Learning Visual Engagement for Trauma Recovery

Svati Dhamija, Terrance E. Boult
University of Colorado Colorado Springs
{sdhamija,tboult}@vast.uccs.edu

Abstract

Applications ranging from human emotion understanding to e-health are exploring methods to effectively understand user behavior from self-reported questionnaires. However, little is understood about non-invasive techniques that involve face-based deep-learning models to predict engagement. Current research in visual engagement poses two key questions: 1) how much time do we need to analyze facial behavior for accurate engagement prediction? and 2) which deep learning approach provides the most accurate predictions? In this paper we compare RNN, GRU and LSTM using different length segments of AUs. Our experiments show no significant difference in prediction accuracy when using anywhere between 15 and 90 seconds of data. Moreover, the results reveal that simpler models of recurrent networks are statistically significantly better suited for capturing engagement from AUs.

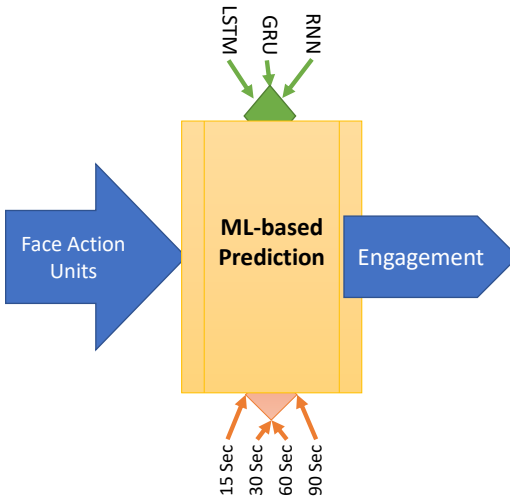

Figure 1. **Machine learning prediction of engagement:** This paper analyzes which deep learning algorithms (LSTM, GRU or RNN) as well as how much facial data (15 to 90 seconds) should be used for ML-based prediction of contextual engagement.

1. Introduction

Engagement is an indispensable part of user experience and interaction with applications including social-assistive robots, gaming, web-interventions, online-learning, etc. [33, 3, 8, 40, 63, 22, 23]. A challenging yet important application of video-based engagement prediction are self-help websites that aim to deliver large-scale high quality health care. Due to the shortage of mental health professionals i.e. 1 professional per 1000 individuals [51], it is not surprising to see rapid growth in internet-driven approaches for mental-health interventions creating awareness, understanding and treatment of symptoms. Prior research indicates that web-interventions improve patient involvement and have the potential to advance patient-centered mental health care delivery [37, 9]. Unfortunately, such self-care websites often suffer from engagement issues and lack the tools to measure user engagement in a reliable and automated manner [45, 24, 41]. Continuous prediction of engagement can help create adaptive websites, detect people at risk and reduce dropouts, thereby having far reaching effects in web-based mental health intervention techniques.

With recent advances in computer vision and deep learning, it is becoming possible to assess the current emotional state of an individual with a webcam [20, 4, 36]. These methods typically analyze facial expression of the subject by performing either frame level or temporal modeling of facial features. The static classifiers classify a frame in the video to one of the facial expression categories and dynamic classifiers take into account the temporal pattern in predicting the displayed facial expression recognition [16].

Engagement is abstract, complex, unstable and context sensitive. According to Sidner *et al.*: *"Engagement is the process by which interactions start, maintain and end their perceived connection to each other during an interaction"* [58]. This paper explores the amount of data needed and the learning algorithm to predict context-sensitive engagement, as shown in Figure 1.

In the domain of student learning application Whitehill *et al.* [59] emphasized the importance of choice of timescales for engagement prediction. They considered three different timescales of 60 seconds, 10 seconds and per

frame annotations (obtained by external annotators) and deduced that shorter durations do not provide enough context for engagement annotation while longer duration clips tend to be harder to evaluate, possibly because of increase in data mixes from different levels of engagement [59]. However, they did not consider any deep-learning techniques for engagement prediction and no facial features. We hypothesize that this observation for manual annotations of engagement by Whitehill *et al*. might have a significant impact on sequential learning for engagement prediction.

Our prior work [19] addressed predicting engagement from video sequences in a trauma recovery intervention. We suggested that temporal patterns are beneficial for face-based engagement prediction, however, we just assumed an LSTM model and also assumed 30 seconds (900 frames) should be used for engagement prediction. The latter assumptions leads us to ask: How much time do we need to observe facial expressions of an individual to infer their engagement level accurately? Detailed analysis for varying video sequence length for engagement prediction is the *first* contribution of this paper.

