NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

# Curriculum Learning Based Sample Selection Using Posterior Probabilities

S. Ziyabari, V. Shah, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University

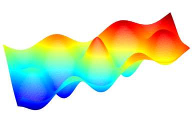
College of Engineering
Temple University


## Abstract

- Neural network architectures are highly sensitive to the random initialization of their parameters, which in turn, means they are highly sensitive to the ordering of the data.
- Performance can vary by over 100% with the ordering of the data, and there is no systematic way to find the optimal ordering of the data.
- Curriculum learning (CL) orders the input data based on the level of difficulty.
- We explore the sensitivity of this bootstrapping process for two hybrid systems used to order the data: HMM-SdA and CNN-LSTM.
- Performance of the CNN-LSTM system using CL is 32.13% sensitivity with 9.90 false alarms per 24 hours, which is very close to the best published results for our manually-tuned system, known as AutoEEG (30.83% Sens. with 7.13 FAs).
- The main benefit of CL is that we can automate the training process and reach near-optimal performance. This reduces the time and cost of creating an optimized system for new applications.

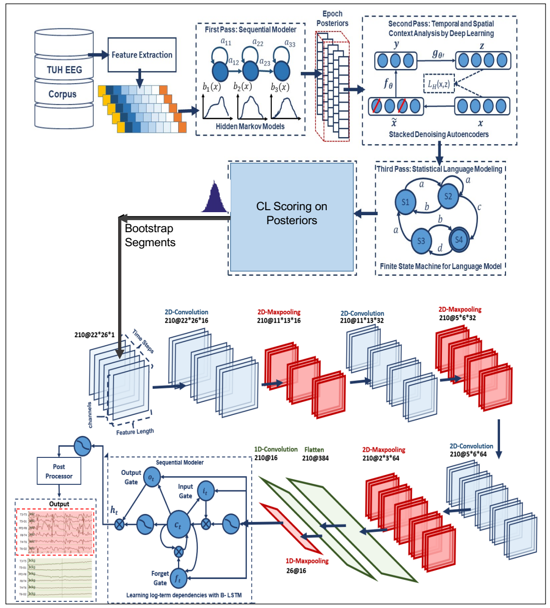

## Deep Learning Convergence Issues

- The performance of a generic deep learning system varies significantly with the ordering of the data. Error rates can often fluctuate by over 100% depending on the ordering of the data.
- It takes significant amounts of time and money to find the optimal ordering of the data, and this ordering is data specific.
- The majority of deep learning systems use optimizers which are variants of gradient descent methods (i.e., SGD, ADAM, AdaGrad).



- Common approaches to deal with such convergence issues are:
  - **Annealing Learning (AL):** Start with a higher learning rate for a few epochs to explore the parameter space and then start reducing the learning rate until convergence is reached.
  - **Snapshot Ensembling (SE):** Use a cyclic learning rate, converging & escaping multiple minima to create multiple models. During the prediction stage, take a majority vote using a manager model.
  - **Stochastic Weight Averaging (SWA):** Average the weights collected by multiple (N-best) models.
  - **Curriculum Learning (CL):** Change the ordering of the data from easy to hard to gradually/smoothly train the model.
- In this study, we focus on CL by reordering EEG epochs after the self prediction stage.

## Deep Learning Architectures for Seizure Detection




### HMM-SdA:

- First pass: uses a simple 3-state left-to-right hidden Markov model topology (HMM) with 8 Gaussian mixture components per state and Baum-Welch training to segment the signal.
- Using HMMs, the recognition task is performed on each channel independently. For each epoch, a supervector is created by concatenating output probabilities of the individual channels.
- Second pass: spatial context is learned using channel-based posteriors which are input to a 3-layer SdA system. PCA is performed prior to feeding data to the SdA system.
- A stochastic grammar is applied to the output of SdA to learn temporal context.

### CNN-LSTM:

- The CNN-LSTM system uses a combination of three 2D CNN layers, one 1D CNN and one 2-layer LSTM network.
- 2D CNN layers are followed by a maxpooling layer and a flattening layer is applied prior to the 1D CNN.
- Exponential Linear Units (ELU) are used as activation functions except for the last layer, which uses a sigmoid.

