Coupled Systems for Modeling Rapport Between Interlocutors

Srijan Sharma[1], Kantha Girish Gangadhara[1], Fei Xu[3], Anne Solbu Slowe[2],
Mark G. Frank[2], Ifeoma Nwogu[1,3]
[1] Computer Science Dept., Rochester Institute of Technology. NY
[2] Communication Dept., University at Buffalo, SUNY
[3] Computer Science and Engineering, University at Buffalo, SUNY
{ss5938,kg2605}@rit.edu; {fxu3,as255,mfrank83,inwogu}@buffalo.edu

2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) | 978-1-6654-3176-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/FG52635.2021.9667067

***Abstract*— This research work explores different machine learning techniques for recognizing the existence of rapport between two people engaged in a conversation, based on their facial expressions. First using artificially generated pairs of correlated data signals, a coupled gated recurrent unit (cGRU) neural network is developed to measure the extent of similarity between the temporal evolution of pairs of time-series signals. By pre-selecting their covariance values (between 0.1 and 1.0), pairs of coupled sequences are generated. Using the developed cGRU architecture, this covariance between the signals is successfully recovered. Using this and various other coupled architectures, tests for rapport (measured by the extent of mirroring and mimicking of behaviors) are conducted on real-life datasets. On fifty-nine ($N = 59$) pairs of interactants in an interview setting, a transformer based coupled architecture performs the best in determining the existence of rapport. To test for generalization, the models were applied on never-been-seen data collected 14 years prior, also to predict the existence of rapport. The coupled transformer model again performed the best for this transfer learning task, determining which pairs of interactants had rapport and which did not. The experiments and results demonstrate the advantages of coupled architectures for predicting an interactional process such as rapport, even in the presence of limited data.**

I. Introduction

While the computational analysis of Internet-scale social network data has enjoyed significant progress in this era of big-data and deep-learning technologies, the same cannot necessarily be said for comprehensively analyzing face-to-face interactions; the types that exist in everyday occurrences, such as in classrooms, counseling sessions, meetings *etc*.

Research in social psychology has that in a well-functioning group setting, individuals with high cohesion and established rapport tend to exhibit different forms of interactional synchrony[1][4]. This includes a *mirroring effect*, which occurs when individuals match each other's facial expressions and facial articulations subconsciously. It also includes the *mimicking effect*, which is a type of mirroring but with time delays. Interactional synchrony often occurs in social settings in order to gain and keep rapport.

In this work, we are interested in exploring different deep sequence learning coupling methods, to model the signal exchanges that occur in face-to-face interactions, and determine whether a pair of interlocutors have established rapport or not during their interaction with each other. We pose this problem as a binary classification one, where the inputs to the classification models are extracted from the dynamics of the faces of the pair of interlocutors.

The specific objectives of this work therefore, are (i) to develop a model that can predict if there is rapport between two individuals; (ii) to ensure that the proposed model generalizes well on different datasets, i.e. model trained on one dataset should be able to achieve acceptable performance on another never-been-seen-before dataset; and (iii) to localize the regions in the data where rapport occurs.

II. Related Prior Work

Neural network based architectures have been applied to modeling interacting sequences, though not necessarily in the context of human dyads comunicating with each other. Liu et al. [11] presented a long short-term memory (LSTM) based system, to model the similarity in pairs of sentences, by combining their hidden states in four directions. This and similar architectures were very computationally expensive as they coupled in all four directions. Sun et al.[13] proposed a coupled recurrent network (CRN), which involved two parallel streams of LSTMs. The hidden state of stream-1 was concatenated with the input signal of stream-2, and vice-versa, for every succeeding recurrent step. Along similar lines, Morais et al.[12] introduced MPED-RNN which used two parallel streams of GRUs. The hidden states were combined in a way similar to CRN and the model was used for anomaly detection in surveillance videos.

Zadeh *et al.* [1] used a dynamic fusion graph (DFG) to dynamically attend to vision, language, and audio for multimodal sentiment classification from videos. Tsai *et al.* [14] improved on the performance of DFG, by using a bank of several transformers on the same modalities for the same task. Although there is a plethora of related works concerning fusion techniques for improved classification, we highlight [1] and [14] as they rely on the inherent alignment of the three modalities for improved classification. Our work also

This material is based upon work supported by the National Science Foundation under Grant No. 1846076.

