ON REDUCING THE CORRELATION OF BOTTLENECK REPRESENTATIONS IN AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image compression is an important image processing task. Recently, there has been more interest in using autoencoders (AEs) to solve this task. An AE has two goals: (i) compress the original input to a low-dimensional space, at the bottleneck of the network topology, using the encoder (ii) reconstruct the input from the representation at the bottleneck using the decoder. Both parts are optimized jointly by minimizing a distortion-based loss which implicitly forces the model to keep only the variations in the input data required to reconstruct the input without persevering the redundancies. In this paper, we propose a scheme to explicitly penalize feature redundancies in the bottleneck representation. To this end, we propose an additional loss term, based on the pair-wise correlation of the neurons, which complements the standard reconstruction loss forcing the encoder to learn a more diverse and richer representation of the input. The proposed approach is tested using the MNIST dataset and leads to superior experimental results.

1 INTRODUCTION

Image compression is an important task in many applications. Recent advances in deep neural networks (Goodfellow et al., 2016) have enabled efficient modeling for the high-dimensional data and led to outperforming traditional compression techniques (Ullrich et al., 2017; Mentzer et al., 2020; Marcellin et al., 2000; Skodras et al., 2001; Rabbani & Jones, 1991) in image compression (Gregor et al., 2016; Toderici et al., 2017; Ballé et al., 2016). Recently, there has been interest in autoencoders (AEs) (Goodfellow et al., 2016) to solve this problem (Ollivier, 2014; Hu et al., 2020; Cheng et al., 2018; Theis et al., 2017; Rippel & Bourdev, 2017) due to their flexibility and easiness to train (Theis et al., 2017; Hu et al., 2020; Jiang et al., 2017; Yang et al., 2020).

AEs (Goodfellow et al., 2016) are a powerful data-driven unsupervised approach used to learn a compact representation of a given input distribution. AEs have been applied successfully in many applications, such as transfer learning (Deng et al., 2013; Zhuang et al., 2015; Kandaswamy et al., 2014), anomaly detection (Beggel et al., 2019; Zhao et al., 2017; Aygun & Yavuz, 2017; Zhou & Paffenroth, 2017), dimensionality reduction (Petscharnig et al., 2017; Thomas et al., 2016; Wang et al., 2014), and compression (Theis et al., 2017; Han et al., 2018; Golinski et al., 2020; Yingzhen & Mandt, 2018). To accomplish these tasks, an AE uses two different types of networks: The encoder $g(\cdot)$, which maps the input image $\boldsymbol{x} \in \mathcal{X}$ to a compact low-dimensional space $g(\boldsymbol{x})$, called the bottleneck representation, and the second part called the decoder $f(\cdot)$, which takes the output of the encoder as input and uses it to reconstruct the original image $f \circ g(\boldsymbol{x})$.

Given a distortion metric D: $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which measures the difference between the original input and the reconstructed input (Baldi, 2012; Deng et al., 2013), AEs are trained in an end-to-end manner using gradient descent (Goodfellow et al., 2016) to minimize the loss L defined as the average distortion over the training data $\{\boldsymbol{x}_i\}_{i=1}^{N}$:

$$\min_{f,g} L\big(\{\boldsymbol{x}_i\}_{i=1}^{N}\big) \triangleq \min_{f,g} \frac{1}{N} \sum_{i=1}^{N} D(\boldsymbol{x}_i, f \circ g(\boldsymbol{x}_i)). \tag{1}$$

Several extensions and regularization techniques have been proposed to augment this loss (Deng et al., 2013; Golinski et al., 2020; Theis et al., 2017; Cheng et al., 2018; Seybold et al., 2019) to improve the performance of the model. AE-based compression approaches (Theis et al., 2017; Hu et al., 2020; Cheng et al., 2018) led to state-of-the art performance. They are able to map the inputs to compressed compact representations at the bottleneck of the AEs and, at same time, are able to reconstruct the original inputs from these compact representations using the decoder.

