End-to-End Learning and Analysis of Infant Engagement During Guided Play: Prediction and Explainability

Marc Fraile
marc.fraile.fabrega@it.uu.se
Department of Information Technology, Uppsala University
Uppsala, Sweden

Christine Fawcett
christine.fawcett@psychology.su.se
Department of Psychology, Stockholm University
Stockholm, Sweden

Joakim Lindblad
joakim.lindblad@it.uu.se
Department of Information Technology, Uppsala University
Uppsala, Sweden

Nataša Sladoje
natasa.sladoje@it.uu.se
Department of Information Technology, Uppsala University
Uppsala, Sweden

Ginevra Castellano
ginevra.castellano@it.uu.se
Department of Information Technology, Uppsala University
Uppsala, Sweden

ABSTRACT

Infant engagement during guided play is a reliable indicator of early learning outcomes, psychiatric issues and familial wellbeing. An obstacle to using such information in real-world scenarios is the need for a domain expert to assess the data. We show that an end-to-end Deep Learning approach can perform well in automatic infant engagement detection from a single video source, without requiring a clear view of the face or the whole body. To tackle the problem of explainability in learning methods, we evaluate how four common attention mapping techniques can be used to perform subjective evaluation of the network's decision process and identify multimodal cues used by the network to discriminate engagement levels. We further propose a quantitative comparison approach, by collecting a human attention baseline and evaluating its similarity to each technique.

CCS CONCEPTS

• **Human-centered computing**; • **Applied computing** → *Psychology*; • **Computing methodologies** → *Computer vision*;

KEYWORDS

infant engagement; end-to-end learning; explainable artificial intelligence; video analysis; social signals; multimodal cues

ACM Reference Format:
Marc Fraile, Christine Fawcett, Joakim Lindblad, Nataša Sladoje, and Ginevra Castellano. 2022. End-to-End Learning and Analysis of Infant Engagement During Guided Play: Prediction and Explainability. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3536221.3556629

ICMI '22, 7-11 Nov 2022, Bangalore

ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556629

1 INTRODUCTION

Infant engagement during play has been shown to be a reliable indicator of learning outcomes in early childhood, as well as a potential tool to detect psychiatric and familial issues. A few examples: gaze following at 12 months correlates with improved language outcomes at 24 months [30]; increased mutual gaze at 5 months correlates with improved visual attention control at 11 months [31]; reduced engagement with social stimuli at 6 months is associated with autism spectrum disorder diagnosis at 24 months [22]; dyadic measures of engagement around 24 months can be used to identify cases of child neglect [27].

A major issue stopping us from applying these results in real-world scenarios is the data collection bottleneck: obtaining engagement information is a laborious process that requires manual annotation by one or more domain experts. The same problem affects the widely used *Facial Action Coding System* (FACS) [13] for emotion analysis, which typically requires a certified coder to carefully consider video capture of an individual's face to record the activation of Facial Action Units (FAU). In recent years, we have seen the rise of publicly available Deep Learning tools that can reliably estimate FAU activations, given clear video capture of adults [29]. Free access to such tools has allowed researchers to build real-time automatic affect analysis solutions where previously only manual annotation used to be possible (e.g., student engagement detection [38, 42, 44]). This begs the question: can we automate infant engagement analysis in the same way? To answer it, we collected a dataset on infant engagement during guided play, and trained an end-to-end video classifier to separate positive and negative samples.

While recent reviews indicate that Deep Learning based methods are becoming increasingly popular in the wider field of *automatic affect recognition* [33], and end-to-end learning methods have attained promising results [40], the number of studies using end-to-end Deep Learning is still small. To the best of our knowledge, the only previous study on end-to-end Deep Learning for infant engagement is our earlier pilot experiment [15]. In the pilot, we successfully trained task engagement classifiers on video data with coarse annotations, and used attention mapping techniques for

subjective network analysis, but did not analyse social engagement nor compare attention maps quantitatively.

Using an end-to-end Deep Learning approach has clear advantages: we can obtain good performance without explicit modelling of the interaction, and we are not limited by feature extraction methods. As an example, the mentioned FAU extraction tools require a clear view of the face, while our method only requires a general side view of the scene. However, a major drawback is *explainability*: it is typically hard to explain the decision-making process of a Deep Learning model, even if we have full access to its internals. This makes it possible for the system to inadvertently depend on unexpected or unwanted correlations. Lapuschkin et al. [26] give a striking example: An image classifier for the PASCAL VOC 2007 dataset classifies horse images correctly if they have a copyright notice – disproportionately common in this category – and fails if the notice is removed. If the same copyright notice is then added to a car picture, the model confuses it with a horse.

To improve model explainability, a host of techniques has been developed. They are collectively known as *Explainable Artificial Intelligence* (XAI), and have been the focus of much research in recent years [1]. In this study, we focused on *local explanations*: given an input sample and an output decision, these methods produce a simplified explanation of the network's decision process. In the context of computer vision, an important family of local explanation methods are *attention maps*. Given the model's prediction, they assign an importance score to each pixel in the input. Many methods have been proposed, with varying computational costs and theoretical justifications. Some well-known examples are Guided Backpropagation [37], Grad-CAM [35], and LRP [3].

Even though these tools have been successfully used for manual exploration of the network on a sample-per-sample basis, there is no consensus on which attention maps are most useful to a human observer, or how to aggregate them over a whole dataset. We addressed these issues by comparing four common techniques, both subjectively – by analysing the known advantages and shortcomings of each method, and discussing examples of network insights revealed by specific attention maps –, and objectively – by collecting hand-authored *human attention maps*, and computing a similarity measure proposed in the *visual saliency prediction* literature [6].

The following is a summary of our contributions:

(1) We show that an end-to-end Deep Learning model can successfully predict task engagement and social engagement of an infant participating in guided play. We do this from a single video feed showing a lateral view of the whole scene, without dedicated facial or postural capture.
(2) We use four machine attention mapping techniques on a selected subset of the samples, and showcase their use in subjective analysis of the network's decision process, highlighting head, body and contextual cues identified by the network as important to discriminate engagement levels.
(3) We collect human attention maps as a ground-truth, and use established similarity metrics to evaluate machine attention mapping methods quantitatively.

2 RELATED WORK

2.1 Automatic Infant Engagement Recognition

In the wider context of *automatic affect recognition*, it is common to estimate facial and postural features, and use those as inputs for a classification algorithm. In particular, two open-source feature extractors have been widely used: OpenFace [4] takes clear facial images, and estimates several facial features (most notably FAU); OpenPose [8] takes unobstructed body images, and estimates an individual's posture. Both tools are Deep Learning models trained on video capture of adults, so caution should be taken when breaking any of their assumptions – that the subject is an adult, or that we have a clear view of the subject.

