Machine Learning Engineer Nanodegree

Capstone Proposal

Alexander Villasoto

April 10, 2020

Proposal

Domain Background

One of the sectors that deep learning has made a tremendous impact is in healthcare[1]. In fact, researches on the medical sector involve deep learning methods in their analyses. Anyone could agree that it is imperative that disease diagnosis early on the treatment is equally if not more important as the treatment itself. Therefore, these advancements when applied to the health sector are really helpful in saving people's lives since one of the benfits of a carefully trained deep learning model is its accuracy and speed in generating results. Truly, this will not only benefit patients but doctors as well [2].

Out of all the diseases that need utmost priority of the medical sector as well as researchers in the field of healthcare, heart disease is the one that needs careful attention. In the United States, heart disease is the leading cause of death for men, women and primary racial and ethnic groups [3]. Alarmingly, one person dies every 37 seconds in the United States from heart disease and about 647,000 Americans die from heart disease each year [3,4]. Fortunately, milestones of deep learning and artificial intelligence have lead several researchers to create fast and accurate models for heart disease detection [1].

One of the ways to do this is through the use of previously-available digitized electrocardiogram (ECG) data and apply deep learning methods to classify specific heart disease [5]. With the already available dataset and publicly available paper tackling heart disease detection, the author decided to to implement a deep learning model that would attain the same level of reported accuracy if not more using a state-of-the-art deep learning framework.

[1] Ahuja A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. PeerJ, 7, e7702. https://doi.org/10.7717/peerj.7702

[2] Aljanabi, Maryam & Qutqut, Mahmoud & Hijjawi, Mohammad. (2018). Machine Learning Classification Techniques for Heart Disease Prediction: A Review. International Journal of Engineering and Technology. 7. 5373-5379. 10.14419/ijet.v7i4.28646.

[3] Heron, M. Deaths: Leading causes for 2017 [PDF – 3 M]. National Vital Statistics Reports;68(6). Accessed November 19, 2019.

[4] Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 [PDF-494K]. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

[5] Kachuee, Mohammad, Shayan Fazeli, and Majid Sarrafzadeh. "ECG Heartbeat Classification: A Deep Transferable Representation." 2018 IEEE International Conference on Healthcare Informatics (ICHI) (2018): n. pag. Crossref. Web.

Problem Statement

The problem revolves around the utilization of a publicly available digitized EKG dataset to create a deep learning model that would attain the same level of the paper's reported accuracy, 93.4% on the MIT-BIH dataset if not more using the PyTorch framework. The author will try and replicate the model definition of the paper with several improvements to prevent overfitting and employ functions for training normalization. The author will implement train -> test -> validation strategies and report the results on a per class basis. Since the problem is multiclass classification, accuracy and loss are the major metrics that the author need to increase and decrease respectively. Thus, the author will make sure that the results are communicated via reproducible document and communicate the environment specifications along the way.

Datasets and Inputs

The author's project is all about creating a model that categorizes and predicts several heartbeat observations. He will use the MIT-BIH Arrythmia Dataset, which is readily available on Kaggle. This project is based on the paper entitled ECG Heartbeat Classification: A Deep Transferable Representation by Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh; Shayan Fazeli being the owner of the dataset that the author is working with. The paper can be found in a publicly available arxiv document.

For the MIT-BIH Arrhythmia Dataset, all the samples are cropped, downsampled and padded with zeroes when necessary to the fixed dimension of 188 as per the Kaggle dataset remark which can be found on the dataset above. The dataset that is provided is divided into separate CSV files, one for training and one for testing. The author is responsible for dividing the training dataset into two to realize a validation dataset (see the project design for more details).

The dataset has five classes, 'N' assigned to 0, 'S' assigned to 1, 'V' assigned to 2, 'F' assigned to 3 and 'Q' assigned to 4. Specifically, classes above refer to the beat annotations enumerated below:

1. N - Normal beat
2. S - Supraventricular premature or ectopic beat (atrial or nodal)
3. V - Premature ventricular contraction
4. F - Fusion of ventricular and normal beat
5. Q - Unclassifiable beat

These classes can be found on the PhysioBank Annotations under the Beat Annotations section.

Additionally, each observation in this particular dataset are actual recorded heartbeats of 47 different subjects at a sampling rate of 360Hz annotated by at least two cardiologists under five classes mentioned above in accordance with the Association for the Advancement of Medical Instrumentation (AAMI) standard. The last column indicates how each heartbeat is classified, that is the classification of a specific beat per the beat annotations above.

