STRATEGIC ATTENTION LEARNING FOR MODALITY TRANSLATION

Jonathan Martinez[1], Ali Akbari[2], Kaan Sel[3], and Roozbeh Jafari[1,2,3]

[1] Departments of Computer Science and Engineering, [2] Biomedical Engineering,
[3] Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

ABSTRACT

Novel wearable sensor modalities, such as bio-impedance (Bio-Z), are being introduced and often provide various advantages over current state-of-the-art in terms of accuracy, sensing coverage, or convenience of wear. The principal challenge, however, lies in the ability to interpret the sensor reading by healthcare providers. In this work, we propose a two-stage deep learning framework that leverages a novel attention mechanism to translate Bio-Z signals to highly interpretable electrocardiogram (ECG) waveforms while also predicting translation uncertainty. Our experiments indicate a 66% improvement in accuracy for 1D-CNN based models to perform competitively with more sophisticated hybrid CNN-LSTM based models in a fraction of the training time while also providing a valid uncertainty measurement.

***Index Terms*—** modality translation, sequence-to-sequence, deep learning, uncertainty quantification

1. INTRODUCTION

Continuous monitoring of physiological parameters is of paramount importance for health diagnosis and prevention [1]. However typical measurement systems could be uncomfortable due to bulky equipment or require a stationary patient. Wearable sensors provide opportunity to capture parameters such as electrocardiograms (ECG), bio-impedance (Bio-Z), or photoplethysmogram (PPG) more conveniently [2]. Novel wearable sensors, for example Bio-Z, are able to enhance wearability and information extraction [3]–[5]. Yet the challenge is the ability of the care providers to interpret these new modalities. The aim in this study is to translate a specific physiological signal modality to other modalities that are more interpretable and familiar to care providers. Since such models are desired for health diagnostic use, it is also critical that the translation model is aware of the quality of its own translation. Therefore, to enhance usability of our system, we propose a translation model capable of quantifying uncertainty or confidence of its own translations.

In this work, we use a specific case study of translating the changes in the Bio-Z signals to highly-interpretable ECG waveforms. Bio-Z is the measure of impedance of biological cells and tissues with respect to a very small electric current flow. The variations in the Bio-Z signal correspond to heart and lung movements, muscle contractions and blood flow [4], [6]. This non-invasive signal can be captured from various locations of the body, including the wrist, allowing the whole Bio-Z sensing system to be integrated in a wearable device such as a smart-watch [7]. However, this is a direct example of a physiological parameter that is convenient to collect from the user but is not easy to interpret by physicians. Thus, we translate it to an ECG which is a well-studied bio-potential signal that is generated by the electrical activity of the heart but requires the sensors to be placed around the heart area, causing problems for the integration of the system on wearable platforms such as smartwatches, or armbands.

Modality translation aims to interpret the relationships between two signals that were generated by distinct processes – for example, image-to-text, text-to-speech, English-to-French, *etc*. Consequently, the mapping between two modalities is typically complex and non-linear making it difficult to model. Despite this, data-dependent methods such as neural network based autoencoders have achieved state-of-the-art performance in modality translation by learning a latent representation that distinctly represents the statistical structure of the input [2]. This was then further extended with attention mechanisms that further focus the analysis related to each component of the target modality onto the most relevant features of the input [8]. Preprocessing techniques can assist such mechanisms by leveraging the properties of specific modalities. For example, noun, verb, adjective, and topic embeddings help encourage learning in sentence translation [9] and images may be segmented based on the objects that reside within [10]. This then embeds each component of the input and output modality to a categorical form which allows flexibility with the amount of precision in the model's translations. However, it is not always possible to perform such modality specific preprocessing techniques, especially, for continuous signals that are not yet highly interpretable, such as the case with the Bio-Z signals.

With respect to measuring uncertainty in translation models, prediction intervals (PIs) or confidence intervals (CIs) are typically constructed. PIs generally estimate upper and lower limits of actual prediction values while CIs focus on probabilistically bounding accuracies [11]. Methods such as the Delta method [12] and the Bayesian method [13] achieve this by analyzing the distributions of neural network parameters, however, this does not directly measure data-dependent uncertainty that are due to noisy or high-variance input instances which were not previously seen in training. Other approaches which do target such data-dependent uncertainty often rely on the distribution of network

predictions, such as with mean variance estimation methods [14], however, this depends on prior familiarity the expected target. This may be resolved with Bootstrap methods [15] that make use of ensembles of neural network predictors by comparing the range of their outputs. However, this requires the management of multiple models and several iterations of execution to take place before uncertainty measurement processes can take place. Although very effective, all of these techniques are rather computationally expensive in runtime.