Narrowing the scope, *automatic engagement recognition* has received a lot of interest in the educational context, due to the body of research suggesting ties to academic performance [16]. Recent editions of the EmotiW challenge [12] have contained a category for engagement recognition in Massive Online Open Course (MOOC) education. The dataset contains clear facial video capture of adult students. All the published participants in the latest (2020) engagement sub-challenge [38, 42, 44] use OpenFace, OpenPose, and pre-trained video networks as feature extractors, and train their own classification algorithm on the obtained features. Automatic engagement detection in school-age children has also been investigated in connection to Autism Spectrum Disorder (ASD). Javed et al. [21] use OpenPose to extract postural and facial data, compute custom features from it, and train a 5-layer convolutional network.

This reliance on a small set of open-source tools comes with limitations. OpenFace has been shown to perform well with facial capture of children as young as 5 [2], but requires a clear frontal view, and does not generalize to infants [18]. Similarly, OpenPose needs retraining for infants [9], and requires a clear view of the body (in our experience, it fails on hard-side views). In contrast, an end-to-end model can be trained on any form of video capture, as long as it contains enough information. We show that a general overview of the interaction from the side is enough to train an engagement classifier. We do so by fine-tuning a pre-trained convolutional network on a small amount of data.

2.2 Machine Attention

Consider an input color image I_{ijc}, where (i, j) are the pixel coordinates (row and column), and c is the color channel[1]. In the context of explainability for Computer Vision classifier models, an *attention map* is a real-valued grid that assigns a relevance score h_{ij}^k to each pixel location (i, j), given a target class k. Let S_k be the model's score for k, and $\mathbb{P}(k) = [\text{softmax}(\mathbf{S})]_k$ its estimated probability. Relevance can be interpreted in many ways. For example, Zeiler et al. [43] place a square occluder at each pixel location and record how $\mathbb{P}(k)$ *decreases* to determine how important an image patch is for the final decision: $h_{ij}^k = 1 - \mathbb{P}\left[k \mid \text{occluder at } (i, j)\right]$. A related concept are *learned attention* mechanisms, in which a similar score is calculated internally by the network and is used to filter out information during inference. The methods we study here have

[1]The descriptions in this section are given in terms of image inputs. The same definitions apply to video by adding a frame index t: I_{tijc}, etc.

been classified as *post-hoc* attention maps [24] to distinguish them from learned attention.

To our surprise, we found very few examples of post-hoc attention maps in the affect recognition literature. Gera et al. [17] train an image classifier to take facial capture as input, and predict one of 8 emotions as output. They use Grad-CAM to perform subjective evaluation of the network's decision process. Prajod et al. [32] train two image classifiers to take facial capture as input, and predict if the subject is in pain as output. They use LRP to perform subjective comparison of the two networks. Based on this, they hypothesise that one network focuses on closed eyes, while the other focuses on visible teeth. They then verify this hypothesis by annotating the dataset and training a linear classifier on the output of each network's last pooling layer, showing that each network's embedding is significantly better for predicting the corresponding facial expression. To the best of our knowledge, the only earlier paper in affect recognition comparing various attention mapping methods is our pilot study [15]. We trained a task engagement classifier on coarse labels, and used several attention mapping techniques to perform subjective evaluation, albeit with a focus on the differences between mapping techniques.

In this paper, we focus on the same four post-hoc techniques covered in the pilot: *gradient saliency*, *guided backpropagation*, *Grad-CAM*, and *guided Grad-CAM*. These were chosen because they are popular, computationally efficient, and relatively easy to implement. Similar to [17, 32], we first perform subjective evaluation, with a focus on comparing the unique characteristics of each method. We further collect a ground truth, and use it to perform quantitative comparison. What follows is a brief description of the four chosen methods.

Gradient saliency was introduced by Simonyan et al. [36] as a seeding tool for object segmentation, and is the simplest method. To compute it, (1) calculate the gradient of the target class score with respect to the input pixels $g^k_{ijc} = \partial S_k / \partial I_{ijc}$ (the *input gradient*); (2) take the absolute value, and take the maximum in the color channel dimension $h^k_{ij} = \max_c |g^k_{ijc}|$ (the gradient *saliency*). This is very simple to calculate with modern machine learning libraries, and has proved useful in its original context. However, it suffers from two issues: (i) it is not *class discriminative*, i.e., it does not change meaningfully based on which class k is targeted; and (ii) it is susceptible to high-frequency noise.

Guided backpropagation [37] is an early attempt to address these issues in networks that use ReLU activation units. It uses a modified backpropagation algorithm: augment ReLU layers when calculating their gradient, discarding negative gradients to focus on *positive evidence* only. The algorithm returns a modified gradient $\hat{g}^k_{ijc}$, and the attention map is again calculated by taking the saliency: $h^k_{ij} = \max_c |\hat{g}^k_{ijc}|$. This results in sparse attention maps that focus meaningfully on regions of interest, but still suffer from a lack of class discrimination.

Grad-CAM [35] is a low-resolution activation mapping method for convolutional networks, specifically designed to be class discriminative. An intermediate representation is selected (typically the last convolutional layer with spatial dimensions), and both its activations A_{ijc} and gradients $g^k_{ijc} = \partial S_k / \partial A_{ijc}$ are calculated. A weight is calculated per channel by taking the gradient mean: $w^k_c = \frac{1}{N} \sum_{i,j} g^k_{ijc}$ (where N is the number of pixels in the targeted layer). The channels are then averaged using the weights: $\hat{A}^k_{ij} = \sum_c w^k_c A_{ijc}$. Finally, only the positive evidence is kept: $h^k_{ij} = ReLU(\hat{A}^k_{ij})$. This method gives meaningfully different results when queried about different target classes k, but can be very coarse: in our networks, the last convolutional layer is only 10×10 px.

Guided Grad-CAM [35] was proposed in the same paper as Grad-CAM, and attempts to solve its low-resolution issue by a simple procedure: (1) upscale the Grad-CAM map to the same dimensions as the input, and (2) multiply the Grad-CAM map with the guided backpropagation map. This marries the benefits of clean locality (guided backpropagation) and class sensitivity (Grad-CAM), but since we are multiplying the maps, it runs the risk of being very sparse.

