E2 GAN: End-to-End Generative Adversarial Network for Multivariate Time Series imputation

Sigurd Vang

Time Series are everywhere!

- Because of technological advances
- In different sectors
  - Energy, finance, health etc…
- Useful for advanced analysis
  - prediction, classification, forecasting
- However…
  - Missing values are a huge problem!

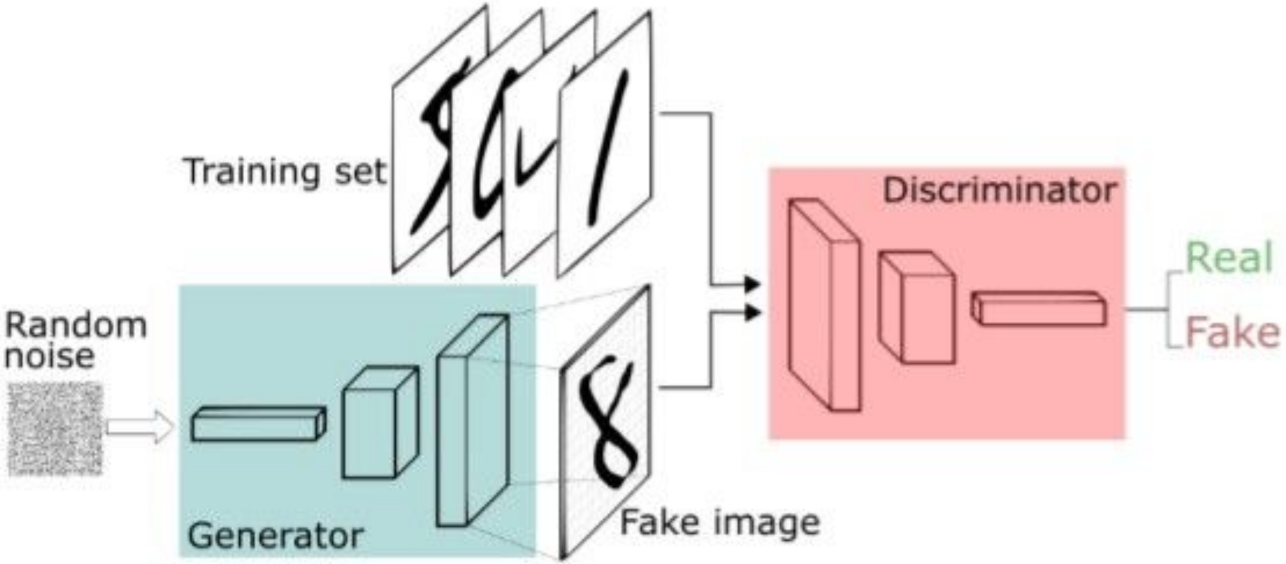
How to deal with missing values in Multivariate Time series?

- Deletion methods
- Imputation methods
  - RNN, KNN, Matrix Factorization
  - Statistical methods
    - Mean, last observed

GANs as a solution

- Generative Adversarial Networks (GANs) have been successful in the image domain
  - Successful in both synthetic image generation and imputation
  - Famous example is DeepFakes
- What are GANs?
  - Generative models, i.e. models that generate synthetic data
  - Comprised of a Generator (who generates synthetic data) and a Discriminator (Classifies data as either real or synthetic)
  - Generator and Discriminator trained in an adversarial loop
  - Essentially a MinMax game

GAN Architecture



GANs in time series imputation

- State of the art performance
- Works as such:
  - Train GAN to generate data based on datasets underlying distribution
  - Generate time series similar to the one that needs imputing
    - Done by noise optimization
  - Impute missing values from generated sample
- Problem:
  - Noise optimization is inefficient and time consuming

Motivation summarised

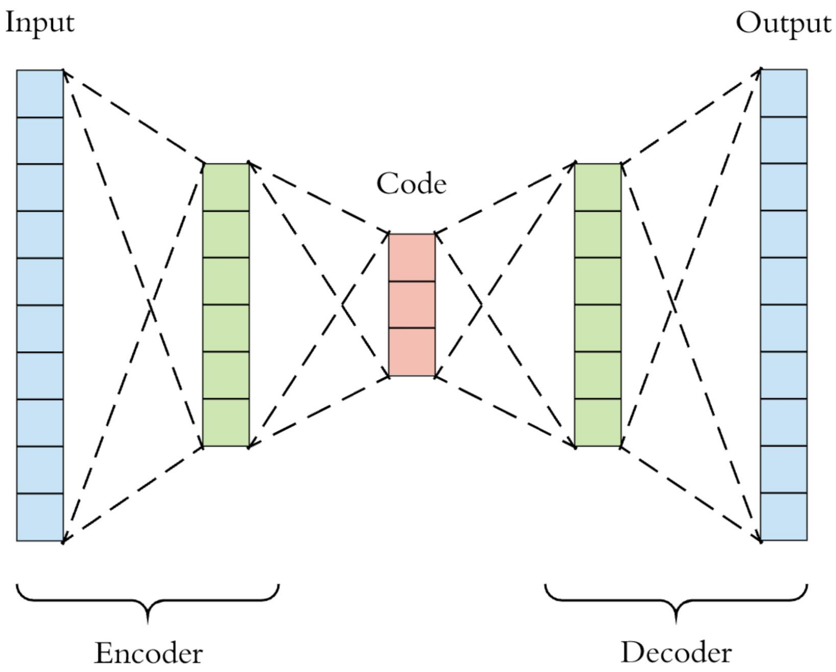
Address the challenge of missing values in multivariate time series by developing an end-to-end GAN-based imputation model, that avoids the noise optimization that hinders the state of the art