In this work, we propose a generalizable two-stage modality translation deep learning model to translate a chest Bio-Z signal to a highly interpretable ECG waveform modality. High precision is achieved through our novel filter-based attention mechanism which isolates learning tasks to a *Morphology Translation Model* and an *Amplitude Correction Model*. In this framework, the first stage learns to translate Bio-Z signals to ECG waveforms with filtered-out amplitude variations, and the second stage learns how to correct this previously learned morphology to its appropriate scale to achieve a precise ECG translation. Our proposed model also contains a natural mechanism to estimate uncertainty of the model through comparison of the outputs of two stages.

Our proposed contributions may be summarized as:

- A generalizable two-stage modality translation deep learning model that leverages a novel attention mechanism best-suited for continuous signals by isolating morphology and amplitude learning
- A novel uncertainty mechanism for translated morphology
- Demonstration of framework effectiveness on a uniquely constructed dataset of Bio-Z and ECG signals

2. PROPOSED FRAMEWORK

The target ECG waveforms that we aim to predict obtain complex characteristics with variations in both amplitude and morphology of the signal. We propose a supervised two-stage deep learning framework to strategically attend to morphology and amplitude learning separately, as shown in Figure 1. The first stage, *Morphology Translation Model*, produces the initial prediction by generating the target morphology of a normalized ECG signal. This is especially important for ECG waveforms since peak and wave location within a window of data describes critical cardiovascular behaviors. The second stage, *Amplitude Correction Model*, maps critical features of the initial prediction to their accurate amplitudes. This is achieved via strategic normalization of the input signals and joint signal analysis. We then compare outputs of the first and second stage to measure data-dependent uncertainty.

2.1. Morphology Translation Model

To solely focus on waveform morphology in the first stage deep model, we filter amplitude features out from the raw input instance vectors X_{raw}, which is Bio-Z signal in our study, and target Y_{raw}, which is the target ECG signal, by independently scaling each window of data. Hence, the

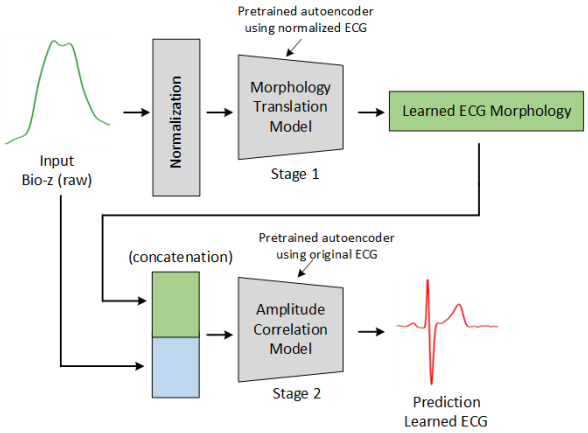

Figure 1. Proposed two-stage translation model.

minimum and maximum value of each first-stage input instance will be 0 and 1 [16]. Then, the scaled Bio-Z input, X_{SC}, will then be passed to the *Morphology Translation Model*, which will aim to predict the corresponding scaled ECG waveform, Y_{SC}, whose amplitude features were also filtered out. This is achieved through a sequence-to-sequence learning autoencoder which first encodes the input Bio-Z signal to a latent representation, H_{SC}, which best represents its critical and distinguishing features. Then, the H_{SC} is decoded to generate the target ECG waveform modality. We implement this with two architectures: a pure one-dimensional convolutional neural network (1D-CNN) encoder with a single dense layer decoder and a hybrid model which procures the decoder with a long short-term memory (LSTM) layer to further analyze the dependency between the extracted features from the convolutional encoder.

2.1.1. One-Dimensional Convolutional Layers

Convolutional neural network (CNN) layers have been largely used to handle spatial information due to their ability to capture complex input features [17]. Our pure 1D-CNN architecture consists of two 1D-CNN layers, one maxpooling layer, and a dense neural network output layer to produce the ECG waveform morphology. The maxpooling layers subsample the extracted features produced by the 1D-CNN layers by sliding over regions of the input and maintaining only that feature with the max value. This is based on the assumption that the max values represent the features with higher activation due to their significance. Lastly, the dense neural network layer completes the morphology translation process by generating the corresponding ECG waveform, $\hat{Y}_{SC}$.

