MULTIMODAL DATA FUSION MODELS PRETRAINED WITH VICREG

An Undergraduate Research Scholars Thesis

by

NICK CHENG

Submitted to the LAUNCH: Undergraduate Research office at
Texas A&M University
in partial fulfillment of the requirements for the designation as an

UNDERGRADUATE RESEARCH SCHOLAR

Approved by
Faculty Research Advisor: Dr. Bobak Mortazavi

May 2023

Major: Computer Science

RESEARCH COMPLIANCE CERTIFICATION

Research activities involving the use of human subjects, vertebrate animals, and/or biohazards must be reviewed and approved by the appropriate Texas A&M University regulatory research committee (i.e., IRB, IACUC, IBC) before the activity can commence. This requirement applies to activities conducted at Texas A&M and to activities conducted at non-Texas A&M facilities or institutions. In both cases, students are responsible for working with the relevant Texas A&M research compliance program to ensure and document that all Texas A&M compliance obligations are met before the study begins.

I, Nick Cheng, certify that all research compliance requirements related to this Undergraduate Research Scholars thesis have been addressed with my Faculty Research Advisor prior to the collection of any data used in this final thesis submission.

This project did not require approval from the Texas A&M University Research Compliance & Biosafety office.

TABLE OF CONTENTS

Page

ABSTRACT

MULTIMODAL DATA FUSION MODELS PRETRAINED WITH VICREG

Nick Cheng
Department of Computer Science
Texas A&M University

Faculty Research Advisor: Dr. Bobak Mortazavi
Department of Computer Science
Texas A&M University

Prediction models can be applied to hospital intensive care units, or ICUs, in order to improve prediction of adverse patient events through the duration of their stay such as mortality. The current field for mortality and length of stay predictions in the ICU consists of mainly single modal models, such as late fusion models like Shukla & Marlin's Interpolation Network, or gaussian process models such as Futoma et al.s' Multitask Gaussian Network [2, 3]. These models create predictions of patient behavior off a single mode of data such as physiological time series data, or clinical text notes. However, they are incapable of leveraging inter-modal patterns where each mode is strongest, which should allow for improved model performance when compared to single modal models. This is especially applicable in a hospital setting, as different modes of time series data are gathered when patients are admitted, such as clinical notes and machine output. Multimodal fusion models for this context have been proposed, and offer a notable performance improvement when compared to their single modal cousins.

Through my research, I tested whether the addition of a pre-training step to multimodal fusion models in a hospital setting improved model performance. This is because a pre-training step will allow the model to leverage the large amounts of unlabeled data that hospitals accumulate daily. The unsupervised step is also expected to increase model performance when transferred

to hospitals with different operating conditions or little labelled data when compared to standard supervised multimodal models. The pretraining technique used is Variance Invariance Covariance Regularization, or VICReg, which relies on maintaining a minimum desired variance during training to prevent unsupervised branch collapse. While VICReg is a technique mainly used for self-supervised image recognition networks, it can be used in this setting as the different modes of data can be considered as augmentations of the patient condition.

After multiple experiments with varying model architectures and VICReg hyperparameters, my results show that VICReg failed to create any noticeable performance benefit when compared to a baseline multimodal model. Despite this outcome, I still believe that a pre-training step, specifically VICReg can be used to boost multimodal fusion model performance, and I will discuss potential steps that can be added to my current experimentation that could create a performance boost. Mainly, a medical multimodal fusion model can see greater benefits through VICReg pre-training if a large and deep model is used, and a hyperparameter search for the VICReg coefficients are conducted.

My work serves to compare and contrast the usage of a pre-training technique from the image recognition field onto a multimodal fusion model from the medical field in order to improve patient care through the use of intelligent systems that can aid workers in the ICU.

DEDICATION

To my lab group and family for supporting me through this journey.

ACKNOWLEDGMENTS

Contributors

I would like to thank my faculty advisor, Dr. Bobak Mortazavi, and graduate advisor, Ryan King, for their guidance and support throughout the course of this research. Special thanks also go to the STMI Lab for providing the hardware resources necessary to train the models to conduct the research for this report. Lastly, thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

The DATA analyzed/used for Multimodal Data Fusion Models Pretrained With VICReg was provided by Physionet. All other work conducted for the thesis was completed by the student independently.