2.3 Comparison with Human Attention

Given the wide array of post-hoc attention mapping techniques available, we would like to have a quantitative *measure of fitness* available, so we can confidently choose the most adequate method to evaluate the network's decision process. Some attempts have been made in the context of image classification. Fong et al. [14] propose an *object segmentation* approach: they check if the brightest pixel in an attention map is contained within the object of interest, as judged by a ground-truth image mask. This, however, does not translate well to our target domain: we are interested in detecting a social construct, rather than an on-screen object. We propose evaluating the network's *human-likeness* instead, by capturing a *human attention* ground-truth and measuring the similarity between human and machine maps.

Human and machine attention maps have been compared in a different context: *visual saliency prediction*. That is, the model's explicit goal is to estimate the amount of time a human observer will spend looking at each part of the image. In this case, annotations are typically captured using eye-tracking technology, which produces a time-series of *focus points*. Models generate a continuous map that aims to separate areas of low and high interest [5]. Evaluation metrics either compare the focus points directly to the output distribution, or generate a distribution based on the focus points and use measure-theoretic comparison tools [6]. While older reviews report classical models and smaller datasets, newer reviews show an increase in dataset size and a move towards convolutional networks for greater performance [7].

Since eye tracking is costly to capture and requires participants to visit the research facilities in person, some authors have focused on mouse-capture based methods to study human attention. This allows for crowd-sourcing the data collection, thus obtaining much larger datasets. Das et al. [11] used Amazon Mechanical Turk to annotate 60 thousand images from the Visual Question-Answering dataset, using a custom annotation tool. Users were presented with a blurred image and a question they had to answer, and could use their mouse to remove the blur from parts of the image. The resulting blur removal mask was used as a human attention distribution. The human attention maps were then compared to a learned attention map by downsampling to a low resolution (14×14px), and using Spearman's rank correlation as a similarity measure.

In this study, we follow the more practical mouse-data approach. We develop an in-house tool to view and label video snippets, and paint over them using the mouse. We rely on a distribution similarity measure to rank attention maps, according to their human-likeness.

3 METHOD

3.1 Data Collection

We recruited 23 infants aged 14 months (11 girls; mean age 14 months and 6 days) from a local list of families who were interested in participating in research with their child. Before the study, the parents were informed about the procedure and signed a consent form. The experiment was approved by the university's ethical committee. Each child participated in three guided play tasks with an adult experimenter, recorded in a single session. During the session, the infant was seated in a high chair at a table, facing the experimenter. One parent was seated behind the child. A Sony Handycam HDR-CX260 video camera (1440 × 1080px, 25fps) recorded the interaction, providing a profile view of the participants.

In the first task (*dolls*), four round boxes were attached to the table. The boxes directly in front of the infant and the experimenter (yellow) contained 10 wooden dolls each. The boxes to the sides of the child (one red, one blue) were empty. The experimenter began by placing a doll in one of the empty boxes. She then invited the infant to join and removed the cover from the infant's doll box. The experimenter placed half of her dolls into one of the boxes one at a time, and then switched to placing them in the other box. The task ended when all the dolls were placed, or at the experimenter's discretion if the child was not participating.

In the second task (*shaker*), the experimenter showed the infant an egg-shaped shaker (musical instrument) and began to shake it at a predetermined tempo (150bpm or 170bpm) for 10 seconds. She then gave the infant an identical shaker, and encouraged joint play for 30 seconds. The experimenter then pretended to drop her instrument on the ground, and changed to the other tempo (170bmp or 150bpm). The task ended after another 30s of joint play.

In the third task (*drum*), the experimenter showed the infant a toy drum and used a drumstick to play at one of the predetermined tempos, as in the previous task. She then gave the infant their own drumstick and encouraged them to join in drumming. After 30 seconds of play, she flipped the drum over and switched to the other tempo. The task ended after another 30s of play.

The collection process resulted in 23 videos (one per child), with a duration of 10min 49s ± 1min 58s, and a total length of 4h 9min. This includes time before, between and after the tasks, as well as a fourth free-play segment not used in this study. Hence, the total time used for this experiment was a fraction of the numbers reported here.

3.2 Engagement Annotation

The annotation process was realised using the ELAN annotation software [41]. ELAN allowed the coders to create separate tracks for each variable, and delineate labeled time spans in each track. Three variables were annotated. The first variable was used to determine the time span corresponding to each of the three guided play tasks. As such, it contained three time-spans, respectively *dolls*, *shaker* and *drum*. The other two variables were binary (*positive condition* along the duration of the designated time-spans, *negative condition* outside the time-spans) and coded the infant's behavior: *task engagement* and *social engagement*. For the context of this study, *task engagement* was defined as playfully interacting with the object of interest, and *social engagement* was defined as visibly paying attention to the experimenter and the intended game. A coding guide was created ahead-of-time to guide the annotation process, and was refined based on annotator feedback.

All 23 collected sessions were annotated. Three coders participated in the process, each one annotating a subset of the data. Five sessions were first used as a pilot study to refine the coding rules, and discuss differences in methodology. For these sessions, data from all three annotators is available. The remaining 18 sessions were randomly divided into three *session blocks* of 6 sessions each. Each coder was assigned a primary block to annotate, and upon completion, rotated to the next block. Thus, each block was annotated by two coders, and for each coder there was one *hold-out block* they had not seen.

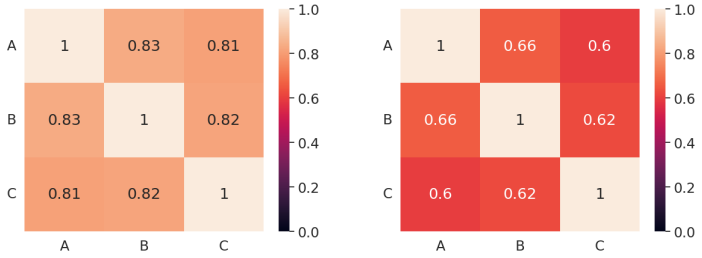
Table 1 summarizes the duration statistics for each task, sampled over all sessions and annotators. Task annotations had a mean duration of 2min 6s, with standard deviation 29s. Statistics requiring common start and end times were computed over the *minimal interval*: the intersection of all available annotations for that task and session. To verify the reliability of this reference interval, Table 1 shows the Intersection Over Union (IOU): the duration of the minimal interval, divided by the duration of the *maximal interval* (union over all available annotations). With an average of 94%, we can conclude there was high overlap among annotators. The minimal interval allows us to estimate the available duration of recorded material for training and analysis: an average of 45min 58s per task, with a total of 2h 17min.

