Project 1 : Auto-encoders

Prakhar Singh
ps32856

Prateek Chaudhry
pc26978

Shivam Garg
sg48957

Abstract

In this project, we build an autoencoder to compress images by approximately 99% using deep convolutional neural networks. The network was trained on diverse data comprising of canonical as well as non canonical images to increase robustness.

1 Introduction

We tackle the problem of compressing images using a deep convolutional autoencoder. An autoencoder is a type of neural network that learns useful encodings of input data in an unsupervised manner. It consists of two parts: (i) an encoder, which maps high dimensional input to a low dimensional encoding and (ii) a decoder, which attempts to reconstruct the original input from this low dimensional encoding.

2 Dataset

We built our dataset by randomly sampling images from multiple sources common in computer vision literature as listed below:

- ImageNet (Deng et al., 2009) - 15000 images
- Places2 (Zhou et al., 2014) - 15000 images
- LVIS (Gupta et al., 2019) - 5000 images
- Labeled Faces in the Wild (Huang et al., 2007) - 2500 images

These data sources were chosen so as to get a good mix of both canonical as well as non canonical images, which is critical for training a robust model.

3 Model Architecture

Our architecture consists of an encoder and decoder, as described in Table 1 and Table 2.

Our encoder is 11 layer deep network with residual connections. These layers can be grouped in terms of reducing layers and convolutional blocks. Each block has a skip connection across it.

The decoder is structured to be a mirror of encoder architecture and follows a similar pattern.

Layer	Output	Details
block_0	256x256x16	3x3, 16
reduce_0	128x128x16	3x3, 16
block_1	128x128x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$
reduce_1	64x64x16	3x3, 16
block_2	64x64x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$
reduce_2	32x32x16	3x3, 16
block_3	32x32x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$

Table 1: Encoder Architecture

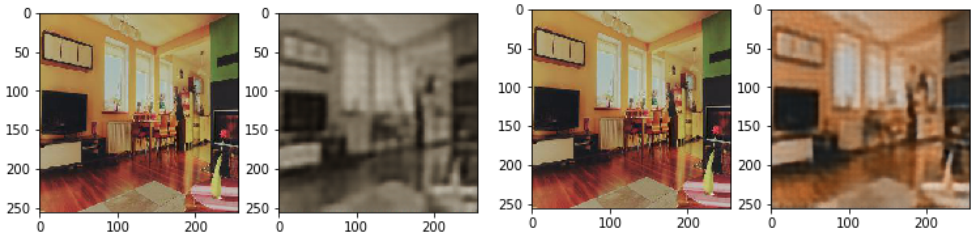
Layer	Output	Details
expand_0	32x32x16	3x3, 16
block_0	32x32x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$
expand_1	64x64x16	3x3, 16
block_1	64x64x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$
expand_2	128x128x16	3x3, 16
block_2	128x128x16	$\begin{bmatrix} 3x3, 16 \\ 3x3, 16 \end{bmatrix}$
expand_3	256x256x16	3x3, 16
block_3	256x256x3	3x3, 3

Table 2: Decoder Architecture

4 Experiments

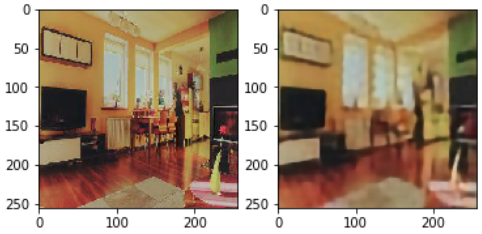
We use pytorch to build our autoencoder and minimize L1 loss. We used adam optimizer with learning rate of 1e-3 and trained for 200 epochs. For experiment purposes, we split data into 2 parts, with training set of 35000 images and remaining 2500 images as dev set. Various architectures were tried as listed in Table 3. The results we report are the per pixel L1 loss averaged across whole dev set, with pixel values represented as floats between 0 and 1. Finally we trained the model on the whole dataset.

In addition to loss we also looked at output images from our decoder which are shown in Fig-



(a) Epoch 1

(b) Epoch 8



(c) Epoch 200

Figure 1: Comparison of input image and reconstructed image across different training epochs

Experiment	Result
4 conv + 4 deconv	0.0523
6 conv + 6 deconv	0.0601
4 conv + 4 deconv + l2 loss	0.0623
Final architecture	**0.0477**

Table 3: Experiments

ure 1. In initial epochs, reconstructed image was quite coarse and without color. We observed color images around 8th epoch. From then on, training proceeded in a slow manner with loss slowly decreasing and finally saturating by 170th epoch.

References

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495.