Funding Sources

No funding was received for this research.

1. INTRODUCTION

Smart systems such as machine learning models can be applied to hospitals to improve patient care in the ICU on tasks such as mortality and length of stay prediction. These tasks are chosen for their high benefit to patient care if predicted correctly, and are trained through the ICU data that hospitals accumulate during regular operation. The data that is fed into these sytems is commonly stored in the form of electronic health records, which contain multimodal data on various measurements taken over the course of the patient's ICU stay like blood oxygen level and clinical notes. Many similar models used for mortality and length of stay prediction only process a single modality of data in isolation when trying to predict patient behavior due to the complexity of the data [2, 4]. However, multimodal models have been shown to outperform single modal models when the content of separate modes of data like clinical text and time series data are fused together into a unified model [2]. The model of interest in this paper will be the Shukla & Marlin multimodal model that uses a novel interpolation network to interpolate physiological time series data, as that data tends to be sparse and irregularly sampled. This model was chosen for its good performance when compared with contemporary models, and for its use of the novel interpolation network.

The focus of this paper will be studying the potential performance benefits gained from adding a pre-training step to the Shukla & Marlin multimodal model, through the use of VICReg. The idea behind adding the pre-training step is twofold as hospitals accumulate a large amount of unlabelled data that normally cannot be used by multimodal models when training, as they require mortality or length of stay labels to train properly. Adding a pre-training step will allow hospitals to use this unlabelled data, and may also make the model more generalizable if used across several different hospitals. Secondly, the pre-training step is also expected to boost classification and regression performance on the mortality and length of stay prediction tasks. This pre-training performance boost through learning a shared latent space has been observed in other fields like image recognition through models that maximize agreement between augmentations such as SimCLR [5].

The pre-training method chosen was VICReg, as it has seen success in the image recognition field through its ability to prevent collapse when pre-training shared latent spaces for self-supervised models. I am specifically interested in VICReg's ability to pre-train branch model architectures with asymmetric branches, such as the clinical text and physiological time series branches of the Shukla & Marlin model. VICReg works through forcing different branches of a model to learn a shared space, which has historically been different augmentations of a base image. Images would be augmented, then both the original and augmented image would be fed into separate branches of a self-supervised network. The network would then have an expander attached at the end in order to facilitate a better latent space, and VICReg would be run across the each batch to minimize the image embeddings which results in more robust model performance.

The key components of VICReg are the variance and covariance terms in the loss applied to the image embeddings, as they force the model to maintain a minimum variance across each of the embedding variables in the batches of images during training. This allows VICReg to minimize the embedding vectors between the base and augmented images without experiencing dimensional collapse during pre-training. VICReg has seen success in pre-training many image recognition networks, and due to its nature of only minimizing the embeddings at the end of the expander outputs, can also be applied to asymmetric networks, like the Shukla & Marlin model which uses separate architectures for the physiological time series branch and clinical text branch.

While VICReg is a technique that has not been historically used for medical models, I believe that VICReg can still be applied here as the separate branches in the multimodal model are analogous to the branches in an image recognition network. Despite the modes of data being different in the Shukla & Marlin model, as long as the final branch embeddings are the same, VICReg can be applied as a pre-training loss as it only requires same sized embeddings. The intuition for learning a shared latent space through VICReg is that the different modes of data for each patient are augmentations of a single ground truth, the patient condition, which are analogous to the augmentations of the images in the image recognition network. The Shukla & Marlin multimodal model is then expected to see a performance boost through using this shared latent space that describes the patient condition, as opposed to learning a separate latent space for the physiological time series and clinical text branches.

Through my research, I will test whether or not adding a pre-training step to a multimodal fusion model improves performance on mortality and length-of-stay prediction tasks. Through my work, I hope to improve patient care through the use of intelligent systems that can aid workers in the ICU.

2. METHODS

The project consisted of two sections, which were selecting and recreating a multimodal model, then applying VICReg as a pre-training step to said model. Selecting the model required researching and comparing the performance of various multimodal models. The model that was selected was the Shukla & Marlin model, and required implementation of the entire model save for the interpolation network, as only partial code was provided by the researchers. The VICReg repository details a TensorFlow implementation, but I had to port it over into PyTorch in order to apply it to the Shukla & Marlin model as the interpolation network was written with keras, which is based off PyTorch.