Inter-rater agreement was measured for each annotator pair. Figure 1 shows the agreement scores (empirical probability of agreement) (1a) and Cohen's Kappa [10] scores (1b). To calculate this, all the sessions annotated by both raters were considered, and both engagement variables were used. For each task, the minimal interval was sampled every 0.1s, resulting in a time series per rater. These time series were then compared using the Python library `statsmodels` [34]. The average Cohen's Kappa over all rater pairs is 0.63, a "substantial" agreement [25]. Note that this is effectively averaged over *social engagement* and *task engagement*. While 0.63 agreement would be considered low in some easier settings, Lemaignan et al. [28] point out that annotating engagement in children is a particularly difficult task, and consider their scores of 0.52 (task engagement) and 0.46 (social engagement) satisfactory in this context. They further point out that ML models can reflect this uncertainty in their output probability distribution, if trained with all the available data. Similarly, Henderson et al. [20] consider a kappa above 0.6 to be satisfactory when annotating engagement in students.

Consistent with Lemaignan's observation, no aggregated ground-truth was created for any of the provided variables. Instead, we decided to capitalize on the plurality of opinions between annotators by devising a random sampling strategy, described in Section 3.3.

	Dolls	Shaker	Drum	Total
Individual Durations	2min 15s ± 46s	1min 58s ± 12s	2min 4s ± 8s	2min 6s ± 29s
Mean Intersection Over Union	93%	94%	95%	94%
Total Duration (Intersection)	47min 40s	43min 46s	46min 27s	2h 17min 53s

Table 1: Duration statistics for each guided play task, over all available annotations. Each coder indicated their own task time-spans. *Individual Durations* shows the mean and standard deviation over all durations, considering each annotation separately. *Mean Intersection Over Union* is calculated per-task as (intersection of available annotations) / (union of available annotations), and averaged. *Total Duration (Intersection)* shows the sum of intersection lengths, computed over all sessions.



(a) Agreement Score **(b) Cohen's Kappa**

Figure 1: Dyadic inter-rater agreement measures calculated on the full annotation set. Coders A, B and C are compared to each other. Each rater's interval annotations were sampled every 0.1s, and the resulting slices compared. 1a shows the agreement score (empirical probability of agreement); 1b shows Cohen's Kappa.

3.3 Classification Algorithm

We trained a separate classifier for each combination of task (dolls, shaker, drum) and variable (task engagement, social engagement). All classifiers shared the same architecture: the *Mixed Convolutions* network `mc3_18` from the `torchvision` package [39]. This was chosen as a compromise between computational cost and reported performance, and because it comes pre-trained on the Kinetics-400 dataset [23].

The Kinetics-400 dataset is a collection of YouTube videos, with 400 classes and at least 400 videos per class. It is a common baseline for pre-training and for reporting performance of video classifier networks [45]. It contains 306,245 clips, each approximately 10 seconds long, for a total of ~ 850 hours. Different clips have different resolutions.

The network `mc3_18` (11.7M parameters) is a variant of the 5-block, 18-layer ResNet architecture [19] with 3D convolutions in blocks 1 and 2, and 2D convolutions in blocks 3 to 5. It can be viewed as a convolutional network that produces a 512-dimensional embedding (the *encoder*), followed by a logistic classifier (the *head*). It expects an input spatial resolution of 112×112 px (or higher).

To ensure proper stratification, we partitioned the data ahead-of-time into five disjoint subsets (folds). Since we expected the data in each session (i.e., data corresponding to the same child) to be highly correlated, we split per session, ensuring that each session video was only used in one of the folds. We used rejection sampling to ensure that all empirical probabilities (per task and variable) in each fold were as close to the whole-dataset values as possible. The last fold was reserved as a test set. The other 4 folds were either used for 4-fold cross-validation (in hyper-parameter searches) or as last-out validation (for the final training). All session recordings were downsampled ahead-of-time to 208×160 px and 3.125fps (1/8th the original framerate).

When considering what is a sample in our dataset, we identified two problems to overcome. First, the available data was very small when compared to typical computer vision datasets (e.g., the total time is over 1,000 times shorter than Kinetics-400). Second, for many sessions we had two (possibly disagreeing) annotations as our "ground truth". A standard approach would have involved dividing the relevant parts of each session into non-overlapping samples, and somehow synthesizing a reference label. However, we identified an option to allow the network to learn from the continuous annotation format, and from the individual opinion of each annotator. Each time a snippet was sampled from a session, an available coder was selected at random. We used that coder's annotations for the relevant task to choose a random offset into the video, and extract a 5-second snippet (15 frames). For each epoch, we sampled each available session 10 times, adjusting the offset range to minimize overlap between consecutive samples.

Upon loading, pixel values were normalized using each fold's mean and standard deviation. At training time, the data augmentation pipeline optionally rotated the image ($\pm 8°$, $p = 0.35$), chose a random 112×112 px crop with scaling and stretching, optionally flipped the image horizontally ($p = 0.5$), optionally adjusted the contrast ($p = 0.35$) and color balance ($p = 0.35$), optionally used a Gaussian blur ($\sigma = 5$px, $p = 0.35$), and optionally added Gaussian white noise ($\sigma \in (0.01, 0.03)$, $p = 0.35$). At testing time, the data augmentation pipeline took a 160×160 px center crop and optionally flipped the image horizontally ($p = 0.5$).

To adapt the pre-trained network to our task, the original multi-class logistic classifier was substituted with a randomly-initialized binary logistic classifier. Reflecting this, the training process was

split in two parts: a *head training* phase to train only the new classifier head, and a *fine-tuning* phase to train the whole network at a lower learning rate. To accelerate the head training phase, we encoded samples and saved them to disk ahead-of-time, by collecting augmented samples for 100 epochs and feeding them to the encoder.

A hyper-parameter search was conducted for each training phase (head training and fine-tuning). The following parameters were explored: learning rate, learning rate decay, L2 penalty, class weights (flat weights vs. linear weights). Each parameter combination was chosen by random sampling, and tested using 4-fold validation with repeated runs, to account for randomness in the initialization procedure. The parameters which produced the best F1 validation score (averaged over folds and repetitions) were then used to train the networks. These values were adjusted by hand if a re-run was considered necessary, based on the observed behavior of the network on the train and validation sets.

3.4 Human Attention Annotation

When running the stratification process, we verified that the test fold contained at least one video for each one of the three session blocks described in Section 3.2. Thus, for each coder there was at least one test video they had not seen before. We chose one such video per coder, and used rejection sampling to find non-overlapping samples, such that the two available annotations agreed on that sample (that is, there was a consensus label). If possible, we retrieved 12 such samples: one for each combination of task, variable, and label value. Due to some children showing very consistent behaviors, this was not always possible, so the final number of samples ranged from 9 to 12 per annotator, for a total of 31 samples. We call this collection of samples the *comparison set*.

