Guided weak supervision for action recognition with scarce data to assess skills of children with autism

Prashant Pandey
IIT Delhi
getprashant57@gmail.com

Prathosh AP
IIT Delhi
prathoshap@iitd.ac.in

Manu Kohli
IIT Delhi
manu.kohli@cognible.tech

Josh Pritchard
IIT Delhi
josh@factari.com

Abstract

Diagnostic and intervention methodologies for skill assessment of autism typically require a clinician repetitively initiating several stimuli and recording the child’s response. In this paper, we propose to automate the response measurement through video recording of the scene following the use of Deep Neural models for human action recognition from videos. However, supervised learning of neural networks demand large amounts of annotated data that are hard to come by. This issue is addressed by leveraging the ‘similarities’ between the action categories in publicly available large-scale video action (source) datasets and the dataset of interest. A technique called guided weak supervision is proposed, where every class in the target data is matched to a class in the source data using the principle of posterior likelihood maximization. Subsequently, classifier on the target data is re-trained by augmenting samples from the matched source classes, along with a new loss encouraging inter-class separability. The proposed method is evaluated on two skill assessment autism datasets, SSBD [23] and a real world Autism dataset comprising 37 children of different ages and ethnicity who are diagnosed with autism. Our proposed method is found to improve the performance of the state-of-the-art multi-class human action recognition models in-spite of supervision with scarce data.

1. Introduction

Autism is a complex neuro-developmental disorder that manifests in children during preschool years [1, 19] as deficits in communication, social skills and stereotypical repetitive behavior. In the last two decades, the prevalence rate of Autism Spectrum Disorder (ASD) has grown by more than 150% [1]. It is well-established that early intervention services modelled on behavior therapies yield the best outcomes for children diagnosed with autism [5] leading to significant societal benefits and cost savings. However, resource and expert scarcity in low resource settings result in delay in the initiation of the treatment process due to lack of identification of the disorder. Families have limited access to diagnostic and evidence-based treatments options because of affordability, lack of insurance support and non-availability of physical infrastructure. In the developing countries, there are entire swaths of children in dire need...
who remain untreated and unseen. Much of this strain and inaccessibility could be alleviated by incorporating technology to assist in early screening and automate the assessment and initial treatment planning process, which is a prerequisite to deliver individualized behavior treatment to children.

Skill assessment processes for autism typically involves invoking instructions to a child, monitoring and recording their responses as they occur. This requires a trained clinician to engage with the child, perform mundane repetitive tasks such as recording the child’s observation and human action responses to fixed set of stimuli. With the advent of tremendously powerful modern Deep learning techniques, one can hope to automate a lot of such tasks bringing affordability and improved access in value chain of autism screening, diagnosis and behavioral treatment activities while reducing the dependence on trained clinicians. Specifically, in this paper, we examine the application of human action recognition from video recordings for tracking the physical behavior of children diagnosed with ASD or otherwise to build cognitive and functional assessments.

2. Motivation

Human action recognition is typically solved in a supervised machine learning setting. Most of the successful models employ Convolutional Neural Networks (CNN) as their backbone. Two stream networks [21], 3-Dimensional (3D) Convolutional Neural Networks [24] and Convolutional long short-term memory (LSTM) networks [4] are some of the state-of-the-art action recognition approaches. The two stream architecture has two CNNs each of them separately trained on image sequences (RGB) and optical flow [9] sequences. The model averages the predictions from a single RGB frame and a stack of multiple optical flow frames after passing them through two CNNs which are pre-trained on large-scale static image datasets. Inflated 3D CNN [2] or 13D is found to be one of the top performers on the standard benchmarks like UCF101 [22] and HMDB51 [14] datasets. Temporal Segment Networks [25] or TSN is another example of two stream networks which has better ability to model long range temporal dependencies.

