International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Extracting Latent Variables and Implementing in Multimodal Variational Autoencoders

Amit Khare
Arya Pratap Singh
Acropolis Institute of Technology and Research, Indore(M.P.)

Abstract **- Significance of Autoencoders are revolutionizing the data we process today. Dimensionality Reduction, Image Compression, Image Denoising, Feature Extraction, Image generation, Sequence to sequence prediction, Recommendation system and what not. This paper focuses on the Dimensionality Reduction part, specifically, extracting the Latent Variables.**

Several techniques of encoding and decoding have been used to extract latent variables, but this paper presents an entirely new technique of using the power of Autoencoders for extraction of Data. Machine learning is about capturing aspects of the unknown distribution from which the observed data are sampled (the data-generating distribution). For many learning algorithms and in particular in manifold learning, the focus is on identifying the regions (sets of points) in the space of examples where this distribution concentrates, i.e., which configurations of the observed variables are plausible. Unsupervised representation-learning algorithms try to characterize the data-generating distribution through the discovery of a set of features or latent variables whose variations capture most of the structure of the data-generating distribution.

1.INTRODUCTION

Affective Computing studies frequently collect rich, multimodal data from a number of different sources in order to be able to model and recognize human affect. These data sources — whether they are physiological sensors, smartphone apps, eye trackers, cameras, or microphones — are often noisy or missing. Increasingly, such studies take place in natural environments over long periods of time, where the problem of missing data is exacerbated. For example, a system trying to learn how to forecast a depressed mood may need to run for many weeks or months, during which time participants are likely to not always wear their sensors, and sometimes miss filling out surveys. While research has shown that combining more data sources can lead to better predictions,as each noisy source is added, the intersection of samples with clean data from every source becomes smaller and smaller. As the need for long-term multimodal data collection grows, especially for challenging topics such as forecasting mood, the problem of missing data sources becomes especially pronounced. While there are a number of techniques for dealing with missing data, more often than not researchers may choose to simply discard samples that are missing one or more modalities. This can lead to a dramatic reduction in the number of samples available to train an affect recognition model, a significant problem for data-hungry machine learning models. Worse, if the data are not missing completely at random, this can bias the resulting model. In this paper we propose a novel method for dealing with missing multimodal data based on the idea of denoising Autoencoders. A denoising autoencoder is an unsupervised learning

method in which a deep neural network is trained to reconstruct an input that has been corrupted by noise. In most cases, noise is injected by randomly dropping out some of the input features, or adding small Gaussianof thosefeatures that are computed using the data from a single modality. We demonstrate that by using a new model, which we call a Multimodal Autoencoder (MMAE), it is possible to accurately reconstruct the data from a missing modality, something that cannot be done with other techniques such as PCA. Further, we show that the MMAE can be trained with additional neural network layers designed to perform classification, effectively leveraging information from both unlabeled and labeled data. We present empirical results comparing MMAE to several other methods for dealing with missing data, and demonstrate that the MMAE consistently gives the best performance as the number of missing modalities increases. Results are shown for the task of predicting tomorrow's mood, health, and stress, using data collected from physiological sensors, a smartphone app, and surveys. The goal of this research is to build a real-world system that can not only help participants predict their future mood and make adjustments to improve it, but also help detect early warning signs of depression, anxiety, and mental illness. However, the data inevitably contain samples with missing modalities, which can easily occur when a participant's smartphone cannot log data, or when sensorhardware malfunctions.

n real world data analysis tasks we analyze complex data i.e. multi dimensional data. We plot the data and find various patterns in it or use it to train some machine learning models. One way to think about dimensions is that suppose you have an data point **x** , if we consider this data point as a physical object then dimensions are merely a basis of view, like where is the data located when it is observed from horizontal axis or vertical axis. As the dimensions of data increases, the difficulty to visualize it and perform computations on it also increases. So, how to reduce the dimensions of a data-

a. Remove the redundant dimensions
b. Only keep the most important dimensions.

Variance is a measure of the variability or it simply measures how spread the data set is. Mathematically, it is the average squared deviation from the mean score.

