Extending the Adversarial Loss Function of Soft-Intro VAE for Stronger Disentaglement

Interdisciplinary Project Report

Matthias Matt
01529399

November 17, 2022

Supervisor: **Hubert Ramsauer**, MSc.
Co-Supervisor: Assistant Prof. Dr.techn. **Manuela Waldner**, MSc.
Domain specific lecture: **Lecture Series: Introduction to Digital Humanities**

Abstract

Disentanglement is hard to achieve in unsupervised representation learning. It can be negatively affected by trying to improve the reconstruction quality of the generated output. To try to alleviate these problems, this project combines two approaches that improve disentanglement and reconstruction quality, specifically β-TCVAE[1] and Soft-Intro-VAE[2]. The hypothesis was that a model that uses a combined loss function of both approaches can retain the positive aspects of both. The results did not confirm the hypothesis and showed no improvement in disentanglement metrics and worse reconstruction results compared to Soft-Intro-VAE[2]. The code for this project is available on GitHub[1].

1 Introduction

Unsupervised representation learning of unstructured data - specifically images - is still an open problem. The goal is to find interpretable dimensions by training a model on a collection of data. Many methods have been proposed to solve this task with varying levels of complexity. It ranges from simple and well-known methods like PCA to complex neural network architectures like GANs.

The focus of this project are *Autoencoders (AE)*, which are neural networks that consist of an encoder and a decoder. The encoder reduces the input to a small vector representation, which is then used to reconstruct the input by the decoder. One application of AEs is image de-noising but they inherently are dimensionality reduction models. Only image data was considered for this project.

An extension of the basic Autoencoders are *Variational Autoencoders (VAEs)* where the intermediate representation from the encoder is forced to be a multivariate normal distribution by introducing a *Kulback-Leibler (KL)* divergence loss term. The KL divergence penalizes the model for clustering points far away in the latent space and therefore provides structure to the latent space that benefits disentanglement. Different factors are encouraged to be encoded in a single dimension of the latent space since using more dimensions will lead to a higher KL divergence. Several methods have been developed to improve upon this architecture.

In summary, disentanglement is an important feature of representation learning and helps interpretability and discoverability of the latent variables.

2 State of the Art

The Autoencoder model architechture itself has been extended in various ways. However, a useful property of the basic architecture is that they can be trained end-to-end with a single pass through the

[1] https://github.com/meffmadd/intro-tc-vae

network. Therefore, I focused on models where only the loss function is modified and the underlying architecture is unchanged.

The loss function of a basic β-VAE is defined as

$$\mathcal{L}_\beta = \sum_{n=1}^{N} (\mathbb{E}_q[\log p(x_n|z)] - \beta KL(q(z|x_n)\|p(z))) \tag{1}$$

and is referred to as evidence lower bound (ELBO) [1]. Here $\mathbb{E}_q[\log p(x_n|z)]$ is the expected log likelihood of an observation x_n based on the learned latent distribution $q(z|x_n)$ (reconstruction loss). The term $KL(q(z|x_n)\|p(z))$ refers to the *Kulback-Leibler* divergence of the learned distribution of the encoder and the the target distribution $p(z) \sim \mathcal{N}(0, 1)$ (regularization/KL loss). The hyperparameter β acts as a weighting parameter to put more emphasis on the regularization (higher value) or reconstruction quality (lower value).

Chen et al. [1] have modified the loss function of the basic VAE loss function by decomposing the KL term into three separate terms. They show that the KL divergence between the learned distribution $q(z|n)$ and the target distribution $p(z)$ can be expressed as:

$$\mathbb{E}_{p(x)}[KL(q(z|x)\|p(z))] = \underbrace{KL(q(z,n)\|q(z)p(x))}_{\text{Index-Code MI}} + \underbrace{KL(q(z)\|\prod_j q(z_j))}_{\text{Total Correlation (TC)}} + \underbrace{\sum_j KL(q(z_j)\|p(z_j))}_{\text{Dimension-wise KL}}, \tag{2}$$

where $q(z,n) = q(z|x_n)p(x_n)$ and $q(z) = \sum_{n=1}^{N} q(z|x_n)p(x_n)$. After this decomposition, only the total correlation term (TC) is scaled with a hyperparameter β instead of the whole KL divergence term to find statistically independent factors. This way, they could improve the disentanglement of the latent space representation compared to traditional VAEs and other methods like InfoGAN [3]. The resulting model is called β-TC-VAE.

