Invariance in Variational Auto-Encoders: Evaluation of Latent Representation for Image Classification

Joseph Mills

2024-11-05

1. Introduction

Model invariance is an important feature in computer vision tasks. A function $f(\boldsymbol{x})$ of an input $\boldsymbol{x}$ is invariant to a transformation $t(\boldsymbol{x})$ if $\boldsymbol{f}(\boldsymbol{t}(\boldsymbol{x})) = \boldsymbol{f}(\boldsymbol{x})$. Number plate recognition from on-street cameras is a computer vision task that highlights the importance of invariance. Despite generally consistent typeface, cameras will capture images at numerous distances and angles. This non-conformity in presentation will require good levels of invariance to ensure the model generalises well.

Figure 1 Example transformations of input image

It is widely known that Convolutional Neural Network's (CNN's, convnets) induce partial invariance to translation through pooling between layers [2,6]. Further CNN development, with the introduction of G-transformations, provide a convolutional layer that largely increases performance on MNIST-rot, the rotated MNIST set [2]. This performance makes them a popular choice for image classification tasks.

Despite the lucrative performance of CNN's, the expensive acquisition of data labelling and collection required to effectively train these networks often makes the process unfeasible. Variational Auto-Encoders (VAE's) [5], however, are a type of generative model consisting of an encoder and decoder. Unlike the discriminative nature of CNN's, VAE's model the underlying probability distributions of the data creating a continuous latent space making VAE's and generative architectures an area of growing interest for classification tasks when extensive data or data labels are unavailable [7].

In this report, VAE's invariance is explored through varying the size of the latent space and through training on transformations of the underlying training data.

2. Methods

2.1 Data

The standard MNIST [3] database and three addition transformations of the train and test set have been used to evaluate the invariance of VAE's (*Table 1*).

Table 1 Data Sets for Training and Evaluation

Data Set	Tag	Rotation	Scale
MNIST-s	*Standard*	0°	1x
MNIST-fr	*Fixed Rotation*	+20°	1x
MNIST-fs	*Fixed Scale*	0°	1.1x
MNIST-rt	*Random Transformation*	-20°→ +20°	0.9x → 1.1x

2.1 VAE

A standard VAE architecture was used (*Figure 1*). The encoder consisted of 1 linear fully connected layer with ReLU activation, and the decoder consisted of 2 linear full connected layer with ReLU and Sigmoid activation functions respectively.

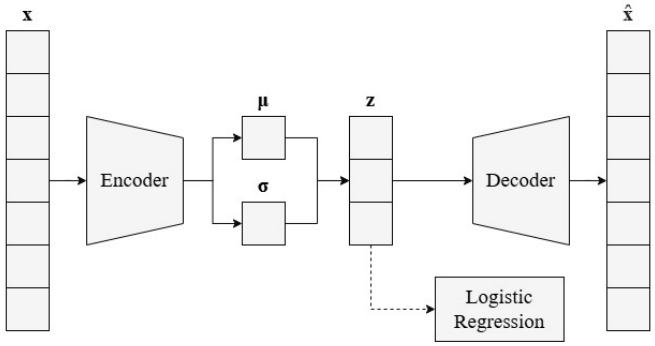

Figure 2 Simplified VAE Methodology

2.2 Logistic regression

A simple logistic regression model was trained on the latent variable output of the encoder after training of the VAE.

3. Results

Table 2 displays the results of the architecture training on MNIST-s and MNIST-rt with varying sizes of latent variables. Unsurprisingly, the model trained on MNIST-s with 10 latent variables performed the best overall on the MNIST-s test set. The model trained on MNIST-rt performed best overall on the three other transformed test sets.

Table 2 VAE Results

	Latent Variables	Test Set Accuracy			
Train Data	N	MNIST-s	MNIST-fr	MNIST-fs	MNIST-rt
MNIST-s	10	**0.8799**	0.6846	0.8228	0.8067
	20	0.8564	0.6721	0.7624	0.7713
	50	0.8547	0.6739	0.7737	0.7837
	100	0.8550	0.6826	0.7778	0.7742
MNIST-rt	10	0.8632	**0.7386**	**0.8238**	0.8070
	20	0.8504	0.7289	0.7734	0.7951
	50	0.8561	0.7334	0.7906	**0.8089**
	100	0.8563	0.7378	0.7965	0.7983

Interestingly, the models that have performed the best, or amongst the best, have the least number of latent variables (10). This could be an effect of the Kullback-Leibler component, used in the reconstruction loss of the VAE, where when working in high-dimensional latent spaces the network learns representations below the network capacity [1]. This is often referred to as overpruning [1].

4. Discussion

The probabilistic nature of the VAE required ensuring a level of reproducibility in running the experiments. Setting the random seed proved difficult and despite achieving this feature when re-initialising the computational kernel, it was evident that the results can still be volatile. In future work it would be prudent to both ensure the seed is set and to also repeat each experiment numerous times to calculate the mean and standard deviation prior to comparing results between the different experiments.

This report shows that baseline VAE's architecture can show a level of invariance to rotations and transformations of scale; exploring and comparing how other model variations and transformations would prove valuable. β-VAE's, for example, perform incredibly well on learning disentangled representations where singular latent variables are sensitive to changes in generative factors [4]. This feature of β-VAE's increases the VAE's ability to be invariant to transformation.

In future work, also, designing the experiment to show VAE's performance on reduced and more complex training data would be highly valuable to subsequently compare the performance to that of CNN's when presented with limited training data.

5. Conclusion

From the experiments completed, training a VAE on randomly transformed data performs far better across various transformed data sets whilst maintaining comparable performance to the VAE trained on the standard training data. By increasing the size of the latent variables in the experiments, the desire was that a greater representation of the data could be learned. However, although this may be true to some degree, the step-size between the latent variables may well have been too large and resulted in overpruning. It may be that the size of the latent variables as a hyperparameter in training the VAE can improve the models' ability to be invariant however it was not possible to evidence this in the experiments, as intended.

6. References

[1] A. Asperti. Sparsity in Variational Autoencoders. In ASPAI, 2019.

[2] T. S. Cohen and M.Welling. Group Equivariant Convolutional Networks. In PMLR, 2016.

[3] L. Deng. The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 2012.

[4] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework . In ICLR, 2017.

[5] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In ICLR, 2013.

[6] S. J. D. Prince. Understanding Deep Learning. MIT Press, 2024.

[7] L. Yang, W. Fan, N. Bouguila. Robust unsupervised image categorization based on variational autoencoder with disentangled latent representations. Knowledge-Based Systems, Volume 246, 2022.