[1]Interactional synchrony refers to how the speech or behavior of two or more people involved in a conversation become more synchronized with each other, and they can appear to behave almost in direct response to one another.

relies on detecting inherent alignment (mirroring/mimicking) between the data from the 2 interlocutors. But these models are extremely data hungry (due to the large number of model parameters), where the authors trained and tested the models on 3 different audiovisual datasets - with 2,199 videos, 10,000 videos and over 23,454 movie videos, respectively. Our collected data consists of 59 videos only!

Other works include Zhao *et al.* [17] which involved a rules-based framework for detecting rapport with virtual agents, Yu *et al.* [16] which also implements a transformer model but analyzes only individual for affect recognition, both [8] and [6] are dialogue-based engines, where their analysis of the sentiment of a conversation is more global than what we consider here. In our work, the interactional synchrony we measure is local and occurs in bursts; and as such, is quite different from the sentiment classification of a conversation.

III. Baseline Methods

In this section, we discuss some baseline methods for modeling interactional synchrony between pairs of signals. The models can be used either in a regression setting (for our artificially generated coupled data in Section V-A) or in a classification setting (for the real-life dyadic datasets in Sections V-B and V-E)

A. Baseline 1: End-to-End System of GRUs (e2eGRU) with Late Fusion

As a baseline to investigate the effectiveness of coupling, we develop a standard system involving two independent uncoupled GRUs, each GRU modeling the stream of data from each participant. The final output embedding h_T^1 and h_T^2 are combined and passed through a fully-connected layer as [illegible] d la [illegible] o

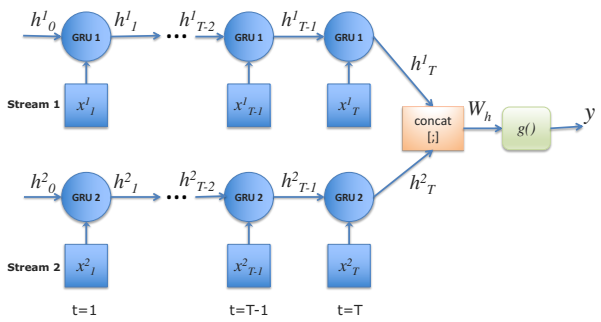

Fig. 1. The e2eGRU architecture for modeling interacting time-series data.

this architecture as end-to-end because by employing back-propagation-through-time, the network can be jointly trained from end-to-end, even though it consists of disparate chains. The overall architecture can be summed up as

$$p(y|X^1, X^2) = g(W_h * [h_T^1, h_T^2]) \quad (1)$$

where y is the prediction from the system; X^1, X^2 represent data streams 1 and 2 respectively; W_h are the weights at the last time step T; h_T^1, h_T^2 represent the last hidden messages along GRU^1 and GRU^2 respectively; $g(\cdot)$ introduces additional nonlinearity in to the system.

Comments:: Although the end-end GRU jointly trains the disparate models to optimize the task at hand, it misses any intricate exchanges that happen a few time steps apart. This model does not take advantage of the benefits of coupling the hidden states from one time step to the next as will be seen in our proposed coupled model in Section IV.

B [illegible] *R*

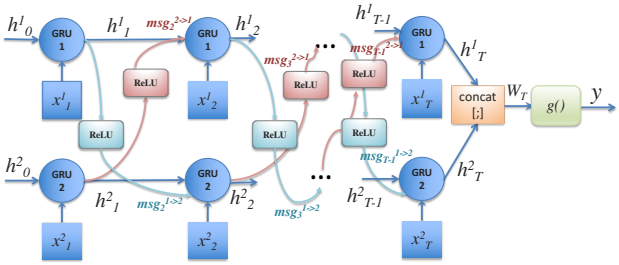

Fig. 2. The MP-RNN style architecture showing only the encoder with its message-passing processes

We briefly discuss the Message-Passing Encoder-Decoder Recurrent Neural Network (MPED-RNN) [12], as its architecture is also related to our proposed approach. Although the entire system is an auto-encoder, we focus only on the encoder component, which consists of two RNN branches that process local and global data separately. Hence, we refer to this version as the MP-RNN architecture. The 2 channels interact via cross-branch message passing at each time step. One main difference between this and many other architectures is how ReLU activation is applied to the hidden state of the opposite stream and then concatenated with the input signal at the next time step. The authors refer to this process as message-passing. The original architecture was used to detect abnormal events in videos of people. The governing equations are given below.

$$\begin{aligned} msg_t^{1\to 2} &= ReLU(W^{1\to 2} * h_t^1) \\ msg_t^{2\to 1} &= ReLU(W^{2\to 1} * h_t^2) \\ h_t^1 &= GRU_1([x_t^1; msg_t^{2\to 1}], h_{t-1}^1) \\ h_t^2 &= GRU_2([x_t^2; msg_t^{1\to 2}], h_{t-1}^2) \\ P(y|X^1, X^2) &= g(W_h * [h_T^1; h_T^2]) \end{aligned} \quad (2)$$

One drawback of this model is that it is not clear how Attention can be applied, due to its cross-branch message passing mechanism.

