Literature Review

Max Parkin
Supervisors: Steven James, Richard Klein, Benjamin Rosman

May 2021

1 Introduction

Image classification and recognition are important components in fields of study where a system's ability to process and understand visual information is required. A frequently encountered challenge in this area is the high dimensionality that is inherent in image data. Large feature vectors are required to represent each image, making them difficult to work with and increasing the computational cost when training and evaluating image classifiers, or when clustering the data into similar groups.

Much of the existing research that addresses this problem focuses on applying various dimensionality reduction techniques to real-world image data sets. Notable examples of such data sets include the ImageNet database and CIFAR-10 that feature real-world objects and animals. These same techniques have also been applied to hyperspectral data sets like HYDICE. Once an image has been reduced to a lower-dimensional form, a classifier can then be used more effectively to learn the decision boundary from a reduced feature set.

While most work has centered around this type of real-life imagery, there has been little application of these techniques to images featuring artificial environments. This begs the question: how accurately can a model classify the learned representation of synthetic imagery (like that from the Minecraft video game [Guss *et al.* 2019] [Johnson *et al.* 2016]) obtained using the same dimensionality reduction techniques that have traditionally been applied to real-world image data.

This review will look at existing work that covers common unsupervised dimensionality reduction techniques as well as state representation learning methods that produce a latent representation of an image. Additionally, research that describes a powerful convolutional neural network (CNN) image classifier at the time that advanced the classification field will be surveyed. Reviewing works that have focused on real-world images will provide a guideline for the techniques that should be applied, and the way they should be applied, when testing how the procedures translate to a synthetic environment.

2 Background and Review of Literature

Image research has become a prolific area of study in domains ranging from computer vision and robotics (where the ability to process and respond to visual stimuli is essential, enabling meaningful interactions with a complex environment) to remote sensing and geographical analysis.

Particular advancements in the field of CNNs and feature learning algorithms for dimensionality reduction, are the impetus to the surge in image analysis popularity. Krizhevsky *et al.* [2012] details the construction and training of a CNN architecture called AlexNet. The network achieved state of the art performance when tested on the 2012 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) data set. It was a major contributor to the increased interest in using CNN for image classification tasks. It is computationally costly to train a CNN on high resolution images. Due to the memory constraints at the time, AlexNet utilised a mutli-GPU approach that split the neurons of the CNN on two different GPUs. This aided in reducing training time and improved the classifier performance.

Two-dimensional RGB images can be represented as $\mathbb{R}^d$ feature vectors, where $d = M \times N \times 3$ and M and N are the number of pixels in a single row and column in the image, respectively. Thus, an image can be thought of as a data point residing in d-dimensional feature space. The dimension of this feature space increases rapidly with small changes in M and N. Therefore, a pertinent challenge in the image research field is the curse of dimensionality phenomenon, where the performance of classifiers and other machine learning algorithms degrades when the dimensionality of the feature space grows. There has been much research in the field of dimensionality reduction where numerous techniques have been employed to reduce the number of input features present in a given data set. This process can also be understood as learning a representation that captures the essence of the underlying structure of the data. The representation should contain fewer input features, thereby improving the feasibility of applying standard classifiers and learning algorithms on the reduced data.

Frequently applied dimensionality reduction techniques include principle component analysis (PCA) and autoencoder neural networks. Both of these techniques serve as unsupervised feature extraction methods. This review will

discuss these methods and the contribution of various authors to this field.

2.1 Principle component analysis

In any dimensionality reduction technique, there is a tradeoff between precision and complexity. Naturally, removing any input features will have an effect on accuracy. However, the methods applied aim to preserve as much information contained in the data as possible, still allowing for adequate analysis and description of such data while simplifying its representation. PCA is a linear dimensionality reduction algorithm that generates a set of 'principle components' – uncorrelated mutually orthogonal unit vectors (these are the normalised eigenvectors of the covariance matrix) that are linear combinations of the initial features. For d-dimensional data, there are d such principle components, where the leading components, sorted by eigenvalue, capture the most amount of variance in the data. In other words, the leading principle components contain the most amount of information. By removing those components that account for minimal variance in the data, we reduce the dimension of our feature space. The original data is then projected onto the basis defined by the remaining principle components.

As one of the older feature learning techniques and relatively straightforward to implement, PCA has frequently been employed in the image classification domain. Rodarmel and Shan [2002] apply PCA to hyperspectral data from the HYDICE and AVIRIS data sets. A large number of bands are present in hyperspectral images so it becomes necessary to perform feature extraction to reduce the processing complexity. They compare the classification accuracy of the reduced PCA bands with the original hyperspectral images and find that PCA produces an acceptable classification performance when using only a small set of PCA bands.

There are a few drawbacks associated with PCA. One of these is the need to standardise the data. More significant, however, is the scalability issues PCA encounters when applied to large data sets, where the procedure can require exhaustive memory usage and has relatively high computational cost. Benito and Peña [2005] introduce a faster PCA approach to dimensionality reduction with high quality image reconstruction rates. The method involves projecting images onto lower-dimensional representations while retaining the images' internal structure. They show that this new method provides an im-

proved reconstruction performance when compared to traditional PCA and N-mode PCA (a higher-order generalisation of PCA) techniques, and is a valid approach for standard dimensionality reduction on high-dimensional image data. Here, the reconstruction tests were carried out on a data set consisting of images of human faces. Again, it is not clear how this would translate to synthetic data.