2.1.2. Long Short-Term Memory

Recurrent neural network (RNN) layers also possess ideal characteristics to handle sequential modeling tasks. Particularly, time-steps of a series are analyzed sequentially before their hidden result is concatenated to the next time-step to be jointly analyzed. This incorporates the notion of retained memory for the model and has proven to great advantage for deep models. For our experiments, we test the

abilities of a hybrid CNN-LSTM architecture which has demonstrated to achieve state-of-the-art performances when combined with CNN based encoders [18] . Therefore, we pass the hidden state, h_{SC}, produced from the 1D-CNN encoder into the LSTM layer before then feeding its resulting hidden state, h_{LSTM}, into the output dense layer described in equation 3 to produce the scaled ECG waveform, $\hat{Y}_{SC}$.

2.2. Amplitude Correction Model

The output of the first stage deep model, $\hat{Y}_{SC}$, will be the independently normalized version of the target ECG waveform which does not possess amplitude features of the target modality. This will be concatenated column-wise to the raw Bio-Z signal state, X_{raw}, which still contains amplitude feature information, to produce a two-dimensional input instance. This re-introduces the previously discarded amplitude information to $\hat{Y}_{SC}$ before it is passed into the second stage autoencoder – the two aforementioned 1D-CNN and hybrid CNN-LSTM based architectures are maintained for this learning task. Therefore, the 1D-CNN layers of the second stage encoder are able to consider the two signals with respect to the other in a sequential fashion, and encode them to a joint latent representation that represents all raw features extracted by the Bio-Z signal (describing both morphology and amplitude characteristics) and the modality morphology translation prediction as a one-dimensional vector. The final decoder can then analyze this to produce the final translation of the target raw ECG modality, $\hat{Y}_{raw}$.

2.2.1. Model Uncertainty

Since modality translation is inherently learned through the *Morphology Translation Model* and should only be rescaled by the *Amplitude Correction Model*, signal morphology should be maintained between Y_{SC} and $\hat{Y}_{raw}$. Thus, the predictions of both stages should be in agreement regarding the shape of the signals, although they have different amplitude characteristics. This can be measured by the Pearson correlation between the two predictions. High correlation indicates agreement between the predictions of the two stages while low correlation indicates disagreement or low confidence of what the true morphology should be. So, when testing, there should be a linear relationship between prediction accuracy and confidence. In runtime, a threshold coefficient, ρ_r, can then be empirically learned to determine when a translation should be ignored or when the input instance is too unfamiliar to the model. This idea of agreement is related to null detection via autoencoders where reconstruction error measured in the test phase is used to determine the input samples that do not belong to the same class as previously seen training samples [19].

3. EXPERIMENTS AND RESULTS

For this study, we used twenty minutes of Bio-Z and ECG signals captured from ten healthy subjects, under IRB approval IRB2017-0086D by Texas A&M University, using the methodology presented in [4]. Both raw ECG and Bio-Z pairings were downsampled to 333Hz before segmenting signals based on one complete heartbeat. This provided us with a dataset of 13,704 instances which was split for training and testing. All translation models were trained for 100 epochs while the two-stage frameworks would split these iterations to 50 for the first stage and 50 for the second stage. This assured that each model was given an equal amount of iterations for training. We measure translation accuracy based on normalized mean squared errors (MSE) and Pearson correlation coefficients between target and translation pairings. Accuracy of the proposed uncertainty metric is also evaluated through its relationship to model performance.

3.1. Translation Quality

Figure 2 shows sample translations performed by the 1D-CNN based two-stage framework. Here we show X_{raw} (green), $\hat{Y}_{SC}$ (blue), Y_{raw} (black), and $\hat{Y}_{raw}$ (red) for six pairings over the time-steps of a window of data for an instance. We observe that morphology is maintained for the second-stage prediction while only amplitude is adjusted. This figure also demonstrates the variety of ECG types in our dataset – standard, possible J point notching, and possible inverted QRS complex. Despite this, all architectures were able to perform accurate translations with differences in performance being due to the amount of precision.