IV. Our Proposed Models

We propose two additional coupling models that focus on exploiting the role of attention (single- and multi-head) in determining the extent to which 2 or more dynamic signals are interacting.

A. The coupled GRU (cGRU)

p
s

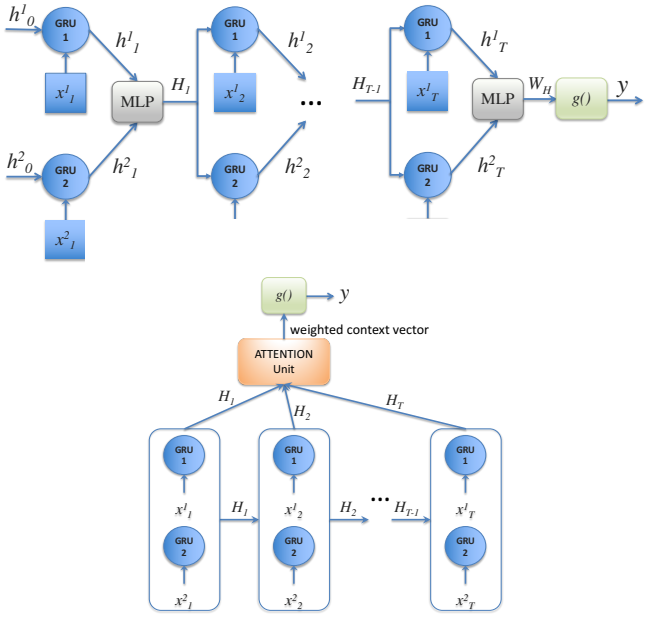

Fig. 3. (Top:) The cGRU architecture where coupling between chains is performed at every time step. (Bottom:) The abstraction of the network from the top image, showing how attention is applied on the coupled hidden messages H_i instead of the individual hidden messages h_i.

Figure 3 summarizes this coupling approach. Rather than simply concatenating the hidden messages from each time step, we feed the hidden output messages from the different streams $\{h_t^k\}; k = 1 \cdots M$ into a multi-layer perceptron (MLP) to couple them and fix the dimensionality of the output hidden messages H^t for M chains (but we only work with 2 chains in this project). The overall architecture can be summed up as:

$$\begin{aligned} H_t &= f(W_H * [h_t^1, h_t^2]) \\ p(y|X^1, X^2) &= g(W_H * H_T) \end{aligned} \tag{3}$$

where the terms are as defined previously and H_T is the aggregated output of the MLP at time step T; $f(\cdot)$ and $g(\cdot)$ are additional nonlinear functions in the system.

Adding an Attention Unit

One main disadvantage of the cGRU implementation is its inability to remember all the relevant information in long sequences, where the system can potentially forget the earlier portions, by the time it has processed the entire the sequence. The attention mechanism was forged to resolve this problem because it directs the focus of the network to pay greater attention to certain factors when processing data. We therefore impose attention on the cGRU as part of the proposed architecture. The coupled hidden states are fed to an attention head where the unit computes a weighted sum of hidden states. The high level network graph is shown in Figure 3b and the equations governing the system are given as:

$$\begin{aligned} e_t = a(H_t); \quad \alpha_t &= \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \quad c = \sum_{j=1}^{T} \alpha_j H_j \\ p(y|X^1, X^2) &= g(c) \end{aligned} \tag{4}$$

The top three equations above define the attention mechanism, which uses a Softmax-like estimator to determine the influence of each coupled hidden message H at time j.

Benefits:

The benefits of this architecture over previously existing ones include: (i) the ease to extend to multiple chains without a significant increase in the number of parameters in the network (as would be the case with MPDED-RNN and other similar architectures); (ii) the ability to add an attention module to the network system using H_t, especially when the number of stream sources is greater than 2. It is not clear how to accomplish this with other networks discussed in the literature; (iii) the improved performance over previously existing methods, when tested with real-life data.