We also just presumed an LSTM was the appropriate sequence learning model [19]. A range of sequence learning algorithms is used in natural language processing, speech recognition, image captioning, visual question answering etc. Recurrent Neural Network (RNN) architectures, Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) are all used extensively for deep-learning. Chung *et al*. [15] performed detailed comparison of LSTM and GRU in the domain of polyphonic music modeling and speech signal modeling and found that GRU to be comparable to LSTM. Jozefowicz *et al*. [35] performed a similar but more extensive empirical analysis of various RNN, LSTM and GRU for language modeling and music modeling. Both variants of RNN architectures i.e. LSTM and GRU were built to address the vanishing and the exploding gradient problem and efficiently learn long-range dependencies [53]. However, detailed comparison of various sequence learning algorithms in the domain of facial video analysis, especially for affective computing has not been actively explored in the past. For face-based engagement prediction, it is not clear whether a large look-back memory or modulated inputs are essential, important or a liability. How do we integrate the temporal transitions in facial expressions over time for visual engagement? Is long-term memory essential for face-based models to predict current state of an individual? Systematic comparison of various deep sequence learning for engagement prediction is the *second* contribution of this paper.

In summary, we evaluate the effect of length of video segments used for performing engagement prediction on deep sequence learning model. We analyze this effect across multiple leading deep sequence learning algorithms to understand the relationship between network complexity and prediction performance. Further, we report the role that context plays in engagement prediction for sequence learning algorithms in the domain of trauma recovery. In the following sections we first explain the related work followed by description of various sequence learning algorithms used in this work. We explain the dataset used followed the experimental analysis comparing the algorithms and time scales.

2. Related Work

In this section we review related research works from multiple related domains such as sequence learning, recurrent neural networks and facial feature extraction.

Visual Engagement Applications: Engagement prediction from facial expressions, gesture, eye gaze tracking etc has been researched for numerous applications like human-robot-interaction [54], dyadic interactions with virtual agents [46] and designing social-robots for autism [49]. Traditional approaches of engagement prediction in domain of affective computing use static representations of individual frames [10, 25, 44]. More recently, deep-sequence-learning models for engagement prediction have been explored, in order to incorporate changing interactions over time and concurrent sequential dynamics [19]. D'Mello *et al*. [44, 48, 21, 1] and Whitehill *et al*. [60, 61] have a longstanding line of research into learning-centered student engagement and the facial expressions associated with it. Recently, correlations between specific fine-grained facial movements and self-reported aspects of engagement, frustration, and learning were identified [25]. Thus, to create engagement prediction model for trauma recovery, we use subtle facial movements or action-units rather than emotional responses as an intermediate representation. Due to sparsity of labels (engagement self-reports), we resorted to considering approximately 15-90 seconds video segment before the self-report was captured.

Facial Features: Work in the domain of engagement prediction, emotion prediction, and facial expression analysis has benefitted tremendously from the recent advances in the domain of facial feature extraction, face tracking, and facial action unit coding. Automated detection of facial action units (AUs) [18, 34, 17] have proved to advance multiple face-based affective computing systems [38]. In our work, we rely on AUs extracted from video frames as an intermediate representation provided as input to our sequence learning models. Recent facial expression recognition systems can recognize several AUs with reasonable accuracies [2, 18, 55]. Finally, there have been multiple notable works in the domain of facial expression and affect analysis that has pushed state of the art in affect recognition beyond six basic emotion categories [56, 47].

Sequence Learning: Sequence classification has long

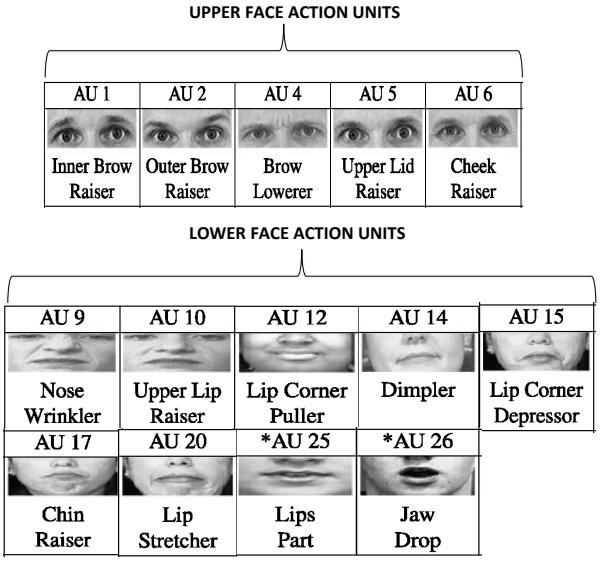

Figure 2. **Intensity-based Action Units from OpenFace:** 14 intensity-based AUs derived from OpenFace that are used for engagement prediction are displayed above. 5 AUs are from upper face and 9 from lower face. Movements of these individual facial actions are used in the form of intensity values varying from 0 (not present) to 5 (present with maximum intensity).

and rich history with numerous machine learning techniques being developed for handling and classifying sequences. Hidden Markov Models (HMM), a class of generative models, have been widely used in speech domain for speech recognition, speech tagging and handwriting recognition [52]. Linear and non-linear variants of Conditional Random Fields (CRF) [39] is often used for structured prediction for labeling or parsing of sequential data and is widely used in the domains of biological sequence processing [6], natural language processing [43] and computer vision [31].Traditional methods such as HMM, CRF and others have been extensively compared with recent deep learning methods for handling sequences by the works of [42, 26], hence we do not perform exhaustive comparison in our work. As noted earlier, engagement with an intervention varies with time, and hence there is need to employ methods that can integrate temporal information in learning framework. Such problems are often modeled as sequence learning problems [29, 28]. To address sequence learning, various methods such as Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), and Long Short-Term Memory (LSTMs) were proposed and employed in wide range of problems [30]. In this work, we explore all three aforementioned algorithms to model face-based engagement prediction.

