The L_P Autoencoder: Disentangling without variation

Rowan Callahan[1]

[1] *Cornell University Department of Biomedical Engineering*

We propose a novel method for autoencoding to encourage discrete representations in the codes generated by the autoencoder. This is achieved by applying an additional loss function that penalizes representations of images that have more than one non zero variable in the latent representation. We show that maximizing both reconstruction loss and this additional loss provides a way to create more interpretable examples when some of the underlying data is discretely distributed. Advantages exist with using an autoencoder for this purpose over a GANN because autoencoders have clear benchmarks that can be used such as reconstruction loss as opposed to GANNs. Finally this method provides a way to view examples of the classes that have been disentangled from each other.

I. INTRODUCTION

Disentangled representations are an important area of machine learning research. Recently much progress has been made with creating disentangled representations of various images. Information maximizing variational autoencoders are one method that is able to provide disentangled representations, however there are a few drawbacks. One of these drawbacks is that the latent representations are not always represented in a format that is easy for humans to understand. Many [1] attempts have been made to factorize these representations in a way that interpretable by humans. However, not all of these methods are simple to understand and it is not always clear why they work[2].

Most of these methods also use either the reparametrization trick in the case of variational auto encoders, or another objective function as is seen in GANNs that does not necessarily have a simple metric to evaluate. We show that class separation can be achieved with zero variational reparametrization and that in some cases this is in fact preferrable.

We present a usage of the L_P norm to encode information. Specifically Where $0 < P < 1$ We also show that while this norm is not convex, the minima of this objective function in an autoencoder hasshifts the latent space of the encoder towards the maximum "correct" variance that is expressed in the information that we are attempting to encode, when paired with a reconstruction loss.

We show that applying this norm allows for an easily differentiable objective function that has many desirable properties. Furthermore we provide code that can be used as a drop in replacement for certain properties of variational autoencoding.

II. INTUITION BEHIND THE METHOD

Methods like variational autoencoders and especially information maximizing variational auto encoders look for disentangled representations of data distributions. However there are often issues with how these different disentangled representations are displayed.

Often the issues with variational auto encoders is that they can find a local minimum where everything is viewed as noise passed a mean of zero [3]. This issue can be solved with Lp norms and sparse coding are a common tool used to address this, but often have issue because they define non-convex optimization problems [4] However, because optimization of Neural networks is already a non convex method, an addition of another non convex optimization is not an issue.

Nevertheless learning disentangled representations remains a difficult task in unsupervised learning. [5] [6] [7]

Our method is very similar to variational encoding in that we are predicting a maximum likelihood estimator for latent variables with an assumption they fall within an expected distribution [8]. However we show that this is not necessary given good choices of parameters.

A. The 2 dimensional case

In the 2 dimensional case one can easily imagine a signal that is coming in to be processed by the neural network. We know that this signal can be represented discretely as two separate classes. We also have two separate latent

variables that we are using for our representation space. When picking a loss function we want something that has the following properties.

1. it is differentiable
2. it penalizes significantly less along the bases of the latent space