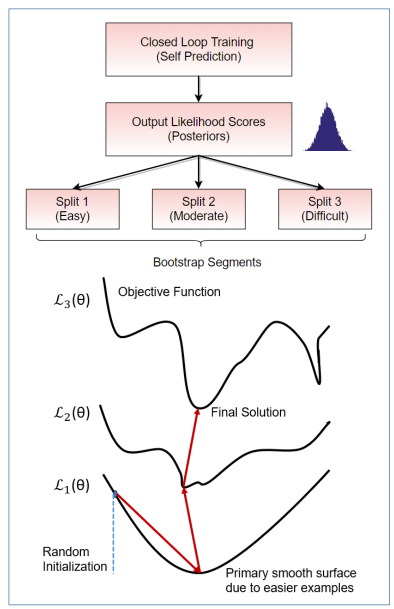
## Curriculum Learning Using Self-prediction

### Approach:

- Curriculum design as a continuation method: we use a heuristic approach to order the data by performing closed loop pretraining using our one of the hybrid architectures Hidden Markov Models (HMM) + Stacked Denoising Autoencoders (SdA).
- Algorithms which are not sensitive to the random initialization such as HMMs, Random Forests and other Bayesian approaches are good fits for the CL data preparation block.
- Posteriors for each 1-second epoch are evaluated based on whether its label was correctly detected and based on the strength of the output probability.
- The training data is then split into multiple segments ordering data from easiest to most difficult. Training the system on these ordered sequences gradually makes error surface less smooth with time.
- After the self-prediction stage, CL is applied to generate the bootstrapping sequences. CNN-LSTM is used as the training architecture used is with an epoch window size of 21 secs.
- Approximately 10% of each bootstrapping segment is replaced with other segments to better generalize the training examples and increase overall entropy of the segment.

### Justification:

- The surface generated by the easier examples are smoother. (i.e. convex). Training models in these smooth spaces is easy and convergence is quick.

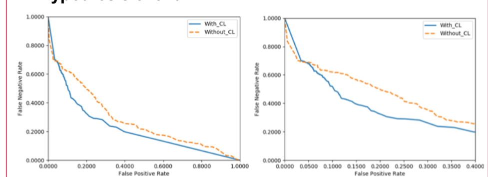



## Results

- The experiments were conducted on TUSZ v1.1.1 – a subset of publicly available TUH EEG Corpus (*www.isip.piconepress.com/projects/tuh_eeg*).
- Performance on TUSZ:

| CNN-LSTM | Sensitivity | Specificity | FAs/24 Hrs. |
|---|---|---|---|
| Without CL | 30.83% | 97.10% | 7 |
| With CL | 32.13% | 95.13% | 10 |

- The results reported here are using Any-Overlap metric (OVLP). Correct detection is considered if hypothesis event fully/partially overlaps with reference events. False positives (FPs) are assigned when there is no overlap between reference and hypothesis event.




- Full DET curve (on the left) compares performance of the CNN-LSTM system with and without CL. Training with CL significantly outperforms our best baseline system for high FPRs.
- The expanded DET curve (on the right) shows that our best baseline system performs better for extremely low FPRs.
- However, over the range [0.0,0.4], CL (AUC = 0.5580) outperforms the baseline (AUC = 0.4432).
- Posterior seizure confidence of the system with CL is higher (μ = 95.16, σ = 4.12) than of the system without CL (μ = 91.27, σ = 4.01)

## Conclusions and Future Work

- The CL approach maximizes learning efficiency by providing the system with increasing rates of prediction accuracy on fewer attempts of training.
- This study suggests that CL can significantly accelerates the training process of deep learning architectures for applications similar to EEG event recognition.
- Future work is focused on the development of automated versions of curriculum learning (i.e. self-paced curriculum learning (SPCL), student-teacher learning) that automate the ordering and the distribution of input data.


## Acknowledgements

- Research reported in this poster was supported by National Human Genome Research Institute of the National Institutes of Health under award number 3U01HG008468-02S1.
- The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.