A custom tool was used to visualize the comparison set, and collect human attention maps. The annotator could view the low-resolution, low-framerate snippet that would have been presented to the network, and could press buttons to decide if the child was engaged or not. They could then navigate frame by frame, and use the mouse to paint over the video. They were instructed to highlight the parts of the video that supported their final decision, focusing on *positive evidence* for their label choice (i.e., the annotations should be *class discriminative*). Figure 2 shows a screenshot of the tool in use.

3.5 Machine Attention Computation

We implemented each of the four studied post-hoc attention mapping techniques (Gradient Saliency, Guided Backpropagation, Grad-CAM, Guided Grad-CAM), and applied them to each sample in the *comparison set*. Note that, while the explanation in Section 2.2 is done in terms of still images with 2D pixel coordinates (i, j) – consistent with the original sources –, all these techniques apply naturally to video data with 3D frame-and-pixel coordinates (t, i, j), using all the available information. In particular, the relevance of motion should be captured by the attention maps.

Since samples in the comparison set correspond to different tasks and variables, these maps were computed using all 6 classification networks obtained during the training process (Section 3.3): for each sample, the corresponding network (same task and variable)

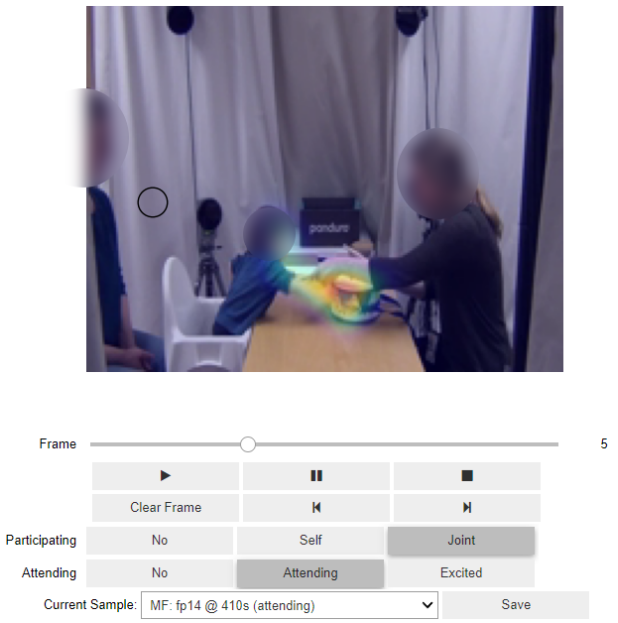


Figure 2: Custom annotation tool used to capture human attention maps. The user can view the video in the same format provided to the network, and can choose an appropriate label. They can then move frame by frame and use the mouse to paint or erase regions of high interest. The black circle is located at the mouse cursor location and indicates the current painting size. The faces of the participants have been anonymized for this illustration.

was used. When defining attention maps, we explained they can target any class in the output. To study *class sensitivity* accross mapping methods, we targeted both *positive* and *negative* labels in each case, resulting in 8 attention maps per sample. Both targets were used for subjective analysis of the network's decision process.

3.6 Attention Comparison

Due to the differences in high-frequency detail between different attention maps, a pixel-for-pixel comparison was discarded, focusing instead on the low-frequency content of the distributions. To that effect, we chose to resize all attention maps to one common spatial size: 32×32px (1/5th of the input resolution). For full-resolution maps, which typically contained more fine-grained detail, downsampling was performed by first applying Gaussian blur ($\sigma = 5$px) and then sampling with stride 5px. For Grad-CAM (original size 10×10px), bicubic upsampling was performed.

Once the size was standardized, the average Earth Mover's Distance per frame (EMD) [6] was used to evaluate the similarity between the human annotation and each machine attention map. The Earth Mover's Distance, also known as the 1st Wasserstein Distance, is a metric on the space of probability distributions. Informally, it can be described as the minimum amount of work that it would take to reshape one distribution into another, if they were piles of earth. We chose to calculate the metric per frame for two reasons: (1) some attention maps have very unequal values between

frames, while the human annotation is near-constant in maximum value and total mass per frame; (2) EMD's computational cost scales very poorly with the size of the map, making it infeasible to run on the whole video.

We calculated the EMD scores in every *full-agreement* sample in the comparison set, that is, every sample that was identically classified by all annotators, and by the network. We did so while targeting the *matching category* (i.e., the true label). This was done to ensure that the hand-painted maps were focusing on the same information as the machine attention maps.

4 RESULTS AND ANALYSIS

4.1 Network Performance

Table 2 summarizes the performance statistics for the top network in each category, as judged by validation F1 score. We have listed the empirical probability q for comparison (calculated for the positive class, over the continuous-time annotations for the whole dataset). As a baseline, the best accuracy achievable by a random classifier is $\max(q, 1-q)$ (achieved by always predicting the most likely class), and the best F1 achievable by a random classifier is $2q/(1+q)$ (achieved by always predicting the positive class). All values above the baseline are marked in black. We can see the *dolls* networks struggle in the test set (despite performing well in the validation set), while all other networks perform successfully. It is worth mentioning that agreement scores between annotators are in the 0.81-0.83 range (see Fig. 1a), so one should not expect accuracies above those values.

4.2 Agreement with the Network

Figure 3 shows aggregate agreement scores (3a) and Cohen's Kappa (3b) between each human annotator and the best performing networks. Unlike Figure 1, which considered all available annotations, and used their original time-interval form, this table uses the test set exclusively, and is calculated by taking video snippets (the samples as they are fed to the network for training and inference). Each session in the test set is sampled for every combination of task and variable, taking as many non-overlapping samples as possible. Labels are calculated for the available annotators with the same approach used for training. The human-human pairs show similar Kappa scores to Figure 1, albeit with a wider range: 0.55-0.71 (Moderate to Substantial agreement). The human-machine pairs are clearly lower: 0.29-0.31 (Fair agreement). Overall, the networks performed clearly better than random choice, but not well enough to substitute a human annotator.

Table 3 shows Cohen's Kappa for each network, averaged over all human-human pairs (column "Human"), and averaged over all human-machine pairs (column "Machine"). We can see *(dolls, social engagement)* and *(drums, social enagement)* failed to obtain better-than-random scores, while other networks performed better (in some cases, close to human levels of agreement).