2.2 Autoencoders

An autoencoder is a non-linear unsupervised approach to representation learning where a neural network transforms input data to a compressed form. The aim is to learn a reduced encoding of the data that serves as the latent space representation of the input. This reduced encoding should maintain the necessary meaningful information that allows the decoder portion of the network to perform a suitable reconstruction of the original data. In this way, the network should learn a latent representation of the image by learning those latent features that adequately describe it. This is often realised by utilising a reconstruction loss function and some regularisation term to avoid overfitting to the training samples. Since an autoencoder is a neural network, it is capable of learning a more complex non-linear representation of the data that can be used to accurately describe the data in a lower-dimensional space.

Latent variables can be used to reduce the dimensionality of the observed data. Thus, by obtaining a latent representation of the input images, where each latent feature is a learned characteristic or property that the data possesses, we can reduce the representational complexity of the input. A recent addition to the class of autoencoders is the variational autoencoder (VAE) proposed by Kingma and Welling [2014]. The idea that they introduce here is that a VAE enables the encoder portion of the network (the feature extractor) to produce an encoding vector, where each learned latent feature represents a probability distribution instead of a single value. We then sample from each of these probability distributions (for each latent variable) to obtain the vector that serves as input to the decoder (reconstructor) portion of the network. If we have a latent variable z that generates an observed variable x, we would like to infer the properties of this latent variable by computing $p(z|x)$. In practice we approximate this distribution with another distribution that is tractable (for example a Gaussian distribution) and feed this approximation to the encoder. The encoder then learns a mapping from

input x to latent variable z and the decoder learns a mapping from the latent representation to a reconstruction of the original data.

Looking at implementations of this VAE, Fournier and Aloise [2019] compare traditional singular value decomposition dimensionality reduction techniques, for example PCA, to an autencoder approach in image classification tasks. The MNIST, Fashion-MNIST, and CIFAR-10 image data sets were used to test the classification performance. These data sets again feature only real-world imagery. To keep the comparison fair, no convolutions were added to the neural networks. They found that PCA had significantly lower accuracy than the autoencoders (a deep autoencoder and a variational autoencoder) at lower dimensions when testing the learned representation on a trained k-nearest neighbours classifier. However, the accuracies were comparable when the dimension of the projection increased. The autoencoders had consistently high classification accuracy at all dimensions tested, but the computational cost associated with training the neural networks resulted in significantly slower training times when compared to PCA. While autoencoders are very good at feature extraction by focusing on reconstruction error, Fournier and Aloise [2019] found PCA to still be a valid and feasible approach to dimensionality reduction for image classification. For this reason it would be worthwhile to test the performance of both an autoencoder and PCA when applied to synthetic imagery.

As with a standard feed-forward neural network, autoencoders can be trained using stochastic gradient descent and the backpropagation algorithm. If we want a latent categorical representation for our data, it would be difficult to train the autoencoder that produces this discrete vector encoding. This is because backpropagation requires a well-defined gradient with respect to the network weights. Jang *et al.* [2017] introduce a continuous distribution called Gumbel-Softmax. Samples from this distribution approximate categorical samples and are differentiable. They find that this gradient estimator outperforms other estimators like DARN, MuProp, and VIMCO on categorical latent variables, and is effective when using a VAE to learn a discrete latent representation of the input data. They used the MNIST data set and focused on output prediction and the training of a generative model to test the estimators' performances. However, Gumbel-Softmax has not yet been utilised when learning latent categorical representations of non-real-world synthetic data.

3 Representation learning

Anand *et al.* [2020] introduce a new unsupervised state representation learning technique that utilises the spatial and temporal features of a visual environment. VAEs, another unsupervised technique discussed above, learn a representation of the input by minimising the reconstruction error. While generally able to produce reasonable results, the technique tends to focus on pixel-level details and so abstract latent features may be ignored. The method that Anand *et al.* [2020] propose, called Spatiotemporal DeepInfomax (ST-DIM), attempts to maximise the mutual information across spatial and temporal attributes of a given environment. In this way, the technique is able to successfully capture small objects and underlying latent features present in visual data when tested on Atari 2600 games. These tasks prove difficult for generative approaches like the VAE. This is a powerful technique that can aid in representation learning when applied to similar artificial environments like Minecraft.

4 Conclusion

This review has outlined some of the common dimensionality reduction and representation learning techniques which the referenced authors have used to compress the dimension of the image input data and learn the latent features that define it. As discussed, these techniques have primarily been applied to real-world image data, for example handwritten digits in MNIST or real-life objects in ImageNet. It is not clear how well the latent representations and the classifiers trained using these learned representations will transfer to artificial environments featuring synthetic objects and features like those in Minecraft. It will thus be valuable to apply these techniques to such a synthetic environment and to evaluate the performance of the trained classifiers.

References

[Anand *et al.* 2020] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R Devon Hjelm. *Unsupervised State Representation Learning in Atari*, 2020.

[Benito and Peña 2005] Monica Benito and Daniel Peña. A fast approach for dimensionality reduction with image data. *Pattern Recognition*, 38:2400–2408, 12 2005.

[Fournier and Aloise 2019] Quentin Fournier and Daniel Aloise. Empirical comparison between autoencoders and traditional dimensionality reduction methods. *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Jun 2019.

[Guss *et al.* 2019] William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. *MineRL: A Large-Scale Dataset of Minecraft Demonstrations*, 2019.

[Jang *et al.* 2017] Eric Jang, Shixiang Gu, and Ben Poole. *Categorical Reparameterization with Gumbel-Softmax*, 2017.

[Johnson *et al.* 2016] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *IJCAI*, 2016.

[Kingma and Welling 2014] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*, 2014.

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[Rodarmel and Shan 2002] Craig Rodarmel and Jie Shan. Principal component analysis for hyperspectral image classification. *Surv Land inf Syst*, 62, 01 2002.