3. Algorithms for Sequence Learning

In recent years, a wide range of sequence learning problems have been modeled using RNNs. They learn a representation for each time step by incorporating observation from the previous step and the current step. This process allows RNNs to preserve information over time using recurrent mechanisms. RNNs were developed by Schmidhuber *et al*. [57] and have gained applications in multiple sequence learning problems such as handwriting recognition, visual question answering, image generation, speech recognition and others. Hochreiter *et al*. [32] proposed LSTMs, a variant of RNNs consisting of memory cell. LSTMs have been widely used in sequence learning problems such as language modeling, image captioning, translation and others. More recently, GRUs were proposed by Cho *et al*. [11] to make each recurrent unit to adaptively capture dependencies of different time scales.

An RNN has connections between units to form a directed cycle allowing it to exhibit temporal dynamic behavior. They have internal memory to process sequences of inputs. LSTM and GRU share the similarities in terms of having gating units that modulate the flow of information inside the unit and balance the information flow from the previous time step and the current time step dynamically. However, GRU does not require separate memory cell for this purpose, making its internal structure simpler.

We model the problem of engagement prediction as a sequence learning problem, where input consists of sequences x_i of AUs computed from facial video data of a particular length. Each sequence is associated with a label y_i which relates to engagement self-report provided by trauma subjects. Our implementation is based on TensorFlow which in turn is based on [27, 64], and we follow their notation. We let subscripts denote timesteps and superscripts denote layers. All our states are n-dimensional equal to the number of AUs tracked, currently 14.

3.1. Recurrent Neural Network (RNN)

An RNN is a neural network that consists of a hidden state $\mathbf{h}$ and an optional output $\mathbf{y}$ which operates on a variable-length sequence $\mathbf{x} = (x_1, \ldots, x_T)$. At each time step t, the hidden state $\mathbf{h}_{\langle t \rangle}$ of the RNN is updated by

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right), \tag{1}$$

where f is a non-linear activation function. f may be as simple as an element-wise logistic sigmoid function and as complex as an LSTM unit [12].

An RNN can learn a probability distribution over a sequence by being trained to predict the next symbol in a sequence. In that case, the output at each timestep t is the conditional distribution $p(x_t \mid x_{t-1}, \ldots, x_1)$. For example, a multinomial distribution (1-of-K coding) can be output using a softmax activation function

$$p(x_{t,j} = 1 \mid x_{t-1}, \ldots, x_1) = \frac{\exp\left(\mathbf{w}_j \mathbf{h}_{\langle t \rangle}\right)}{\sum_{j'=1}^{K} \exp\left(\mathbf{w}_{j'} \mathbf{h}_{\langle t \rangle}\right)}, \quad (2)$$

for all possible symbols $j = 1, \ldots, K$, where $\mathbf{w}_j$ are the rows of a weight matrix $\mathbf{W}$. By combining these probabilities, we can compute the probability of the sequence $\mathbf{x}$ using

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \ldots, x_1). \quad (3)$$

From this learned distribution, it is straightforward to sample a new sequence by iteratively sampling a symbol at each time step.

3.2. Gated Recurrent Units (GRU)

GRUs were proposed by Cho *et al*. to enable RNNs to adaptively capture time dependencies at different timescales. GRUs lie in between RNN and LSTM in terms of complexity. Similar to RNN, the hidden state of GRU is updated at each time step in order to be linearly interpolation between previous hidden states.

The activation h_t^j for j^{th} unit of the GRU at time t is a linear interpolation between previous activation previous activation h_{t-1}^j and candidate activation function $\tilde{h}_t^j$:

$$h_t^j = (1 - z_t^j) h_{t-1}^j + z_t^j \tilde{h}_t^j \quad (4)$$

where z_t^j is the update gate which is computed by

$$z_t^j = \sigma(W_z x_t + U_z h_{t-1})^j \quad (5)$$

The candidate activation function is computed similar to that of traditional recurrent unit

$$\tilde{h}_t^j = tanh(W x_t, + U(r_t \odot h_{t-1}))^j \quad (6)$$

where $\odot$ is element-wise multiplication and r_t is set of reset gates which is computed using

$$r_t^j = \sigma(W_r x_t + U_r h_{t-1})^j \quad (7)$$

When t_t^j is close to 0 implies reset gate reading first input from the sequence allowing it to forget the previously computed state.