Reproducibility:

Reproducing the proposed technique is fairly straightforward, where two or more RNN variants (GRU, LSTM, etc) are each set up to represent the data from a stream of data. After each time step, the hidden outputs of each RNN are coupled via an MLP unit, resulting in one "aggregated-hidden" message H_t at time t. The messages at the last time step are then concatenated and fed into a non-linearity and finally to the classifier/regressor as the task may be. The aggregated messages are fed into an Attention unit. The entire system is jointly trained using standard back-propagation through time.

B. The coupled Transformer (cXf)

The transformer architecture was introduced by [15] to improve on the attention mechanism, accounting for longer dependencies with a multi-head self-attention mechanism. Unsupervised learning using transformer-based encoder-decoder mechanisms was first introduced in the bidirectional encoder representations from transformers (BERT) [5] for Q&A language task.

For this coupled Transformer (cXf) model, we build on the previously existing work by attempting to align or couple two time series signals using individual transformers, and then couple the signals using a third one. Where the previous models aimed at locally coupling signals at each time step, this coupling occurs at a more global level.

Figure 4 summarizes our next proposed coupling approach. Synonymous with the hidden messages from previously described models, we couple the latent representations from each signal (the outputs of the individual transformers) using a third coupling transformer. The query (Q), key (K),

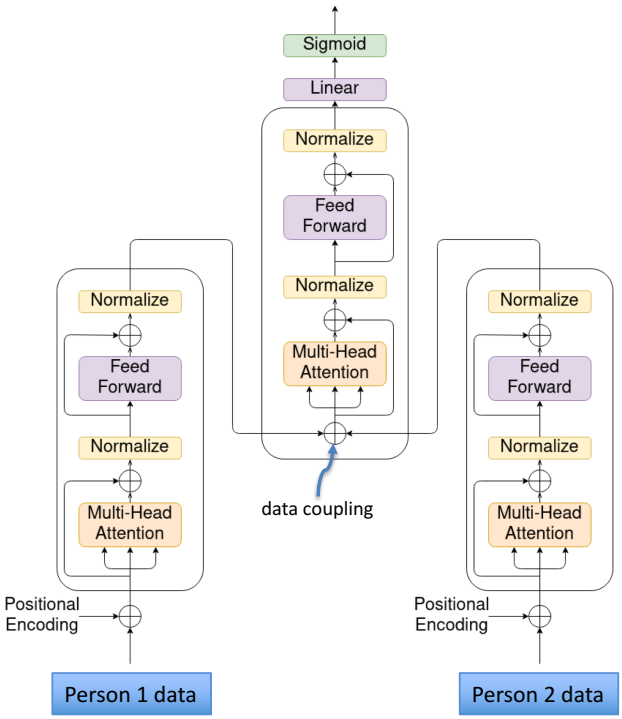

Fig. 4. The cXf architecture where 2 data streams are represented by two encoding transformers and the coupling between them is achieved by the decoder.

and values V for each of the individual transformer encoders are calculated as in Equation (5).

$$\begin{aligned} K_f &= linear\_key(input_f) \\ Q_f &= linear\_query(input_f) \\ V_f &= linear\_value(input_f) \end{aligned} \tag{5}$$

where, *linear_key*, *linear_query*, and *linear_value* are three individual linear layers learned for the input features (input$_f$), where, $f \in \{Person1data, Person2data\}$.

The attention mechanism for one of the transformer encoder layer is shown in Equation (6).

$$Attention(K_f, Q_f, V_f) = softmax(\frac{Q_f K_f^T}{\sqrt{d_k}})V_f \tag{6}$$

The output from each encoder layer is then input to the final block that attends to the information collectively from both data streams Consider y_f the output from each of the add and normalization blocks of the encoder, then the input to the coupling transformer can be written as (7).

$$input_{coupling} = [y_{f_1} : y_{f_2}] \tag{7}$$

The output of the final coupling transformer is then fed to a linear layer and finally to the SoftMax rapport classifier.

V. Experiments, Data and Results

Although most dynamic processes in nature are coupled, there is still such a shortage of the availability of such interacting datasets, especially ones that involve minor delays as can be observed in mimicry. Even more so, there is a lack of stochastic models for artificially simulating such coupled processes in order to study them more effectively.

