous random variable constructed from a uniform sweep of 120 waveform *duty cycle* values $\in [0.2, 0.8)$. The Cartesian product of the two data generating factors forms the set of observations. The challenge in disentangling this dataset arises from the fact that both the *frequency* and *duty cycle* of a waveform affect the length of the "on" period of each wave. We visualize the complete latent space of different algorithms and evaluate the reliability of the algorithms in extracting the correct number of ground truth data generating factors using this dataset.

dSprites is a disentanglement dataset released by the authors of β-VAE - it is comprised of a set of $737,280$ images of white 2D shapes on a black background. The Cartesian product of the type of shape (categorical: square, ellipse, heart), scale (continuous: 6 values), orientation (continuous: 40 values), x-position (continuous: 32 values), and y-position (continuous: 32 values) forms the independent ground truth of the dataset. We measure the β-VAE disentanglement metric score for different algorithms using this dataset.

CelebA is a dataset comprised of $202,599$ images of the faces of $10,177$ different celebrities. Associated with each image are 40 different binary attribute labels such as *bangs*, *blond hair*, *black hair*, *chubby*, *male*, and *eyeglasses*. We measure the TAD score of different algorithms using this dataset.

Empirical Fixed Point Results

In section 3, we indicate that higher values of $\lambda \in (0, 1)$ should result in a statistically independent NashAE latent space, and that redundant latent elements will not be learned. This is supported by observations of the fixed point of the optimization process for all experiments with nonzero λ: as λ is increased, the number of dead latent representations increases, and the average R^2 correlation statistic between latent representations and their predictions decreases.



(a) $\lambda = 0.0$. The average R^2 statistic is 0.52, and 32 latent representations are learned.

(b) $\lambda = 0.1$. The average R^2 statistic is 0.09, and 28 latent representations are learned.

(c) $\lambda = 0.2$. The average R^2 statistic is 0.04, and 22 latent representations are learned.

Fig. 2: Visualization of true latent representations (*x-axis*) vs predicted latent representations (*y-axis*) on the CelebA dataset

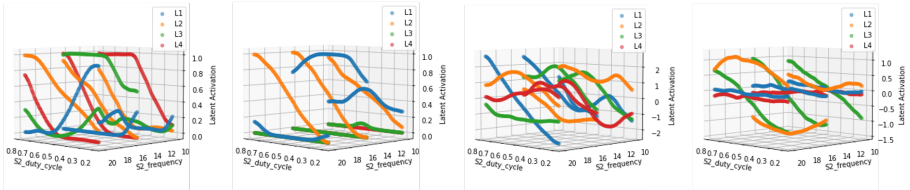
Figure 2 depicts each of the 32 true latent representations vs their predictions for 1000 samples of the CelebA dataset after three different NashAE networks have converged. When $\lambda = 0$ (standard AE), all latent elements are employed towards the reconstruction objective, and the predictions exhibit a strong positive linear relationship with the true latent variables (average R^2 is 0.52). When $\lambda = 0.1$, only 28 latent representations are maintained, and the average R^2 statistic between true latents and their predictions becomes 0.09. The 4 unused latent representations are each isolated in a dead zone of the sigmoid non-linearity, respectively. When λ is increased to 0.2, only 22 latent representations are maintained and the average R^2 statistic decreases even further to 0.04. Note also that the predictions become constant and are each equal to the expected value of their respective true latent feature. This is consistent with the conditional expectation of each variable being equal to its marginal expectation everywhere, and it indicates that no useful information is given to the predictors towards their regression task.

Beamsynthesis Latent Space Visualization

Figure 3 depicts the complete latent space generated by encoding all 360 observations of the Beamsynthesis dataset for the different algorithms and their baselines with a starting latent size of $m = 4$.

A standard AE latent space (leftmost) employs all latent elements towards the reconstruction objective, and their relationship with the ground truth data generating factors, *frequency* (categorical) and *duty cycle* (continuous), is unclear. Similarly, when a standard VAE (center right) converges and the μ component of the latent space is plotted for all observations, all latent variables are employed towards the reconstruction objective, and no clear relationship can be established for the latent variables.