4.3 Subjective Analysis of Machine Attention

Figure 4 shows all four attention mapping techniques acting on the same frame. The sample under consideration belongs to the task *shaker* and the target variable is *social engagement*. All annotators agreed that this is a negative sample: the child is *not engaged* (negative). The relevant network also classified the sample as negative. In this figure, the attention methods are targeting the *matching class*: *negative* or *not engaged*. We can observe the properties we discussed in Section 2.2: Gradient Saliency is noisy but somewhat informative, focusing on the infant's head; Guided Backpropagation is sparse and focuses on the two participants, with special attention to the shaker (and possibly similar oval shapes, like the infant's ear); Grad-CAM is low-resolution but focuses more clearly on the child; Guided Grad-CAM is the sparsest and potentially most informative – in this case, focusing on the infant's head alone.

Figure 5 shows the same frame and the same methods yet again. However, in this case the target is the *opposite* class: *positive* or *engaged*. As discussed in Section 2.2, Gradient Saliency and Guided Backpropagation are not *class sensitive*: there is no discernible difference between the matching and opposite targets. However, Grad-CAM and Guided Grad-CAM shift the attention from the infant to the experimenter. Assuming we can trust the explanation, and considering that the prediction target is the infant's emotional state, focusing on the experimenter could be considered *contextual* information. Depending on our goals, this could be seen as a failure of the network. However, we will see that human annotators similarly rely on contextual cues when annotating – in their case, the cross-relation between social and task engagement.

4.4 Subjective Analysis of Human Attention

Figure 6 shows four consecutive frames from a manually annotated attention map for the task *dolls* and the target variable *social engagement*. Using the custom tool described in Section 3.4, annotator A classified the sample as positive, and painted areas of interest in every frame to support their decision. We can see that they considered the child's gaze and arm movement to be important factors. Unlike the network example in Section 4.3, the annotator did not highlight the experimenter as relevant contextual information. However, the focus on the arms was added because they indicate *task engagement*. Given that this was not the target variable, the arm annotation constitutes contextual information: the child is seen to engage with the experimenter even if they temporarily break eye contact, because they are still actively participating in the activity. In this case, the judgement of one variable affects the judgment of the other.

Compared to the machine attention maps displayed in Figures 4 and 5, the mouse-painted maps have a distinctive lack of detail, without reaching the coarseness of Grad-CAM. The observed disparity between different attention maps motivates our choice to blur and down-sample before performing a quantitative comparison.

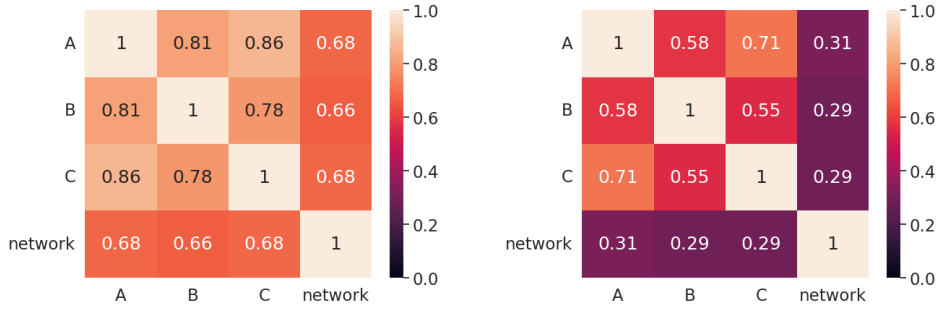
4.5 Quantitative Comparison of Attention Maps

Figure 7 shows the original frame, two machine attention maps, and human attention, before and after being resampled to a common resolution (see Section 3.6). Visual inspection suggests that we have successfully mapped heterogeneous methods to a similar detail scale, while preserving each method's identity.

Table 4 shows the EMD mean and standard deviation for each machine attention method, ranked by mean EMD (lower is better). Matching human expectations, we see that both *locality* and *class sensitivity* are graded positively, with Guided Grad-CAM selected

Task	Variable	Empirical Prob.	Val. Acc.	Val. F1	Test Acc.	Test F1
Dolls	Task Engagement	0.60	**0.93**	**0.94**	**0.64**	0.64
	Social Engagement	0.72	**0.85**	**0.92**	0.68	0.81
Shaker	Task Engagement	0.41	**0.68**	**0.72**	**0.72**	**0.61**
	Social Engagement	0.56	**0.93**	**0.94**	**0.74**	**0.80**
Drum	Task Engagement	0.53	**0.80**	**0.82**	**0.88**	**0.90**
	Social Engagement	0.70	**0.85**	**0.92**	**0.76**	**0.86**

Table 2: Statistics for the best network in each category (as judged by validation F1 score): empirical probability of the positive class (calculated over the whole dataset), validation accuracy, validation F1 score, test accuracy, and test F1 score. Bold numbers indicate scores above the theoretical maximal score of a random classifier. Since there is no unified ground truth, these numbers cannot be expected to reach 100%.



(a) Agreement Score **(b) Cohen's Kappa**

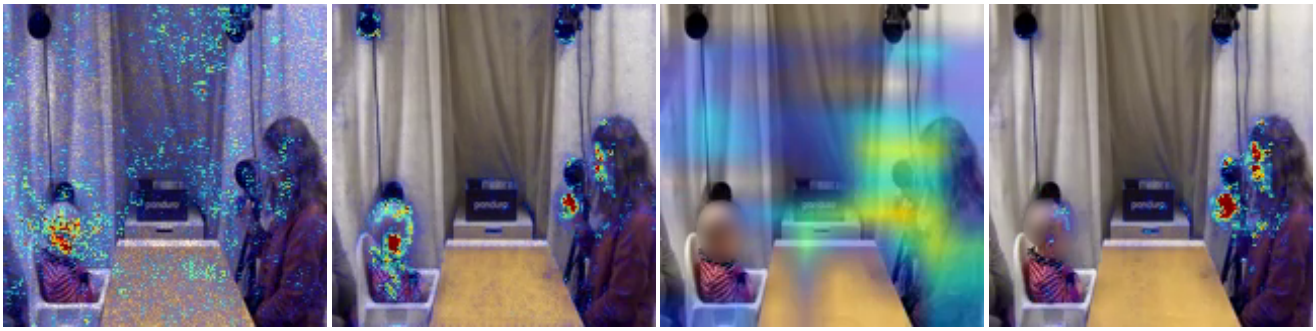
Figure 3: Dyadic inter-rater agreement measures calculated on the test set, and including the network as one of the raters. Human annotators A, B and C are compared to each other and to the network. 3a shows the agreement score (empirical probability of agreement); 3b shows Cohen's Kappa.