In this formulation, when the reset gate is close to 0, the hidden state is forced to ignore the previous hidden state and reset with the current input only. This effectively allows the hidden state to *drop* any information that is found to be irrelevant later in the future, thus, allowing a more compact representation.

3.3. Long Short-Term Memory (LSTM)

Let $h_t^l \in \mathbb{R}^n$ be a hidden state in layer l at time-step t. Let $T_{n,m} : \mathbb{R}^n \to \mathbb{R}^m$ be an affine transform from m to n dimensions, i.e. $T_{n,m}x = Wx + b$ for some W and b). Let $\odot$ be element-wise multiplication and let h_t^0 be an input data vector at time-step t. We use the activations h_t^L to predict y_t, since L is the number of layers in our deep LSTM.

The LSTM has complicated dynamics that allow it to easily "memorize" information for an extended number of time-steps using *memory cells* $c_t^l \in \mathbb{R}^n$. Although many LSTM architectures that differ in their connectivity structure and activation functions, all LSTM architectures have explicit memory cells for storing information for long periods of time, along with weights for updating the memory cell, retrieving it, or keeping it for the next time step. The LSTM architecture used in our experiments is given by the following equations [29], as implemented in TensorFlow basic LSTM cell:

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{2n,4n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}$$
$$c_t^l = f \odot c_{t-1}^l + i \odot g$$
$$h_t^l = o \odot \tanh(c_t^l)$$

where sigm is the sigmoid function, sigm and $\tanh$ are applied element-wise, i, f, o, c, h are the input gate, forget gate, output gate, cell activation vector and hidden vectors, respectively. In this work, we assume the length of sequence is known a-priori and hence use single layer LSTM with static RNN cells.

4. Experiments

We now describe the EASE dataset, followed by the AU computation procedure to extract intermediate feature representation and the methodology for cross-validation testing using LSTMs, GRUs and RNNs. Lastly, we elaborate the training and testing data used for varying time segments with all three sequence learning algorithms.

4.1. EASE Dataset

The EASE data comprises of audio, video and physiological recordings of 110 participants. The broader study was divided into three sessions/visits in the form of a Randomized Control Trial (RCT). Each participant was assigned two out of the six modules in each visit. The first two visits were restricted to "Relaxation" (RX) and "Triggers" (TR) modules and in the third visit the participants were free to choose from the remaining four modules. The

relaxation module presents the user with video demonstrations of various exercises like breathing, muscle relaxation, etc. The triggers module presents educational material to the user about trauma symptoms and prevention. Subject inclusion criteria were determined by the work of Benight *et al*. [5]. During these sessions, a LogiTech webcam with a resolution of 640x480 at 30 fps was placed on top of the monitor recording the participants face video along with audio and physiological data. Although the EASE data is richer in terms of its multi-modal nature, for this work we focus our attention primarily on facial video data captured by webcam placed on monitor. The participants provided self-reports about their engagement level with the task on a scale of 1 to 5, where 1 is "Very Disengaged" and 5 "Very Engaged". As mentioned earlier, each participant came in for three sessions/visits. The first two (controlled) visits are used for experiments in this paper.

The EASE dataset used for all experiments in this work was de-identified and is available from the webpage [1]. In this section, we briefly describe the data used for our evaluation. For more detailed description, we refer readers to the work in [19].

4.2. Action units features:

As noted earlier, the collected dataset consisted of a large number of face videos and engagement self-reports. While the raw video is not available the site provides de-identified AU data that was extracted from the videos. While there are number of software available for extracting facial landmark points and facial action units (e.g. [18, 34]) they provide data using the recent work of OpenFace proposed by Baltrušaitis *et al*. [2]. It is an open-source tool which has shown state-of-the-art performance on multiple tasks such as head-pose, eye-gaze, and AU detection. For our work, we primarily focus on facial action units. In [19], the AUs extracted from OpenFace consisted of both intensity-based and presence-based AUs, making a total of 20 feature dimensions. For our experiments in this work, we focus only on intensity-based AUs to reduce the effect of combining categorical and numeric features. This reduces our input feature dimensions from 20 to 14 as shown in Figure 2. Intensity-based AUs are generated from OpenFace on a 0 to 5 point scale (not present to present with maximum intensity). The list of AUs used in this paper are as follows: Inner Brow Raiser, Outer Brow Raiser, Brow Lowerer , Upper Lid Raiser, Cheek Raiser, Nose Wrinkler, Upper Lip Raiser, Lip Corner Puller, Dimpler, Lip Corner Depressor, Chin Raiser, Lip Stretcher, Lips Part, Jaw Drop. Figure 3 shows the changes in Outer brow Raiser AU02 intensities as a function of frame number.