By controlling the size the bottleneck, one can explicitly control the dimensionality of the codes and the compression rate (Hu et al., 2020; Theis et al., 2017). However, a low size of the bottleneck increases the complexity of the task of the decoder risking a higher distortion rate. This trade-off forces the model to keep only the variations in the input data required to reconstruct the input without persevering the redundancies and noise within the input (Baldi, 2012; Cheng et al., 2018). This is achieved implicitly by minimizing the reconstruction error, i.e., distortion D.

In this paper, we propose to model the feature redundancy in the bottleneck representation and minimize it explicitly. To this end, we propose augmenting the loss L using the sum of the pair-wise correlations between the elements of the bottleneck. In the context of neural networks, it has been shown that reducing the correlation improves generalization (Cogswell et al., 2016), which has been successfully applied for network pruning (Kondo & Yamauchi, 2014; He et al., 2019; Singh et al., 2020; Lee et al., 2020). In this work, we argue that in the context of autoencoders, we can explicitly penalize the pair-wise correlations between the features at the bottleneck and, thus, avoid redundancy and yield more diverse compressed representations of the input images. The contributions of this paper can be summarized as follows:

- We propose a scheme to avoid redundant features in the bottleneck representation of the autoencoders.
- We propose to augment the loss of autoencoders to explicitly penalize the pair-wise correlations between the features and learn diverse compressed codes from the images.
- The proposed penalty acts as an unsupervised regularizer on top of the encoder and can be integrated into any autoencoder-based model in a plug-and-play manner.

2 Reducing the pair-wise correlation within the bottleneck representation

AEs are a special type of neural networks trained to achieve two objectives: (i) to learn to compress an input signal into a low-dimensional space, (ii) to learn to reconstruct the original input from the low-dimensional representation. This is achieved by minimizing the reconstruction loss over the training samples, which implicitly forces a concise 'non-redundant' representation of the data. In this paper, we propose to augment the reconstruction loss to explicitly minimize the redundancy, i.e., correlation, between the features learned at the bottleneck. Given a training data $\{\boldsymbol{x}_i\}_{i=1}^N$ and an encoder $g(\cdot) \in \mathbb{R}^D$, the correlation between the i^{th} and j^{th} features, g_i and g_j, can be expressed as follows:

$$C(g_i, g_j) = \frac{1}{N} \sum_n (g_i(\boldsymbol{x}_n) - \mu_i)(g_j(\boldsymbol{x}_n) - \mu_j), \tag{2}$$

where $\mu_i = \frac{1}{N} g_i(\boldsymbol{x}_n)$ is the average output of the i^{th} neuron. Our aim is to minimize the redundancy of the bottleneck representations which corresponds to minimizing the pair-wise covariance between different features. Thus, similar to (Cogswell et al., 2016), we augment the standard loss $L\big(\{\boldsymbol{x}_i\}_{i=1}^N\big)$ as follows:

$$\begin{aligned} L\big(\{\boldsymbol{x}_i\}_{i=1}^N\big)_{aug} &\triangleq L\big(\{\boldsymbol{x}_i\}_{i=1}^N\big) + \alpha \sum_{i \neq j} C(g_i, g_j) \\ &= \frac{1}{N} \sum_{i=1}^N D(\boldsymbol{x}_i, f \circ g(\boldsymbol{x}_i)) + \alpha \sum_{i \neq j} \Big(\frac{1}{N} \sum_n (g_i(\boldsymbol{x}_n) - \mu_i)(g_j(\boldsymbol{x}_n) - \mu_j) \Big), \end{aligned} \tag{3}$$

where α is a hyper-parameter used to control the contribution of the additional term in the total loss of the model. L_{aug} is composed of two terms, the first term depends on both the encoder and decoder part to ensure that the AE learns to reconstruct the input, while the second term depends only on the encoder and its aim is to promote the diversity of the learned features and ensures that the encoder learns less correlated non-redundant features.

Intuitively, the proposed approach acts as an unsupervised regularizer on top of the encoder providing an extra feedback during the back-propagation to reduce the correlations of the encoder's output. The proposed scheme can be embedded into any autoencoder-based model as a plug-in and optimized in a batch-manner, i.e., at each optimization step, we can compute the covariance using the batch samples. Moreover, it is suitable for different learning strategies and different topologies.