Task	Variable	Human	Machine
Dolls	Task Engagement	0.497	0.395
	Social Engagement	0.790	0.000
Shaker	Task Engagement	0.734	0.176
	Social Engagement	0.524	0.219
Drum	Task Engagement	0.529	0.422
	Social Engagement	0.382	0.000

Table 3: Average Cohen's Kappa over all human-human pairs (column "Human") and over all human-machine pairs (column "Machine"), separated by (task, variable) pair (i.e., per network). Some networks failed, while others performed at almost-human level.

(a) Gradient Saliency **(b) Guided Backpropagation** **(c) Grad-CAM** **(d) Guided Grad-CAM**

Figure 4: All four attention mapping methods, displayed at the same frame. Task *shaker*, target variable *social engagement*. All annotators marked this sample as *not engaged*. The network correctly classified the sample as negative. The target class for these maps is the *matching class* (negative). Faces anonymized for the illustration.



(a) Gradient Saliency (b) Guided Backpropagation (c) Grad-CAM (d) Guided Grad-CAM

Figure 5: Same frame and same mapping methods as in Figure 4. In this case, the target class is the *opposite class* (positive). Faces anonymized for the illustration. There is no discernible difference in Gradient Saliency nor Guided Backpropagation, but we can see that the focus changes to the experimenter in the *class-sensitive* methods Grad-CAM and Guided Grad-CAM.

Figure 6: Four consecutive frames from a manually annotated attention map (task: *dolls*, variable: *social engagement*). All annotators labeled the sample as positive. Faces anonymized for the illustration.

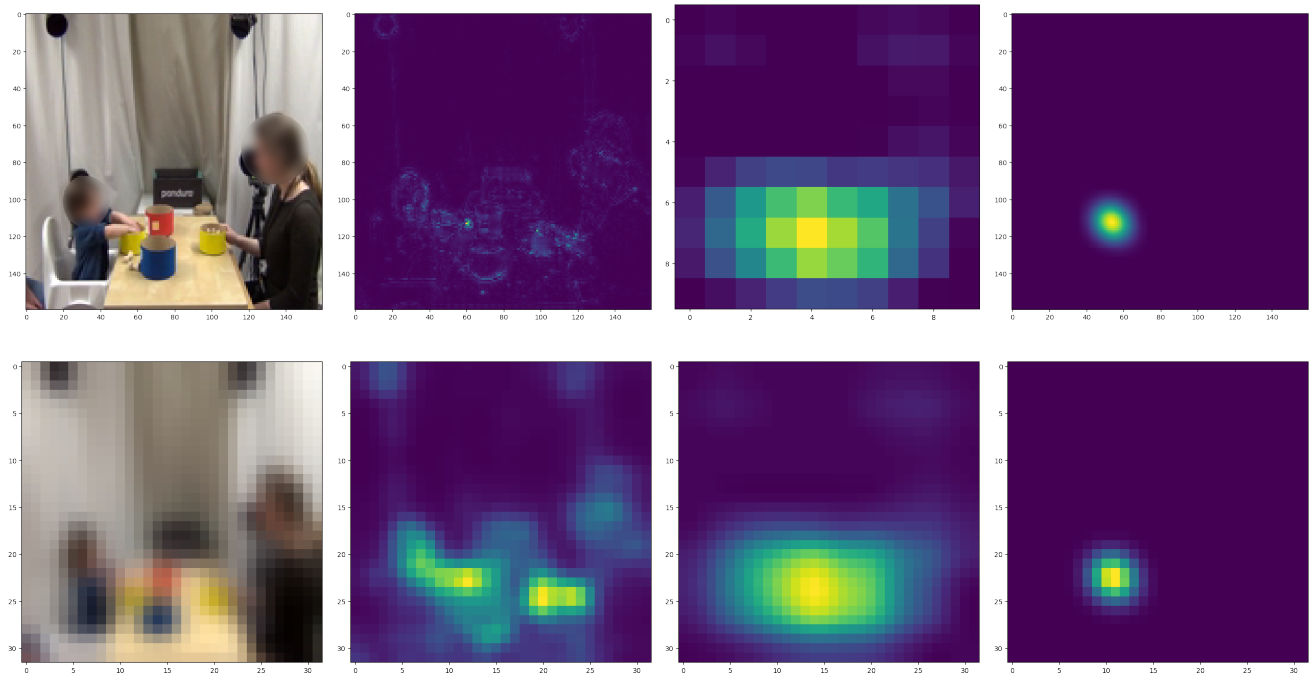


Figure 7: Left to right: original frame, Guided Backpropagation, Grad-CAM, and human attention. Top row: original resolution (160×160px except for Grad-CAM, 10×10px). Bottom row: homogeneous resolution (32×32px). Faces anonymized for the illustration.

Method	Guided Grad-CAM	Guided Backpropagation	Grad-CAM	Gradient Saliency
Mean	7.19	8.35	9.60	10.27
Standard Deviation	3.66	1.93	2.30	1.35

Table 4: Attention mapping methods ranked by increasing mean EMD (lower is better), calculated over all full-agreement samples in the comparison set. The EMD mean and standard deviation for every method are listed under the method's name.

as the best method and Gradient Saliency as the worst. Notice, however, that the standard deviation has a similar magnitude as the difference between means.

5 CONCLUSIONS

In this paper, we have shown that end-to-end Deep Learning models can learn to classify the affective states of an infant during guided play, specifically their task engagement with the toy at hand, and social engagement with the experimenter and the intended activity. We achieved this with very little data for Deep Learning standards: 23 videos, totalling 4 hours, of which only around 50 minutes were used in each training session – several orders of magnitude smaller than standard video datasets. Furthermore, we achieved this with a single video feed showing a general view of the interaction from a side angle, which would be unusable by standard feature extraction tools. The networks we trained showed varying degrees of agreement with human annotators – from bad as chance, to human-like performance. It appears the dominating factor is the task: some interaction scenarios are easier than others. We expected the classification of social engagement to be intrinsically more difficult than that of task engagement, but this was not supported by the results.

We have also shown how careful consideration of the available data can help mitigate the lack of training examples. Part of our pipeline uses standard solutions for this problem: pre-training on bigger datasets, strong data augmentation, good data stratification. But another part is tailored to the video domain: using interval annotations to obtain a continuum of snippets we can sample. Keeping in mind that samples from the same video are likely to show strong correlation, this technique greatly increases the effective number of samples at our disposal. The considerations about data extend to the targeted ground truth: when faced with disagreeing coders, we can avoid synthesizing a joint annotation, and let the network learn from every individual's perspective.