If an adversarial game is played with $\lambda = 0.2$ (center left), the correct number of latent dimensions is extracted, and each nontrivial latent representation aligns with just one data generating factor. In this case, $L1$ level-encodes the *frequency* categorical data generating factor, and $L2$ encodes the *duty cycle* continuous data



NashAE $\lambda = 0.0$ NashAE $\lambda = 0.2$ β-VAE $\beta = 1$ β-VAE $\beta = 100$

Fig. 3: Visualizations of the learned latent space for the different algorithms on the Beamsynthesis dataset

generating factor with a consistent interpretation. The unused neurons remain in a dead zone of the sigmoid non-linearity.

β-VAE with $\beta = 100$ (rightmost) can disentangle the 360 observations in a similar fashion. In this example, $L2$ level-encodes the *frequency* categorical data generating factor, and $L3$ encodes the *duty cycle* continuous data generating factor with a consistent interpretation. The unused neurons each have approximately 0 variance in their μ component and an approximately constant value of 1 in their learned variance component.

Although all algorithms are capable of extracting a disentangled representation of the ground truth data generating factors, there is a stark difference in the reliability of the methods in extracting the correct number of latent variables when the starting latent space size m is changed. Reliability in this aspect is critical, as the dimensionality of the independent data generating factors is often an important unknown quantity to recover from new data. To determine this unknown dimensionality, one should start with a latent space size m which is larger than the number of latent factors that should be extracted.

Table 1: Average absolute difference between the number of learned latent dimensions and the the number of ground truth factors for different starting latent space sizes m on the Beamsynthesis dataset. Lower is better, and the lowest for each latent space size configuration (m) are in bold. The results are averaged over 8 trials

Method	$m = 4$	$m = 8$	$m = 16$	$m = 32$
NashAE $\lambda = 0$	1.375	5.75	13.125	28.75
NashAE $\lambda = 0.2$	**0**	**0**	**0.25**	1
NashAE $\lambda = 0.3$	**0**	**0**	0.375	**0.5**
β-VAE $\beta = 1$	2	6	14	28.875
β-VAE $\beta = 50$	1.375	2	2.5	3.5
β-VAE $\beta = 100$	**0**	1	1.25	3.125
β-VAE $\beta = 125$	1.375	1.5	1.875	3.25
β-TCVAE $\beta = 1$	1.875	5.25	9.625	18.625
β-TCVAE $\beta = 50$	0.75	1.25	1	**0.5**
β-TCVAE $\beta = 75$	0.375	0.5	1	0.875
β-TCVAE $\beta = 100$	0.75	0.625	1.125	0.625
FactorVAE $\beta = 50$	0.5	1.25	0.875	0.625
FactorVAE $\beta = 75$	0.5	1	1.25	**0.5**
FactorVAE $\beta = 100$	0.25	0.75	1	0.75
FactorVAE $\beta = 125$	0.875	0.75	0.625	0.75

Table 1 depicts the results of an experiment in which all hyperparameters are held constant except the starting latent size m as each of the algorithms

are trained to convergence on the Beamsynthesis dataset. Each entry in the table is the average absolute difference between the number of learned latent representations and the number of ground truth data generating factors (2 for Beamsynthesis), collected over 8 trials. Both NashAE $\lambda = 0$ and β-(TC)VAE $\beta = 1$ learn far too many latent variables, and β-VAE $\beta = 125$ tends to learn too few latent variables when $m = 4$ and $m = 8$. NashAE $\lambda = 0.2$ and NashAE $\lambda = 0.3$ perform very well in comparison, keeping the average absolute difference less than or equal to one in all configurations of m. β-TCVAE and FactorVAE perform second-best overall, tending to learn too many latent variables. The results indicate that NashAE is the most consistent in recovering the correct number of data generating factors. See the supplementary material for a similar experiment with the dSprites dataset and details on how learned latent representations are counted.