2.1 Multimodal Fusion Model Selection

From the many current existing multimodal models for ICU mortality/length-of-stay prediction, the Shukla & Marlin interpolation model was selected to be the baseline model as it uses a late fusion architecture that allows a minimizing branch loss, VICReg, to be used as a pre-training step. The original paper details a late fusion and early fusion design, but this paper will only be referring to the late fusion design as it most closely mimics the self-supervised image recognition networks that the pre-training technique was designed to be used with [2,3]. The focus of the Shukla & Marlin model is a novel interpolation step that allows it to outperform similar multimodal models and single modal models alike. The ICU data used, MIMIC-III, can be thought of as irregularly sampled and sparse multimodal time series data due to the nature of how its recorded throughout a patient's stay. While most medical multimodal models can create predictions off of the sparse data directly, interpolating the data and time-aligning it has been proven to improve model performance on both mortality and length-of-stay prediction tasks [1, 2]. The interpolation step consists of a series of semi-parametric layers before the final prediction network where the univariate time series input is transformed into intermediate interpolants, then merged into a fixed sized output array of smooth trends, transients, and intensities [2]. As the interpolation network is a separate entity from the prediction network, it creates a distinct autoencoder loss during the pre-training and fine-tuning steps that can be monitored along with the prediction and pre-training losses, allowing for the interpolation network to be tuned individually.

The interpolated output of the Shukla & Marlin model was then fed into a single GRU layer which represented the prediction network, in order to create the time series branch of the baseline model. Any combination of recurrent layers can be used for the prediction network and a single GRU was chosen as it generated decent results in the original paper while lowering resource costs of the final multimodal model [2].

The text representation for the clinical text branch was performed through modeling each visit's clinical notes as a bag-of-words representation, which was then vectorized by a TF-IDF of a vocabulary of the top 6000 most frequent words in the training note set. The vectorized text was then passed through two dense layers to create the text series branch of the baseline model. The time series and clinical text branches were then combined using a single dense layer to create the final fusion baseline model. As the TF-IDF vectorizer has to be populated with the clinical notes before it can be used, the vectorizer output differs depending on the notes given, allowing the model to be more easily transferred to different hospitals.

The Shukla & Marlin model was not able to be recreated with the same architecture parameters exactly as described in their paper, due to hardware and training time limitations. Specifically, the Shukla & Marlin paper experimented with RNNs and multiple GRUs, while I chose to use two fully connected layers on the text branch, and a single GRU on the time series branch. I ran into difficulty when trying to instantiate a larger model, and opted to test on a smaller model in order to move forward with my thesis. The use of a smaller model may have impacted the effectiveness of pre-training, which I will go into detail in my results section. Despite these differences, the recreated model gave a similar AUCROC and accuracy as the one described in the paper, so I believe that my recreated model serves as a decent baseline for comparing the effects of VICReg on multimodal fusion models.

2.2 VICReg Explanation

VICReg is a method for pre-training self-supervised image recognition networks while preventing collapse [3]. Although it has not been historically used on time series data for medical models, it can be applied to my model as both share a branch structure that the technique can be applied to. Self-supervised image networks pre-train by passing batches of augmented and non-augmented images through separate branches, and minimizing the distance between the output embedding vectors in order to increase model robustness and boost classification performance. Collapse occurs when the branches of the model continually predict the same nonsense values despite different images being passed through the network, and results in the pre-training step failing to create any noticeable performance improvement. VICReg prevents collapse during the pre-training step by forcing the embedded output vectors to maintain a minimum variance and covariance while minimizing their distance. The regularization is explained through the below formula, which generates a loss value given the batch outputs of each run.

$$L(Z, Z') = \lambda * s(Z, Z') + \mu * (v(Z) + v(Z')) + \nu * (c(Z) + c(Z')) \tag{1}$$

When VICReg is used to train image recognition networks, the variables Z and Z' represent the batches of original images and augmented images that are fed through the branch network. For this research, those variables represent the time series data, and clinical text notes for each patient visit in the ICU as we assume that the different modalities of data are augmentations on the ground truth of patient condition. s represents the mean squared distance between the branch outputs for each batch. v represent a hinge loss on the standard deviation of a batch of embeddings, and is defined by the following equation.