3 Experimental results

We test the proposed approach using the MNIST dataset (LeCun et al., 1998), which is a handwritten digit dataset composed of 10 classes. MNIST images are 28×28 pixels, which results in 784-dimensional vectors. The dataset has 50000 samples for training and 10000 for testing. We use the last 10000 training samples as a validation set to optimize the hyper-parameter α in equation 3.

For the autoencoder model, we use a simple architecture. The encoder is composed of two intermediate fully-connected layers composed of 128 and 64 neurons, respectively. The final output of the encoder is composed of n neurons, where n is the size of the bottleneck. Similarly, the decoder part takes the encoder's output, maps it to an intermediate layer of 64 neurons, then 128 neurons, and outputs a 784-vector. In all the layers, we use Leaky ReLU (LeakyReLU) (Maas et al., 2013) activation except for the final AE output, where sigmoid activation is used.

For the training, we use Adam as our optimizer with a learning rate of 5×10^{-4} and the binary cross-entropy loss as our standard training loss L. The number of epochs and the batch size are set to 100 and 128 is in all experiments, respectively. The hyper-parameter α is selected from $\{0.001, 0.005, 0.01\}$ using the validation set. The results for different bottleneck sizes are reported in Table 1. We repeat each experiment three times and we report the mean and standard deviation of root-mean-square error (RMSE) errors on the test for the different approaches. We note that the proposed approach consistently boosts the performance of the autoencoder and yields lower errors compared to training with standard loss only.

Table 1: Average and standard deviation of root-mean-square error (RMSE) of different approaches on the MNIST dataset

	Standard loss	Ours
$784 \to 8$	0.1289 ± 0.0002	0.1284 ± 0.0003
$784 \to 4$	0.1688 ± 0.0003	0.1676 ± 0.0001
$784 \to 2$	0.1974 ± 0.0010	0.1969 ± 0.0003

Figure 1 shows the projection of the test data in 2D produced by the autoencoder using the standard MSE loss and MSE augmented with our approach. We note that the extra feedback provided by the proposed regularizer changes the embedding of the data and yields different codes for the input images. Moreover, we note that explicitly penalizing the redundancies at the bottleneck yields a compact representation of the classes. This is clear especially for the fourth and fifth classes.

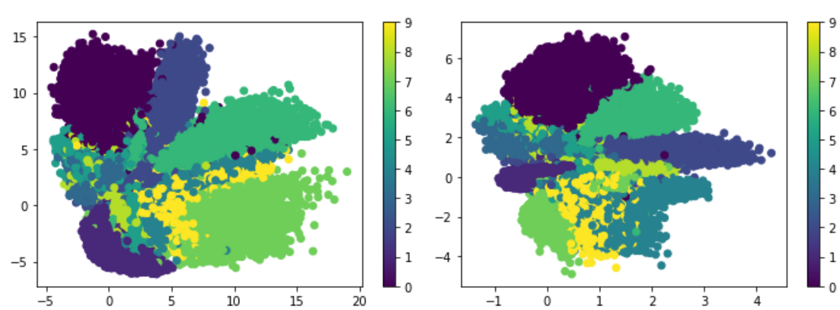

Figure 1: 2D projection of MNIST data using an autoencoder trained with mean square error (left side) and trained with mean square error augmented using our approach (right side).

4 Conclusion

In this paper, we propose a schema for modeling the redundancies at the bottleneck of an autoencoder. We propose to complement the loss with an extra regularizer, which explicitly penalizes the pair-wise correlation of the neurons at the encoder's output and, thus, forces it to learn more diverse and compact codes for the input images. The proposed approach can be interpreted as an unsupervised regularizer on top of the encoder and can be integrated into any autoencoder-based compression model in a plug-and-play manner.

Future directions include extensive testing of our approach with the different compression distortion metrics and different quantization techniques such as compressive autoencoders.