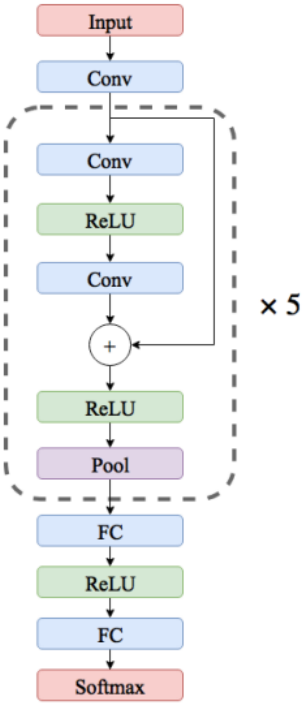
Solution Statement

Adhering to the problem statement, the author would create a reproducible data product that delineates the end-to-end approach in solving the problem, from exploratory data anaysis to model evaluation. Pertinent observations will be communicated every step of the way. The final product will be a model that accurately discern between multiple heart disease classes using PyTorch, a state-of-the-art deep learning framework. The author choose PyTorch deep learning framework to be used in realizing the multiclass classification model for heart disease categorization due to its flexibility in creating the model definitions as well as its ability to enable end-to-end pipeline in solving deep learning problems. That is, one can use PyTorch to directly communicate with native and commonly-used custom Python data structures (e.g. numpy arrays). Also, there are several cloud offerings like Amazon SageMaker and IBM Watson that directly supports the said framework for model creation and deployment via an accessible API with ease.

Benchmark Model

As you can see, the author chose to implement a deep learning algorithm instead of using classical machine learning algorithms in this project. For one thing, the paper which is the basis for this project which generously discussed their own model implementation is a deep learning one. Another reason is the nuances of the signals that need to be distinguished by the model are too complicated for classical machine learning algorithms to find separations for. Finally, it is also an opportunity for the author to test his knowledge of building deep learning models that can effectively applied to a real-world problem.

The said paper used several 1-dimensional convolutional layers with RELU activations on the learning layer and linear fully-connected layer at the end (see the figure). The model also had softmax activation at the very end to enable categorical cross entropy as model criterion and negative log likelihood loss for the loss function. The paper reported 93.4% accuracy across all test observations. This is what the author is trying to achieve. An accuracy on par if not more than the reported accuracy of the paper.



Evaluation Metrics

As for the criterion, the author will use categorical cross-entropy with negative log-likelihood loss as loss function and adaptive learning rate (ADAM) for model optimizer. All of which will be done using PyTorch helper functions readily available for the author to use. To test whether the model is improving, he will use validation loss as metric. This is the same as the accuracy in reverse, accounting for the overall performance of the model by minimizing the loss. The author will also assess the model performance by recording the training and validation loss across all observations. Every time the validation loss decreases, he will save the best model to be used for the test set. The author will train the model for 50 epochs with batch size of 32 and will save the model with the lowest validation score or after he deems that the validation loss is in its converged state. The choice for the loss function, optimization scheme and metric are pretty much arbitrary and are not discussed on the paper but are fairly common to be used in multiclass classification tasks for deep learning. Finally, he will make the final check of accuracy across all classes and observations in the test set using the best model saved.

Project Design

Project design describes the step-by-step procedures in creating the data product. These procedures span from exploratory data analysis section to model evaluation.

Exploratory Data Analysis

Like any other reproducible data analysis document, the author will explore the MIT-BIH dataset and will visualize how each classes differ with each other through appropriate visualizations. The author will include exploration of observation distributions and explain motivations for necessary feature engineering practices that need to be implemented.

Data Wrangling

Given the dataset specification, the author will proceed to data wrangling process where the author makes sure that the dataset, given its predictors (187 values of each signal) is fit for the modeling stage. He will also make sure that each of the categories mentioned are equally distributed, implementing feature creation, feature engineering and resampling strategies when necessary as response to the prior data exporation. The author will make sure that these strategies will not distort the original observations as well as their class representations. Since we are talking about EKGs in a form of signal, strategies like amplification and stretching signals ever so slightly just to have features created for better modeling is one way to go. Another way is to add Gaussian noise to each beat observation by an acceptable amount of deviations. Either way, the author will record the performance of the model that uses either of the strategies mentioned. After the aforesaid strategies, the author will then divide the dataset into two parts, training set which is 80% of the total observations and 20% for the validation.

Model Building

Based on the model definition described on the paper, the author will build two models, testing both of the resampling strategies mentioned above. One is the direct copy of the model definition described in the paper and the other is an attempt to improve model performance by adding improvements to prevent overfitting and employ functions for training normalization. We use PyTorch framework in both cases.

Model Training

For the model training, the author will train both models with strategies presented on the Evaluation Metrics section above. In every decrease of the validation loss, the model that lead to that decrease will be saved which will then be used for model evaluation. We record the average loss per batch and show the progression of the training process every validation batches. In this phase, we will use the splitted training set that was generated during the data wrangling stage.

Model Evaluation

In model evaluation, the testing dataset will be used to finally test the good model between two model definitions. The author will report the accuracy per class labels as well as the overall accuracy for both models created and trained in the model building and model training sections of the project. Also in this section, the author explains future directions of the project as well as the ideas for model deployment.