Personalized estimation of engagement from videos using active learning with deep reinforcement learning

Ognjen Rudovic, Hae Won Park, John Busche, Björn Schuller, Cynthia Breazeal, Rosalind W. Picard

Angaben zur Veröffentlichung / Publication details:

Rudovic, Ognjen, Hae Won Park, John Busche, Björn Schuller, Cynthia Breazeal, and Rosalind W. Picard. 2019. "Personalized estimation of engagement from videos using active learning with deep reinforcement learning." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 16-17 June 2019, Long Beach, CA, USA*, edited by Larry Davis, Philip Torr, Song-Chun Zhu, William Brendel, and Mohamed R. Amer, 217–26. Piscataway, NJ: IEEE. https://doi.org/10.1109/cvprw.2019.00031.

Personalized Estimation of Engagement from Videos Using Active Learning with Deep Reinforcement Learning

Ognjen (Oggi) Rudovic[1,3], Hae Won Park[1], John Busche[1],
Björn Schuller[2,3], Cynthia Breazeal[1] and Rosalind W. Picard[1]
[1]Massachusetts Institute of Technology, USA, [2]Imperial College London, UK, [3]Augsburg University, Germany
{orudovic,haewon,jcbusche,cynthiab,roz}@mit.edu, bjoern.schuller@imperial.ac.uk

Abstract

Perceiving users' engagement accurately is important for technologies that need to respond to learners in a natural and intelligent way. In this paper, we address the problem of automated estimation of engagement from videos of child-robot interactions recorded in unconstrained environments (kindergartens). This is challenging due to diverse and person-specific styles of engagement expressions through facial and body gestures, as well as because of illumination changes, partial occlusion, and a changing background in the classroom as each child is active. To tackle these difficult challenges, we propose a novel deep reinforcement learning architecture for active learning and estimation of engagement from video data. The key to our approach is the learning of a personalized policy that enables the model to decide whether to estimate the child's engagement level (low, medium, high) or, when uncertain, to query a human for a video label. Queried videos are labeled by a human expert in an offline manner, and used to personalize the policy and engagement classifier to a target child over time. We show on a database of 43 children involved in robot-assisted learning activities (8 sessions over 3 months), that this combined human-AI approach can easily adapt its interpretations of engagement to the target child using only a handful of labeled videos, while being robust to the many complex influences on the data. The results show large improvements over a non-personalized approach and over traditional active learning methods.

1. Introduction

Engagement is a process where multiple parties establish, maintain, and agreeably end their perceived connection during a joint interaction [41]. The ability of socially situated intelligent robots to perceive and estimate users' engagement is critical for enabling timely, naturalistic and affect-sensitive interactions with users, which makes them suitable educational and therapeutic companions [4, 34, 32]. Accurately recognizing the state of user engagement enables such systems to deliver just-in-time interactions necessary to achieve the intervention goals [14]. One of the fundamental challenges in engagement estimation is in the wide spectrum of how people elicit engagement and how it is represented in computational engagement models. Traditional approaches so far have used nonverbal engagement cues such as gaze patterns, body pose, prosody, facial expressions, proxemics, and task-context behaviors such as providing input to the interaction task to build non-parametric engagement state classifiers [6, 35, 34, 38]. However, such models rarely work equally well for every individual, especially when there is a large variation between and within target individuals. Because of this, models learned from data of training subjects usually underperform when tested on previously unseen subjects. This calls for new modeling approaches that can deal effectively with individual differences, thus, moving from "one-size-fits-all" toward personalized models for engagement estimation.

While models of personalization from image data have been researched in several previous contexts (e.g., self-reported pain analysis [29] and robot-assisted therapy for children with autism [36]), they are designed for static images, thus, they do not provide a principled way of dealing with video data. Furthermore, when faced with hours of video data, which is typical in real-world human-robot interactions, it is critical for the model to be able to "actively" select those instances that the model is uncertain about so that they can further be analyzed (e.g., by human experts) and used to personalize target models (in our case, the models for engagement estimation). For this, modeling frameworks such as active learning (AL) [40] and reinforcement learning (RL) [42], provide a principled means for learning an optimal data labeling and classification policy. However, most of the existing frameworks for AL using deep RL are designed for static modeling tasks, such as image classification, and are not directly applicable to videos. On the other hand, several works proposed using AL for action de-