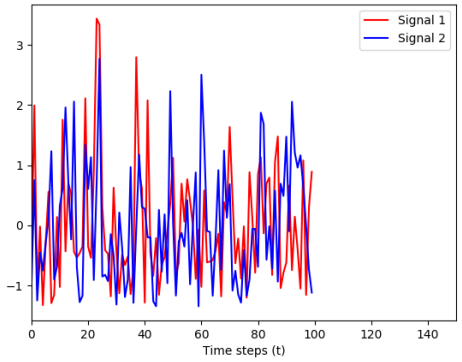

Fig. 5. Generated synchronized sequences with covariance of 0.5.

A. *Experiment 1: Testing on controlled synthetic data*

To address the lack of highly controlled interacting time-series data in the community, we implemented a theoretical method discussed by [9] for generating coupled Gaussian stochastic processes, with arbitrary correlation functions (not necessarily auto-correlated or cross-correlated). In this technique, a Fourier filtering method takes in two uncorrelated sequences of random numbers and returns two correlated ones based on a given correlation value.

We elaborate on the generation method to allow interested researchers to readily generate similar data or extend the process. In addition to the generation process described in Algorithm 1, we will release the code we developed for generating the simulated synchronized data.

Algorithm 1 Data generating process

procedure DataGen($len, C, f_1(t,x), f_2(t,x), delay$)
  $S_{xx} = fft(C_{xx})$; $S_{yy} = fft(C_{yy})$; $S_{xy} = fft(C_{xy})$
  $u = SampleGaussian(len)$; $v = SampleGaussian(len)$
  $U = fft(u)$; $V = fft(v)$
  $\beta = 1$; $\alpha = \overline{\beta} + \cos^{-1}\left(\frac{S_{xy}}{\sqrt{Sxx * Syy}}\right)$
  $A_q = \sqrt{S_{xx}} * \cos(\alpha)$; $B_q = \sqrt{S_{xx}} * \sin(\alpha)$;
  $C_q = \sqrt{S_{yy}} * \cos(\alpha)$; $D_q = \sqrt{S_{yy}} * \sin(\alpha)$
  $X = A_q * U^T + B_q * V^T$
  $Y = C_q * U^t + B_q * V^T$
  $x^* = |ifft(X)|$
  $y^* = |ifft(Y)|$
  $x = x^*[1 + delay :]$
  $y = y^*[1 : -delay]$
  Return $[f_1(1, x_1), ..., f_1(len, x_{len})]$,
    $[f_2(1, y_1), ..., f_2(len, y_{len})]$
end procedure

We successfully simulated 10,000 pairs of correlated sample data (split into 8000/1000/1000 for train/test/validation respectively) with correlation factors in the range $[0, 1]$ each having a sequence length of 100. The correlation factors were sampled from a uniform distribution. An example of the output pair of synchronized 100-length sequences generated by the process is shown in Figure 5, where the sequences

were generated with a covariance of 0.5 and the prediction made by the cGRU network for this example was **0.48**.

1) Are coupled systems useful for evaluating synchronized signals? : The first test performed was to establish the effectiveness of the coupling architecture. To accomplish this, we trained both an end-to-end GRU and the coupled GRU architecture using the negative log-likelihood (NLL) regression loss. The training inputs to the two networks were the 8,000 pairs of simulated data and the targets were their known correlation factors (between 0 and 1); we validated and tested with 1,000 designated pairs respectively. The goal here was to recover the numerical correlation factor (or extent of interaction), given a pair of input test signals. This would allow us to evaluate the effectiveness of coupling hidden states, when assessing interacting signals.

Figure 6 shows the plots of the predicted-versus-actual correlation values for the two architectures, empirically demonstrating that the coupled GRU significantly outperformed its non-coupled counterpart when estimating the original covariance between the two input signals.

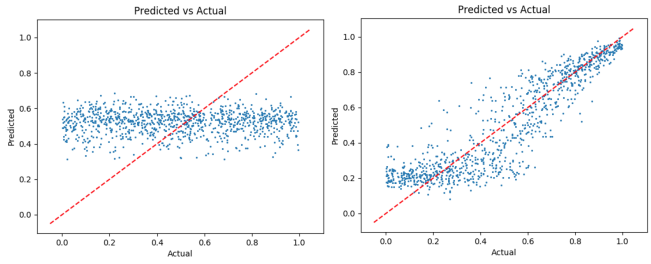

Fig. 6. Image on the left shows a plot of predicted vs. actual correlation values, for 1000 test sequences, as estimated by the end2end GRU; the image on the right shows the results as estimated by the coupled GRU. The associated mean errors were 0.25 and 0.0892 respectively.