Recently, Daniel & Tamar [2] introduced Soft-Intro VAE to improve the reconstruction results of VAEs. Their contribution is to introduce a min-max game between the encoder and decoder, which is very similar to a GAN with its generator and discriminator architecture.

The adversarial game between the encoder and decoder is set up by introducing different loss functions for each:

$$\mathcal{L}_{E_\phi} = \text{ELBO}(x) - \frac{1}{\alpha}\exp(\alpha\text{ELBO}(D_\theta(z))) \tag{3}$$

$$\mathcal{L}_{D_\theta} = \text{ELBO}(x) + \gamma\text{ELBO}(D_\theta(z)) \tag{4}$$

In this game, the encoder and decoder have different objectives. Both are incentivized to improve the overall ELBO of the input. However, in addition, the decoder's objective is to also minimize its own output of a generated sample from a random latent vector (ELBO($D_\theta(z)$)). Conversely, the encoder should maximize the ELBO of the decoder output for the random sample (again ELBO($D_\theta(z)$)), or in other words, it tries to fool the decoder. With this setup, better reconstruction results can be accomplished. For the evaluation, like the original implementation, the values for the hyperparameters are set to $\alpha = 2$ and $\gamma = 1$. The β hyperparameter was equal for the different ELBO terms.

3 Soft-Intro-TCVAE

The goal of this project was to combine the loss functions of TC-VAE and Soft Intro-VAE to determine if the benefits of each are retained. The hypothesis is that by using the decomposed KL loss of TC-VAE in the ELBO terms of the adversarial loss functions of Soft Intro-VAE this could be accomplished in a straightforward manner by simply changing the ELBO loss functions of Soft-Intro-VAE to the modified ELBO$'$ of β-TCVAE. We, therefore, end up with the following formulation of the loss functions:

$$\mathcal{L}_{E_\phi} = \text{ELBO}'(x) - \frac{1}{\alpha}\exp(\alpha\text{ELBO}'(D_\theta(z))) \tag{5}$$

$$\mathcal{L}_{D_\theta} = \text{ELBO}'(x) + \gamma\text{ELBO}'(D_\theta(z)) \tag{6}$$

4 Evaluation

To evaluate the performance of the models, all variations of VAEs (β-VAE, β-TCVAE, Soft-Intro-VAE and Soft-Intro-TCVAE) are trained on the *ARC Ukiyo-E faces* [4] and *dSprites* [5] datasets. The first was chosen to test a model's ability to learn complex images with intricate details. It contains images of faces extracted from Japanese woodblock prints. The latter is popular for evaluating representation learning approaches and contains images of small white shapes over a black background. The images are generated from known factors that correspond to an identifiable attribute in the generated image (e.g. position or size of the shapes etc.) which should be learned by the model. There are 5 ground truth factors that are used to create the images. These factors are shape, size, rotation, X-position and Y-position.

For the size of the latent space was 48 dimensions for the *ARC Ukiyo-E faces* [4] and 10 dimension for the emphdSprites [5] dataset. For the *ARC Ukiyo-E faces* [4] dataset mean squared error (MSE) was used as the reconstruction loss function. With *dSprites* [5] binary cross-entropy loss was used since the values of the output are 0, 1, so the reconstruction is reduced to a classification problem.