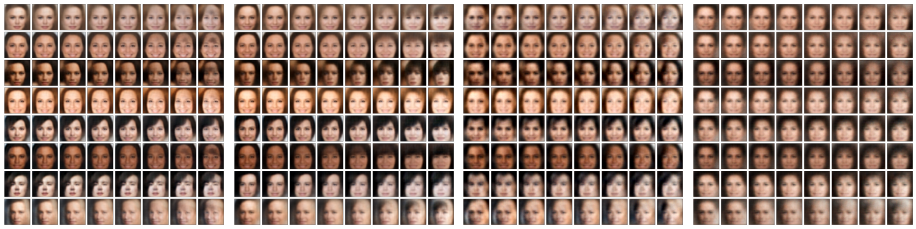
β-VAE Metric on dSprites

Table 2 reports the disentanglement score of each algorithm averaged over 15 trials - please refer to Higgins et al. [10] for more details on the metric. In general, the standard AE (NashAE, $\lambda = 0$) and standard VAE (β-VAE, $\beta = 1$; β-TCVAE, $\beta = 1$) performed the worst on the β-VAE disentanglement metric. As λ and β are increased, the disentanglement score of NashAE and β-VAE increases to over 96%. We do not observe the difference between NashAE and β-VAE in top performance on this metric to be significant, so both are in bold. In general, β-TCVAE performed slightly worse on this metric than β-VAE and NashAE, achieving just over 95%. We observed that increasing λ or β beyond these values leads to poorer performance for all algorithms. All algorithms achieve higher disentanglement scores on some initializations than others, but no *outliers* are removed from the reported scores (as is done in [10]). Overall, the results indicate that NashAE scores at least as high as those of β-VAE and β-TCVAE algorithm on the β-VAE metric.

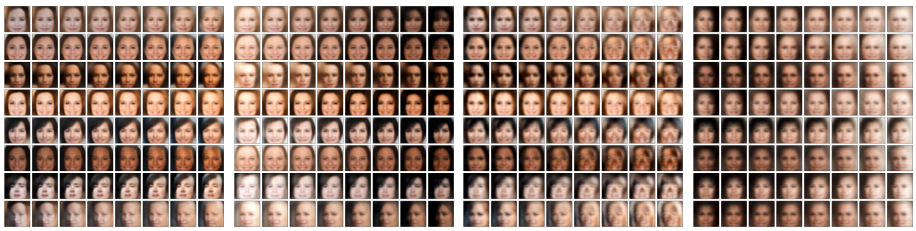
Table 2: β-VAE Metric Scores on dSprites averaged over 15 trials. Higher is better, and the highest scores of all models and hyperparameter configurations are in bold. Optimal λ values for disentanglement are different for this dataset because the dSprites image data is not normalized, following the precedent of previous works [10]

NashAE	$\lambda = 0.0$	$\lambda = 0.001$	$\lambda = 0.002$
	91.41%	92.58%	**96.57%**
β-VAE	$\beta = 1$	$\beta = 4$	$\beta = 8$
	84.63%	93.68%	**96.21%**
β-TCVAE	$\beta = 1$	$\beta = 2$	$\beta = 4$
	84.64%	95.01%	93.95%

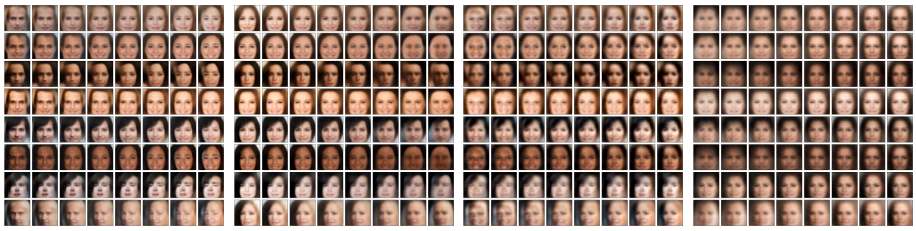
Latent Traversals and TAD Metric on CelebA