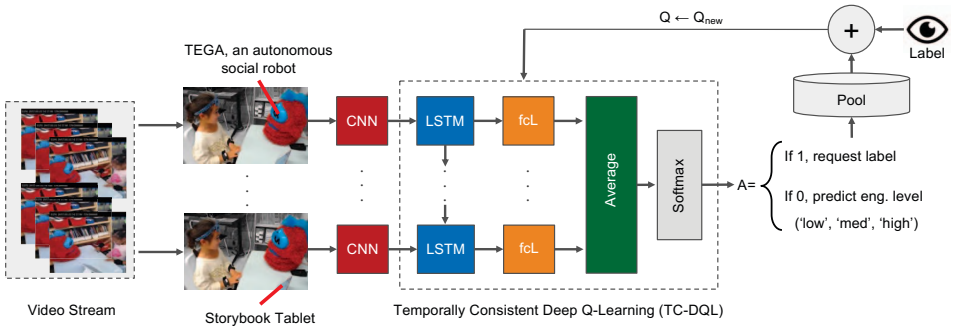

Figure 1: Overview of the proposed approach. The input is a video stream of the child-robot interactions segmented into fixed size intervals (5 seconds, divided into 10 time steps). The image frames are first passed through a pre-trained CNN (ResNet), and then used as input to the proposed Temporally Consistent Deep Q-learning (TC-DQL) for video labeling/classification policy learning. This is modeled using an LSTM cell unrolled over the time steps, and followed by fully connected (linear) layers (fcL). The outputs of these fcL are then averaged over time and passed through a softmax layer, which outputs the optimal action for the target video: whether to ask for the video label or estimate the engagement level of the child. If the label is requested, the target video is stored in a data pool for further labeling (e.g, by a human expert). After each session of the child-robot interactions, the engagement classification and labeling policy is personalized to the target child using the previously and newly labeled videos from the data pool, by updating the parameters Θ of the Q-function.

tection [43, 3], action recognition [17], and action localization [7] from videos. However, these works rely on heuristic AL strategies and not data-driven AL through deep RL, as done here. More importantly, they do not attempt personalized AL from video data, which is the focus of this work.

To this end, we propose a novel deep AL approach for determining whether a robot requires a new label to maintain a high confidence level of target-user-specific engagement estimation. Our method starts by learning an offline label request and classification policy (the group policy). This policy is subsequently personalized to the target user based on the requested video data for labeling during the inference stage (i.e., as the new data of target users become available). In the approach presented here, the queried videos are labeled in an offline manner by the human expert. In the future, a similar approach could be used to allow a robot to autonomously request new labels during an interaction, e.g., by asking questions such as "do you want to keep playing?", when it perceives a user's engagement as low, or when it is uncertain about its estimates.

This work brings together ideas from personalized machine learning [36], and deep RL [47], to formulate a personalized active learning approach for efficient labeling and classification of user's engagement from videos. The contributions of this work can be summarized as follows: (i) We propose a novel approach for automated estimation of engagement levels (low, medium, high), as coded by human experts, directly from videos of child-robot interactions in real-world conditions. (ii) We introduce a novel deep RL architecture, named Temporally Consistent Deep Q-learning (TC-DQL), that provides principled means for learning the Q-function for RL from videos by leveraging temporal dependencies between image frames (see Figure 1). This is in contrast to existing AL frameworks based on active-one-shot learning (AOSL) and RL (e.g., [47, 28, 39]) that deal with static image classification and assume that image labels are always available. (iii) We propose a novel algorithm for personalized policy learning that enables the model to adapt its engagement interpretations to each child in a sequential manner using only a handful of human-labeled videos of the child-robot interaction sessions. This largely reduces the human labeling effort, which is time and labor intensive. We evaluate this approach using video recordings of 43 kindergarten children, being part of a new child-robot storytelling interaction dataset [34]. We show that by personalizing the model policy for requesting the video labels, we can largely improve the engagement estimation for children in the dataset. We also show that this approach outperforms the traditional and non-personalized AL strategies, when using the same budget for requesting the video labels.

2. Related Work

A large body of work in human-robot interaction (HRI) explored the use of various affective and social cues, such as gaze patterns, body pose, prosody, facial expressions, proxemics, and physiological information (e.g., skin conductance), as well as task behaviors to infer about a user's engagement state. These can be divided into those that detect the presence of a set of the engagement cues or interaction events [35, 34, 15], or use supervised classifiers trained with social, physiological, or task-based interaction

features [6, 38, 9]. Such approaches require expert engineering of input features and cannot deal with large feature dimensions efficiently, e.g., when pixel values from face images are used as input. To address this, [33] proposed a deep learning approach for engagement estimation from face images. However, the traditional "one-size-fits-all" models usually do not work well when the data is highly heterogeneous (e.g., due to the differences in facial expressions/body gestures as a result of individual engagement styles).

Recently, several works proposed models for personalized estimation of engagement in HRI. [36] proposed a multi-modal deep learning for engagement estimation that combines body, face, audio and autonomic physiology data of children with autism during therapy sessions with a humanoid robot. Similarly, [37] proposed a deep learning architecture for engagement estimation from face images, by adapting the target approach to different cultures and individuals. However, these models are static and do not deal with video data. By contrast, the proposed approach deals with raw video data in a principled manner, and is able to learn an efficient policy for video labeling and engagement estimation in a personalized fashion.

The approach proposed here is highly related to AL frameworks [40]. Central to the AL framework is the query strategy used to decide when to request a label for target data. The most commonly used query strategies include uncertainty sampling, entropy, or query-by-committee [40]. Furthermore, more advanced query strategy have been proposed to adapt deep neural network classifiers based on the uncertainty of the network output (e.g., [21, 45, 23]). Yet, the candidate query strategies still must be specified by a human. More recent works (e.g., [2, 20]) proposed AL "by learning" using the notion of meta-learning [26]. Despite their success in various learning tasks, these models still approximate the learning strategy via a pre-defined set of basic AL strategies (e.g., uncertainty sampling or entropy).

Instead of using heuristic strategies, recent deep AL approaches (e.g., [28, 13, 47, 44, 11]) have adopted a data-driven approach that learns a model-free AL off-line policy using RL [42]. For instance, [47] proposed a model where an agent makes a decision whether to request a label or make a prediction. The agent receives a reward related to its decision: a positive reward is given for correct predictions, and negative rewards for incorrect predictions or label requests. This is achieved by Q-learning modeled using the notion of deep RL [30]. However, this static RL approach is designed for problems such as image classification on the Omniglot dataset [25]. Its main goal is to adapt the prediction model to new tasks, using a minimum number of queries. This problem has also been addressed by the recent AOSL frameworks [39, 31, 22], that use meta-learning to adapt quickly to new tasks from a few examples of new classification categories.