For this project, the size of the *dSprites* [5] dataset was reduced by limiting the number of possible values for some factors. The angle of rotation was fixed to 4 values (0°, 90°, 180° and 270° respectively). The horizontal and vertical positions were also reduced from 32 to 10 possible values.

4.1 Evaluation Metrics

To empirically evaluate the model performance, several established metrics were used.

4.1.1 β-VAE Metric

Higgins et al. [6] introduced a novel evaluation metric in their β-VAE paper. It tests the accuracy of a simple classification model (logistic regression) to identify the known factors based on the latent space representation. Various pairs of batches are sampled so that one ground-truth factor is equal between them (i.e. the two batches show the same object shape in case of *dSprites*[5]). If some dimensions in the representations correspond to the random factor index then they will be equal in the representations as well which will be easy to predict by the logistic regression model. The range of the metric is between 0 and 1 where 1 is best.

4.1.2 Mutual Information Gap (MIG)

Chen et al. [1] also present an evaluation metric for disentangled representation learning. It is based on the mutual information between the factors and the latent representation. The difference between the two highest mutual information values is calculated for each factor. These values are then averaged to give the final metric. This means that the metric expects the model to express a single factor only in a single latent dimension. If two dimensions are used to express a given factor the mutual information for both dimensions will be high and the difference is going to be close to zero. The range of the metric is between 0 and 1 where 1 is best. However, since there were twice as many dimensions in the latent space than ground truth factors the best possible value in this scenario is 0.5.

4.1.3 Modularity and Explicitness

Another popular metric is the Modularity and Explicitness score which was devised by Ridgeway Mozer [7].

Modularity: Similar to the MIG metric the mutual information between the latent dimensions and factors is calculated. Ideally, a single latent dimension will have a single corresponding factor. For each latent dimension m we take the maximum mutual information with a factor and set the rest to 0. This would be the ideal state since one latent dimension only contributes to one factor. The metric is then calculated by averaging the deviations of the actual vector to the ideal vector. The range of the metric is between 0 and 1 where 1 is best.

Explicitness: This metric is similar to the β-VAE metric. This metric also tests the ability of a classifier (again logistic regression) to recover the factor value from the latent dimensions. However, this is done directly without any modifications to the latent representations. The range of the metric is between 0 and 1 where 1 is best.

