Analyzing Information Loss of Invertible Variational Autoencoders

Jake Callahan, Taylor Paskett

February 20, 2021

Abstract

We want to understand the inherent information loss of invertible variational autoencoders. We find that ...

1 Introduction

As we gain access to more and more unstructured data, unsupervised learning methods become more and more vital. Among existing unsupervised deep learning models, variational autoencoders have become prominent. An autoencoder uses an unlabelled dataset to create more compactly encoded representations of the dataset. This is done by introducing a bottleneck to the architecture, as the following image (courtesy of [3]) depicts:

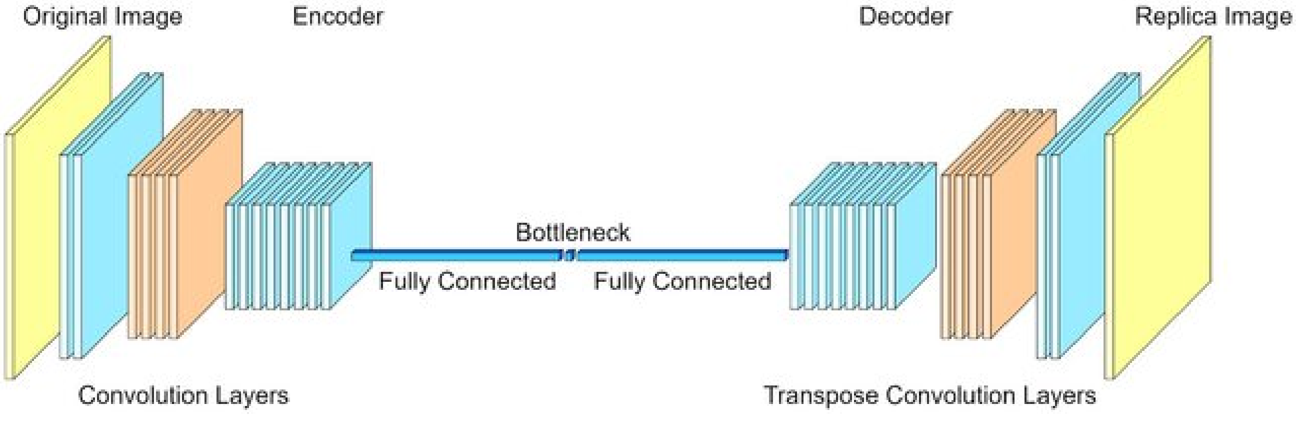

Figure 1: Illustration of Bottleneck in Autoencoder

Traditionally, autoencoders are trained by requiring that the original image and the replica image are nearly the same, measured using a reconstruction loss [1]. After training, we can extract an encoded version of a test image by inputting the test image to the encoder and extracting the value at the bottleneck layer.

More recently, information-theoretic concepts such as mutual information and entropy have been used to better understand autoencoders [5]; and in some cases, information-theoretic concepts can even be used in place of gradient descent to train neural networks [2].

Since information theory helps us understand and train autoencoders, we are interested in studying the information-theoretic properties of a new type of autoencoder that is built using invertible neural networks (INNs).

2 Invertible Autoencoders

We examine the work of Nguyen, Ardizzone, and Köthe, in their paper, ”Training Invertible Neural Networks as Autoencoders” [4].

In the following figure from Nguyen et al's work, they explain the fundamental building block of INNs: the invertible coupling layer. Just like a ResNet is built by stacking residual blocks, INNs are built by stacking invertible coupling layers. Theoretically, an INN can be built to any depth by stacking any number of invertible coupling layers.

Although traditional neural networks are universal function approximators, INNs are limited to approximating bijective functions. The question you might be asking, then, is this: if INNs can only approximate bijective functions, how can an INN become an autoencoder? Since the point of an autoencoder is to encode data in a smaller-dimensional latent space, a bijection won't work, right?