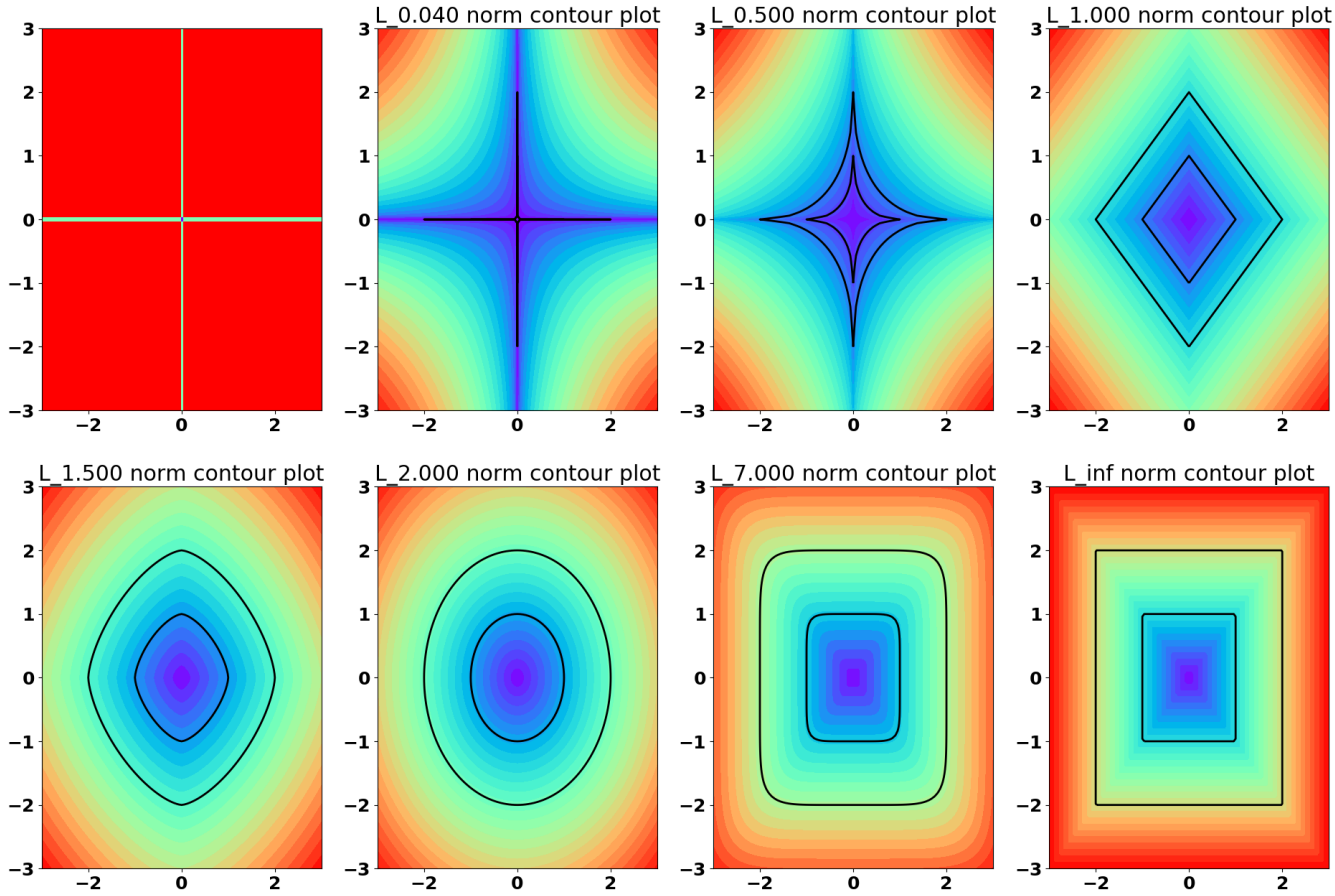

FIG. 1. various Lp norms

In this paper we pick the fractional P, $0 < P < 1$ norm because while it not only has all of the following desireable properties, it is also able to be relaxed and tightened depending on the P parameter that is chosen by the user. We use this norm because it penalizes all variation that does not align with one of the axes in the latent dimensions Z. This means that for each image that is picked, the minima are all inside the contours that are shown in 1.

For datasets where most of the data is described as being part of one of N many discrete classes the L_P norm actually improves represenatation above simple using reconstruction loss.

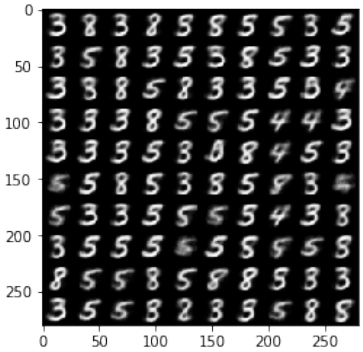

FIG. 2. Example of output after training with 2 latent variables

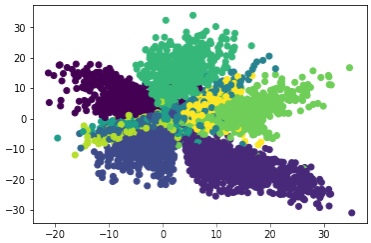

FIG. 3. separation with 2 latent variables, each color is a separate class

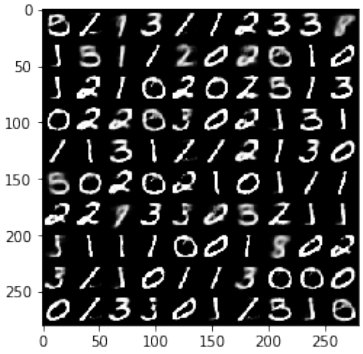

FIG. 4. Example of output after training with 2 latent variables

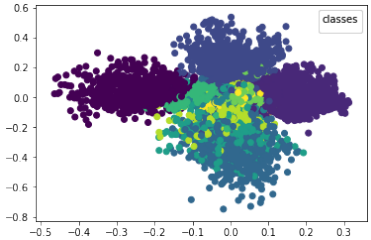

FIG. 5. separation with 2 latent variables, each color is a separate class

B. Expected maximum variance of data containing discrete classes

Given balanced classes of discrete variables, variables that are either zero or one, it is possible to think of encoding strategies that work. One popular example is one hot encoding where only one element in a vector is one and all of the other elements are zero. Minimizing the P "norm" can provide an approximation of this by forcing only one of the values to have an absolute value that is greater than zero. It is possible to force positive values but this is not necessary to get our desired effect.