[1] http://vast.uccs.edu/~sdhamija/

4.3. Model-tuning and Cross-Validation pipeline:

Because we want consistent data for different timescales we consider only those sequences that had a minimum of 90 seconds before response which reduces the dataset slightly from that used in [19]. The total number of segments extracted from the EASE dataset for "Triggers" were 443 and for "Relaxation" 347. For all experiments in this paper, we use 20-fold cross-validation. For all 20 fold experiments, the training data consists of 421 segments in training set and 23 segments for the testing set for TR module. Similarly, in RX module, our training set consists of 330 segments for training and 17 segments for testing. Each segment was accompanied by an engagement self report.

To validate the slightly smaller subset, we replicated the LSTM engagement classifier using the same time window of 30 seconds as used in [19] and their reported LSTM parameters (with tensor defaults for those not reported). Using this data, our results, using only 14 AUs, is slightly higher than the values reported in their paper, validating our subset and approach.

In our experiments, for the all three classifiers LSTM , GRU and RNN, we train the deep networks to minimize the post softmax cross-entropy loss of the predicted and actual class. During the training process, we use Adam optimizer with a learning rate of 0.01 and fix training at 10 epochs, which was after validation error stabilized. The batch size used was 100 samples irrespective of context i.e. for both Relaxation and Triggers. The forget-bias of the LSTM was set to 1.0 to remember all prior input weights. Prior work on forget bias of LSTMs also suggests that setting it to 1.0 bridges the gap between LSTM and GRU [35]. All gate activations were *tanh* activations for all three learning algorithms. The gradient computation for *tanh* is less expensive and converges faster.

Next, we extracted varying segment lengths of intensity-based AUs preceding the engagement self-report for 15, 30, 60 and 90 seconds each. We train separate contextual LSTMs, GRUs and RNNs for each engagement prediction model (Relaxation – RX and Trigger – TR).

5. Results and Evaluation

In this section, we discuss in detail the results obtained for engagement prediction across variety of time windows and sequence learning algorithms along with its task specificity. The results are summarized in Table 1 and Table 2.

Table 1 shows the results for LSTM, GRU and RNN with 30 seconds of input data i.e. 900 frames. LSTMs trained with intensity-based AUs with 30 seconds data from the engagement segments serve as baseline for GRU-30 and RNN-30 predictions. We obtain 52.30 $\pm$ 16.94% average prediction accuracy across 20-folds on TR module

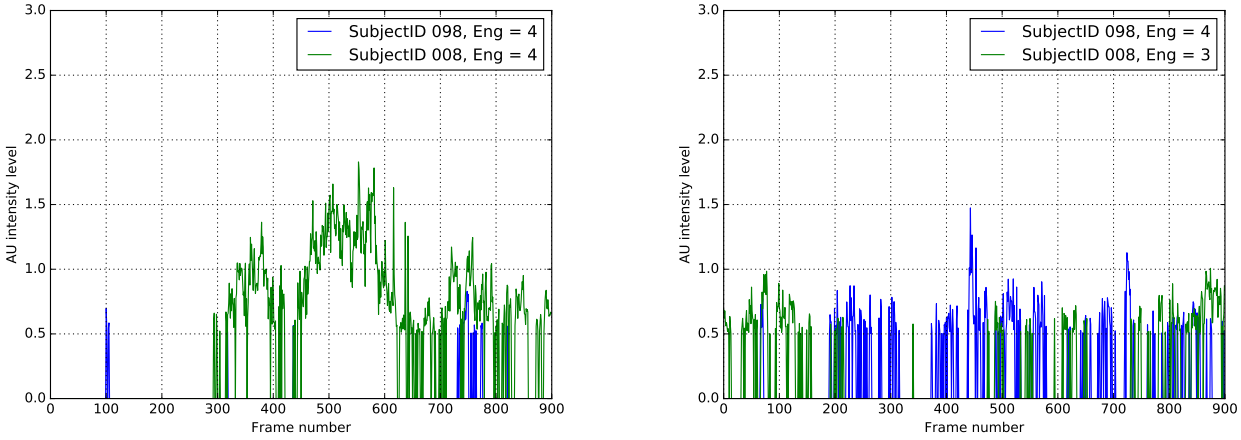

Figure 3. The above figure shows the inherently noisy signals from action unit transitions for "Outer Brow Raiser". Each plot shows 30 seconds of AU data for two different subjects with engagement level show in the legend. The left plot is from Triggers module, where the subjects have exactly the same self-reported engagement, yet the AU intensity is quite different. The AU signal on the right is from the Relaxation module which shows similar signals for subjects with different engagement. Note that facial expressions for same subjects vary in different contexts of Relaxation and Triggers. This is a challenging prediction problem.

	Triggers	Relaxation
Baseline LSTM-30 [19]	50.7 ±11%	39.1 ± 8.8%
LSTM - 30	52.30 ± 16.94%	39.44 ± 10.51%
GRU - 30	53.90 ± 17.37%	43.48 ± 11.26%
RNN - 30	54.32 ± 17.74%	44.24 ± 13.09%

Table 1. The first row are results from [19] shows engagement prediction results for their LSTM at 30 seconds. The baseline used 20 AUs, compared to our 14 AUs and had a few more segments for each context. The second, third and forth row shows our results for LSTM, GRU and RNN for over 20 fold validation, with all three using exactly the same data/folds. It can be seen that RNNs are the most accurate for both contexts of Relaxation and Triggers. GRU performance is significantly better from LSTMs at 30 seconds with p=.03 for Relaxation at 95% confidence intervals using a 2-tailed paired test. RNN-30 is significantly better from LSTM-30 for both Relaxation and Triggers at p=.05 and p=.016 respectively.