In summary, the main difference of this work from prior work on RL [28, 13, 47] and AL [2, 20], is that prior efforts are devised for static inputs such as image frames, and not videos. Furthermore, the standard AL frameworks and those that "learn how to learn" require the labeling budget to be pre-specified, while in our case, this is learned from training data. Moreover, most of these approaches are designed for stream-based AL. This, in turn, requires the models to be updated after each query. This is impractical for two reasons. First, it is computationally demanding to update the perception modules after each video query (in our case, every 5 sec). Second, to maintain a naturalistic and engaging interaction, the robot should avoid asking too often a human expert or the user to provide the correct engagement label. Instead, we propose an approach where the robot first stores the videos it is uncertain about, in which case no engagement estimation is made. After the interaction, a human expert is asked to provide her feedback for those videos, which are then used in an off-line manner to personalize the data labeling/engagement estimation policy by optimizing it for future interactions with the target child.

3. Preliminaries

3.1. Problem Statement and Notation

In our learning setting, we use video recordings of child-robot interactions [34], described in Section 5.1. Formally, we denote our dataset as $\mathcal{D} = \{d_1, .., d_i, .., d_C\}$, where d_i comprises video recordings of child i, and C is the number of children. The data of each child are segmented into a maximum of $M = 8$ different sessions (one session per week) as $d_i = \{S_{i,j}\}_{j=1,..,M}$, but the number of sessions may vary per child (e.g., when a child did not attend a session). Furthermore, each child's session contains K video clips of that child, denoted as $S_{i,j} = \{v^1_{i,j}, .., v^k_{i,j}, .., v^K_{i,j}\}$, where K may vary per child. Lastly, each video clip comprises $v^k_{i,j} = \{X, y\}$, where $X = [x_1, .., x_T] \in \mathcal{R}^{250\times250\times T}$ is a window of T image frames (size 250×250 pixels) associated with the target label $y = \{0, 1, 2\}$, corresponding to the child's engagement level (see Section 5 for details). Given these data, we address it as a multi-class image sequence classification problem, where our goal is two-fold: (i) to predict the target label given a window of image frames, and (ii) to actively select the data of each child so that our prediction model can iteratively be personalized to that child as the sessions progress.

3.2. Action Recognition from Video Data

To classify each video clip, we use a Long Short-Term Memory (LSTM) [19] model, which enables long-range learning of time-feature dependencies between image frames. This has shown great success in tasks such as action recognition [10, 1] and speech analysis [16, 12].

Each LSTM cell has hidden states augmented with nonlinear mechanisms that allow the network state to propagate without modification, be updated, or be reset, using simple learned gating functions. More formally, a basic LSTM cell can be described with the following equations:

$$\begin{aligned} &\hat{g}^f, \hat{g}^i, \hat{g}^x, \hat{c}_t = W_x \cdot x_t + W_h \cdot h_{t-1} + b \\ &g^f = \sigma(\hat{g}^f), g^i = \sigma(\hat{g}^i), g^x = \sigma(\hat{g}^x) \\ &c_t = g^f \odot c_{t-1} + g^i \odot \tanh(\hat{c}_t), \quad h_t = g^x \odot \tanh(c_t), \end{aligned} \tag{1}$$

where $\hat{g}^f, \hat{g}^i, \hat{g}^x$ are the forget gates, input gates, and output gates respectively, $\hat{c}_t$ is the candidate cell state, and c_t is the new LSTM cell state. W_x and W_h are the weights mapping from the observation (x_t) and hidden state (h_{t-1}), respectively, to the gates and candidate cell state, and b is the bias vector. $\odot$ represents element-wise multiplication; $\sigma(\cdot)$ and $\tanh(\cdot)$ are the sigmoid and hyperbolic tangent functions respectively [47]. To model the window of T image frames, we adopt an architecture resembling that of the Long-term Recurrent Convolutional Network (LRCN) [10] framework, proposed for fully supervised action recognition. In this approach, each instance of the unrolled LSTM cell receives a temporally ordered image frame from a video clip (where each image is first passed through a CNN network). Then, their output-state values h_t are passed through fcL_t, and averaged across time. Finally, a sigmoid function is applied to obtain the target label y^*, as depicted in Figure 1.

Note that more advanced deep architectures can be used to model video data, as done in fully supervised learning settings (e.g., [5, 46, 8, 10]). Within our approach, these architectures can be considered as more effective feature extractors (Section 4.1). However, the focus of this work is on the learning of the personalized policy for data labeling and adaptation of the engagement classifier in a data efficient manner, and using the notion of deep RL.