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} max(0, \gamma - \sqrt{Var(z^j) + \epsilon}) \tag{2}$$

In the equation, d is the dimension of the output embedding vector, γ is the minimum standard batch deviation we want to enforce while pre-training, and ϵ is a constant to prevent instabilities when calculating variance loss. In my experiments, I used the γ and ϵ value that the original paper settled on, although I intend to research the effects of varying them in future experiments [3]. If I wanted to force my multimodal model to learn a better defined latent space during the pre-training step, I could have achieved this through increasing the γ value, which would require the model to enforce a higher batch embedding variance. The importance of varying γ is also supported by the fact that the values the original paper found were meant to be used when pre-training image recognition networks, which use data with different characteristics than the medical data used in this paper. If the medical data used naturally tends to create embedding vectors with larger variances than the image data, then it may be beneficial to use a larger γ value, and vary it accordingly throughout subsequent experiments.

c represents the sum of the squared off diagonals coefficients of the covariance matrix of a batch of embeddings. The importance of VICReg lies in the variance and covariance terms, as they prevent model collapse while pre-training through forcing the embedded outputs to vary across each batch. VICReg also allows a model to learn a meaningful latent space during the pre-training process, which has been proved to create an improved representation space after pre-training [3]. For this reason, the batch size is an important hyperparameter, as a larger batch requires the model to maintain a standard deviation over more values during the pre-training process which results in a more meaningful latent and final representation space.

The hyperparameter coefficients λ, μ, and ν control the importance of variance, covariance, and invariance in the final loss. Different combinations of these coefficients can greatly affect the final fine-tuned model performance, and I go over some of the coefficient combinations that I experimented with in the results section. The researchers from the original paper found the most success when giving larger coefficients to invariance and variance as compared to covariance, and also discovered that using the wrong set of coefficients would still result in model collapse. As such, these coefficient combinations are extremely important and require a hyperparameter search in order to boost a multimodal model through a VICReg pre-training step. While I would have liked to run a full grid search, or bayesian hyperparameter search for the coefficients, I was not able to finish this step within the time constraints of the program, and instead opted to experiment with the coefficient combinations that the original paper found the most success with.

2.3 VICReg Implementation

The VICReg pretraining step was added to the selected multimodal model through replacing the final connected prediction layer of the model with an expander network. VICReg was then ran across the outputs of the expander network as a loss function during the pretraining phase. VICReg is expected to minimize the distance between the embedded outputs of the text and time series branches while avoiding collapse, through the variance and covariance terms in the loss. This results in the model learning a shared latent space between the clinical notes and time series data before fine-tuning on labelled mortality data.

There are also several key differences between my implementation of VICReg and the original paper that could potentially impact the benefits of my pretraining step. The original VICReg paper uses an expander of three densely connected layers with 8000 hidden units each, but I was unable to pretrain a model of this size. Instead, I tested different expander sizes and configurations and settled on an expander network of two connected dense layers of size 1024, as this gave me the best balance between pretrain time, and pretrain loss. The original paper also uses the LARS optimizer when pretraining, but I was unable to implement this due to library execution errors related to how the model passes data when training. The importance of the LARS optimizer is through learning a different learning rate for each layer of the model it trains, which is especially important when working with extremely deep networks, such as the image recognition networks that VICReg was originally created to improve on. For example, the benchmark that VICReg uses for its performance improvement was trained on a ResNet-50 backbone, which consists of 50 layers, which is significantly deeper than my model which uses three layers. As such, I believe that my use of the Adam optimizer over LARS does not significantly impact the results obtained from pretraining with VICReg.

While VICReg is a technique created for use on self-contrastive networks, specifically for image recognition, and the multimodal model is not a true self-contrastive model, each branch of the multimodal model can be interpreted as an augmentation of the patient condition. In this sense, the multimodal model can be treated like a self-contrastive network with each branch operating on an augmentation of the ground truth. Therefore, any performance benefits that are created by applying VICReg to multimodal models are a result of the model learning to leverage common information shared between the different branches of data.

2.4 Training Details

The MIMIC-III dataset was used to train, test, and compare the effects of the pre-training step. MIMIC-III is a collection of physiological data collected at the Beth Israel Deaconess Medical Center from 2001 to 2012, and was chosen for its public availability, and use in other medical models allowing for easy comparison in model performance. The baseline and pre-trained models were then asked to perform the mortality prediction task on this dataset.