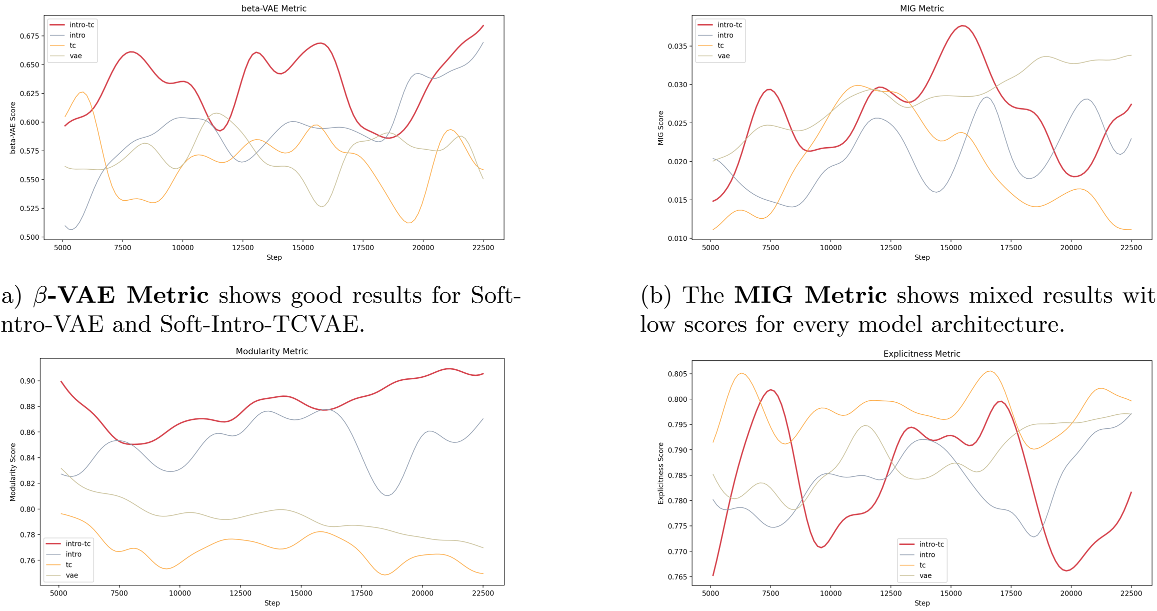
5 Results

The results are presented in this section. For the *dSprites* dataset, we can calculate the disentanglement metrics since we have the original ground-truth factors available to us. For the *ARC Ukiyo-E faces* dataset we will focus on visual inspection of the reconstruction results.

The models were trained for 200 epochs for both datasets and with equal weights for reconstruction and KL loss terms. This leads to a balanced weighting of the loss terms during later stages of training. Specifically, the loss weights were set to 8.0 for both the reconstruction and regularization loss and the following results are from a single training run.

5.1 dSprites

5.1.1 Metrics



(a) β**-VAE Metric** shows good results for Soft-Intro-VAE and Soft-Intro-TCVAE.

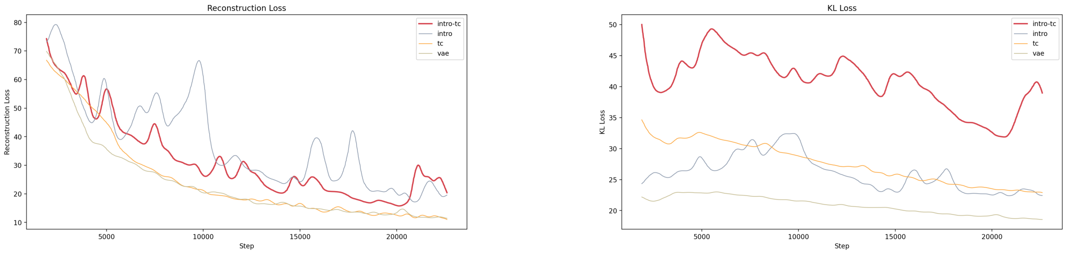
(b) The **MIG Metric** shows mixed results with low scores for every model architecture.

(c) The **Modularity metric** favors the models with introspective loss functions with Soft-Intro-TCVAE showing the best overall results.

(d) The **Explicitness metric** scores Soft-Intro-TCVAE lowest with large variability during training.

Figure 1: Evaluation metrics for the different models. Soft-Intro-TCVAE does not outperform the other models and performs very similarly to Soft-Intro-VAE.