3.3. RL for Data-labeling Policy Learning

RL [42] is a framework that can be used to learn an optimal data labeling policy $\pi(s_i)$. Given a video (v_i), the policy takes a state (s_i) and outputs an action (a_i) by maximizing an optimal action-value function $Q^*(s_i, a_i)$. This function is at the heart of RL, and it specifies the expected sum of discounted future rewards for taking action a_i in state s_i and acting optimally from then on:

$$a_i = \pi^*(s_i) = \arg\max_{a_i} Q^*(s_i, a_i); \tag{2}$$

The optimal Q function is given by the Bellman equation:

$$Q^*(s_i, a_i) = \mathbb{E}_{s_{i+1}}[R_i + \gamma \max_{a_{i+1}} Q^*(s_{i+1}, a_{i+1}) | s_i, a_i], \tag{3}$$

where $\mathbb{E}_{s_{i+1}}$ indicates an expected value over the distribution of possible next states s_{i+1}, R_i is the reward at the current video i given state s_i (image features) and action a_i, and γ is a discount factor, which incentivizes the model to seek reward in fewer time steps. Recently, [47] proposed to use this approach for AL from the Omniglot image dataset, where the model's actions a_i are defined as binary states: 1 when the true label y_i^* is requested, and 0 when there is no request, in which case the model makes a prediction y_i for the target image category. We adopt this definition of the action space in our RL model (Section 4.2).

4. Methodology

We propose a deep learning approach for learning an optimal data-labeling policy and engagement estimation from fixed-sized video segments. The proposed deep architecture has two blocks designed for: (i) the extraction of deep features from video frames using pretrained CNNs, and (ii) the learning of the Q-function of the RL model implemented using an LSTM cell and fcL followed by a softmax layer (see Figure 1). The role of the latter is to learn simultaneously the group-policy for data labeling and multi-class engagement classification from input videos of the training children. During inference, for a new child, the group policy is first used to select videos that the model is uncertain about. These are then used to personalize the policy and engagement estimator to that child in an iterative fashion.

4.1. Deep Features

The image frames X from a target input video v[1] are passed individually through a pre-trained CNN network. Specifically, we applied the ResNet-50 [18] architecture, pre-trained on the ImageNet dataset [24]. We used all 50 network layers (convolutional and dense) but the last (i.e., the softmax layer) to obtain the network activations as our deep features, $\hat{X} = \{\hat{x}_1, \ldots, \hat{x}_T\} \in \mathcal{R}^{2048 \times T}$, where 2048 is the size of the output layer (conv_5x). Recent works have showed that these features, capturing edges, corners, shapes and other data representations, work well as general-purpose deep features for image classification tasks [27]. We also applied data augmentation by rotating the images in order to account for different camera-views.

4.2. Group-policy Learning

We start by learning the group-policy π_g for making the decision when to query a video label and when to estimate the engagement level. For this, we use the deep features from the training videos $v = \{\hat{X}, y\}$ to learn the Q function of the RL model (Section 3.3). Specifically, given a video v_i, the active learner may choose an action a_i of either requesting the true label y_i ($r_i = 1$), or providing its estimate y_i^* ($r_i = 0$). If the engagement label is requested, the model receives a negative reward to reflect that obtaining video labels is costly. On the other hand, if the model

[1] For notational simplicity, we drop the dependence on $\{i, j, k\}$

decides to estimate the child's engagement level, the model receives positive reward if the estimation is correct; otherwise, it receives negative reward. This is encoded by the following RL reward function, also used in [47]:

$$R_i = \begin{cases} R_{req}, \text{if } r_i = 1 \\ R_{cor}, \text{if } r_i = 0 \wedge y^* = y \\ R_{inc}, \text{if } r_i = 0 \wedge y^* \neq y \end{cases} \quad (4)$$

This reward drives the learning of the Q-function that we use to learn the target policy. We approximate the Q-function using an LSTM cell, with 128 hidden units, unrolled in time for $t = 1, .., T$. The outputs of these LSTMs simulate the states of the Q function, but at the frame level. The states for the input video are denoted by $H_i = \{h_{i,1}, .., h_{i,T}\} \in \mathcal{R}^{128 \times T}$. Subsequently, each set of states $h_{i,t}$, $t = 1, .., T$ is passed through a linear fcL (128×4) with parameters $\{W_l, b_l\}$, for each image frame mapping the states to a 4-D action space $\hat{a}_{i,t}$ (see below). This is followed by the parameter-free averaging layer:

$$\hat{a}_i = \sum\nolimits_{t=1}^{T} \hat{a}_{i,t} \leftarrow W_l \cdot h_{i,t} + b_l. \quad (5)$$

The discrete action space is then obtained as:

$$a_i = [r_i, \ \vec{y}_i] \leftarrow \text{softmax}(\hat{a}_i), \quad (6)$$

and is a one-hot encoding of the request and engagement labels as $a_i = [ask, low, med, high]$. For instance, if a label is requested, $a_i = [1, 0, 0, 0]$; otherwise, for engagement level $y = high$, we obtain $a_i = [0, 0, 0, 1]$. To leverage the memory-augmentation property of the LSTM models, whenever the label is requested, it is provided in the next iteration of the model learning by augmenting the deep features of the next video (v_{i+1}), being the input to $LSTM_{i+1}$:

$$\begin{array}{l} LSTM^t_{i+1} \leftarrow [\hat{x}^t_{i+1}, \ \vec{y}_i], \text{if } r_i = 1 \\ LSTM^t_{i+1} \leftarrow [\hat{x}^t_{i+1}, \ \vec{0}], \ \text{if } r_i = 0 \end{array}, t = 1, .., T \quad (7)$$

This strategy is commonly used in AOSL methods (e.g., see [2, 39, 47]) designed to adapt faster and more easily to new tasks in an online fashion. Since we deal with a single task (i.e., engagement estimation), in our experiments, this only resulted in more stable and faster model convergence during training. However, it did not affect the model's estimation performance. Furthermore, as described in Section 4.3, during the inference stage, we do not have access to the previous video label, even if there was a request.

