Disentangled Variational Auto-Encoders for Explaining ECG Beat Embeddings

Tom Van Steenkiste, Dirk Deschrijver, and Tom Dhaene

Ghent university - imec, IDLab, Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium, tomd.vansteenkiste@ugent.be

Abstract. Electrocardiogram signals are often used in medicine. An important aspect of analyzing this data is identifying and classifying the type of beat. Advancements in neural networks and deep learning have led to high classification accuracy. However, adoption of such models into clinical practice is limited due to the black-box nature of the classification method. In this work, the use of variational auto encoders to learn human-interpretable encodings of the beat types is analyzed.

Keywords: interpretable model · ECG beat classification · deep learning · disentangled variational auto encoder.

1 Introduction

Electrocardiogram (ECG) measurements are used throughout all branches of medicine. Deep learning classifiers have shown high accuracy in classifying ECG beats. Nonetheless, the adoption of such neural network models into clinical practice is limited by the lack of model interpretability and trustworthiness.

To reduce the complexity of neural networks, dimensionality reduction techniques such as Auto Encoders (AE) have been proposed. These techniques force the model to use a lower-dimensional embedding of the data. However, there can still be complex interactions across individual components of the embedding leading to difficulties in interpretation.

To disentangle such interactions, disentangled variational auto encoders (β-VAE) were introduced [1]. Such models are capable of learning disentangled generative embeddings by forcing the model to represent the information in as few dimensions as possible, by using a probabilistic interpretation of the embedding. In this work, the use of such a β-VAE is investigated for creating an interpretable and explainable ECG beat embedding that can subsequently be used in a classification system.

2 Experimental Setup

The variational AE (VAE) transforms regular AE models into a probabilistic method. The embedding layer of the AE is replaced with two vectors Z_μ and

Z_σ, and a sampler drawing random samples from the distribution $\mathcal{N}(Z_\mu, Z_\sigma)$. Independence and interpretability of the embedding dimensions is stimulated by the addition of the KL-divergence D_{KL} to the loss function of the model which encourages the embedding to consist of independent and standard normally distributed dimensions. This effect can be enhanced in β-VAE via the addition of a hyperparameter β to balance the reconstruction loss with D_{KL}. In this work, an embedding is created using a standard AE and a β-VAE on data of normal and paced beats from the MIT-BIH Arrhythmia dataset. Both models are trained with an embedding size of 10.

3 Results and Discussion

The models are both able to learn an embedding which can separate the beat types. Whereas the AE model uses all ten available dimensions, the β-VAE model only has two active dimensions. Perturbing one of the dimensions in the AE model results in an unrecognizable beat due to large correlations and dependencies across the different dimensions. On the other hand, perturbing one of the active dimensions of the β-VAE, reveals base shapes with a smooth transition across the entire embedding space as shown in Fig. 1.

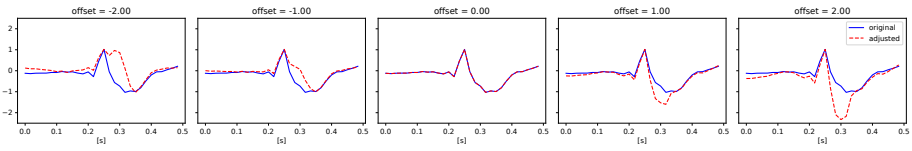

Fig. 1: Smooth transition of the decoder output when perturbing an active dimension of the β-VAE.

4 Conclusion

By extending deep learning models to include a β-VAE embedding, representative beat patterns can be identified leading to interpretable, explainable and independent embedding dimensions. The resulting model is no longer black-box and beats can be represented as combinations of learned independent base beats.

References

1. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework (2016)