The objective of this work is to apply human action recognition algorithms to evaluate the responses of the children with autism, to measure stimulus in the area of imitation, listener response and motor skills. To accomplish this using already established methods requires large amounts of expert annotated data corresponding to the particular classes of actions to be recognized. The process of data collection and annotation is non-trivial because of non-availability of expert annotators, lack of co-operation from children and is also very time consuming. Despite these limitations, on the other hand, there is abundant availability of large-scale public datasets that contain thousands of annotated video clips corresponding to hundreds of action classes. Further, human action classes share a lot of similarity (in a well-defined feature space like Optical flow) in-spite of being disjoint as shown in Figure 1. For example, intuitively, the action classes ‘playing piano’ and ‘typing the keyboard’ can be thought similar in a suitable feature space. Motivated by these observations, in this paper we address the following question - Given a target data distribution with few annotated samples, can the availability of a large-scale annotated source dataset with ‘similar’ attributes as the target data, be leveraged, to improve the generalization abilities of classifiers built on target data? Specifically, we attempt to weakly supervise the task of human action recognition in a guided way using large-scale publicly available datasets. We build robust classifiers on video data with small number of annotated samples for skill assessment.

3. Related work

Children’s social behavior is decoded by measuring their degree of engagement during adult-child interactions and their ability to produce key behaviors like smile and gestures [20]. Eye-tracking data along with deep neural networks [12] can be leveraged for diagnosis of ASD. Attention based ASD screening [3] utilizes photo-taking and image-viewing modalities. Child-robot interactions [18] are used to curate large scale video dataset with variable actions on which 3D human pose estimation methods are adapted to evaluate action and emotion recognition. Most of these methods rely on measuring specific traits in children. Even though these cues are important, the screening or diagnosis is sub-optimal as they do not involve assessment of listener response and measurement of imitation, gross, fine and complex motor skills. Acquisition of the such skills is important to understand various abilities of children like language development, playing and interacting with others, mental imagery and perception [7]. All the existing action recognition based skill assessment approaches for autism assume large number of samples in the training data. State-of-the-art action recognition models [2] [6] [15] [25] [29] are deep neural networks which overfits easily to the dataset with fewer samples leading to poor generalization. Few-Shot Learning (FSL) action recognition methods [8] [10] [11] [17] [26] [27] [28] are characterised by different labels between source and target but a similar feature space. FSL evaluates on novel classes with limited training examples but these novel classes are sampled from the same dataset. The lack of domain shift between the base and novel classes is unrealistic. The large scale public datasets are not good candidate to be source in FSL as there is significant domain shift that exists between source and target data.
4. Proposed method

For a given a video, there exists transformations such as optical-flow, that are non-unique mappings of the video space. This suggests that given multiple disjoint set of action classes, there can be spaces (as flow) where a given pair of action classes may lie ‘close’ albeit they represent different semantics in the RGB-space. For example, the optical flow characteristics of a ‘baseball-strike’ class and ‘cricket-slog’ class could be imagined to be close. Further, it is also the case that, there exists large-scale, open datasets (e.g., Kinetics [13]) that encompasses a large number of annotated videos for several action classes. Thus, if one can find the classes in the open datasets that are ‘close’ to a given class in the data of interest, then the videos from the open dataset can be potentially used for augmentation resulting in regularization. In the subsequent sub-sections, we will formalize the aforementioned idea and describe a procedure to find the closer classes and use it for data augmentation.

Let \( \mathcal{X} \) denote the sample space encompassing the elements of transformed videos (e.g., Optical flow). Let \( \mathbf{P}^s(x_s) \) and \( \mathbf{P}^t(x_t) \) be two distributions on \( \mathcal{X} \) called the source and target distributions respectively. Suppose a semantic labeling scheme is defined both on \( \mathbf{P}^s(x_s) \) and \( \mathbf{P}^t(x_t) \). That is, let \( \mathcal{Y}_s = \{y^1_s, y^2_s, ..., y^N_s\} \) and \( \mathcal{Y}_t = \{y^1_t, y^2_t, ..., y^M_t\} \) be the source and target class labels that are assigned for the samples of \( \mathbf{P}^s(x_s) \) and \( \mathbf{P}^t(x_t) \) respectively which in-turn defines the joint distributions \( \mathbf{P}^s(x_s, y_s) \) and \( \mathbf{P}^t(x_t, y_t) \). \( N \) and \( M \) are the respective number of source and target classes. Let \( D_s = \{(x_s, y_s)\} \) and \( D_t = \{(x_t, y_t)\} \) denote the tuples of samples drawn from the two joint distributions \( \mathbf{P}^s \) and \( \mathbf{P}^t \), respectively. Suppose a parametric discriminative classifier (Deep Neural Network) is learned using \( D_t \) to obtain estimate of the conditional distribution \( \mathbf{P}^t_\theta(y_t|x_t) \) where \( \theta \) represent the parameters of the neural network. With these notations, we consider the case where the cardinality of \( D_t \) is much less than that of \( D_s \) implying that the amount of supervised data in the case of target distribution is much less than that of the source distribution. In such a case, \( \mathbf{P}^t_\theta(y_t|x_t) \) trained on \( D_t \) is deemed to overfit and hence doesn’t generalize well. If there exists a \( y^p_s \in \mathcal{Y}_s \) that is ‘close’ to \( y^q_t \in \mathcal{Y}_t \), then samples drawn from \( \mathbf{P}^s(x_s| y^p_s = y^p_s) \) can be used to augment the class \( y^q_t \) for re-training the model \( \mathbf{P}^t_\theta(y_t|x_t) \). In the subsequent sub-section, we describe a procedure to find the ‘closest’ \( y^p_s \in \mathcal{Y}_s \), given \( y^q_t \in \mathcal{Y}_t \) and a model \( \mathbf{P}^t_\theta(y_t|x_t) \) trained on \( D_t \).