Once we defined the action space and the input features, the parameters of the Q-function are optimized by minimizing the Bellman loss on each training video v_i:

$$L^{(i)}_B(\Theta) = [Q_\Theta(s_i, a_i) - (R_i + \gamma \max_{a_{i+1}} Q_\Theta(s_{i+1}, a_{i+1}))]^2, \quad (8)$$

which encourages the model to improve its estimate of the expected reward at each training iteration. We do this over $i = 1, \ldots, N$ video instances from the training children, and over a number of training episodes. The loss minimizaton is performed in an end-to-end fashion[2], where Θ are the parameters of the proposed deep Q-function (see Figure 1). However, this unconstrained parametrization of the action space may lead to (i) slow convergence of the loss function, and more importantly, (ii) a high number of requests, which in practice may easily exceed the available budget for obtaining the data labels. To this end, we constrain the action space by adding a surrogate cross-entropy loss directly on the logits of the action space as follows:

$$L^{(i)}_X = \log\left(soft\max(\hat{a}_i \odot \ [0 \ 1 \ 1 \ 1])\right) \cdot [0 \ \vec{y}_i]^T, \quad (9)$$

where we mask the label request through the vector element-wise multiplication $\odot$. The the newly introduced loss function that optimizes the group-policy π_g is then:

$$\min_\Theta \sum_{i=1..N} L(\Theta) = \min_\Theta \sum_{i=1..N} [L^{(i)}_B(\Theta) + \alpha L^{(i)}_X], \quad (10)$$

where the cross-entropy loss is paramater-free, and α balances the trade-off between the two losses. Note that when α is high, we obtain the model similar to the LRCN [10], i.e., we obtain a supervised model. Conversely, when $\alpha = 0$, our model is a generalization of the active-one-shot RL approach [47] to video data. We name this new model the temporally consistent deep Q-learning (TC-DQL).

4.3. Personalized-policy Learning

The learned group-policy can be applied to videos of previously unseen children. However, this policy may be suboptimal due to the highly diverse styles of engagement expressions across the children, in terms of their facial expressions, head movements, body gestures and positions, among others. These may vary from child to child not only in their appearance (e.g., facial) but also dynamics during the engagement episodes. To account for these individual differences, we personalize the policy to each child. Specifically, for a new child, we assume we have access to multiple interaction sessions over a period of time. Then, we start with the group-level policy to provide initial engagement estimates, but also to select "difficult" videos that need to be expert-labeled and used to personalize the policy to the target child. The main premise here is that with a small number of human-labeled videos, the parameters of the group-level policy can easily be adapt to the target child.

As an example, consider a child interacting with a robot: we estimate the engagement levels from video segments as the interaction proceeds. For the current video segment v_i,

[2] We freeze the parameters of the input CNNs.

the engagement label is obtained as:

$$y_i^* = \arg\max(a_i \odot [0\ 1\ 1\ 1]^T), \quad (11)$$

where for the first session (S_1), we use the TC-DQL model parameters Θ_0 from the group policy π_g to obtain $a_i = [r_i, \vec{y}_i]$. Note that during this inference stage, the model still may request the label; however, since the labels are not available at that point, we introduce a masking layer, as in Eq. 11. This also returns the $\vec{0}$ to the LSTM units, thus, no memory augmentation is enabled. On the other hand, to store the "difficult" videos for expert-labeling after the session is completed, we use the action-request bit obtained before the masking layer, i.e., $r_i^* = a_i(1)$.

The requested videos from an ongoing session (and all previous sessions of that child) are stored in a video pool v_r. Once the session is completed, the stored videos are labeled by an expert, and used to update the model policy for further data requests and engagement classification for the target child. In a general case, for videos from session S_j, this is performed through the following updates of the model parameters:

$$\hat{\Theta}_j \leftarrow \min_{\Theta_{j-1}} \sum_{v_r \in \{S_j\}} [L_B^{(v_r)}(\Theta) + \alpha L_X^{(v_r)}], \quad (12)$$

where the loss function is defined in Eq. 10, and the model parameters are initialized using those from the previous session S_{j-1}, $j = 1, .., M$. Therefore, the new parameters Θ_j are optimized only using the labels for the requested videos. This results in a new personalized policy π_t for target child, which is updated after each session is completed. Currently, for each new child, the learning of the personalized policy starts from the group-policy learned during training.

5. Experiments

5.1. Dataset

We used the child-robot storytelling interaction dataset of 43 children between the ages of 4–6 years recruited from 12 local kindergarten classrooms (55% female, age $\mu = 5.36 \pm 0.62$ years) [34]. In each interaction, the robot and the child took turns telling stories to each other. We used video recordings from the bird's-eye view of the interaction (see Figure 1). On average, each robot story lasted for about 15 minutes, and each child interacted with the robot for 6–8 sessions over three months. From the videos of this dataset, we sampled ∼7.2K 5-second video clips (@30fps) annotated by three expert psychologists in terms of visual cues of engagement levels as low engagement/disengagement ($y = 0$), mid-engagement ($y = 1$), and high engagement ($y = 2$). We averaged these annotations and used them as the ground truth for the engagement estimation task.