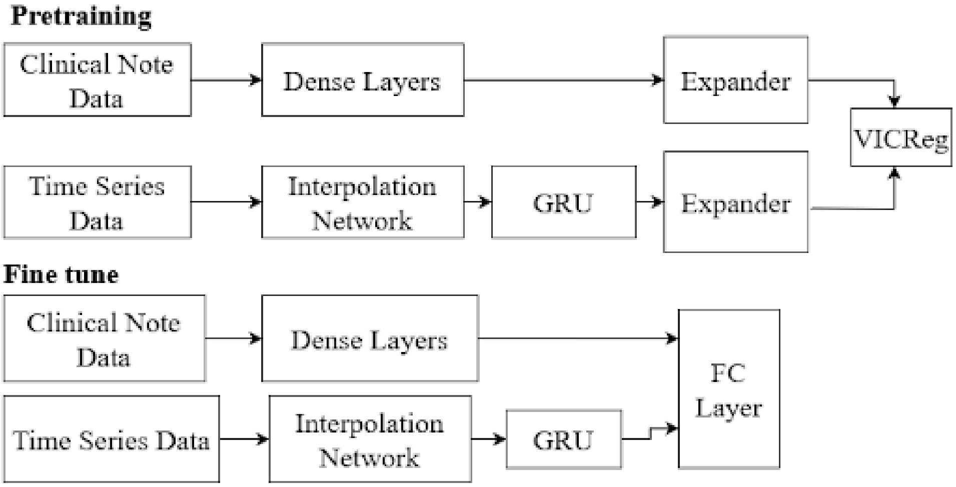

Figure 1: Overview of data flow through model architecture

To observe the effects of the pre-training step on the multimodal model, a baseline metric had to be established. For the baseline performance, physiological time series data and clinical notes were sampled from the MIMIC-III dataset. The dataset was split into a train and test split, with mortality labels being used as the classification task. The time series data was passed into the interpolation network in order to time align the data, and extract the smooth, transient, and intensity components as described in the Shukla & Marlin paper [2]. The text data would be vectorized using a TF-IDF featurizer that has a vocabulary of the 6000 top words from all text in the train set. The vectorized text would then be passed into a prediction network to generated the text embeddings. The physiological time series and text embeddings would then be combined through a final prediction network to create the baseline multimodal model. The finished model would then be evaluated on the test set to generate a baseline metric for mortality classification accuracy that the pre-trained model would be compared to.

For the pre-trained performance, the final prediction layer of the multimodal model was removed and replaced with an expander network. The original VICReg paper experimented with using shared expanders and separated expanders and found better performance through the use of a shared expander. The physiological time series data and clinical text data would then be passed through the model identically to the baseline, except the mortality labels would not be passed as the pre-training loss function is unsupervised. The expander network would then be discarded, and an identical prediction network as the baseline would be added. The training data would then be re-passed through the model, with labels this time, and trained until convergence. The pre-trained model would then be evaluated on the test set to compare with the baseline model.

3. RESULTS

Results were obtained by passing the MIMIC-III data through the modified multimodal model pre-trained VICReg, and then repeating this process with the baseline multimodal model. The mortality accuracy and AUCROC from the runs would then be compared to measure whether VICReg pretraining created any noticeable performance boost. The layer architecture for VICReg was chosen through testing multiple hidden layer sizes, and selecting one that allowed for pre-training to be completed in a reasonable amount of time. A siamese expander was also tested in comparison to separate expanders for each branch, as the original VICReg implementation used a siamese expander when working with self-contrastive networks, although the pre-training technique can be applied to any branch outputs of the same embedding regardless of whether the expander weights are shared.

3.1 Dataset Processing

For each pre-trained vs baseline comparison, the MIMIC-III data was initially put through a 80/20 train and test split. The train and test data would then be passed through the Shukla & Marlin model to generate baseline metrics. To generate the pre-trained metrics, the train data without labels would be passed through the baseline model with the expander networks attached and trained for a set amount of epochs. The expander networks would then be removed, and the same final prediction network from the baseline network would be reattached. The train data would then be passed through the new model and fine-tuned until convergence. The test data would then be passed through the baseline and pre-trained models and compared to observe the effects of the pre-training.