References

R Can Aygun and A Gokhan Yavuz. Network anomaly detection with stochastically improved autoencoder based models. In *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pp. 193–198. IEEE, 2017.

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27, pp. 37–49, 02 Jul 2012.

Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimization of nonlinear transform codes for perceptual quality. In *2016 Picture Coding Symposium (PCS)*, pp. 1–5. IEEE, 2016.

Laura Beggel, Michael Pfeiffer, and Bernd Bischl. Robust anomaly detection in images using adversarial autoencoders. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 206–222. Springer, 2019.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Deep convolutional autoencoder-based lossy image compression. In *2018 Picture Coding Symposium (PCS)*, pp. 253–257. IEEE, 2018.

Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations*, 2016.

Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pp. 511–516. IEEE, 2013.

Adam Golinski, Reza Pourreza, Yang Yang, Guillaume Sautiere, and Taco S Cohen. Feedback recurrent autoencoder for video compression. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. *arXiv preprint arXiv:1604.08772*, 2016.

Jun Han, Salvator Lombardo, Christopher Schroers, and Stephan Mandt. Deep generative video compression. *arXiv preprint arXiv:1810.02845*, 2018.

Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2019.

Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. Learning end-to-end lossy image compression: A benchmark. *arXiv preprint arXiv:2002.03711*, 2020.

Feng Jiang, Wen Tao, Shaohui Liu, Jie Ren, Xun Guo, and Debin Zhao. An end-to-end compression framework based on convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3007–3018, 2017.

Chetak Kandaswamy, Luís M Silva, Luís A Alexandre, Ricardo Sousa, Jorge M Santos, and Joaquim Marques de Sá. Improving transfer learning accuracy by reusing stacked denoising autoencoders. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1380–1387. IEEE, 2014.

Yusuke Kondo and Koichiro Yamauchi. A dynamic pruning strategy for incremental learning on a budget. In *International Conference on Neural Information Processing*, pp. 295–303. Springer, 2014.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Seunghyun Lee, Byeongho Heo, Jung-Woo Ha, and Byung Cheol Song. Filter pruning and re-initialization via latent space clustering. *IEEE Access*, 8:189587–189597, 2020.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*. Citeseer, 2013.

Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek. An overview of jpeg-2000. In *Proceedings DCC 2000. Data Compression Conference*, pp. 523–541. IEEE, 2000.

Fabian Mentzer, Luc Van Gool, and Michael Tschannen. Learning better lossless compression using lossy compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6638–6647, 2020.

Yann Ollivier. Auto-encoders: reconstruction versus compression. *arXiv preprint arXiv:1403.7752*, 2014.

Stefan Petscharnig, Mathias Lux, and Savvas Chatzichristofis. Dimensionality reduction for image features using deep learning and autoencoders. In *Proceedings of the 15th international workshop on content-based multimedia indexing*, pp. 1–6, 2017.

Majid Rabbani and Paul W Jones. *Digital image compression techniques*, volume 7. SPIE press, 1991.

Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pp. 2922–2930. PMLR, 2017.

Bryan Seybold, Emily Fertig, Alex Alemi, and Ian Fischer. Dueling decoders: Regularizing variational autoencoder latent spaces. *arXiv preprint arXiv:1905.07478*, 2019.

Pravendra Singh, Vinay Kumar Verma, Piyush Rai, and Vinay Namboodiri. Leveraging filter correlations for deep model compression. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 835–844, 2020.

Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The jpeg 2000 still image compression standard. *IEEE Signal processing magazine*, 18(5):36–58, 2001.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

Spencer A Thomas, Alan M Race, Rory T Steven, Ian S Gilmore, and Josephine Bunch. Dimensionality reduction of mass spectrometry imaging data using autoencoders. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7. IEEE, 2016.

George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5306–5314, 2017.

Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 490–497, 2014.

Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. *arXiv preprint arXiv:2006.04240*, 2020.

Li Yingzhen and Stephan Mandt. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pp. 5670–5679. PMLR, 2018.

Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1933–1941, 2017.

Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017.

Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

5 Appendix

You may include other additional sections here.