4.1. Guided weak supervision

Videos lie in a very high dimensional space and are of variable length in general. Thus, standard vector distance metrics are not feasible to measure the closeness of two video objects. Further, the objective here is to quantify the distance between the classes as perceived by the discriminative model (classifier) \( \mathbf{P}^t_\theta(y_t|x_t) \) so that the data augmentation is sensible. Thus, we propose to use the maximum posterior likelihood principle to define the closeness between two classes. Let \( \mathcal{X}_{(y_s=y^p_s)} = \{x_{s1}, x_{s2}, ..., x_{sa}\} \) denote the samples drawn from \( \mathbf{P}^s(x_s| y^p_s = y^p_s) \). Now \( \mathbf{P}^t_t(y_t|x = x_{sj}), j \in \{1, 2, ..., l\} \) denotes the posterior distribution of the target classes \( y_t \) given the \( j^{th} \) feature vector from the source class \( \mathcal{X}_{(y_t=y^q_t)} \). With this, a joint posterior likelihood \( L_{y_s|x_s} \) of a class \( y^p_s \) can be defined as observing the target classes given a set of features \( \mathcal{X}_{(y_s=y^p_s)} \) drawn from a particular source class \( y^q_t \).

Mathematically,

\[
L_{y_s|x_s} = \mathbf{P}^t_t(y_{t1}, y_{t2}, ..., y_{tl}|x_{s1}, x_{s2}, ..., x_{sa})
\]

where \( x_{sj}, j \in \{1, 2, ..., l\} \), are from the class \( y^p_s \). If it is assumed that \( x_{sj} \) are drawn i.i.d., one can express Eq. 1 as,

\[
L_{y_s|x_s} = \prod_{j=1}^{l} \mathbf{P}^t_t(y_{tj}|x_{sj})
\]

This is because, the parameters \( \theta \) of the discriminator model created using \( D_t \) are independent of \( \mathcal{X}_{(y_s=y^p_s)} \), and are fixed during the evaluation of \( L_{y_s|x_s} \), which implies that \( y_{t1}|x_{s1} \) is independent of \( x_{sj} | \forall j \neq j \) thus leading to Eq. 2. The posterior likelihood in Eq. 2 can be evaluated for every target class \( y_t = y^q_t \), \( q \in \{1, 2, ..., M\} \), denoted by \( L_{y_s=y^q_t|x_s} \) called the target-class-posterior likelihood corresponding to the features from source class \( y^q_t \) under the learned target classifier \( \mathbf{P}^t_\theta(y_t|x_t) \). Mathematically,

\[
L_{y_s=y^q_t|x_s} = \prod_{j=1}^{l} \mathbf{P}^t_\theta(y_{tj} = y^q_t|x_{sj})
\]

With this definition of the target-class posterior likelihood, we define the matched source class \( y^p_s | y^q_t \in \mathcal{Y}_s \) to a given target class \( y^q_t \) as follows:

\[
y^p_s | y^q_t = \text{argmax}_{\mathcal{Y}_s} L_{y_s=y^q_t|x_s}
\]