	Triggers	Relaxation
LSTM - 15	52.30 ± 16.94%	39.44 ± 10.51%
LSTM - 60	52.30 ± 16.94%	39.44 ± 10.51%
LSTM - 90	52.30 ± 16.94%	39.44 ± 10.87%
GRU - 15	53.90 ± 17.37%	43.48 ± 11.26%
GRU - 60	51.84 ± 18.28%	42.82 ± 11.50%
GRU - 90	51.84 ± 18.28%	42.82 ± 11.50%
RNN - 15	54.32 ± 17.74%	44.24 ± 13.09%
RNN - 60	54.32 ± 17.74%	44.24 ± 13.09%
RNN - 90	54.32 ± 17.74%	44.24 ± 13.09%

Table 2. The table above displays the results of different time-windows for all three sequence learning algorithms LSTM, GRU and RNN. The data is relatively stable over different time scales, with only the decrease in accuracy between from GRU-15 to GRU-60 being even weakly statistically significant (p=.10). Algorithm performance continues to be different, e.g. RNN-90 outperforms LSTM-90 significantly for both TR and RX at (p=.05 and p=.02 respectively).

and 39.44 ± 10.51% average prediction accuracy for RX module (margins of error correspond to standard deviation computed over 20-folds). We noticed significant improvements in performance for RX module in a 2-sided tailed paired t-test (p=.03) with average prediction accuracy of 43.48 ± 11.26% for GRU-30, whereas TR accuracy increased to 53.90 ± 17.37% there is statistically weak evidence (p=.14). RNN-30 and GRU-30 were not statistically different at p=.7 for RX and p=.6 for TR. RNN-30, however, was statistically significantly better than LSTM-30 for TR (p=.05) and RX (p=.016) highlighting the importance of the choice of sequence learning algorithm. The significant increase in accuracy from LSTM-30 to RNN-30 is likely due to the reduced number of parameters from LSTM to RNN, hence reducing overfitting with the simpler RNN model.

The results for varying time segment lengths are presented in Table 2. Notice that the average 20-fold accuracy for RNN and LSTM doesn't vary with segment length. Interestingly, for GRUs the average accuracy drops slightly for TR from 53.90 ± 17.37% to 51.84 ± 18.28% and for RX from 43.48 ± 11.26% to 42.82 ± 11.50%. There is statistically weak evidence p=.10 for TR depending on sequence length. This drop in accuracy is may just be random variations, or it may be due to limited scale ranges in a GRU i.e. its cells cannot accumulate large values even if there is strong presence of a certain feature [50]. LSTM and RNN

on the other hand have consistent accuracy for both RX and TR irrespective of the segment lengths. While each is self-consistent there is a difference between them; a paired test of the per-fold accuracy of LSTM-90 with RNN-90 shows significant difference with p=.02 for RX and p=.05 for TR. The LSTM comprising of a forget gate, can adaptively ignore certain inputs and keep a set of weights that never get passed to the activation function. This can be a problem while predicting current state of engagement for a subject, since it remembers all transitional states to predict the final outcome. RNN's on the other hand can look back only a few steps which appears to helps it predict current engagement levels better.

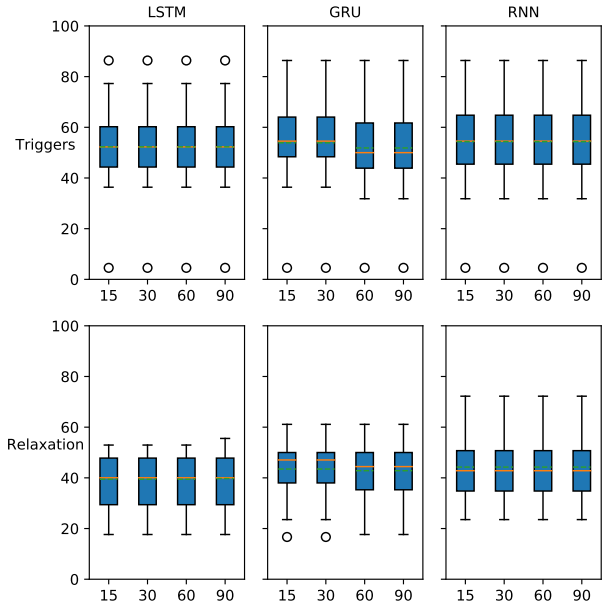

Figure 4. The figure above shows the box and whisker plots for LSTM, RNN and GRU in different contexts of Relaxation and Triggers. Each subplot in shows the Accuracy (%) on the y-axis and the Time Windows (seconds) on the x-axis. Each boxplot is generated from the 20-fold validation results, showing mean (green line), median (orange line), 25% (solid box) and 75% (whisker lines) percentile bands and outlier (circles). Notice that RNN's and LSTMs perform consistently across different temporal depths whereas GRU decreases from 30 to 60 for Triggers.