Related works

- Deletion method
- Imputation method
    - Statistical and machine learning based
- GANs in image domain
- GANs in time series domain

Background for Method: Autoencoders

- Machine learning model that learns the identity function of data
- Why?
  - Creates sparse representation of data



Overview of E2 GAN

- Generator is an autoencoder rather than a decoder
  - Uses recurrent cells in encoder and decoder
  - Input incomplete time series x
  - encodes it into latent representation z
  - decodes it into complete time series x'
  - Idea is to impute missing values in reconstruction of z
  - Uses Mean Squared Error to force x' close to x
- Discriminator is a recurrent neural network, that tries to distinguish between incomplete time series x, and complete time series x'

Recurrent Cells

- Recurrent cells in encoder, decoder and discriminator are GRUI units
    - Gated Recurrent Unit for Imputation
    - Much like regular GRUs...
        - ...But takes into account the values that are missing from time series, so as to decay historical values

Additional elements of the architecture

- Noise added to inputs to the autoencoder
  - Beneficial as one trains the decoder to “repair” i.e. impute the values of input
- Discriminator outputs probability of authenticity of time series

Experiments

- E2 GAN tested on two real world Datasets
  - PhsyioNet
    - 80% missing medical dataset
    - Records patient data over 48 hours on ICUs
    - Imputation methods tested by training classifier on imputed data
  - KDD
    - Meteorological dataset with air quality and weather data collected over a year
    - 15% missing data
    - Models are compared on two tasks
      - Imputation task: reconstruction error on various percentages of missing data
      - Downstream task: classifier/regression accuracy on imputed data

Baseline imputation methods:

- **Statistical imputation methods**: We simply replace the missing values with **zero** value, **mean** value and **last observed** value.
- **Matrix Factorization (MF) imputation** [Acar *et al.*, 2010]: MF method is used to factorise the incomplete matrix into low-rank matrices and fill the missing values.
- **KNN** [Hudak *et al.*, 2008]: The missing values are replaced by using the k nearest neighbor samples.
- **MICE** [White *et al.*, 2011]: Multivariate Imputation by Chained Equations (MICE) fills the missing values by using iterative regression model.

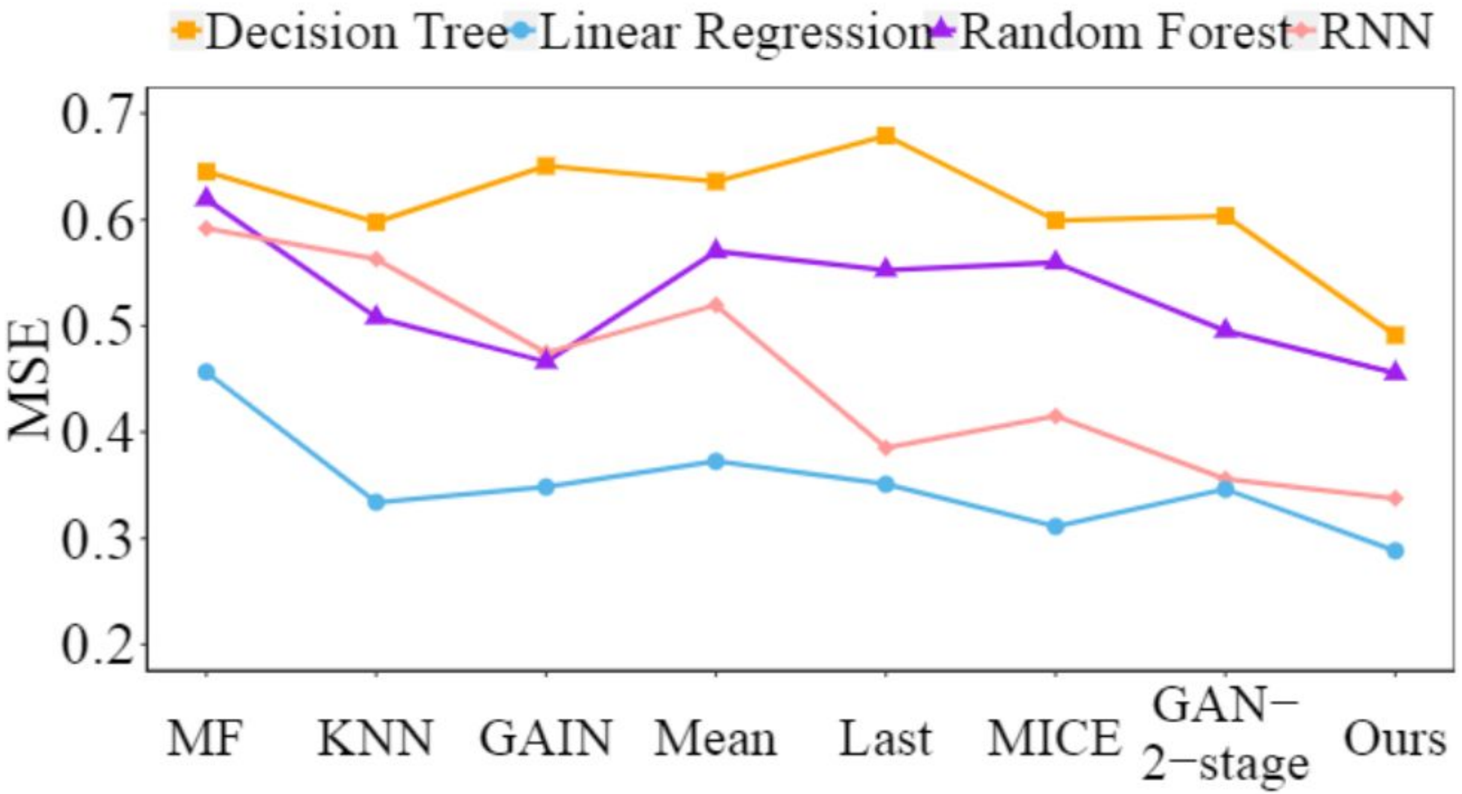
Baseline methods continued...