Table I shows the average normalized MSE (NMSE) between $\hat{Y}_{SC}$ and Y_{raw} for single-stage and two-stage implementations of 1D-CNN and CNN-LSTM based architectures. This primarily expresses the precision of amplitude learning. The MSE is normalized in that all instances were scaled with respect to the whole dataset before training and testing. So, the minimum value in the whole dataset is 0 while the maximum value in the whole dataset is 1. While the more sophisticated single-stage CNN-LSTM hybrid significantly outperforms single-stage 1D-CNN, our proposed two-stage framework enabled 1D-CNN to outperform and be competitive with both the single-stage and two-stage CNN-LSTM. This is especially significant since 1D-CNN can be trained on the scale of minutes while the hybrid CNN-LSTM case tends to take up to a few hours on a

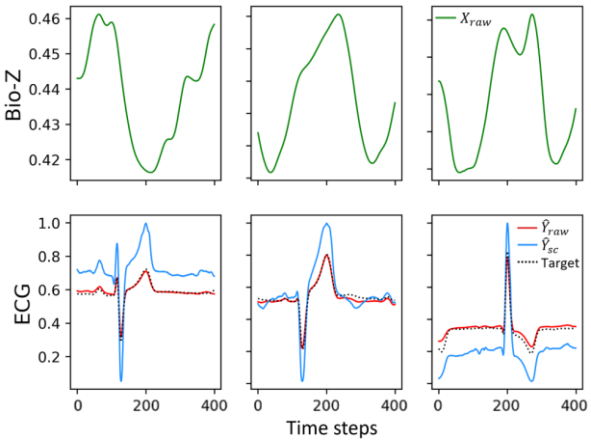

Figure 2. Two-stage 1D-CNN framework translations.

TABLE I NMSE FOR DIFFERENT ARCHITECTURES

Method	NMSE
1D-CNN Autoencoder	1.52×10^{-3}
CNN-LSTM Autoencoder	7.72×10^{-4}
Two-Stage 1D-CNN Autoencoder	7.66×10^{-4}
Two-Stage CNN-LSTM Autoencoder	8.74×10^{-4}

standard computer. This plays a big factor in model selection when training for personalization on an embedded system that may have strict computational constraints.

To more directly analyze the quality of waveform morphology translation, Table II shows the average Pearson correlations between each input and target pairing. Generally, if the critical peaks and foots of the ECG waveform occur in the same time-steps of the translation and the target, then the correlation score will be highest. Again, our proposed framework has the greatest impact on 1D-CNN. Although it was not able to outperform the one-stage or two-stage settings for hybrid CNN-LSTM, translation performance was very competitive. As aforementioned, LSTM layers possess characteristics which make them better suited for sequence modeling compared to pure 1D-CNN layers which also contributes to the marginal impact that our proposed two-stage framework impresses on hybrid CNN-LSTM. However, this also greatly supports the impact of our proposed two-stage framework showing that it may enable an efficient learning model to match the complex learning abilities of a more sophisticated model in a fraction of the time. This is ideal for the embedded system scenario.

TABLE II CORRELATION RESULTS FOR DIFFERENT ARCHITECTURES

Method	Correlation
1D-CNN Autoencoder	0.855
CNN-LSTM Autoencoder	0.898
Two-Stage 1D-CNN Autoencoder	0.876
Two-Stage CNN-LSTM Autoencoder	0.887

3.1. Uncertainty Analysis

We analyze our proposed uncertainty mechanism by plotting its metric (y-axis) against the quality of morphology translation (x-axis) as measured in the previous section. Figure 3 displays this comparison for the two-stage 1D-CNN, and we can see that there is a relatively linear relationship so that when our proposed uncertainty is measured at $\propto$, then translation accuracy is also measured around $\propto$, which shows the ability of our model to predict the quality of its own translations. We can also empirically determine that a potential value for the threshold value, ρ_r, previously discussed in Section 2.2.1, may be approximated to 0.8.

Figure 4 shows an example of uncertainty measurements, r_u, produced by our proposed two-stage framework for 1D-CNN. The top three plots show the positive effects when uncertainty is low as it is indicated by high correlations over

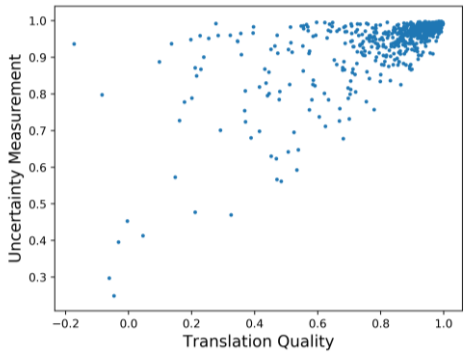

Figure 3. Uncertainty metric versus translation quality for two-stage 1D-CNN.