6. Discussion and Future Work

We evaluated major deep sequence learning algorithms for engagement prediction in the domain of trauma recovery. We used Facial Action Unit as the feature representation and performed detailed analysis with RNN, LSTM and GRUs. As noted earlier in the previous section, we notice algorithms with simpler internal structure like GRUs and RNN perform consistently better than LSTMs. Why does this happen? In the LSTM unit, the amount of the memory content that is seen, or used by other units in the network is controlled by the output gate. The ability to balance forgetting old and new information and squashing of the cell state is important in LSTM architecture. On the other hand the GRU exposes its full content without any control. Coupling the input and the forget gate avoids the problem of block output growing in unbounded manner, thereby making output non-linearities encoded in LSTM less important. Given that GRUs overall have less parameters, it is likely that the amount of training data available might allow GRUs to perform better compared to LSTM, but the data does not show a statistically significant difference. With the motivation of reducing the potential for overfitting, we also performed another experiment where we compare different amounts of training data. We observed that GRU-15 (see Table 2) performs significantly better than GRU-90. This suggests that merely increasing the temporal training data is detrimental, potentially because of mixing engagement levels as suggested by Whitehill *et al*. [59]. However, the best performance was with an RNN, the simplest model, and the RNN did not show any difference with different time windows. Future work will explore just how short a time frame can be used in an RNN.

In many machine learning applications, especially in medical/health applications, it is very time-consuming and expensive to acquire large amounts of annotated training data. While the sensory data may appear to have long sequences that could be used, and there is a prevailing intuition that "more data is better", this work shows that often shorter sequences are at least as good if not better. When annotations for training data are relatively limited, simpler units like GRU or RNN should be considered.

For these experiments, we relied on well known facial representation (FAUs) because of their use in decades of studies from psychology, affective computing and computer vision. However, with expressive power of the recent deep learning methods, researchers are exploring computing higher level inferences like facial expressions, directly from pixel intensities obtained from the video [7], but would require access to the raw video data. In the experiments we conducted so far, we used off-the-shelf AU detector as a feature representation. The AU detector was trained using existing datasets from computer vision and affective computing community. Moving forward, we could explore adapting pre-trained generic AU detectors on EASE dataset using techniques from unsupervised domain adaptation [62, 65]. Recently, Selective Transfer Machine were proposed as a transfer learning method, where an SVM classifier for AU detection is personalized by attenuating person-specific biases [13]. The authors of selective transfer machine accomplished state-of-the art results by learning the classifier and re-weighting the training samples that are most relevant to the test subject during inference. Other advances in this

field have also exploited the AU co-occurence by employing hybrid network architectures of Convolutional Neural Networks (CNN) and RNNs to jointly model AU detection and prediction [66, 67, 14]. There is potential of integrating the aforementioned methods to create more robust AU detectors that can likely increase accuracy of contextual engagement models by jointly training for engagement predictions.

References

[1] R. S. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser. Better to be frustrated than bored: The incidence, persistence, and impact of learners? cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241, 2010.

[2] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[3] P. Baxter, C. De Jong, R. Aarts, M. de Haas, and P. Vogt. The effect of age on engagement in preschoolers' child-robot interactions. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 81–82. ACM, 2017.

[4] A. Beckford. Applications of emotion sensing machines. *Exp Psychol*, 25(1):49–59.

[5] C. Benight, K. Shoji, CarolynYeager, A. Mullings, S. Dhamija, and T. Boult. Changes self-appraisal and mood utilizing a web-based recovery system on posttraumatic stress symptoms: A laboratory experiment. In *International Society for Traumatic Stress Studies*. ISTSS, 2016.

[6] A. Bernal, K. Crammer, A. Hatzigeorgiou, and F. Pereira. Global discriminative learning for higher-accuracy computational gene prediction. *PLoS computational biology*, 3(3):e54, 2007.

[7] R. Breuer and R. Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017.

[8] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny. The development of the game engagement questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, 45(4):624–634, 2009.

[9] R. A. Calvo, K. Dinakar, R. Picard, and P. Maes. Computing in mental health. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3438–3445. ACM, 2016.

[10] O. Celiktutan, E. Skordos, and H. Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 2017.

[11] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[13] W.-S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013.

[14] W.-S. Chu, F. De la Torre, and J. F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016.

[15] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[16] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding*, 91(1):160–187, 2003.

[17] M. Cox, J. Nuevo-Chiquero, J. Saragih, and S. Lucey. Csiro face analysis sdk. *Brisbane, Australia*, 2013.