We also computed the mean of the absolute percent error: $\mu_e = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$.
where Y_i is the covariate value used to generate the i^{th} pair, $\hat{Y}_i$ is the model's prediction on the same i^{th} pair and N is the total number of testing pairs.

For the same dimensionality of hidden layers (64), the same number of epochs (50) and the same learning rate (0.001), the test errors from the e2eGRU and cGRU were **0.25** and **0.0892** respectively.

Observation: With this experiment, we have shown that a coupled network is capable of modeling the inherent interactions between signals and coupling does indeed improve the ability to evaluate the interactivity between pairs of input signals.

B. *Experiment 2: Testing on Real-life Interviewing Data*

The data for this aspect of testing was obtained from a study involved in the larger context of deception detection where we were interested in detecting whether a pair of interlocutors have established rapport between them or not during an interview.

Data collection involved the interviewing process between a retired police officer and an interviewee. The interviewer

Fig. 7. Sample video frame showing the interviewing process between a retired police officer (right hand side) and the interviewee on the left hand side.

was instructed to build rapport with the subject for the rapport scenarios, while they had on their "game-face" in the non-rapport scenarios. During the interview, each participant's face was recorded with a separate video camera and a third camera captured both the interviewer and interviewee simultaneously in order to preserve spatial information for body posture analysis (future work). A total of fifty-nine ($N = 59$) interviews from the study were used for this rapport analysis. Thirty of the videos involved interviewers building rapport with the subjects while the remaining twenty nine did not.

The lengths of the videos ranged from 6 to 10 minutes per conversational pair. The videos were then sampled at 30 frames per second, so that a 7-minute video would yield 12,600 frames. Because these are too long for a temporal model, each video was divided into non-overlapping segments of 100 frames. All segments from a video were assigned with same overall label of the video. We split the 59 interview videos into 45/6/8 for training, validation and testing respectively and the splits were performed at the video level, not segment-level. We shuffled the splits and repeated the test ten times.

We extracted **17 AUs** over time, on the frontal facing videos for each participant and applied OpenFace [2], for facial action unit detection. We also extracted **68 landmarks** via DLib[10] from each participants face on every frame. These served as the dynamic input features to the models we tested.

The goal here was to classify the interaction between a conversational pair into has_rapport or no_rapport interaction - a binary classification problem.

1) Recognizing rapport in the interview dataset: The four networks we evaluate here are the 2 baselines: (i) the e2eGRU (from Section III-A), (ii) the MPED-RNN described in Section III-B; and the three newly proposed methods: (iii) cGRU and (iv) cGRU with Attention, both in Section IV-A; and lastly, (v) coupled transformer (cXf) model described in Section IV-B . Each of the networks handled the same two streams of data, whose inputs were either 17 AUs or the 68 face landmark points extracted from the interview videos. To

Fig. 8. Face-to-face condition in the virtual human project from USC

significantly augment the data, we used an overlapping 100-width sliding window to generate the segments from which the input features were extracted.

Training scheme: We split the 59 interview videos into 45/6/8 for training, validation and testing respectively. We trained the end-to-end GRU, coupled GRU, coupled GRU with attention, MPED-RNN and coupled transformer to predict rapport between two conversing individuals, on the AB Rapport dataset. We used cross entropy as our loss function and applied the Adam optimizer with a different learning rates for each model. Learning rates were selected with trial and error to give the best training performance for that model.

We trained all our models with two input configurations, one with landmarks as the input and the other with action units. All the configurations were trained and tested 10 times. Each run was started with randomly initialized weights, with no transfer learning from last run. For each run, training set, validation set and test set were randomly selected. We calculated the average accuracy for each model with each input configuration. The mean accuracy is tabulated in Table I, column AB Rapport. The delta between the mean and extremum is tabulated in "Delta"column. To visually compare the different models and variants of input data (landmarks or action units), the results are shown in Figure 9.

C. *Experiment 3: Generalizability: Transfer Learning to a Never-before-seen Dataset*

The third experiment was designed to study how well the proposed models could generalize to a dataset not seen before. This publicly available data collected in 2007 by Gratch *et al.* [7], was part of a larger virtual human study designed to investigate the importance of *contingent feedback*[2] in creating feelings of rapport.