5.2. Data Processing and Evaluation Setting

For training/testing, we computed the CNN features for each frame (scaled to 250×250) from target videos, and averaged them within 1 second intervals, with an overlap of 0.5 seconds. This resulted in 10 temporally coherent deep feature vectors (Section 8), which were then fed into the LSTM cell of the TC-DQL model. We report the average F1 score and accuracy (ACC) for the 3-class engagement classification task. For the RL methods that provide a mechanism to request a label, we also report the portion of the requested videos, and precision (PR), defined as the number of correctly classified videos from those that were not requested by the model. To evaluate the models, we split the children into: training (18), validation (4) and test (21).

For the proposed approach, we investigated different network architectures by changing the number hidden states in the LSTM units ($h = 32, 64, 128$ and 256), and $n = 128$ was selected as the best on the validation set. The size of the fcL was set to 128×4, as the size of the action vector a was set to 4. For each training iteration, we used a batch size of 20 episodes, with each episode containing 30 randomly selected videos. For the Bellman loss in Eq. 8, we set the epsilon greedy action to $\epsilon = 0.05$, with discount factor $\gamma = 0.8$, and used the Adadelta optimizer with a learning rate $lr = 0.2$. If not said otherwise, the reward values were set to: $R_{req} = -0.05$, $R_{cor} = 1$ and $R_{inc} = -1$. We evaluated several versions of our model: TC-DQL – trained with cross-entropy loss only (this setting has the same architecture as the LRCN [10] model, previously proposed for supervised activity recognition), TC-DQL that uses the Bellman loss only ($\alpha = 0$, see Eq. 12). For the combined loss (Bellman+xentropy), we used TC-DQL with $\alpha = 3$, as it performed the best on the validation set.

5.3. Compared Methods

We compared our approach with several baselines. As fully supervised models, we used the CNN features and the temporally unrolled LSTM cell (see Figure 1), which output states ($10 \times 128 = 1280$) were then fed to a linear fcL (1280×3) and a softmax layer outputting probabilities for the engagement levels (LSTM-S). We also compared to a traditional deep network that performs a majority voting from the frame-level engagement estimates (DNN-MV). For this, the CNN output for each frame was passed to a linear fcL (2048×128) followed by a soft-max layer for per-frame estimation, and the majority vote to obtain the video label. Since our approach is most similar to the deep RL in [47], we applied a straightforward extension of this model to video data. Namely, to obtain the fixed-size representation of the video, each frame was passed through a fcL (2048×128), resulting in 1280-D feature vectors, as in DNN-MV. These features were used as input to the LSTM cell (we used 200 hidden states as in [47]), followed by an-

other fcL (200×4) and a softmax layer. We denote this model as LSTM-RL. During the evaluation of test videos, we return zero-labels, as in our TC-DQL, when requests are made. This is because, in contrast to [47], in our evaluation setting, the true labels are not available during inference.

5.4. Results

Table 1 shows the performance of different methods on test children. In the case of RL-based models, this correspond to the case when the group-policy, learned from videos of training children, is applied to test children.[3] We note that the static DNN-MV approach fails to reach the performance of its temporal counterparts. We also note that the fully-supervised LSTM-S approach shows slight improvement in terms of F1/ACC over the LSTM-RL approach. However, the latter uses much less video labels during training (16%). On the other hand, the proposed deep architecture (TC-DQL), when evaluated in the fully supervised setting, i.e., using the cross-entropy loss only, outperforms the LSTM-S by 2% in terms of all three performance measures.

When the combined loss (Bellman + cross-entropy) is used in the proposed TC-DQL ($\alpha = 3$), this model outperforms TC-DQL without the cross-entropy regularization. We attribute this to the fact that the action space of the latter is unconstrained, resulting in a less dicriminative model. Also, the lack of the proposed regularization results in a higher number of video requests (19% vs. 4%). This evidences the importance of constraining the action space, when the requests and the labels are modeled jointly in the output of the RL model. By changing the negative reward in the TC-DQL, we noticed a drop in its F1/ACC, even though the number of requests for the labels increased. A similar trend can be observed during the training stage. This is depicted in Figure 2, showing the models' performance on videos of the training children. Overall, we note that among these models, the TC-DQL with the proposed combined loss reaches the highest performance, while minimizing label requests. Note that for the TC-DQL ($\alpha = 3$), we set the request bit in the action space to 0 for the first 1800 iterations, effectively giving a warm start to the model.

We next compare TC-DQL with the LSTM-S model that is personalized to target children using video labels queried using traditional (heuristic) AL strategies. Namely, we evaluate the following AL strategies: the entropy (ENTR), the least confidence (LCONF) and the smallest margin (SMAR). To investigate the models performance under similar conditions, the budget of the heuristic AL strategies is set to the average number of requests by TC-DQL (REQ). Table 2 shows the results of the personalized models on test children and per child-robot interaction session. For the first session, the group policy (INIT) is used (i.e., the non-personalized model). After each session, the queried video labels are used to personalize the policy to the target child, as described in Section 4.3, and such model is evaluated on all video data[4] from the next session of the target child. Likewise, the LSTM-S is personalized to the target child by fine-tuning the model parameters using the queried video labels. To avoid overfitting during model personalization, the parameters of both models were fine-tuned using only 10 iterations of the Adadelta optimizer ($lr = 0.02$). On average, the heuristic AL strategies outperform the RND data selection in LSTM-S in terms of ACC.

[3]Because of the highly imbalanced labels of engagement levels, the reported results are obtained on the balanced set.