Bangs NashAE $\lambda = 0.0$; Lat16 (0.788) | ***Bangs*** NashAE $\lambda = 0.2$; Lat2 (0.831) | ***Bangs*** β-VAE $\beta = 1$; Lat3 (0.701) | ***Bangs*** β-VAE $\beta = 100$; Lat19 (0.724)



Blond NashAE $\lambda = 0.0$; Lat28 (0.820) | ***Blond*** NashAE $\lambda = 0.2$; Lat12 (0.807) | ***Blond*** β-VAE $\beta = 1$; Lat18 (0.765) | ***Blond*** β-VAE $\beta = 100$; Lat11 (0.768)



Male NashAE $\lambda = 0.0$; Lat5 (0.664) | ***Male*** NashAE $\lambda = 0.2$; Lat15 (0.697) | ***Male*** β-VAE $\beta = 1$; Lat6 (0.645) | ***Male*** β-VAE $\beta = 100$; Lat6 (0.632)

Fig. 4: Traversals of latent features corresponding to the highest AUROC score for the ***bangs***, ***blond***, and ***male*** attributes for the different disentanglement algorithms. Each latent representation with maximum AUROC score and its corresponding score are reported

We include traversals of latent representations that have the highest AUROC detector score for a small set of attributes on CelebA in Figure 4. In each case, we start with a random image from the dataset and hold all latent representations constant except the one identified to have the highest AUROC score for the attribute of interest. We vary that representation evenly from its minimum to

its maximum (as observed across 1000 random samples) and decode the resulting latent representation to generate the images reported in Figure 4.

Note that in all cases, employing disentanglement methods (NashAE $\lambda > 0$ or β-VAE $\beta \gg 1$) leads to a visual traversal that intuitively matches the attribute that the latent representation is a good detector for. Furthermore, the visual changes are significant and obvious. Contrarily, when there is no effort to disentangle the representations ($\lambda = 0$ or $\beta = 1$), the relationship between the representation's high AUROC score and its traversal visualization become far less clear. In some cases, the traversal does not make meaningful change or even causes odd artifacts during decoding. We hypothesize that this is due to redundant information being shared between the latent features, and changing just one may have either no significant effect or the combination will be "out of distribution" to the decoder, leading to unnatural decoding artifacts. The idea that standard latents hold redundant information is supported by Figure 2, where the predictors establish a high average R^2 value on CelebA when $\lambda = 0$. We employ the TAD metric to quantify disentanglement on the CelebA dataset. Table 3 summarizes the TAD results and number of captured attributes for each of the algorithms averaged over three trials. An attribute is considered *captured* if it has a corresponding latent representation with an AUROC score of at least 0.75. The resulting scores indicate that the NashAE consistently achieves a higher TAD score, suggesting that its latent space captures more of the salient data characteristics (determined by the labelled attributes). Furthermore, NashAE achieves high scores over a broad range for $\lambda \in (0, 1)$. β-TCVAE performs second best, achieving a TAD score of 0.446 when $\beta = 15$, yet it does not capture as many attributes as NashAE. In general, β-VAE and FactorVAE tend to capture fewer attributes and score lower TAD scores, suggesting that their latent spaces capture fewer of the salient data characteristics.

Table 3: TAD Scores on CelebA (averaged over 3 trials). Higher TAD scores are better, and the highest average score is in bold