PCA finds a new set of dimensions (or a set of basis of views) such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them. It means more important principle axis occurs first. (more important = more variance/more spread out data)

2.METHOD

The MMAE was developed to ameliorate the likely problem where a number of contiguous features from the same modality go missing at once. We start by normalizing all of the features to be in the range [0, 1]. We then represent a missing modality by filling all features from that modality with the special value −1. It is important to use a special value to indicate missing data that must be filled, rather than fill with a value such as 0 which could actually occur in the real data. To train the MMAE, we first use samples that have data from every modality to provide the ground truth noise-free X. At training time, for every sample X, we compute Xe by adding noise using two methods. First, we add simple masking noise to 5% of the features. Second, we randomly select one or more modalities and set all of the feature values for that modality to −1; essentially, masking entire modalities at once. The model is then trained to reproduce X from Xe. Effectively, this means that the model must learn to predict reasonable values for the missing modality from the rest of the features. For example, it may use the participant's physiology and location patterns to predict her survey responses, such as how much time she spent in class, or whether she drank caffeine. After training the autoencoder portion of the network with the clean, unsupervised examples for which all sensors are available, we then begin a second phase of training for classification. main interest are modalities of high complexity. We consider models based on variational autoencoders (VAEs, Kingma & Welling, 2014; Rezende et al., 2014). Standard VAEs learn a latent representation $z \in Z$ for a set of observed variables $x \in X$ by modelling a joint distribution $p(x, z) = p(z)p(x|z)$. In the original VAE, the intractable posterior $q(z|x)$ and conditional distribution $p(x|z)$ are approximated by neural networks trained by maximising the ELBO loss taking the form

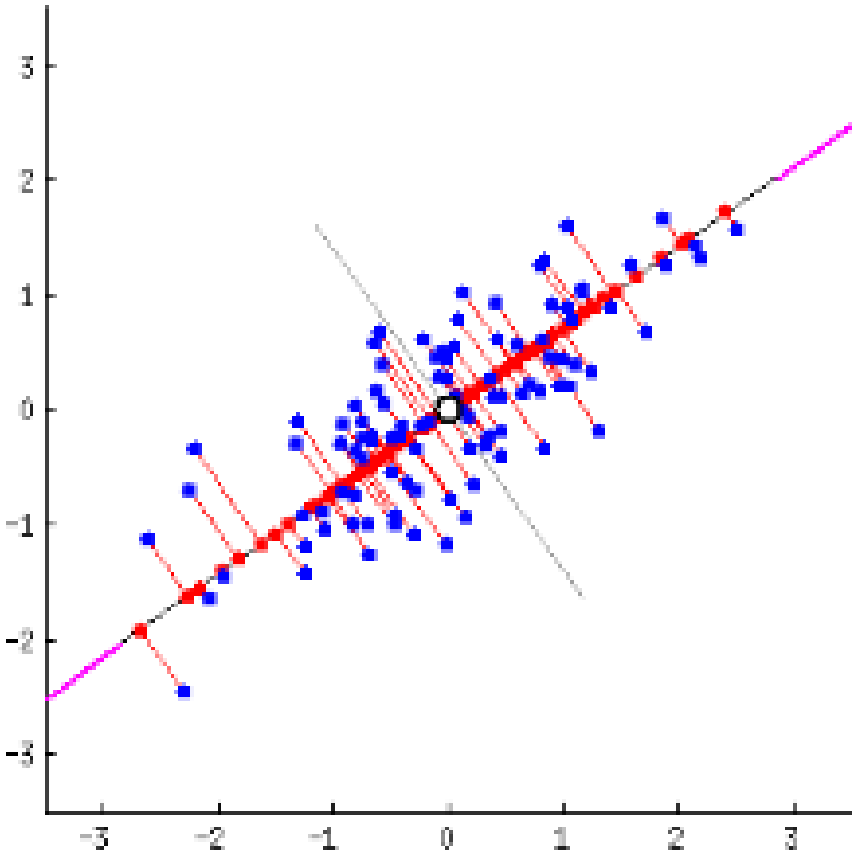
$$L = E_{q(z|x)} [\log p(x|z)] - D_{KL}(q(z|x) \,k\, N(0, I)) \quad - (1)$$

with respect to the parameters of the networks modelling $q(z|x)$ and $p(x|z)$. Here $D_{KL}(\cdot \,k\, \cdot)$ denotes the KullbackLeibler divergence. Bi-modal VAEs that can handle a missing modality extend this approach by modelling $q(z|x1, x2)$ as well as $q1(z|x1)$ and $q2(z|x2)$, which replace the single $q(z|x)$. Multimodal VAEs may differ in 1) the way they approximate $q(z|x1, x2)$, $q1(z|x1)$ and $q2(z|x2)$ by neural networks and/or 2) the structure of the loss function, see Figure 1. Typically, there are no conceptual differences in the decoding, and we model the decoding distributions in the same way for all methods considered in this study. Suzuki et al. (2017) introduced a model termed JMVAE (Joint Multimodal VAE), which belongs to the class of approaches that can only learn from the paired training samples (what we refer to as the (fully) supervised setting). It approximates $q(z|x1, x2)$, $q1(z|x1)$ and $q2(z|x2)$ with three corresponding neural networks and optimizes an ELBO type loss of the form –