[4]No balancing of the labels for the engagement levels is performed during this inference step.

Table 1: Comparison of different models. The performance measures are obtained by running the group policy on the test subjects, and are reported in $\%$.

Model	F1	ACC	PR	REQ
DNN-MV	37	39	39	100
LSTM-S	41	43	43	100
LSTM-RL	40	38	42	16
TC-DQL (xentropy)	43	45	45	100
TC-DQL ($\alpha = 0, R_{inc} = -1$)	38	37	**47**	19
TC-DQL ($\alpha = 3, R_{inc} = -1$)	**45**	44	**47**	**4**
TC-DQL ($\alpha = 3, R_{inc} = -2$)	43	42	48	9
TC-DQL ($\alpha = 3, R_{inc} = -3$)	41	40	47	17

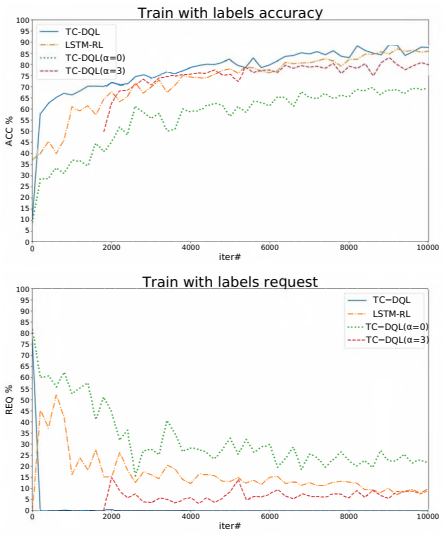

Figure 2: The models' performance on training data during learning stage. As the number of training iterations increases, the ACC of the models increases, while their label requests decrease, as expected. The proposed approach reaches the lowest number of requests for $\alpha = 3$.

Table 2: Comparisons of the proposed TC-DQL approach and LSTM-S, personalized per child and interaction session. We apply the initial model (INIT), i.e., the group policy (Θ_0), to S1, and the personalized policy to consequent sessions S2-S8, where the labels for the queried videos from the previous sessions were used to personalize the models. For both models, we investigate the performance when the random (RND) data selection was used to personalize the policy. For LSTM-S, the AL strategies: entropy (ENTR), the least confidence (LCONF) and the smallest margin (SMAR), were used to query the video labels. For TC-DQL, we report the results of the personalized policy based on the requested data (RL). The last row shows the average percentage of the requested videos (REQ). The numbers in bold depict the best performance per model.

Model		**ACC [%]**									**F1 [%]**								
		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**AVE.**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**AVE.**
LSTM-S	**INIT**	65	59	61	60	62	60	48	67	60	31	40	36	37	38	40	32	35	35
	RND	65	**67**	69	65	67	64	55	68	65	31	**46**	40	35	40	41	**41**	33	38
	ENTR	65	61	65	**70**	73	65	**57**	**75**	66	31	42	**43**	37	38	43	40	35	39
	LCONF	65	66	62	69	72	69	52	**75**	66	31	40	39	**41**	37	42	**41**	35	38
	SMAR	65	66	**70**	66	**75**	**70**	55	73	**68**	31	45	41	36	**48**	**47**	37	**36**	**40**
TC-DQL	**INIT**	72	57	66	61	74	64	56	71	65	32	42	33	40	43	47	35	31	39
	RND	72	67	60	64	72	70	63	81	69	32	38	39	32	**46**	40	41	41	40
	RL	72	**70**	**74**	**70**	**82**	**75**	**65**	**85**	**74**	32	**45**	**42**	**46**	45	**50**	**48**	**46**	**44**
	REQ [%]	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6	2.2	0.6	4.8	6.3	14.9	5.2	7.6	11.5	6.6

On the other hand, only the SMAR approach outperforms RND in the case of LSTM-S and F1 score. From these results, we note that there is no a one heuristic AL strategy that is optimal across all the sessions, i.e., there is not a single strategy that achieves a consistent improvement over the random queries. On the other hand, we note that in the proposed TC-DQL approach, the requests are in most cases more informative than RND queries, leading to consistent improvements over the INIT and RND strategies. Overall, TC-DQL outperforms LSTM-S with the best heuristic AL strategy (SMAR). This is achieved with an average number of requests of 6.6%, thus, only ~ 10 5 sec videos per child.

Figure 3 depicts the improvements in ACC (over the group-policy – TC-DQL-INIT) per target child. As can be noted, in many cases, the model does not request labels. There are also cases where the model fails to capitalize on the updated policy, resulting in lower performance than when the group-policy is used (e.g., for child ID= 14 and 9). This behavior occurs in cases when the model overfits during the policy personalization. For this reason, there are no consistent improvements over the sessions of the same child. This is also because the child's position as well as the engagement behaviour patterns can vary largely from session to session. However, the model achieves large improvements (reaching 100% for the child with ID= 9), on a majority of the children. This clearly shows the benefits of using the personalized policy for video labeling/engagement classification, over the group-level policy.

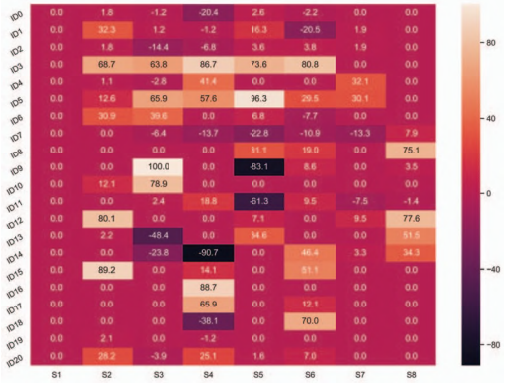

Figure 3: The relative improvement in terms of ACC[%] by the TC-DQL personalized-policy over the group-policy.