5.1.2 Losses

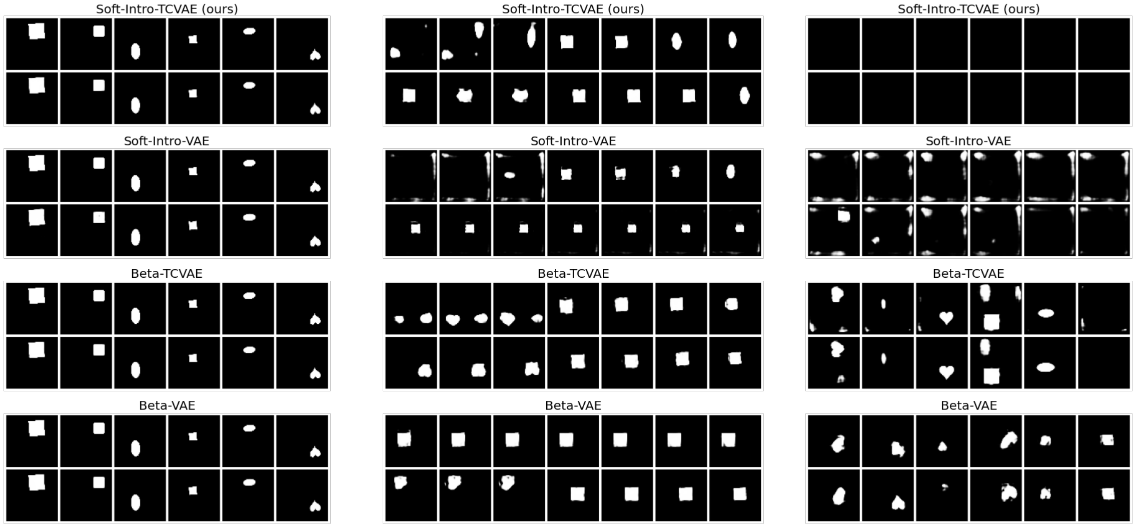


(a) **Reconstruction loss**

(b) **KL loss**

Figure 2: Loss functions of the models during training. Unscaled losses are used to compare the TC models to the non-TC models. As we can see, high KL loss does not directly translate to worse disentanglement metric scores. This makes it hard to compare the quality of the latent space when no ground-truth factors are available as is the case with the *ARC Ukiye-E faces* [4] dataset. The models with TC loss have higher losses since they only optimize for the decomposed loss term.

5.1.3 Reconstruction results



(a) Comparison between sample images (first row) and their reconstructions (second row). The results of the different models are nearly indistinguishable from each other.

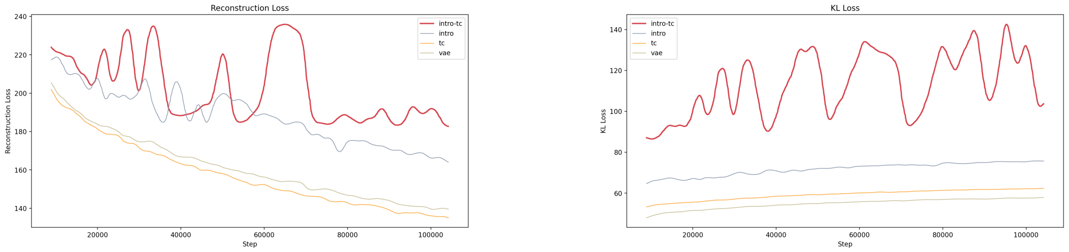
(b) Latent traversals of two latent vectors that were sampled from two images. The specific dimension that is interpolated was selected based on visual inspection and the specific results cannot be directly compared between the models.

(c) Out-of-sample reconstruction of a random vector with small and large standard deviations (first and second row) shows a large difference between the models. β-VAE and β-TCVAE show significantly better results than their adversarial counterparts with Soft-Intro-TCVAE showing no output in the sampling range.

Figure 3: Reconstruction results for sample images, latent traversals and out-of-sample latent vectors.