- **GRUD** [Che *et al.*, 2018]: GRUD can be used to impute missing values. We use it as one of the baselines.
- **GAN-2-stage** [Luo *et al.*, 2018]: This method uses a GAN based two-stage method to impute missing values. We call this method as "GAN-2-stage".
- **GAIN** [Yoon *et al.*, 2018]: GAIN is another GAN based imputation method that uses a hint vector to impute the missing values.
- **BRITS** [Cao *et al.*, 2018]: This method is one of the state-of-the-art methods that uses bidirectional recurrent network to impute time series.

MSE of imputation reconstruction on KDD

Miss -ing	Last	Mean	KNN	MF	MICE	GAIN	GAN-2 -stage	E^2GAN
10%	.614	.374	.465	.382	.468	.378	.355	**.334**(5.9%)
20%	.701	.578	.604	.598	.573	.557	.532	**.523**(1.7%)
30%	.812	.686	.640	.633	.662	.635	**.599**	.606(−1.2%)
40%	.808	.681	.685	.676	.678	.664	.652	**.650**(0.3%)
50%	.788	.747	.723	.710	.727	.693	**.653**	.657(−0.6%)
60%	.807	.801	.750	.722	.740	.732	.714	**.709**(0.7%)
70%	.885	.835	.783	.782	.825	.772	.751	**.747**(0.5%)
80%	.933	.827	.824	.791	.919	.798	.776	**.763**(1.7%)

Prediction results of regressors on imputed datasets



AUC score of mortality classifiers on imputed datasets of PhysioNet

Method		Zero	Mean	Last	MF	KNN	MICE	GAIN	GAN-2-stage	E^2GAN
SVM	Poly	0.7378	0.6774	0.7709	0.3226	0.6773	0.6773	0.7605	0.7725	**0.7892**(2.2%)
	Linear	0.6436	0.6582	0.6672	0.6583	0.6582	0.6584	0.7185	0.7185	**0.7464**(3.9%)
	Sigmoid	0.5285	0.7887	0.7452	0.7886	0.7885	0.7890	0.7967	0.7921	**0.8070**(1.3%)
	RBF	0.5000	0.8043	**0.8213**	0.8045	0.8044	0.8043	0.8178	0.8157	0.8201(−1.4%)
/	RF	0.6937	0.6906	0.7443	0.7074	0.7003	0.6882	0.7302	0.7546	**0.7998**(5.7%)
/	LR	0.6586	0.6620	0.6701	0.6846	0.6120	0.6113	0.7122	0.7012	**0.7677**(9.4%)
/	RNN	0.7659	0.8423	0.8362	0.8495	0.8534	0.8521	0.8431	0.8603	**0.8724**(1.4%)

Table 2: The AUC score of mortality prediction by different classification models trained on datasets that are imputed by different methods.

Additional empirical results

Discovered that E2 GAN performs sufficiently faster than the previous state of the art GAN method, using noise optimization

In conclusion

- The paper proposes a novel end-to-end model, called E2 GAN, for imputing missing values in MTS
- Novelty lies within the generator of the GAN, which utilises a autoencoder, that imputes incomplete time series.
- It produces state of the art imputation, whilst being more efficient than previous solutions

Sources

- https://miro.medium.com/max/1000/1*Q7sZcfRj2M64GDD1ncvoCA.jpeg
- https://miro.medium.com/max/3524/1*oUbsOnYKX5DEpMOK3pH_lg.png
- https://ijac.org/Proceedings/2019/0429.pdf