A camera was placed in front of each speaker and a third one was attached to the ceiling to record the speaker and listener. The speakers were told to narrate a story while the listener (a confederate) listened. We used only the condition where both the speaker and listener were humans (as opposed to other conditions involving virtual characters), since this most closely matched our previous data. Figure 8 shows an example of a face-to-face session during data collection. This condition consisted of 20 pairs of conversants.

After the experiment, the speakers completed a post-interview questionnaire which consisted of various questions on a 10-point rapport scale. For positive rapport, we used the cases where both participants indicated on the post-interview questionnaire that they experienced rapport ≥ 6 with their partner. Other cases were treated as no rapport. We then converted the scale into rapport(1)/non-rapport(0), to comply with our previously trained models.

We applied the models we trained from the previous exercise (from the AB deception interviews) to investigate how well they generalize. No additional training was performed on this dataset; we used the available data only for inference.

Similar to the previous experiment, we extracted both landmarks data and action units from the faces. Each video was split into non-overlapping 100 frame segments. All 20 pairs of interactants were tested using the pre-trained networks and the classification results are given in Table I, column ICT Rapport.

TABLE I

VALIDATION RESULTS FOR THE AB RAPPORT DATASET AND THE TRANSFER LEARNING PERFORMANCE ON USC ICT RAPPORT DATASET.

*TRAINING THIS MODEL WAS SO SLOW, THAT WE COULD ONLY PERFORM A SINGLE RUN.

	Input Type	AB Rapport	Deltas	ICT Rapport
GRU	Landmarks	67.44	12.54	59.18
	Action Units	53.42	7.27	59.549
cGRU	Landmarks	57.5	1.83	51.81
	Action Units	62.98	10.97	60.549
cGRU w/ Attention	Landmarks	72.5	7.98	59.75
	Action Units	74.15	5.97	63.98
MPED-RNN*	Landmarks	70.4	NA	44.0
	Action Units	79.14	NA	53.37
Transformer	Landmarks	**81.97**	6.64	**67.22**
	Action Units	75.71	2.93	62.277

Observation: From the results of this and the previous experiment, we show that (i) our proposed coupled transformer model can recognize rapport better than the other networks, in the presence of real data which can be noisy, having interactions that might be delayed relative to each other, or occur sporadically. (ii) our coupled systems generalize well to data that has never been seen before by the networks and the coupled transformer model also performs the better than all the others.

D. *Experiment 4: Execution Time Comparison*

We bench-marked the time needed to train one iteration for each the models we are investigating. For running the benchmark, we used a batch size of 256. The execution time of all the models are tabulated in table II.

[2]The authors defined contingency as nonverbal feedback by the listener, such as facial expressions, nods or posture shifts that are tightly coupled to what the speaker is doing at the moment. Non-contingent feedback is defined as the listener feedback similar in frequency and characteristics to the contingent feedback, but not coupled with the speaker.

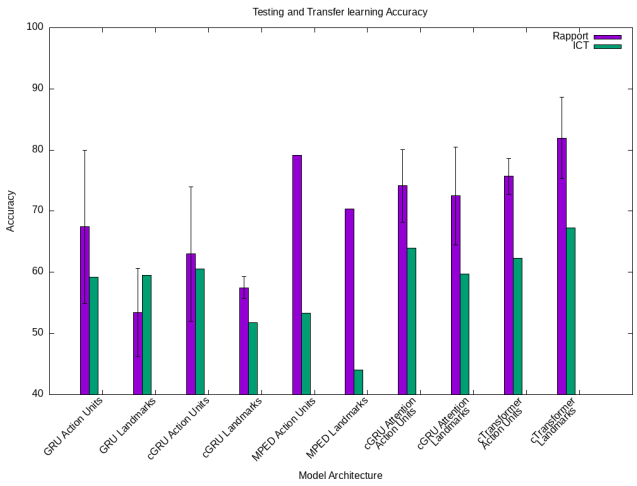

Fig. 9. A visual comparison of the different models with different inputs tested on the Rapport datasets

TABLE II

TIMING COMPARISON FOR DIFFERENT MODELS

Model	Feature	Time
GRU	Landmarks	7.8ms
	Action Units	9.3ms
cGRU	Landmarks	97ms
	Action Units	92ms
cGRU w/ Attention	Landmarks	93ms
	Action Units	97ms
MPED-RNN	Landmarks	351ms
	Action Units	364ms
cXf	Landmarks	32ms
	Action Units	40ms

Observation: From Table II, we observe that there is no clear advantage in choosing landmarks over action units as far as the execution time is concerned for the models. Of the five bench-marked, the end-to-end GRU has the lowest execution time, probably due to the simplicity of model. The coupled transformer takes more time than the GRU, but is significantly faster than both cGRU and cGRU with attention. This is due to the lack of recurrence, which simplifies training. Coupled GRU and coupled GRU with attention have similar execution times because of the similarity in their architectures. Message Passing Encoder Decoder RNN is the slowest as this architecture requires a significant amount of memory copies in its implementation; this slows down the model considerably.