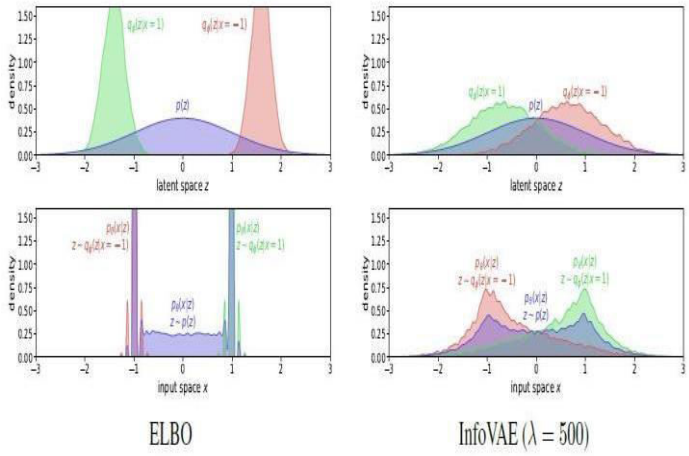
$$L = E_{q(z|x1,x2)} [\log p1(x1|z) + \log p2(x2|z)] - D_{KL}(q(z|x1, x2) \,k\, N(0, I)) - D_{KL}(q(z|x1, x2) \,k\, q1(z|x1)) - D_{KL}(q(z|x1, x2) \,k\, q2(z|x2)) \quad - (2)$$

3.IMPLEMENTATION

While using MSE is easy and most common, we found that using a cross-entropy (CE) reconstruction loss reliably led to better results for the MMAE than using MSE. We created an image reconstruction dataset based on MNIST digits (LeCun et al., 1998). The images were These regions are considered as different input modalities. In the above notion of "AND" and "OR" tasks we implicitly assume an be guessed from only one part of the image makes the new MNIST-Split benchmark a mixture of an "AND" and an "OR" task.This is in contrast to the MNIST-SVHN task described below, which can be regarded as an almost pure "OR" task. Since cross- entropy is appropriate for binary values, before applying this loss we first normalized all of our features to the [0,1] range. In addition, we experimented with implementing the MMAE as a Variational Autoencoder (VAE) [20], which constrains the features in the embedding to follow K independent Gaussian distributions. This makes it more likely that a random embedding sampled from a K-dimensional multivariate Gaussian with mean 0 and variance 1, will actually correspond to a plausible sample when passed through the decoder; in other words, it makes it possible to generate new samples by interpolating in the embedding space. While this ability to generate realistic-looking samples of data is interesting, we conducted experiments using the VAE version of our MMAE and found it did not improve reconstruction or classification performance. To assess the MMAE, we compared it to two other dimensionality reduction techniques: PCA, and a supervised feature selection technique in with the features with the highest ANOVA F-value with the classification label in the training data were selected. We constrained each method to reduce the original 343 features to 100 dimensions to enable fair comparison; this allowed the PCA to capture 98% of the variance in the data, assuring a fair comparison. We also compared MMAE to four ways of dealing with missing data, including discarding the data when training the model, filling it with a special value like -1, filling it with the average for that feature, and filling it using a PCA reconstruction. PCA reconstruction of missing data was conducted by applying the inverse transformation learned by PCA to the 100-dimensional principle components vector. We also compared the MMAE's classification performance to three other machine learning algorithms including Support Vector Machines (SVM), Logistic Regression (LR), and a feedforward neural network (NN). For all models we performed a grid search over possible hyperparameter settings, optimizing for performance on the validation set.

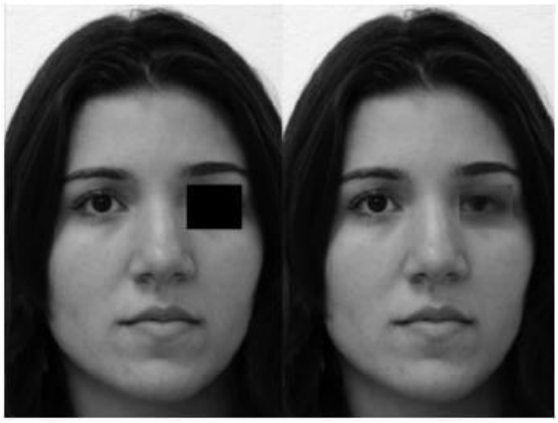


Fig(1) shows the variance in the linear regression validation through the PCA



Fig(2) shows the density of input and latent space through dimension reduction in ELBO and InfoVAE.

