ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Missing Image Data Imputation using Variational Autoencoders with Weighted Loss

Ricardo Cardoso Pereira[1], Joana Cristo Santos[1], José Pereira Amorim[1,2],
Pedro Pereira Rodrigues[3] and Pedro Henriques Abreu[1] *

[1] Centre for Informatics and Systems of the University of Coimbra
Department of Informatics Engineering
Pólo II, Pinhal de Marrocos, 3030-290, Coimbra, Portugal

[2] IPO-Porto Research Centre - CI-IPOP
Rua Dr. António Bernardino de Almeida, 4200-072, Porto, Portugal

[3] Center for Health Technology and Services Research
Faculty of Medicine of the University of Porto
Alameda Prof. Hernâni Monteiro, 4200-319, Porto, Portugal

Abstract. Missing data is an issue often addressed with imputation strategies that replace the missing values with plausible ones. A trend in these strategies is the use of generative models, one being Variational Autoencoders. However, the default loss function of this method gives the same importance to all data, while a more suitable solution should focus on the missing values. In this work an extension of this method with a custom loss function is introduced (Variational Autoencoder with Weighted Loss). The method was compared with state-of-the-art generative models and the results showed improvements higher than 40% in several settings.

1 Introduction

Missing data is a recurrent problem when dealing with real-world contexts. As many machine learning methods are unable to handle incomplete data, there is a need to impute the missing values with new plausible ones resembling the underlying complete information. Imputation becomes increasingly important in domains such as medical imaging where the number of samples is typically small and the cost of acquiring new ones is high, which makes ignoring incomplete images not a solution. A recent trend in imputation is the use of generative models, such as Generative Adversarial Networks (GAN) [1] or Variational Autoencoders (VAE) [2]. These methods try to generate new observations based on the available data, and they produce good results particularly for the Missing Completely At Random (MCAR) mechanism [1], where the missingness does not depend on any variables[1]. However, the use of VAEs has not been explored for missing image data imputation. Moreover, this method focus on the reconstruction of the whole image giving equal importance to observable and missing values. In this paper we propose a variant of the VAE, called Variational Autoencoder with Weighted Loss (VAE-WL), which has a custom loss function

*This work was supported in part by the Portuguese Foundation for Science and Technology (FCT) Research Grants SFRH/BD/149018/2019 and SFRH/BD/136786/2018.

[1] For more information on missing data mechanisms consult [3, 4].

that prioritizes the reconstruction of the missing values. Using this method we tackle the imputation of missing values under MCAR in images. We compare our approach to other state-of-the-art generative methods and the results show clear improvements in the imputation quality, achieving in several settings error improvements above 40%.

2 Related Work

In recent years some directions have been explored to perform the imputation of missing data in images. State-of-the-art methods are based on deep learning algorithms, such as Autoencoders and Generative Adversarial Networks (GAN).

Autoencoders and their denoising variants are able to extract a clean output from a noisy input. Mattei and Frellsen [5] used an approach based on an importance weighted auto-encoder (IWAE) to perform single or multiple imputations of incomplete images. The resulting dataset was later used in a classification task and achieved 98.683% of accuracy, an increase of 0.02% from the complete dataset. L. Gondara [6] applied a denoising autoencoder built using convolutional layers. This approach achieves good denoising performance for small sample sizes, which are typical on medical image databases. The increase from 300 to 720 samples only improved the mean SSIM score from 0.89 to 0.90.

GANs have been applied in the field of imaging due to its capacity to reconstruct an image and to perform efficiently when complete data is unavailable. Shang et al. [7] developed a novel approach for View Imputation with a GAN (VIGAN). This approach is able to integrate knowledge from the domain mappings and the view correspondences to effectively recover a missing view/modality. Yoon et al. [1] proposed a method for imputing missing data named Generative Adversarial Imputation Nets (GAIN). This method outperformed several state-of-the-art imputation techniques such as Multiple Imputation by Chained Equations, MissForest, Matrix Completion, Autoencoders, and the Expectation-Maximization algorithm, presenting a difference from 0.0062 to 0.0182 in RMSE values between the GAIN and the best performing imputation technique for each dataset.

Autoencoders and GANs represent the most recent algorithms to perform imputation of missing image data and present promising results in comparison to more traditional methods. However, the Variational Autoencoder (VAE) variant has not yet been used for image imputation, being this a novel aspect of the presented work.

3 Proposed Approach

A Variational Autoencoder (VAE) is a variant of the Autoencoder family that has generative capabilities. While the basic Autoencoder simply learns a compressed representation of the input data in a unsupervised way, the VAE learns the parameters of a probability distribution representing that data, namely the mean and variance of a Gaussian curve. By sampling from these learned parameters, the model is able to generate new data with the same characteristics [8].