Note that the definition of \( L_{y_s=y^q_t|x_s} \) is specific to a source-target class pair and therefore all \( x_{sj} \) in the objective function of the optimization problem in Eq. 4, comes from a particular source class. Thus, one can employ the discriminative classifier trained on the target data to find out the ‘closest’ matching source class as the one that maximizes the posterior likelihood observing that class as the given target class under the classifier. Since every class in the joint distribution can be looked as a ‘mode’ and the goal here is to match the classes (‘modes’) in the joint distribution of the source and target distributions, we call this procedure, Guided weak supervision (GWS). Figure demonstrates the idea of mode matching through examples. Optical flow frames of the target and the source classes have
similar visual properties indicating the closeness. Once the matched source class is determined to every given target class, a set of source classes matched is defined as \( \mathcal{Y}_s^* = \{ y_{s1}^*, y_{s2}^*, \ldots, y_{sM}^* \} \). Now, the discriminative classifier \( \Phi_t^s \) can be re-trained on the samples from the source dataset corresponding to \( \mathcal{Y}_s^* \) in a supervised way with class labels being the corresponding \( y_{t}^* \). This procedure thus increases the quantity and variety of the training data for \( \Phi_t^s \).

4.2. Directional Regularization

The procedure of mode matching described in the previous sub-section effectively changes the semantic meaning of the matched source classes to the semantic meaning of the target classes. Thus, it is possible to train a classifier on the source data to discriminate between the matched source classes \( \mathcal{Y}_s^* = \{ y_{s1}^*, y_{s2}^*, \ldots, y_{sM}^* \} \). Suppose such a classifier is denoted by \( \Phi_t^s(y_s^*|x_s) \), where \( \phi \) are the model parameters. We assume that \( \Phi_t^s(y_s^*|x_s) \) and \( \Phi_t^s(y_t|x_t) \) have the same architectural properties. Also, it is assumed that the source dataset is larger in size and more diverse compared to the target dataset\(^1\). This implies that the \( \Phi_t^s(y_s^*|x_s) \) has better generalization abilities compared to \( \Phi_t^s(y_t|x_t) \) (This fact is corroborated empirically in the experiment section).

We propose to leverage this fact in improving the generalization capabilities of \( \Phi_t^s(y_t|x_t) \) using \( \Phi_t^s(y_s^*|x_s) \). Further, during the training of \( \Phi_t^s(y_s^*|x_s) \) with samples from \( \mathcal{Y}_s^* \), it is desirable that the separation that is achieved between the classes in \( \mathcal{Y}_s^* \) under the classifier \( \Phi_t^s(y_s^*|x_s) \) is ‘preserved’ during the training of \( \Phi_t^s(y_t|x_t) \) with samples from \( \mathcal{Y}_s^* \). We propose to accomplish the aforementioned properties by imposing a regularization term during the training of \( \Phi_t^s(y_t|x_t) \). Specifically, we propose to push the Eigen directions of the parameter matrix of \( \theta \) towards that of the parameter matrix of \( \phi \). Note that \( \phi \) is fixed during the training of \( \Phi_t^s(y_t|x_t) \). Intuitively, this implies that the significant directions of the target parameters should follow that of the source parameters\(^2\).

Mathematically, let \( M_\theta \) and \( M_\phi \) be two square matrices formed by re-shaping (without any preference to particular dimensions) the parameters \( \theta \) and \( \phi \), respectively. We perform an Eigen-value decomposition on \( M_\theta \) and \( M_\phi \), to obtain the Eigen vector matrices \( E_\theta \) and \( E_\phi \), respectively. Let \( \hat{E} \) denote the truncated versions of \( E \) with first \( k \) significant (a model hyperparameter) Eigen vectors. Under this setting, we desire the Eigen directions \( E_\theta \) and \( E_\phi \) to be aligned. Mathematically, if they are perfectly aligned, then

\[
\hat{E}_\theta^T \hat{E}_\phi = I_k
\] (5)

where \( I_k \) is a \( k \)-dimensional identity matrix and \( T \) denotes the transpose operation. Thus, any deviation from the condition laid in Eq. 5 is penalized by minimizing the Frobenius norm of the deviation. We term this as directional regularization (DR) denoted as \( L_{DR} \) given by the following equation:

\[
L_{DR} = \| \hat{E}_\theta^T \hat{E}_\phi - I_k \|_F
\] (6)

where \( \| \cdot \|_F \) denotes the Frobenius norm of a matrix. Note that this regularizer is on \( \theta \) imposed during the training of \( \Phi_t^s(y_t|x_t) \) ensuring that the directions of separating hyperplanes of the target classifier is encouraged to follow those of the source classifier trained with the matched classes.