In fact our ideal latent variable distribution is the "P norm" where p is equal to zero. This distribution has already been previously characterized and is equivalent to the gaussiandistribution for p=2 [9] In this norm shown in 1 all of the values concentrate along the when it is taken to zero. This means that only one of the variables has a non zero value. fitting the latent representation into this shape is equivalent to encoding most of the information of each variable to one single axis. This type of encoding has also been found to be useful in k means clustering.

This paper takes two relaxations of this idea in order to get a workable method that gains some of the desirable properties of this "norm". First we relax our norm so that we allow for some small values along the axes. We can show that the Kullback-Leibler divergence is still quite small between the latent space distribution that we care about the l_0 norm.

III. RESULTS

IV. USAGE AND APPLICATION

We define the L_P norm as follows for all vectors $\hat{x}$ where

$$\hat{x} \in \mathbb{R}^y, \; y \in \mathbb{N}$$

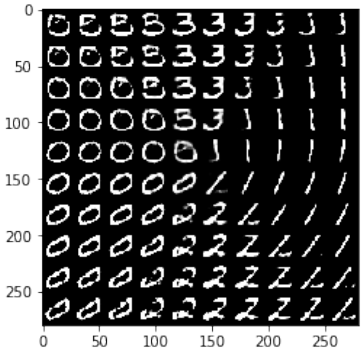

FIG. 6. P normed latent traversal

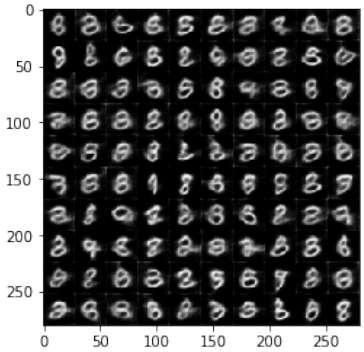

FIG. 7. Non P normed traversal

The L_P norm is written as

$$L_P = ||x_1 ... x_y||_p$$

where $||x_1 ... x_y||$ is defined as

$$||x_1^p \ ... \ x_y^p|| \ \forall \ p > 0$$

In the following experiments we define our loss with respect to the reconstruction loss and the fractional P norm of the latent variables in our model.

we define our loss as follows

$$\hat{z} = latent - variables$$

$$\hat{y} = input - tensor$$

$$\hat{x} = output - tensor$$

$$\mathcal{L}_{combined} \ = \ \alpha \ ||z||_p \ + \mathcal{L}_{reconstruction}$$

where the reconstruction loss can be defined by the user a common loss that is used as in **??**

$$\mathcal{L}_{reconstruction} = (\hat{x} - \hat{y})^2$$

During experimentation we often find that it is necessary to set a very low alpha parameter the larger the latent space, as the L_P norm penalizes points that lie on the edge extremely highly, and this is unimportant at the beginning of trianing where parameters do not need to be so finely tuned. In fact we find with large α parameters set reconstruction loss never ends up going as a priority is set instead on reducing the L_P loss, and instead the solution is often trapped in local minima.

ACKNOWLEDGMENTS

We thank Juan Felipe Beltran for his assistance in editing and writing this manuscript.

V. REFERENCES

[1] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK," p. 22, 2017.

[2] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," *arXiv:1804.03599 [cs, stat]*, Apr. 2018. arXiv: 1804.03599.

[3] S. Zhao, J. Song, and S. Ermon, "InfoVAE: Information Maximizing Variational Autoencoders," *arXiv:1706.02262 [cs, stat]*, June 2017. arXiv: 1706.02262.

[4] H. Ji, Y. Quan, and Z. Shen, "0 norm based dictionary learning by proximal methods with global convergence," p. 8.

[5] A. Ruiz, O. Martinez, X. Binefa, and J. Verbeek, "Learning Disentangled Representations with Reference-Based Variational Autoencoders," *arXiv:1901.08534 [cs]*, Jan. 2019. arXiv: 1901.08534.

[6] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating Sources of Disentanglement in VAEs," p. 11.

[7] H. Kim and A. Mnih, "Disentangling by Factorising," *arXiv:1802.05983 [cs, stat]*, Feb. 2018. arXiv: 1802.05983.

[8] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv:1312.6114 [cs, stat]*, Dec. 2013. arXiv: 1312.6114.

[9] F. Sinz, S. Gerwinn, and M. Bethge, "Characterization of the p-generalized normal distribution," *Journal of Multivariate Analysis*, vol. 100, pp. 817–820, May 2009.