4.RESULT



Fig(3) shows the retrieved values of reconstructed pixel.

An autoencoder is an unsupervised learning technique in which a deep neural network is trained to reproduce an input X based on the reconstruction error between X and the network's output X0 ; e.g. if using squared reconstruction error, the model would be trained to optimize the following loss function:

$$L(X, Y) = ||X - Y||^2$$

The MMAE was developed to ameliorate the likely problem where a number of contiguous features from the same modality go missing at once. We start by normalizing all of the features to be in the range [0, 1]. We then represent a missing modality by filling all features from that modality with the special value −1. It is important to use a special value to indicate missing data that must be filled, rather than fill with a value such as 0 which could actually occur in the real data. To train the MMAE, we first use samples that have data from every modality to provide the ground truth noise-free X. At training time, for every sample X, we compute Xe by adding noise using two methods. First, we add simple masking noise to 5% of the featuresSecond, we randomly select one or more modalities and set all of the feature values for that modality to −1;essentially, masking entire modalities at once.The model is then trained to reproduce Xfrom Xe. Effectively, this means that the model must learn to predict reasonable values for the missing modality from the rest of the features. For example, it may use the participant's physiology and location patterns to predict her survey responses, such as how much time she spent in class, or whether she drank caffeine. After training the autoencoder portion of the network with the clean, unsupervised examples for which all sensors are available, we then begin a second phaseof training for classification.

5.Discussion and Conclusion

We have described a new method for restoring missing sensor data, which is frequently lost in multimodal, realworld data collection settings. Empirical results demonstrate that the MMAE can accurately reproduce data from a lost modality, while other methods such as PCA cannot. The MMAE offers valuable new advantages for Affective Computing researchers who would like to train unbiased models on noisy data, accurately cluster noisy samples, or make robust predictions in the face of real-world data loss. The MMAE has potential benefits in terms of providing enhanced flexibility and privacy to users of a mood prediction system. Because it can make accurate mood predictions even when data is lost, it could allow users to opt-out of providing data for all modalities. This could be particularly enticing to certain users, e.g. those who are uncomfortable wearing sensors throughout the day, or those who are concerned about privacy issues surrounding location data. The MMAE also provides an effective feature reduction method that may enhance privacy; the embeddings learned by the MMAE can be used to provide roughly equal classification performance to the raw features, meaning that the raw features would not have to be stored once the embeddings are computed. The embeddings could potentially allow the highly sensitive personal data collected from this study to be shared with other researchers in a non-identifiable way. SVAE resembles VAEVAE in the need for additional networks besides one encoder per each modality and the structure of ELBO loss. It does, however, solve the problem of learning the joint embeddings in a way that allows to learn the parameters of approximated q(z|x1, x2) using all available samples, i.e., both paired and unpaired. If q(z|x1, x2) is approximated with the joint network that accepts concatenated inputs, as in JMVAE and VAEVAE (b), the weights of q(z|x1, x2) can only be updated for the paired share of samples. If q(z|x1, x2) is approximated with a PoE of decoupled networks as in SVAE

REFERENCES –

1. Wentao Xue, Zhengwei Huang, Xin Luo, and Qirong Mao, "Learning speech emotion features by joint disentangling-discrimination," in Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE, 2015, pp. 374–379.

2. Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Fruhholz, and Bj ¨ orn ¨ Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 500–504, 2017.

3. Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," Scientific reports, vol. 6, 2016.

4. O Shcherbakov and V Batishcheva, "Image inpainting based on stacked autoencoders," in Journal of Physics: Conference Series. IOP Publishing, 2014, vol. 536, p. 012020.

5. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

6. Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 341–349.

7. A. Sano, Measuring College Students Sleep, Stress and Mental Health with Wearable Sensors and Mobile Phones, Ph.D. thesis, MIT, 2015.

8. B. Dacorogna. Introduction to the Calculus of Variations. World Scientific Publishing Company, 2004.

9. Karol Gregor, Arthur Szlam, and Yann LeCun. Structured sparse coding via lateral inhibition. In NIPS'2011, 2011.

10. G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. Neural Computation, 18:1527–1554, 2006.

11. Aapo Hyv¨arinen. Estimation of non-normalized statistical models using score matching. J. Machine Learning Res., 6, 2005.

12. Aapo Hyv¨arinen. Some extensions of score