*E. **Experiment 5:** Synchrony Detection by Visualizing Transformer Attention*

In this test, we are interested in localizing the regions of high synchrony in a video segment labeled as having rapport. To accomplish this, first we only test on time segments that have the rapport label. Next, we select samples that gave the lowest losses during testing, as these were the samples the model was most confident of its predictions.

Figure 11 illustrates some frames from such a sample having rapport and the corresponding Attention map is displayed in Figure 10. The actual video (along with others are provided in the Supplementary material) shows a lot of overlapping nodding and smiling actions between the pair.

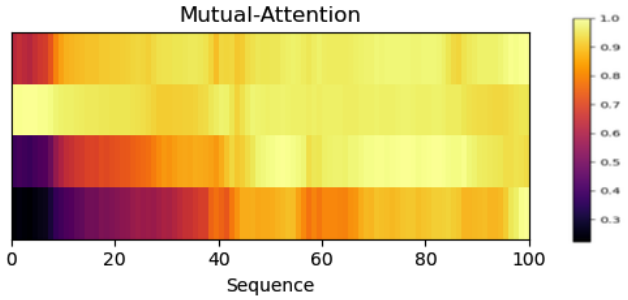

Fig. 10. Attention intensities from the 4-head coupling transformer. The bottom two heads are from the Interviewer and the top 2 from the Interviewee. Attention is high for all rows where synchrony is strong.

Fig. 11. Top: example of strong synchrony instances from interviewee (frames 7851-71); bottom: same overlapping instance from interviewer (frames 7851-63)

VI. DISCUSSION AND CONCLUSION

We have demonstrated that even in the presence of very limited data, we are able to computationally detect a complex social phenomenon such as rapport between 2 interactants, by measuring for the presence of synchrony using their facial expression data. We accomplished this with two novel deep-learning based technologies where a system of two or more neural networks can receive inputs from multiple data streams (one network per stream) and couple them either via the hidden outputs from each network, or by using a Transformer. Unlike the MP-RNN model, our proposed manners of coupling signals allows for working with of an arbitrary number of input data streams, without the system exploding in the number of parameters or increasing in the hidden message dimensions.

We tested our models on synthetically coupled data and on noisy, real-life face data sets where we successfully detected rapport. We then transferred the trained models to data collected over 13 years prior (poorer resolution images, different experimental conditions, subjective measures of rapport) and still successfully detected rapport between pairs of conversants. We have showed that our transformer-based coupled system is most robust and consistently gives good performance on real-life datasets.

REFERENCES

[1] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics - Long Papers*, ACL, pages 2236–2246. Association for Computational Linguistics, jul 2018.

[2] T. Baltrusaitis, P. Robinson, and L. P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 354–361, Dec 2013.

[3] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *cvpr*, volume 97, page 994, 1997.

[4] T. L. Chartrand and J. A. Bargh. The chameleon effect: the perception–behavior link and social interaction. *J. Pers. Soc. Psychol*, 76(2):893–910, 1999.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[6] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[7] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *International workshop on intelligent virtual agents*, pages 125–138. Springer, 2007.

[8] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018.

[9] T. Jamali and G. Jafari. Method for generating two coupled gaussian stochastic processes. *arXiv preprint arXiv:1602.04697*, 2016.

[10] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[11] P. Liu, X. Qiu, and X. Huang. Modelling interaction of sentence pair with coupled-lstms. *arXiv preprint arXiv:1605.05573*, 2016.

[12] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] L. Sun, K. Jia, Y. Shen, S. Savarese, D. Y. Yeung, and B. E. Shi. Coupled recurrent network (crn). *arXiv preprint arXiv:1812.10071*, 2018.

[14] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[16] H. Yu, L. Gui, M. Madaio, A. Ogan, J. Cassell, and L.-P. Morency. Temporally selective attention model for social and affective state recognition in multimedia content. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1743–1751, 2017.

[17] R. Zhao, T. Sinha, A. W. Black, and J. Cassell. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233. Springer, 2016.