3.2 Baseline Run

The baseline run of the recreated Shukla & Marlin model gave an AUC of 0.85, and a mortality classification accuracy of 0.92. I found these results to be comparable to the metrics given in the original paper, which leads me to believe that my recreated model is accurate to the original and that any performance differences I find between my pretrained and baseline models should be generalizable to other medical multimodal models.

3.3 Performance with Shared Expander

A single siamese expander of size 512 was used to the replace the final prediction layer of the baseline multimodal fusion model in order to implement VICReg pretraining. The model was then trained until for 30 epochs on the train data, and then fine tuned until convergence on the mortality labels. The addition of the single expander and pretraining did not create any noticeable performance benefit when comparing AUROC or accuracy.

3.4 Performance with Separate Expanders

Two expander networks consisting of two fully connected dense layers of size 1024 replaced the final prediction layer of the baseline model during pretraining. The pre-trained model was then trained for 30, then 80 epochs on the train data, then fine-tuned on the mortality labels. Both the 30 and 80 epoch models consistently slightly outperformed the baseline accuracy across 5 random runs with the 80 epoch model outperforming the 30 epoch model. However, the AUCROC of both pre-trained models dropped, with the 80 pre-trained epoch model dropping more AUCROC. This suggests that the model is losing information as it pre-trains, and is a sign that the pre-training loss is failing to regularize and prevent collapse. This issue of AUCROC lowering shows up in multiple further experiments, and likely must be remedied by performing a hyperparameter search over the VICReg coefficients. A search was unable to be performed due to time constraints, as such, I used the coefficients from the VICReg paper but a better combination will likely prevent the AUC from dropping as pre-training increases.

3.5 Performance with different VICReg hyperparameters

A pre-trained model that used the VICReg hyperparameters from the original paper was created, and allowed to train for 1000 epochs with model weights being saved every 100 epochs. The pre-trained models were found to occasionally outperform the baseline model, but performance gains were very minimal 0.5% and were inconsistent as training continued. AUCROC was also found to slightly decrease while pre-training, although it stopped lowering around the 500 epoch mark and stayed stable to the final 1000th epoch. The drop in AUCROC suggests that the model started forgetting the weights adjusted by pre-training, which is likely due to a bad set of variance loss coefficients being used. These results show that the quality of VICReg pre-training is sensitive to the coefficients of the loss function, and a lack of performance results likely stems from an underperforming set of coefficients used. The original paper found the best pre-training performance from using a high variance and covariance coefficient for a shared expander, but the multimodal model can also use a set of two different VICReg loss terms as it has a separate expander for the clinical text and physiological time series branches. I experimented with different sets of VICReg loss coefficients, and found that in the worst cases like penalizing with an exceptionally high variance, the model would make noticeable drops in AUCROC when being evaluated after the pre-training.

3.6 Final Thoughts

The pre-trained models seem to slightly perform than the baseline models when compared over several test runs, although a significant performance improvement was not able to be found. Through my testing, I've also noticed that the pretrained models occasionally tend to exhibit less deviation across testing folds when compared to the baseline models. This may suggest that the pre-training is still affecting the model performance beneficially even if most of the pre-training is forgotten as pre-training time lengthens, and that a larger benefit can be acquired through training a larger model with a larger expander network and better coefficient hyperparameters for VICReg.

4. CONCLUSION

While the addition of the VICReg pretraining to the multimodal fusion model failed to create any noticeable improvements, there are multiple further steps that can be taken to create a performance improvement. Namely, the architecture of the expanders can be modified to better leverage the time series nature of each branch, and the individual branch models can also be modified to better leverage their data modalities. Due to the time constraints of the URS program, I was not able to implement these changes. However, I believe that the pre-trained model will outperform the baseline once the changes are fully tested.