Thus the final objective function during the re-training of the target classifier is as follows:

\[
L_{Total} = -\sum_{j=1}^{1} y_{ij} \log \hat{y}_{ij} + \| \hat{E}_\theta^T \hat{E}_\phi - I_k \|_F
\] (7)

\(^1\)This is a reasonable assumption because the source dataset mostly comprises large publicly available datasets annotated by human experts.

\(^2\)It is not desirable that all the parameters are exactly the same because in that case, the model will overfit on one type of dataset, either source or target.
Table 1: Distribution of the training samples in accordance with subject’s ethnicity and age across various Autism classes. Total samples include few clips with camera facing the clinician as well.

| Autism class     | Ethnicity | Age ≤ 5 years | Age > 5 years | Total samples |
|------------------|-----------|---------------|---------------|---------------|
| Move the table   | Asian     | 13            | 12            | 19            |
| Touch ear        | Caucasian | 8             | 7             | 14            |
| Lock hands       | Asian     | 8             | 7             | 12            |
| Touch head       | Caucasian | 14            | 13            | 23            |
| Touch nose       | Asian     | 14            | 12            | 20            |
| Rolly polly      | Caucasian | 11            | 9             | 18            |
| Tapping          | Asian     | 14            | 13            | 26            |
| Arms up          | Caucasian | 9             | 7             | 16            |

where $\hat{y}_{tj}$ is the predicted target class. Thus in summary, given a small amount of data from the target distribution, the proposed method (1) trains a classifier on the target samples, (2) estimates the closest classes from the source distribution to all the target classes using the target classifier, (3) trains a new (relatively robust) classifier on the samples from the source distribution with re-labeled source classes (matched with the target classes), (4) uses the samples of the matched source classes to re-train the target classifier along with directional regularization, which forms the final model for the target data. This entire procedure is pictorially depicted in Figure 2.

4.3. Implementation

The idea of mode-matching detailed in the previous section assumes that the source and the target distributions are ‘similar’ under a certain feature transformation. The raw video data in the RGB space does not adhere to such assumptions because of the variability of the content/scene/subjects etc. However, transformations such as optical flow [9] masks most of the non-motion related information exaggerating the motion information. In this domain, it is reasonable to assume that the source and target action classes are ‘similar’ (Refer Figure 1 for an example). We match each target Autism class to a source class using GWS on the baseline Autism model $P_{t\theta}(y_t|x_t)$. Since a probabilistic softmax layer is used at the output of the classifier, the matched source class, for every given target class, can be simply taken to be that source class whose samples get labeled as the given target class most of the times as compared to all other source classes. The optimization problem in Eq. 4 can be approximated as follows:

$$y^*_s|y^q_t = \arg\max_{y_s} \mathcal{L}_{y_t=y^q_t|x_s}$$

$$= \prod_{j=1}^{l} P_{t\theta}(y_{tj} = y^q_t|x_{sj})$$

$$\approx \arg\max_{y_s} \left[ \text{Count} \left( j \in \{1,2,...,l\} \right) \left( y^q_t = \arg\max_{y_t} P_{t\theta}(y_{tj}|x_{sj}) \right) \right]$$

We propose to use state-of-the-art action recognition models such as I3D [2] and TSN [25] that employ optical flow streams. Baseline target Autism model $P_{t\theta}(y_t|x_t)$ is obtained by training Autism data using an I3D or TSN architecture by initializing their weights with pre-trained Kinetics/ImageNet models. The mode matching and re-training with the directional regularization loss are performed only on the flow stream without altering the RGB stream. For implementation details, please refer to the supplementary material.

5. Autism Dataset

The target Autism data consists of 37 subjects (‘child’ and ‘subject’ are used interchangeably). Five of the subjects have Caucasian origin and the rest are Asians. Their ages range from 2-14. During assessment, a clinician performs the functional assessment by probing a child for age-appropriate listener response and imitation skills by invoking an instruction response and expecting a child to respond through a human action. We deliberately chose eight representative human action responses invoked through either listener response or imitation instruction for our experiments. Specifically, the action classes selected are - ‘Move the table’ and ‘Arms up’ for gross motor skill assessment, ‘Lock hands’ and ‘Tapping’ for fine motor skills, ‘Rolly polly’ for complex motor action, ‘Touch nose’, ‘Touch head’
and ‘Touch ear’ for distinguishing identification of different parts of the body. We chose these particular actions since the presence of age-appropriate fine and gross motor skills demonstrate neuro-typical development of a child, as well as providing a clear picture of when the development is atypical [7].

