# Improving Image Autoencoder Embeddings with Perceptual Loss[⋆]

Gustav Grund Pihlgren, Fredrik Sandin, and Marcus Liwicki

EISLAB, Luleå University of Technology, Luleå, Sweden
firstname.lastname@ltu.se

**Abstract.** Recently, Deep Perceptual Loss (DPL) has been used to improve autoencoders. However comparisons of autoencoders trained with the standard pixel-wise loss to those trained with DPL for feature embedding are lacking. This work mitigates this by training autoencoders with both losses on three different image datasets. The value of the embeddings is tested by training a host of Multi-Layer Perceptrons for prediction using the embeddings as input. The results show that training with DPL leads to embeddings that enable more accurate prediction. The experiments are available online. [1]

**Keywords:** Autoencoder, Deep Perceptual Loss, Feature Extraction

## 1 Introduction

The autoencoder is a type of neural network used for feature learning that has been in use for decades. Autoencoders work by having an encoder part that tries to embed the data into a lower dimensionality latent space, and a decoder that tries to reconstruct the data from the embeddings. Autoencoders are commonly trained with pixel-wise loss, which is some function of the sum of differences between the predicted pixels and their corresponding target pixels. A problem with pixel-wise loss is that it does not take into account the higher-order structures in the data. The use of loss based on Deep Perceptual Loss (DPL) mitigates this problem [3]. With DPL the predicted image and the target image are not compared directly. Instead an additional neural network, called the loss network, is used to extract features from both images. The loss is then calculated as some function of the sum of differences between the features of the predicted image and the corresponding features of the target image. DPL has been successful for better image generation. This work investigates if these benefits translate to the embeddings created by the autoencoders. This is done by training autoencoders (AE) and variational autoencoders (VAE) with both methods and comparing the embeddings for how useful they are for downstream supervised Multi-Layer Perceptrons (MLP). This is done for three datasets: A collection of images from LunarLander-v2 of OpenAI Gym [1] (*LLv2*), *STL-10* [2], and *SVHN* [4]. *LLv2* has lander position and the other datasets have image class as supervised labels.

[⋆] Complete manuscript, by the same authors, has been accepted for presentation and publication at IJCNN 2020

[1] https://github.com/guspih/Perceptual-Autoencoders

## 2 Experimental setup

The loss network used in the experiments is AlexNet pretrained on ImageNet. Features for calculating DPL are extracted after the second ReLU layer. The DPL is calculated as the Mean Square Error (MSE) between the feature extraction with the original image as input and the feature extraction with the reconstructed image as input. Pixel-wise loss is calculated as simply MSE between the pixels of the original image and the reconstruction.

AEs and VAEs are trained with several different number of dimensions of the latent space. For each unique autoencoder one model is trained with pixel-wise loss (called AE and VAE) and one with DPL (called P. AE and P. VAE). The trained autoencoders are then used to embed the data for supervised training into the latent space. For each autoencoder a host of MLPs of varying architectures were trained to predict the label of the original images when given the embedding as input.

## 3 Results

Table 1 shows the performance on the test sets for the MLP on each dataset with the lowest validation loss. The performance on *LLv2* is the average L1 norm distance between the predicted lander position and actual lander position, and accuracy for the other datasets. Fig. 1 shows how well a decoder trained with pixel-wise MSE can reconstruct images from embeddings of encoders trained with different losses.

| Dataset | AE | VAE | P. AE | P. VAE |
|---|---|---|---|---|
| LLv2 | 13.22 | 12.70 | 1.44 | **1.15** |
| STL-10 | 45.0% | 42.4% | 63.4% | **64.8%** |
| SVHN | 81.9% | 82.7% | **84.0%** | 81.2% |

**Table 1.** The performance of the MLPs with lowest validation loss.

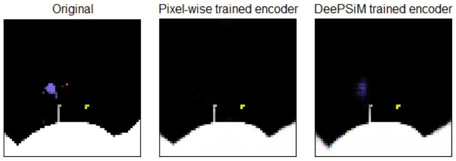

**Fig. 1.** Reconstructions based on pixel-wise and DPL trained embeddings.

## References

1. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym (2016)
2. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223 (2011)
3. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in neural information processing systems. pp. 658–666 (2016)
4. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011)