[18] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[19] S. Dhamija and T. Boult. Exploring contextual engagement for trauma recovery. *CVPR Workshop on Deep Affective Learning and Context Modelling*, 2017.

[20] A. Dingli and A. Giordimaina. Webcam-based detection of emotional states. *The Visual Computer*, 33(4):459–469, 2017.

[21] S. D'Mello, P. Chipman, and A. Graesser. Posture as a predictor of learners affective engagement. In *Proceedings of the 29th annual cognitive science society*, volume 1, pages 905–910. Citeseer, 2007.

[22] S. D'Mello, E. Dieterle, and A. Duckworth. Advanced, analytic, automated (aaa) measurement of engagement during learning. *Educational Psychologist*, 52(2):104–123, 2017.

[23] S. D'Mello, A. Olney, C. Williams, and P. Hays. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of human-computer studies*, 70(5):377–398, 2012.

[24] T. Dunne, L. Bishop, S. Avery, and S. Darcy. A review of effective youth engagement strategies for mental health and substance use interventions. *Journal of Adolescent Health*, 2017.

[25] J. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*, 2013.

[26] A. Graves. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer Science & Business Media, 2012.

[27] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[28] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. *ICASSP*, 2013.

[29] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[30] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 2016.

[31] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–II. IEEE, 2004.

[32] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[33] Y.-H. Hung and P. Parsons. Assessing user engagement in information visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1708–1717. ACM, 2017.

[34] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

[35] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

[36] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 2017.

[37] S. Kipping, M. I. Stuckey, A. Hernandez, T. Nguyen, and S. Riahi. A web-based patient portal for mental health care: benefits evaluation. *Journal of medical Internet research*, 18(11), 2016.

[38] F. D. la Torre and J. Cohn. Facial expression analysis. *Visual Analysis of Humans - Springer*, pages 377–409, 2011.

[39] J. Lafferty, A. McCallum, F. Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[40] I. C. Landa-Avila and M.-L. Cruz. Engagement in a virtual reality game with gesture hand interface. an empirical evaluation of user engagement scale (ues). In *International Conference of Design, User Experience, and Usability*, pages 414–427. Springer, 2017.

[41] M. E. Levin, S. C. Hayes, J. Pistorello, and J. R. Seeley. Web-based self-help for preventing mental health problems in universities: Comparing acceptance and commitment training to mental health education. *Journal of clinical psychology*, 72(3):207–225, 2016.

[42] Z. C. Lipton, J. Berkowitz, and C. Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[43] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.

[44] H. Monkaresi, P. Bosch, R. Calvo, and S. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Trans. on Affective Computing*, 2017.

[45] C. Morrison and G. Doherty. Analyzing engagement in a web-based intervention platform through visualizing log-data. *Journal of medical Internet research*, 16(11), 2014.

[46] Y. I. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 139–148. ACM, 2010.

[47] M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011.

[48] J. O. Nigel Bosch, Sidney K D?Mello, R. S. Baker, and V. Shute. Using video to automatically detect learner affect in computer-enabled classrooms? 2010.

[49] A. G. Pour, A. Taheri, M. Alemi, and A. Meghdari. Human–robot facial expression reciprocal interaction platform: Case studies on children with autism. *International Journal of Social Robotics*, pages 1–20, 2018.

[50] A. Pulver and S. Lyu. Lstm with working memory. *arXiv preprint arXiv:1605.01988*, 2016.

[51] J. Purtle, A. C. Klassen, J. Kolker, and J. W. Buehler. Prevalence and correlates of local health department activities to address mental health in the united states. *Preventive medicine*, 82:20–27, 2016.

[52] L. Rabiner and B. Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[53] R. Rana. Gated recurrent unit (gru) for emotion classification from noisy speech. *arXiv preprint arXiv:1612.07778*, 2016.

[54] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 375–382. IEEE, 2010.

[55] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):944–958, 2015.

[56] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015.

[57] J. Schmidhuber. A local learning algorithm for dynamic feedforward and recurrent networks. *Connection Science*, 1(4):403–412, 1989.

[58] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005.

[59] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions.

[60] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. Faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Trans. on Affective Computing*, 5(3):86–98, 2014.

[61] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.

[62] S. Yang, O. Rudovic, V. Pavlovic, and M. Pantic. Personalized modeling of facial action unit intensity. In *International Symposium on Visual Computing*, pages 269–281. Springer, 2014.

[63] C. M. Yeager. *Understanding Engagement with a Trauma Recovery Web Intervention Using the Health Action Process Approach (HAPA) Framework*. PhD thesis, University of Colorado Colorado Springs., 2016.

[64] W. Zaremba, I. Sutskever, and O. Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[65] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe. Learning personalized models for facial expression analysis and gesture recognition. *IEEE Transactions on Multimedia*, 18(4):775–788, 2016.

[66] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2207–2216, 2015.

[67] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit and holistic expression recognition. *IEEE Transactions on Image Processing*, 25(8):3931–3946, 2016.