Video clips (amount to about 1481 in number) were recorded in a semi-structured environment with the clinician facing the child, and three synchronized cameras were placed to record the videos. The first camera faced the clinician, the second faced the child and the third was positioned laterally to both the clinician and the child. Figure 3 depicts representative frames of some action classes during training sessions for different camera positions. The videos were annotated by trained clinical psychologists. The response of the child for a particular stimulus is taken as a human action response classification problem to be measured using a Deep learning model. Please refer to the supplementary material for distribution of video samples across different classes.

6. Experiments and Results

Two large-scale publicly available human action recognition datasets namely Kinetics [13] and HMDB51 [14] are used as the source datasets while the data described in the previous section (termed as the Autism dataset) is the target dataset. The task is of standard 8-class classification on the target Autism data with classes as described in the previous section.

6.1. Baselines

All the accuracy numbers reported on the baselines are averaged over a 5-fold cross validation. Two state-of-the-art Deep learning models namely the Inflated 3D CNN (I3D) [2] and Temporal Segment Networks (TSN) [25] are selected as the architectural backbones. For mode matching experiments, I3D is used in conjunction with Kinetics and TSN with HMDB51, respectively.

Our baselines are I3D and TSN which are pre-trained on Kinetics-600 and HMDB51 respectively. Next, using the baseline model, the Kinetics and HMDB51 classes are mode matched to Autism classes using GWS. Table 2 shows matched classes on both the source datasets. It is apparent from Table 2 that GWS maps semantically similar actions from Kinetics and HMDB51 to the Autism actions, confirming the proposed hypothesis.

| Autism         | Kinetics          | HMDB51        |
|----------------|-------------------|---------------|
| Move the table | Pushing cart      | Push          |
| Touch ear      | Tying necktie     | Sit up        |
| Lock hands     | Playing trombone  | Shake hands   |
| Touch head     | Blowdrying hair   | Shoot ball    |
| Touch nose     | Putting on eyeliner| Eat          |
| Rolly polly    | Playing hand clapping games | Flic flac |
| Tapping        | Playing drums     | Chew          |
| Arms up        | Jumping jacks     | Fall floor    |

Table 2: Action classes from the source datasets (Kinetics and HMDB51) matched to target (Autism) classes.

Figure 4 shows the variation of test accuracy with different amounts of source data from Kinetics and HMDB51 mode matched classes augmented with Autism classes. Baseline models are re-trained with varying amounts of mode matched Kinetics or HMDB51 samples (equivalent to 5% (~75 samples) to 30% (~450 samples) of all the Autism samples). In all of these baseline models, when the size of augmented source data is increased and the model is re-trained, the baseline accuracy increases till the percentage of the augmented source data is comparable (in terms of number of samples) with the Autism data. It is seen that when the source data dominates the Autism data, the accuracy drops. This is expected because when the source distribution dominates, the classifier tends to overfit on it. However, with GWS along with DR, not only the rise in test accuracy is more, the drop in accuracy after peaking is smoother as compared to GWS without DR. This implies that DR offers more tolerance towards the augmented source data which allows the performance to increase further. Our approach outperforms all baseline Autism models.
of I3D and TSN with comparable source samples. When we re-train the baseline Autism I3D classifier by augmenting random samples from the source(without GWS), the test accuracy drops to 32% thereby showing importance of GWS.

t-SNE [16] plots for the 8-dimensional embeddings from penultimate layer of I3D and TSN are obtained for a baseline models with and without GWS and DR as shown in Figure 5. With our approach, the inter-class separability of samples has increased while the intra-class separability has decreased so the model tends to be more confident in its predictability of Autism classes. In the next set of experiments, we iteratively re-train the target model with GWS and DR. That is, in every new iteration we discard the mode matched source samples from the previous iteration keeping the number of samples similar in every iteration (equivalent to the number of Autism training samples) and re-train the target model with new set of samples from the matched modes classes. Figure 6 shows the variation in accuracy when the new source samples are augmented with Autism data in every iteration. In Figure 6a, the test accuracy increases from the baseline for initial iterations with GWS and DR. With GWS only (I3D+GWS), the accuracy drops after the 2nd iteration. If GWS is applied with DR on the baseline (I3D+GWS+DR), it continues to increase even after the 2nd iteration but starts to dip from 3rd iteration on-wards which can be ascribed to overfitting on the source data. As with the previous cases, we observe more tolerance of DR with newer samples as compared to GWS. Newer samples are accepted with lesser surprise in DR which enhances the generalizability and performance. Similar behavior is observed with TSN as shown in Figure 6b.