Although a VAE can be used in its original form to perform missing data imputation tasks, its loss function is not the most suitable for this purpose [2]. As Equation 1 shows, the default VAE loss function contains two terms: the first is the reconstruction error and the second is a regularizer. Moreover, $q(z|X)$ is the encoder output, $p(X|z)$ is the decoder output, X is the input data and z is the new sampled data from the learned distribution.

$$L(X) = -E_{z \sim Q(z|X)}[logp(X|z)] + KL(q(z|X) \parallel p(z)) \tag{1}$$

The reconstruction error is the basis (and often the only term) of every loss function used with neural networks. Some of the most frequent used functions here are the Mean Squared Error or the Binary Cross-Entropy for scenarios with only two possible outcomes. This term is essential for the decoder to learn how to reconstruct the data. On the other hand, the regularizer from the second term is the Kullback-Leibler divergence between the encoder and decoder distributions. This term is needed to ensure that the latent space is well structured, meaning that similar input data should be represented by similar representations of the latent space [8]. When considering the use of a VAE for missing data imputation, this loss function poses two problems. First the reconstruction error gives the same importance to the available values and to the missing ones. Although this error should consider all the data to ensure a complete learning of the network, for imputation purposes the reconstruction of the missing values should have a heavier weight on this process. Second, considering the importance of both terms from the loss function, in imputation tasks it is admissible to lose some of the structure from the latent space to ensure a better reconstruction of the missing values [9]. In other words, the Kullback-Leibler divergence may have a smaller impact on the learning process, which will lead to better reconstructions and, as consequence, better imputation results.

To address these issues we propose in this work the Variational Autoencoder with Weighted Loss (VAE-WL), consisting in a VAE with an extension of the default loss function that is presented in Equation 2. In this new function, the reconstruction error is split between the data containing missing values (X_{mv}) and the data that is complete (X_{av}), assigning a heavier weight to the first one through a coefficient $\gamma > 1$. Moreover, the Kullback-Leibler divergence is penalized by using another coefficient β within the range $[0, 1[$.

$$\begin{aligned} P_E(X) &= E_{z \sim Q(z|X)}[logp(X|z)] \\ L(X) &= -(P_E(X_{av}) * \boldsymbol{\gamma}\, P_E(X_{mv})) + \boldsymbol{\beta}\, KL(q(z|X) \parallel p(z)) \end{aligned} \tag{2}$$

An example of the impact of the proposed changes in the VAE-WL loss function is presented in Figure 1. The first image is an original character from the MNIST[2] dataset and the second one is the same image with 50% of its pixels

[2] Available at `http://yann.lecun.com/exdb/mnist/`.

missing completely at random. The last two images represent the imputation with a regular VAE and with the VAE-WL (respectively). The use of the new loss function shows obvious improvements in the image reconstruction. Notice that in both imputation scenarios the VAEs have the same architecture and hyperparameters and were trained with the same data.

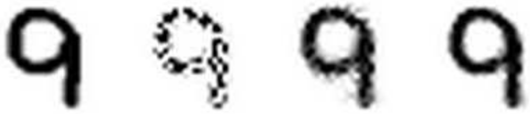

Fig. 1: Example of image imputation. From left to right: the 1st image is the original one, the 2nd image has 50% of its pixels missing completely at random, the 3rd image was imputed with a regular VAE, and the 4th image was imputed with the VAE-WL.

4 Experimental Results

In order to properly evaluate the impact of the proposed approach in an imputation task, an experiment was conducted to compare the VAE-WL with a regular VAE. Also, the Generative Adversarial Imputation Nets (GAIN) [1] method was also considered in the study, being this another generative state-of-the-art model for missing data imputation.

Regarding the VAEs, the used architecture was obtained through experimentation and its main aspects are presented in Figure 2: the encoder has two convolutional layers with 32 filters, a kernel size of three, ReLu as the activation function and a stride length of two (which avoids the use of max pooling layers); the encoder also has two fully connected layers with 392 and 196 units, which also use ReLu as the activation, while the layers for the mean and variance have 32 units; the train used the optimization algorithm Adam with a learning rate of 0.001, batches of 64 images and a maximum of 200 epochs; to avoid overfitting each layer uses the L2 regularizer and is followed by a dropout layer with a 20% rate; and finally the decoder presents the inverse architecture of the encoder.

Regarding the parameters for the VAE-WL loss function, after some experimentation with stable results they were define as $\gamma = 5$ and $\beta = 0.1$. Both VAEs use Binary Cross-Entropy for the reconstruction error of the function.

The experiment considered three datasets: MNIST, CBIS-DDSM Mass and Calcification[3]. The first is a well-known dataset for benchmarking in image related works, and it contains 70000 greyscale handwritten digits with a size of 28 by 28 pixels. The second and third datasets contain greyscale scanned mammography studies with 1696 and 1872 images, respectively, which where resized to 64 by 64 pixels. These two datasets were used because they represent a domain of medical imaging where missing values are frequent. All datasets

[3] Available at `https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM`.