4.1 Expander Architecture

A set of two connected dense layers was used as the expander network for the pretraining, although the original VICReg paper details the user of a larger multi-layered linear expander. The paper also experiments with a non-linear expander, which could result in a better learned latent space when compared to linear expanders, which could allow for the model to better learn shared information between branches during pre-training. The expanders used in the original VICReg paper also consisted of 8000 hidden units each, while the ones I used only consisted of 1000 hidden units. The lack of a large output size likely resulted in my model not pre-training a well-defined latent space, which could have led to the lack of performance gains during the fine-tuning step through the inability to pre-train effectively. To fully understand the role of the expanders on the performance of the final pre-trained model, a shared expander model and a model with separate expanders for each branch should be tested. The original paper also uses the LARS optimizer which scales well with deeper models, while I used the ADAM optimizer for my experiments as I previously mentioned I ran into implementation conflicts and errors when trying to use LARS. The separate learning rates per layers that LARS uses may create a better pre-training with a deeper expander network when compared with the single learning rate that I was using throughout the model.

4.2 Branch Architecture

The time series branch of the model consisted of the Shukla & Marlin interpolation network which fed into a single GRU. The pre-training step could be improved by using multiple GRUs instead, as a deeper model may experience a more effective pre-training as it has a larger latent space to work with. For the text branch, I used a TF-IDF vectorizer for the clinical text prediction branch that fed into two fully connected dense layers. Use of a GRU or some similar deeper network over the single layer in the text branch could allow for a more effective pretraining, as the current model may be too shallow and not pre-training enough weights to be effective when the fine-tuning step occurs. Another benefit of using a GRU is that the input would consist of time-aligned notes, while my current experiments read all the notes during a visit as a single bag-of-words model before vectorizing and passing to the text prediction network. The pre-training step may be more effective if both branches use time-aligned data, rather than only the time series branch, as this may force the pre-training to learn a shared latent space between the text and time series data at each time step, rather than a shared latent space between all clinical text and the time series data. For this reason, a prediction network of multiple GRUs in the text branch may have created a better pre-training than the TF-IDF and fully connected layer network that I experimented with. This intuition of using a deeper model is also supported by the original VICReg paper, as it was pre-trained with a ConvNet backbone of 50 layers, which is much deeper than my current model architecture.

4.3 VICReg hyperparameters

VICReg uses a different coefficient for each of the three loss terms, and the coefficients chosen greatly impact the final fine-tuned model evaluation performance. I was not able to run a full hyperparameter search with the time I was given to complete this thesis, although I found decent performance from the coefficient values that were given in the original VICReg paper, which used large values for variance and covariance when compared to invariance. This creates a pre-training step that slowly minimizes the embedding distance during each pre-train epoch, while carefully preventing collapse. However, this may not be the best strategy when working with clinical text and time series data, as a more effective pre-training could be achieved through using a separate set of coefficients for each branch. This is because VICReg was created to operate on augmentations of a base image, and the variance and covariance for each branch are comparable as they share the same ground truth, but this may not be the case for different modes of hospital data. To solve this, a full hyperparameter search for the VICReg loss coefficients must be done, with a separate pair of variance and covariance terms for each branch preferred. From my experiments, I observed that the pre-trained models tended to slightly outperform the baseline model consistently. However, my pre-trained models also had lower AUCROCs which signify a lack of good hyperparameter coefficients, which lead to model forgetting during the pre-training phase.

4.4 Conclusion

Application of variance pretraining to multimodal models can be expected to yield performance benefits, but the model architecture of each branch and pre-training hyperparameter selection selection is extremely important. This is because a shallow branch architecture can result in an ineffective pretraining, and a shallow expander prevents an effective latent space that contains the shared information from being learned. A large batch size is also required, as VICReg is dependent on maintaining a minimum variance for each embedding variable across each batch of data passed during the train steps. If all these constraints are satisfied, the pretraining step is expected to be effective towards boosting multimodal model performance.

5. REFERENCES

[1] Satya Narayan Shukla and Benjamin M. Marlin. Interpolation-Prediction Networks for Irregularly Sampled Time Series. In ICLR, 2019.

[2] Satya Narayan Shukla and Benjamin M. Marlin. Integrating Physiological Time Series and Clinical Notes with Deep Learning for Improved ICU Mortality Prediction. In arXiv, 2021.

[3] Adrien Bardes, Jean Ponce and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In ICLR, 2022.

[4] Joseph Futoma, Sanjay Hariharan, Katherine Heller, Learning to detect Sepsis With a Multitask Gaussian Process RNN Classifier. In stat.ML, 2017.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton, A Simple Framework for Contrastive Learning of Visual Representations In PMLR, 2020.