5.2 ARC Ukiyo-E faces

5.2.1 Losses

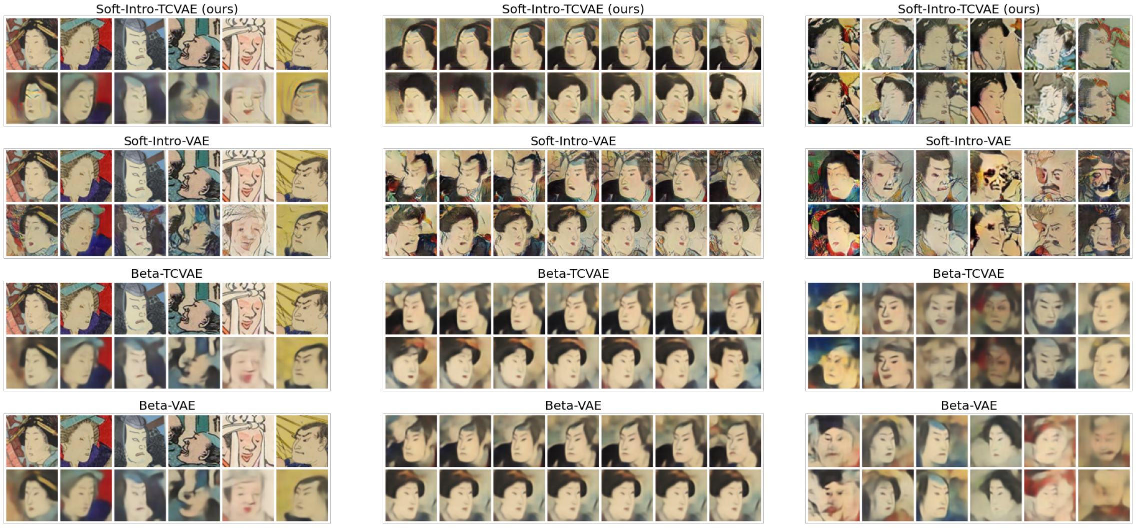


(a) **Reconstruction loss**

(b) **KL loss**

Figure 4: Loss functions of the models during training. Unscaled losses are used to compare the TC models to the non-TC models. The adversarial models have higher reconstruction loss which is contrary to their original goal. KL loss is very unstable for Soft-Intro-TCVAE.

5.2.2 Reconstruction results



(a) Comparison between sample images (first row) with their reconstructions (second row). Visually, Soft-Intro-VAE has significantly better results than the other models. This is contrary to the higher reconstruction loss values during training. Soft-Intro-TCVAE has blurry reconstructions similar to β-VAE and β-TCVAE and failed to retain the reconstruction quality of Soft-Intro-VAE.

(b) Latent traversals of two latent vectors that were sampled from two images. The specific dimension that is interpolated was selected based on visual inspection and the specific results cannot be directly compared between the models.

(c) Out-of-sample reconstruction of a random vector with small and large standard deviation (first and second row). Interestingly, Soft-Intro-TCVAE has comparable results with Soft-Intro-VAE for random vectors.

Figure 5: Soft-Intro-VAE has the best reconstruction results and Soft-Intro-TCVAE failed to retain the advantages of an adversarial loss function.

6 Conclusion

The goal originally set out for the project could not be accomplished. The disentanglement scores of Soft-Intro-TCVAE are very similar to Soft-Intro-VAE without any noticeable improvement. The reconstructions show no high-frequency details and are more similar to β-VAE and β-TCVAE in this respect. During training, Soft-Intro-TCVAE was more unstable compared to the other methods which might explain the results.

The results for *dSprites* [5] are very similar for all models. Surprisingly, the adversarial models had worse reconstruction results for the latent traversals and random vectors. Further hyperparameter tuning might alleviate this problem. Overall, β-TCVAE showed the best results for *dSprites* [5].

Soft-Intro-VAE outperforms the other models for the *ARC Ukiyo-E faces* [4] dataset. The reconstruction loss is higher than β-VAE and β-TCVAE models, however, this is very likely because of the higher frequency details in the reconstructions of Soft-Intro-VAE. Soft-Intro-TCVAE shows similar results to β-VAE and β-TCVAE which means that for complex datasets the combined loss loses any advantage of the adversarial loss function of Soft-Intro-VAE.

References

[1] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018.

[2] T. Daniel and A. Tamar, "Soft-introvae: Analyzing and improving the introspective variational autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4391–4400, June 2021.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[4] Y. Tian, T. Clanuwat, C. Suzuki, and A. Kitamoto, "Ukiyo-e analysis and creativity with attribute and geometry annotation," in *Proceedings of the International Conference on Computational Creativity*, 2021.

[5] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dsprites: Disentanglement testing sprites dataset." https://github.com/deepmind/dsprites-dataset/, 2017.

[6] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[7] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," *CoRR*, vol. abs/1802.05312, 2018.