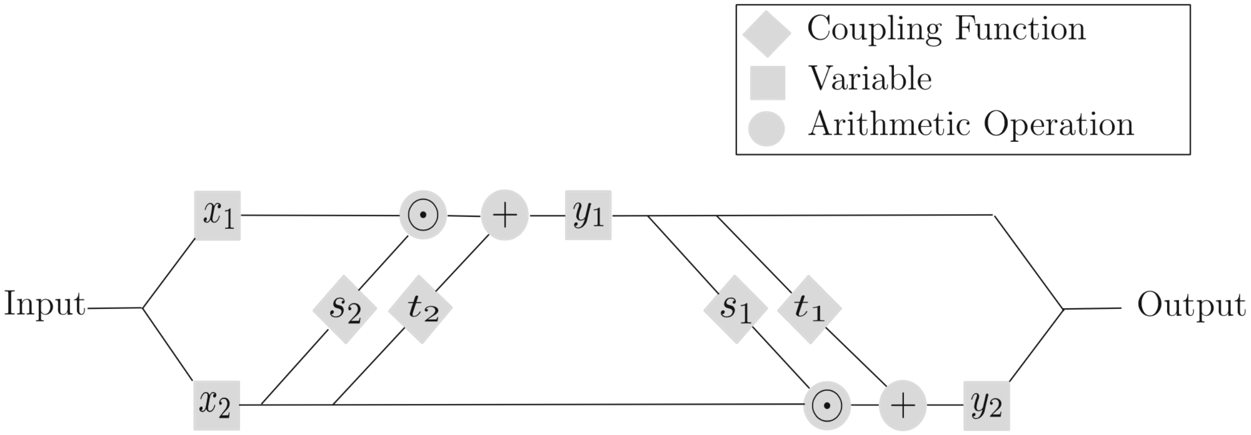

Fig. 1. Visualization of the invertible coupling layer

The main contribution of Nguyen, et al. was to come up with a creative solution to this problem. They do so by ”zero-padding” the output of the INN. For example, suppose our inputs have a dimension of 1000, and we wish to create a bottleneck of size 20. To train our INN, we give it the input x, obtaining $\text{INN}(x) = y$. Then, we replace all but the first 20 entries of y with zero, call this new output $\hat{y}$. Finally, we run the inverse model: $\text{INN}^{-1}(\hat{y}) = \hat{x}$. Then, we seek to minimize the reconstruction loss

$$\mathscr{L}(x, \hat{x}).$$

In so doing, we have effectively created a bottleneck. The zero-padded $\hat{y}$ has only 20 nonzero entries, and these entries can be thought of as our latent space encoding.

Nguyen et al. obtained very promising results. Testing using MNIST, CIFAR-10, and CelebA, they showed that, compared to traditional autoencoders,

1. INNs required less training epochs;
2. INNs could achieve smaller reconstruction losses across most bottleneck sizes;
3. INNs required less trainable parameters than their classic autoencoder counterparts.

The reason for 3 is primarily because the INN architecture is constant with respect to the bottleneck size, where classic autoencoders must change the number of trainable parameters depending on the bottleneck size. In their discussion of their results, they hypothesize the following

> We already established, that if a DNN does not learn a bijective function, information loss occurs during the forward process making the inverse process ambiguous. The INN solves this ambiguity problem by introducing latent variables z containing all the information lost during the forward process.... Therefore, we hypothesize that INNs have no intrinsic information loss contrary to DNNs and the findings of Yu et al. do not apply to INNs. In other words, INNs are not bound to a maximal number of layers (depth) after which only suboptimal results can be achieved.

This is a bold hypothesis, and we will investigate it using information theory. Primarily, we seek to answer the following questions:

- How much information entropy is gained or lost at each layer of the INN?

3 Methods, Results, Discussion

4 Conclusion

References

[1] Carl Doersch. Tutorial on variational autoencoders, 2021.

[2] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2019.

[3] Lukas Hollenstein, Lukas Lichtensteiger, Thilo Stadelmann, Mohammadreza Amirian, Lukas Budde, Jürg Meierhofer, Rudolf M. Füchslin, and Thomas Friedli. Unsupervised learning and simulation for complexity management in business operations. In *Applied Data Science*, pages 313–331. Springer International Publishing, 2019. doi: 10.1007/978-3-030-11821-1_17. URL https://doi.org/10.1007/978-3-030-11821-1_17.

[4] The-Gia Leo Nguyen, Lynton Ardizzone, and Ullrich Köthe. Training invertible neural networks as autoencoders. In *Lecture Notes in Computer Science*, pages 442–455. Springer International Publishing, 2019. doi: 10.1007/978-3-030-33676-9_31. URL https://doi.org/10.1007/978-3-030-33676-9_31.

[5] Shujian Yu and Jose C. Principe. Understanding autoencoders with information theoretic concepts, 2019.