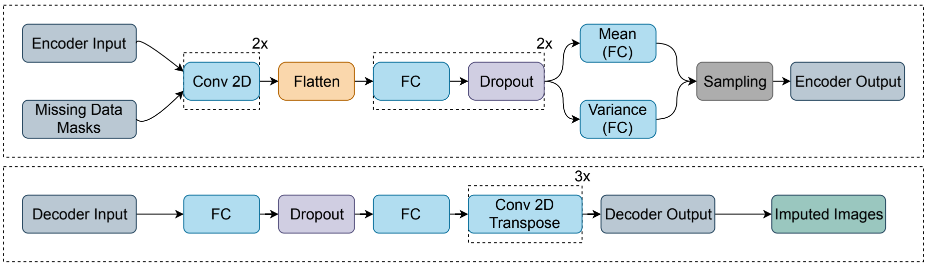

Fig. 2: Architecture of the VAEs used in the experiments. The top rectangle represents the encoder and the bottom one the decoder.

where normalized within $[0, 1]$ and split in train, validation and test sets with 60%-20%-20% proportions. The experiment considered four missing rates (20%, 30%, 40% and 50%), with the missing values being assigned randomly to each image (following therefore the MCAR mechanism) and pre-imputed with zero. To mitigate bias and stochastic behaviors, each method was executed 30 times (the average was used as the final result) with the datasets being shuffled in each run. The imputation results were assessed through the Mean Absolute Error (MAE) metric calculated with the original images and the imputed ones.

The results obtained from the experiment are presented in Table 1. The VAE-WL outperforms the standard VAE and the GAIN method in all experimented scenarios. This allows for the conclusion that the imputation of missing values with the VAE-WL is in fact better than the state-of-the-art generative models. Moreover, the VAE-WL presents stable results across the different missing rates, with insignificant error increases in higher rates (the same behavior is observed in the regular VAE). On the other hand, the GAIN method presents in general worse results for smaller missing rates. An analysis of the percentage results for the VAE-WL shows average improvements of 43%, 12% and 13% for the MNIST, CBIS-DDSM Mass and Calcification datasets when comparing with a standard VAE, and 47%, 34% and 23% when comparing with the GAIN method.

5 Conclusion

In this article an extension of the Variational Autoencoder (VAE) is proposed, called Variational Autoencoder with Weighted Loss (VAE-WL). It uses a custom loss function that is more suitable for the imputation of missing values. The method was experimented with three image datasets (MNIST, CBIS-DDSM Mass and Calcification) and compared with two other state-of-the-art generative models: a regular VAE and the GAIN method. The VAE-WL outperformed both models in all scenarios, achieving improvements over 40% in some settings.

In the future the method will be tested with more datasets containing colored images, and other missing mechanisms besides MCAR will be addressed.

Table 1: Results from the experiment. The first three columns present the MAE values for the used methods. The last two columns present the percentage improvement of the VAE-WL compared with the VAE and the GAIN, respectively. The best results for each combination of dataset with missing rate are bolded.

			VAE-WL	**VAE**	**GAIN**	**↑ % VAE**	**↑ % GAIN**
MNIST		**MR 20%**	**0.036**	0.063	0.091	43%	61%
		MR 30%	**0.036**	0.064	0.066	44%	45%
		MR 40%	**0.037**	0.065	0.064	44%	43%
		MR 50%	**0.039**	0.067	0.064	42%	39%
CBIS-DDSM	**Mass**	**MR 20%**	**0.044**	0.051	0.084	13%	47%
		MR 30%	**0.045**	0.051	0.067	12%	32%
		MR 40%	**0.046**	0.052	0.055	11%	17%
		MR 50%	**0.046**	0.052	0.075	11%	39%
	Calc.	**MR 20%**	**0.046**	0.054	0.078	16%	42%
		MR 30%	**0.047**	0.054	0.066	12%	29%
		MR 40%	**0.048**	0.054	0.055	10%	13%
		MR 50%	**0.047**	0.054	0.051	12%	8%

References

[1] Jinsung Yoon, James Jordon, and Mihaela Schaar. GAIN: Missing Data Imputation using Generative Adversarial Nets. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5675–5684, 2018.

[2] John T McCoy, Steve Kroon, and Lidia Auret. Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit. *IFAC-PapersOnLine*, 51(21):141–146, 2018.

[3] Miriam Seoane Santos, Jastin Pompeu Soares, Pedro Henriques Abreu, Hélder Araújo, and João Santos. Influence of Data Distribution in Missing Data Imputation. In *International Conference on Artificial Intelligence in Medicine in Europe*, pages 285–294, 2017.

[4] Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Fonseca Costa, Jastin Pompeu Soares, João Santos, and Pedro Henriques Abreu. Generating Synthetic Missing Data: A Review by Missing Mechanism. *IEEE Access*, 7:11651–11667, 2019.

[5] Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4413–4423, 2019.

[6] L. Gondara. Medical Image Denoising Using Convolutional Denoising Autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246, 2016.

[7] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. Vigan: Missing view imputation with generative adversarial networks. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 766–775. IEEE, 2017.

[8] Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of the 20th Conference on Computational Natural Language Learning*, pages 10–21, 2016.