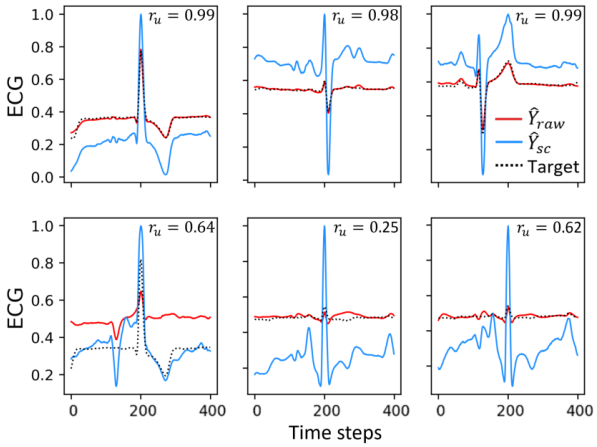

Figure 4. Demonstration of our proposed uncertainty metric.

0.98. For this case, we can see how morphology is only scaled between the translations of $\hat{Y}_{SC}$ and $\hat{Y}_{raw}$. On the contrary, the bottom three plots show the negative effects when uncertainty is high as it is indicated by low correlations below the aforementioned ρ_r value of 0.8; particularly, morphology is relatively lost in critical regions of the waveforms.

5. CONCLUSION

We proposed a two-stage modality translation framework that leverages a novel attention mechanism to produce highly interpretable ECG waveforms from less informative, yet conveniently collected, Bio-Z signals while producing a valid uncertainty metric. Our framework improved 1D-CNN accuracy by 66%, enabling it to achieve competitive translations in a fraction of the training time compared to hybrid CNN-LSTM. This is especially significant when training for personalization on embedded systems where resources are heavily constrained.

6. ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health, under grant 1R01EB028106-01 and National Science Foundation, under grants CNS-1738293 and CNS-1734039. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations.

REFERENCES

[1] S. Majumder, T. Mondal, and M. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 1, p. 130, 2017.

[2] I. Sutskever, O. Vinyals, and Q. V Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[3] B. Ibrahim, D. A. Hall, and R. Jafari, "Bio-Impedance Spectroscopy (BIS) Measurement System for Wearable Devices," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2017.

[4] K. Sel, J. Zhao, B. Ibrahim, and R. Jafari, "Measurement of chest physiological signals using wirelessly coupled bio-impedance patches," in *Engineering in Medicine and Biology (EMBC)*, 2019.

[5] B. Ibrahim and R. Jafari, "Cuffless Blood Pressure Monitoring rom an Array of Wrist Bio-impedance Sensors using Subject-Specific Regression Models: Proof of Concept," *IEEE Trans. Biomed. circuits Syst.*, 2019.

[6] B. Ibrahim and R. Jafari, "Continuous blood pressure monitoring using wrist-worn bio-impedance sensors with wet electrodes," in *IEEE Biomedical Circuits and Ssytems Conference*, 2018.

[7] B. Ibrahim, J. McMurray, and R. Jafari, "A wrist-worn strap with an array of electrodes for robust physiological sensing," in *Engineering in Medicine and Biological Society*, 2018, vol. 1, pp. 4313–4317.

[8] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1243–1252.

[9] L. Wu, F. Tian, L. Zhao, J. Lai, and T.-Y. Liu, "Word attention for sequence to sequence text understanding," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[11] T. Heskes, "Practical confidence and prediction intervals," in *Advances in neural information processing systems*, 1997, pp. 176–182.

[12] J. Xu and J. S. Long, "Using the delta method to construct confidence intervals for predicted probabilities, rates, and discrete changes," *Stata J.*, 2005.

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[14] D. A. Nix and A. S. Weigend, "Estimating the mean and variance of the target probability distribution," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, 1994, vol. 1, pp. 55–60.

[15] J. Franke and M. H. Neumann, "Bootstrapping neural networks," *Neural Comput.*, vol. 12, no. 8, pp. 1929–1949, 2000.

[16] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, 2001.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156--3164.

[19] A. Akbari and R. Jafari, "An Autoencoder-based Approach for Recognizing Null Class in Activities of Daily Living In-the-wild via Wearable Motion Sensors," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3392–3396.