Method	TAD	# Attributes
NashAE $\lambda = 0$	0.235	5.33
NashAE $\lambda = 0.1$	0.362	4
NashAE $\lambda = 0.2$	**0.543**	5
NashAE $\lambda = 0.8$	0.474	5
β-VAE $\beta = 1$	0.158	3.67
β-VAE $\beta = 50$	0.287	2.67
β-VAE $\beta = 100$	0.351	2.33
β-VAE $\beta = 250$	0.307	2
β-TCVAE $\beta = 1$	0.165	3.33
β-TCVAE $\beta = 8$	0.359	4
β-TCVAE $\beta = 15$	0.446	4.33
β-TCVAE $\beta = 25$	0.403	3.67
β-TCVAE $\beta = 50$	0.362	3.67
FactorVAE $\beta = 1$	0.188	3
FactorVAE $\beta = 8$	0.208	2.33
FactorVAE $\beta = 15$	0.285	3
FactorVAE $\beta = 50$	0.276	3
FactorVAE $\beta = 75$	0.148	1.33

5 Discussion

We have shown with our quantitative experiments that NashAE can reliably extract disentangled representations. Furthermore, qualitative latent traversal inspection indicates that the latent variables of NashAE which are the best detectors for a given attribute indeed visually reflect independent traversals of the attribute. Hence, the adversarial covariance minimization objective presented in this work promotes learning of clarified, interpretable representations in neural networks. We believe that improvements in neural network interpretability can aid engineers in diagnosing and treating the current ailments of neural networks such as security vulnerability, lack of fairness, and out-of-distribution detection.

Future work will investigate more sophisticated latent distribution modeling and to make NashAE a generative model. This could further boost NashAE's disentanglement performance and provide deeper insight with information-theoretic approaches. It could be interesting to apply the adversarial covariance minimization objective to clarify the representations of DNNs for image classification.

6 Conclusion

We have presented NashAE, a new adversarial method to disentangle factors of variation which makes minimal assumptions on the number and form of factors to extract. We have shown that the method leads to a more statistically independent and disentangled AE latent space. Our quantitative experiments indicate that this flexible method is more reliable in retrieving the true number of data generating factors and has a higher capacity to align its latent representations with salient data characteristics than leading VAE-based algorithms.

Acknowledgements This research is supported, in part, by the U.S. Department of Energy, through the Office of Advanced Scientific Computing Research's "Data-Driven Decision Control for Complex Systems (DnC2S)" project. Additionally this research is sponsored by the Artificial Intelligence Initiative as part of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL). This research used resources of the Experimental Computing Laboratory (ExCL) at ORNL.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

This research is further supported by US Army Research W911NF2220025 and the National Science Foundation OIA-2040588.

References

1. Alvi, M., Zisserman, A., Nellåker, C.: Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
2. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599 (2018)
3. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. Advances in neural information processing systems **31** (2018)
4. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 2180–2188 (2016)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. He, Z., Rakin, A.S., Li, J., Chakrabarti, C., Fan, D.: Defending and harnessing the bit-flip based adversarial weight attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14095–14103 (2020)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
10. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework (2016)
11. Hu, Q., Szabó, A., Portenier, T., Favaro, P., Zwicker, M.: Disentangling factors of variation by mixing them. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3399–3407 (2018)
12. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Kulkarni, T.D., Whitney, W., Kohli, P., Tenenbaum, J.B.: Deep convolutional inverse graphics network. arXiv preprint arXiv:1503.03167 (2015)
15. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848 (2017)
16. Li, D., Zhao, D., Zhang, Q., Chen, Y.: Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. IEEE Computational Intelligence Magazine **14**(2), 83–98 (2019)
17. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
18. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)

19. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
20. Lu, X., Tsao, Y., Matsuda, S., Hori, C.: Speech enhancement based on deep denoising autoencoder. In: Interspeech. vol. 2013, pp. 436–440 (2013)
21. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv preprint arXiv:1511.05644 (2015)
22. Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., Glorot, X.: Higher order contractive auto-encoder. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 645–660. Springer (2011)
23. Ross, A.S., Doshi-Velez, F.: Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In: Thirty-second AAAI conference on artificial intelligence (2018)
24. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
25. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
26. Torfi, A., Shirvani, R.A., Keneshloo, Y., Tavaf, N., Fox, E.A.: Natural language processing advancements by deep learning: A survey. arXiv preprint arXiv:2003.01200 (2020)