We have analysed four common post-hoc attention mapping methods: Gradient Saliency, Guided Backpropagation, Grad-CAM and Guided Grad-CAM. We have calculated attention maps for a *comparison set*, and discussed their differences when performing example-based subjective analysis – in our experience, the dominant use case in the literature. Through this approach, we observed head, body and contextual cues identified by the network as important to discriminate engagement levels. We also observed the *(shaker, social engagement)* network using contextual information: in negative samples, it correctly focuses on the infant to determine *not engaged*. But, when asked about evidence for *engaged*, it focuses on the researcher.

Finally, we have provided a numerical comparison of the different post-hoc mapping methods. For this, we collected a human-annotated attention map baseline for the comparison set, and used the Earth Mover's Distance to evaluate the *human-likeness* of each technique. Our results indicate that Guided Grad-CAM is closest to human attention, while Gradient Saliency is furthest. Previous literature has shown that both gaze tracking data and explicit mouse-painting can be successfully used to create the ground truth. In this paper we preferred the mouse-painting technique, which can capture *class sensitivity*.

6 OPEN ACCESS

The code for this project can be accessed at https://github.com/MarcFraile/infant-engagement

ACKNOWLEDGMENTS

Deep thanks to Elisabeth Wetzer and Mengyu Zhong for the many hours spent annotating the dataset.

This work was partly funded by the Centre for Interdisciplinary Mathematics, Uppsala University, and the Swedish Research Council (grant n. 2020-03167).

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Huda Alsofyani and Alessandro Vinciarelli. 2021. Attachment Recognition in School Age Children Based on Automatic Analysis of Facial Expressions and Nonverbal Vocal Behaviour. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 221–228.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[5] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti. 2013. Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE international conference on computer vision*. 921–928.

[6] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.

[7] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. 2016. Where should saliency models look next?. In *European Conference on Computer Vision*. Springer, 809–824.

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[9] Claire Chambers, Nidhi Seethapathi, Rachit Saluja, Helen Loeb, Samuel R Pierce, Daniel K Bogen, Laura Prosser, Michelle J Johnson, and Konrad P Kording. 2020. Computer vision to automatically assess infant neuromotor risk. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 11 (2020), 2431–2442.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[12] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon. 2020. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 784–789.

[13] Paul Ekman, Wallace V. Friesen, and Joseph C. Hager. 2002. *The Facial Action Coding System.*

[14] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision.* 2950–2958.

[15] Marc Fraile, Joakim Lindblad, Christine Fawcett, Nataša Sladoje, and Ginevra Castellano. 2021. Automatic analysis of infant engagement during play: An end-to-end learning and Explainable AI pilot experiment. In *Companion Publication of the 2021 International Conference on Multimodal Interaction.* 403–407.

[16] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[17] Darshan Gera and S Balasubramanian. 2020. Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. *arXiv preprint arXiv:2009.14440* (2020).

[18] Zakia Hammal, Wen-Sheng Chu, Jeffrey F Cohn, Carrie Heike, and Matthew L Speltz. 2017. Automatic action unit detection in infants using convolutional neural network. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 216–221.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[20] Nathan Henderson, Wookhee Min, Jonathan Rowe, and James Lester. 2021. Enhancing multimodal affect recognition with multi-task affective dynamics modeling. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII).* IEEE, 1–8.

[21] Hifza Javed, WonHyong Lee, and Chung Hyuk Park. 2020. Toward an automated measure of social engagement for children with autism spectrum disorder—a personalized computational modeling approach. *Frontiers in Robotics and AI* (2020), 43.

[22] Emily JH Jones, K Venema, R Earl, R Lowy, K Barnes, A Estes, G Dawson, and SJ Webb. 2016. Reduced engagement with social stimuli in 6-month-old infants with later autism spectrum disorder: a longitudinal prospective study of infants at high familial risk. *Journal of neurodevelopmental disorders* 8, 1 (2016), 1–20.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[24] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia* 23 (2020), 2086–2099.

[25] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[26] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* 10, 1 (2019), 1–8.

[27] Chloë Leclère, Marie Avril, S Viaux-Savelon, N Bodeau, Catherine Achard, Sylvain Missonnier, Miri Keren, R Feldman, M Chetouani, and David Cohen. 2016. Interaction and behaviour imaging: a novel method to measure mother–infant interaction using video 3D reconstruction. *Translational Psychiatry* 6, 5 (2016), e816–e816.

[28] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one* 13, 10 (2018), e0205999.

[29] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 325–347.

[30] Michael Morales, Peter Mundy, Christine EF Delgado, Marygrace Yale, Daniel Messinger, Rebecca Neal, and Heidi K Schwartz. 2000. Responding to joint attention across the 6-through 24-month age period and early language acquisition. *Journal of applied developmental psychology* 21, 3 (2000), 283–298.

[31] Alicja Niedźwiecka, Sonia Ramotowska, and Przemysław Tomalski. 2018. Mutual gaze during early mother–infant interactions promotes attention control development. *Child Development* 89, 6 (2018), 2230–2244.

[32] Pooja Prajod, Tobias Huber, and Elisabeth André. 2022. Using Explainable AI to Identify Differences Between Clinical and Experimental Pain Detection Models Based on Facial Expressions. In *International Conference on Multimedia Modeling.* Springer, 311–322.

[33] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. 2019. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing* 12, 2 (2019), 524–543.

[34] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference.*

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision.* 618–626.

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[37] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

[38] Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, and Shaun Canavan. 2020. Recognizing emotion in the wild using multimodal data. In *Proceedings of the 2020 International Conference on Multimodal Interaction.* 849–857.

[39] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 6450–6459.

[40] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[41] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006).* 1556–1559.

[42] Jianming Wu, Bo Yang, Yanan Wang, and Gen Hattori. 2020. Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction. In *Proceedings of the 2020 International Conference on Multimodal Interaction.* 777–783.

[43] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision.* Springer, 818–833.

[44] Bin Zhu, Xinjie Lan, Xin Guo, Kenneth E Barner, and Charles Boncelet. 2020. Multi-rate attention based gru model for engagement prediction. In *Proceedings of the 2020 International Conference on Multimodal Interaction.* 841–848.

[45] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. 2020. A Comprehensive Study of Deep Video Action Recognition. *arXiv preprint arXiv:2012.06567* (2020).