### 6.2. Bias in the training data

Table 3 shows the results for the proposed method under different kinds of dataset biases. The results in the second column in this table are accuracies when the training data has only Asian subjects and test data has Caucasian subjects. Third column are test accuracies when the training data has Caucasian subjects and test data has Asian subjects. In the fourth column, the training data has subjects that are 5 years or below and test accuracies are recorded with test data having subject above 5 years in age. The last column are test accuracies when the training data has subjects above 5 years in age and test set has subjects that are 5 years or below.

| Model         | Asian | Caucasian | ≤ 5 yrs. | > 5 yrs. |
|---------------|-------|-----------|----------|----------|
| I3D           | 58.2% | 34.6%     | 55.1%    | 41.2%    |
| TSN           | 53.7% | 30.2%     | 51.7%    | 38.1%    |
| I3D+GWS+DR    | 63.7% | 40.2%     | 59.8%    | 46.5%    |
| TSN+GWS+DR    | 57.0% | 33.8%     | 55.4%    | 40.9%    |

Table 3: Performance of GWS and DR with a specific bias in the Autism training dataset. It is seen that in all the cases, the proposed approach offers better performance over the baselines.

### 6.3. GWS and DR under different settings

Table 4 shows test accuracy scores on baseline Autism model under six different settings. We sampled examples from second best modes to augment with Autism data using GWS and DR. The results in the second column are the test accuracy scores using second best modes. It can be seen that the performance is better than the baseline model albeit less than the performance of model with the first best modes. It is however apparent that even second best modes preserve the closeness in the optical flow space. All the experiments are executed by re-training optical flow stream of the baseline models. Besides optical flow, if we re-train the RGB stream as well, it is seen that the performance of the classifiers deteriorates as shown in third column of Table 4. This ascertains the fact that the notion of closeness is valid only for optical flow samples. Next, we handpicked similar action classes (like ‘Washing hands’ in Kinetics matched to ‘Lock hands’ in Autism data, etc.) from Kinetics and HMDB51 datasets and applied our approach on the baseline model. Fourth column of Table 4 records test accuracy scores with these handpicked classes (or modes). The results are comparable to our approach where we employ baseline models to find similar classes or actions using GWS. Hence, irrespective of the metric used to find similarity (either using human intelligence or GWS), the performance is much better than the baselines when they are
a) Baseline model on I3D.
b) Baseline model on I3D with GWS+DR.
c) Baseline model on TSN.
d) Baseline model on TSN with GWS+DR.

Figure 5: t-SNE plots of embeddings of penultimate layers of baseline I3D and TSN Autism models with and without GWS and DR. It is clearly seen that the inter-class separability has increased and clusters are more dense after GWS and DR.

| Model                | Second best modes | Flow + RGB stream | Handpicked modes | Combined modes | I3D→TSN | TSN→I3D |
|----------------------|-------------------|-------------------|------------------|---------------|---------|---------|
| I3D+GWS              | 73.5%             | 67.5%             | 74.2%            | 73.8%         | -       | 73.2%   |
| TSN+GWS              | 71.4%             | 65.8%             | 71.9%            | 71.2%         | 71.7%   | -       |
| I3D+GWS+DR           | 74.4%             | 68.1%             | 75.3%            | 74.8%         | -       | 74.4%   |
| TSN+GWS+DR           | 72.4%             | 66.7%             | 72.8%            | 72.7%         | 73.1%   | -       |

Table 4: GWS and DR under different settings on baselines (69% for I3D and 68% for TSN) - It is seen that (a) GWS or GWS+DR with second best modes too leads in better performance (b) re-training the RGB stream is detrimental since there is no similarity in the RGB space (c) GWS or GWS+DR with hand-picked modes also results in improvement in accuracy (d) re-training with samples from matched modes classes from different datasets results in performance enhancement (e) & (f) Cross neural architecture GWS and DR - Increase in accuracy from the baselines indicate that similar actions in the optical flow space retain their meaning irrespective of the neural architectures.

a) Baseline I3D with Kinetics samples.
b) Baseline TSN with HMDB51 samples.