6. Conclusions

We introduced a novel approach for personalized estimation of engagement directly from videos of child-robot interactions using the notion of active learning and deep RL. We demonstrated that this approach outperforms related RL and traditional AL strategies under the same budget for requesting video labels. We also showed that the proposed approach provides a data efficient personalized policy that allows the model to easily adapt to a new child, improving the engagement estimation using a small number of video labels provided by the human expert. This combined human-AI approach has potential to improve the robot perception of user engagement, and personalized video analysis.

Acknowledgements

This work was supported by a National Science Foundation grant IIS-1734443 and in part by gift funding from Futurewei Technologies, Inc. The work of O. Rudovic was funded by EC H2020 Marie Curie Action – Individual Fellowship no. 701236 (EngageMe), and the work of B. Schuller by EC H2020, no. 688835 (DE-ENIGMA). We also thank J. L. Xu, D. Veljkovic, B. Milicevic, and J. J. Huang, for their useful insights about the model, and help with data processing and implementation details.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, pages 29–39. Springer, 2011. 3

[2] P. Bachman, A. Sordoni, and A. Trischler. Learning algorithms for active learning. 2017. 3, 5

[3] S. Bandla and K. Grauman. Active learning of an action detector from untrimmed videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1833–1840, 2013. 2

[4] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. Social robots for education: A review. *Science Robotics*, 3(21):eaat5954, 2018. 1

[5] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi. Action recognition with dynamic image networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2799–2813, 2018. 4

[6] D. Bohus and E. Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, pages 2–9. ACM, 2014. 1, 3

[7] F. Caba Heilbron, J.-Y. Lee, H. Jin, and B. Ghanem. What do i annotate next? an empirical study of active learning for action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 199–216, 2018. 2

[8] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 4

[9] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. W. McOwan. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 119–126. ACM, 2009. 3

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 3, 4, 5, 6

[11] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, and P. Abbeel. Rl ˆ2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016. 3

[12] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 483–487, 2013. 3

[13] M. Fang, Y. Li, and T. Cohn. Learning how to active learn: A deep reinforcement learning approach. *arXiv preprint arXiv:1708.02383*, 2017. 3

[14] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt. Interacting in time and space: Investigating human-human and human-robot joint action. In *IEEE RO-MAN*, pages 252–257, 2010. 1

[15] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 2

[16] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772, 2014. 3

[17] M. Hasan and A. K. Roy-Chowdhury. Context aware active learning of activity recognition models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4543–4551, 2015. 2

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[20] W.-N. Hsu and H.-T. Lin. Active learning by learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2659–2665. AAAI Press, 2015. 3

[21] C. Käding, E. Rodner, A. Freytag, and J. Denzler. Active and continuous exploration with deep neural networks and expected model output changes. *arXiv preprint arXiv:1612.06129*, 2016. 3

[22] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015. 3

[23] K. Konyushkova, R. Sznitman, and P. Fua. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4228–4238, 2017. 3

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[25] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 3

[26] C. Lemke, M. Budka, and B. Gabrys. Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1):117–130, 2015. 3

[27] B. Liu, X. Wang, M. Dixit, R. Kwitt, and N. Vasconcelos. Feature space transfer for data augmentation. *CVPR*, 2018. 4

[28] M. Liu, W. Buntine, and G. Haffari. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 334–344, 2018. 2, 3

[29] D. L. Martinez, O. Rudovic, and R. Picard. Personalized automatic estimation of self-reported pain intensity from facial expressions. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2318–2327. IEEE, 2017. 1

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 3

[31] D. C. Mocanu. Synopsis of the phd thesis-network computations in artificial intelligence. In *30th International Teletraffic Congress (ITC30)*, 2018. 3

[32] S. Mohammed, H. W. Park, C. H. Park, Y. Amirat, and B. Argall. Special issue on assistive and rehabilitation robotics. *Autonomous Robots*, 41(3):513–517, 2017. 1

[33] O. M. Nezami, L. Hamey, D. Richards, and M. Dras. Deep learning for domain adaption: Engagement recognition. *arXiv preprint arXiv:1808.02324*, 2018. 3

[34] H. W. Park, I. Grover, S. Spaulding, L. Gomez, and C. Breazeal. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the Thirty Third AAAI Conference on Artificial Intelligence*, AAAI'19. AAAI Press, 2019. 1, 2, 3, 6

[35] 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. 1, 2

[36] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 2018. 1, 2, 3

[37] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. Castello Ferrer, B. Schuller, and R. Picard. Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018. 3

[38] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 305–312. ACM, 2011. 1, 3

[39] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016. 2, 3, 5

[40] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010. 1, 3

[41] 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. 1

[42] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998. 1, 3, 4

[43] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *Advances in Neural Information Processing Systems*, pages 28–36, 2011. 2

[44] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, and M. Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016. 3

[45] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 3

[46] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 4

[47] M. Woodward and C. Finn. Active one-shot learning. *NIPS, Deep Reinforcement Learning Workshop*, 2016. 2, 3, 4, 5, 6, 7