Figure 6: Performance on I3D and TSN with iteration over the source samples with GWS and DR. The accuracy increases through iteration by augmenting with newer mode matches samples in every iteration although overfitting on source data occurs from 3rd iteration.

augmented and re-trained with similar classes with or without DR. In the final experiment, we combined corresponding top mode matched classes from Kinetics and HMDB51 (like ‘Pushing cart’ from Kinetics is combined with ‘Push’ from HMDB51, etc.) and used these samples for GWS and DR as shown in the fifth column of Table 4. The performance is comparable with our approach when the modes are not combined which implies that the idea of similarity is preserved across different datasets as well. Fifth and sixth columns report test accuracies for the case when the modes matched using one architecture is used to augment the classifier built on another architecture. Fifth column of Table 4 are the test accuracy scores when baseline TSN is re-trained with augmented Kinetics modes extracted from GWS on I3D. Similarly, sixth column has test accuracy scores when baseline I3D is re-trained with HMDB51 modes extracted from TSN. The scores are still better than the baseline Autism model (69% for I3D and 68% for TSN) which indicates that similar actions in the optical flow space retain their meaning irrespective of the neural architectures used as a backbone.

6.4. Comparisons with state-of-the-art

Table 5 reports test accuracy for GWS and DR on baseline I3D and TSN. It is consistently observed that with our approach (GWS or DR or GWS+DR), the performance is better than the baseline I3D and TSN and state-of-the-art action recognition models using Autism dataset and SSBD.
Additionally, it is seen that DR offers additional accuracy benefits over GWS.

| Model                      | Autism Dataset | SSBD  |
|----------------------------|----------------|-------|
| TSN (Pretrained with HMDB51) | 68.0% 87.4%  |       |
| I3D (Pretrained with Kinetics) | 69.3% 91.2%  |       |
| ECO (Pretrained with Kinetics) | 61.4% 80.1%  |       |
| TSM (Pretrained with Kinetics) | 69.8% 90.5%  |       |
| R(2+1)D (Pretrained with IG-65M) | 68.4% 88.3%  |       |
| TSN+DR                     | 70.1% 89.2%  |       |
| I3D+DR                     | 71.3% 92.8%  |       |
| TSN+GWS                    | 71.6% 90.3%  |       |
| I3D+GWS                    | 74.3% 93.6%  |       |
| TSN+GWS+DR                 | 72.5% 91.4%  |       |
| I3D+GWS+DR                 | 75.1% 95.7%  |       |

Table 5: Comparison with state-of-the-art models on Autism dataset and SSBD.

7. Conclusion

In this paper, we proposed a method for improving the generalization abilities of a classifier designed for human action recognition trained on sparse data. Specifically, leveraging the semantic similarities of the action classes in the optical flow space, we proposed a generic method called Guided weak supervision (GWS) to augment and re-train a classifier on the target data with samples from a large-scale annotated dataset, along with a novel loss function termed as directional regularization (DR) that would result in performance enhancement. We demonstrated the efficacy of the proposed method for screening, diagnosis and behavioral treatment for ASD. With the proposed framework, we can integrate and automate complete value chain of screening to treatment for children with ASD by capturing behavioral treatment progress data temporally and remotely. The other future direction is to employ the techniques of GWS and DR for generalized supervised learning tasks beyond the proposed use case of Autism.

References

[1] Data & statistics on autism spectrum disorder. https://www.cdc.gov/ncbddd/autism/data.html 2019. Accessed: 2019-06-30.
[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.
[3] Shi Chen and Qi Zhao. Attention-based autism spectrum disorder screening with privileged modality. In The IEEE International Conference on Computer Vision (ICCV), October 2019.
[4] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor 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.
[5] Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenson, Jamie Winter, and Geraldine Dawson. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 54(7):580–587, 2015.
[6] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12046–12055, 2019.
[7] Emma Gowen and Antonia Hamilton. Motor abilities in autism: a review using a computational context. Journal of autism and developmental disorders, 43(2):323–344, 2013.
[8] Michelle Guo, Edward Chou, De-An Huang, Shuran Song, Serena Yeung, and Li Fei-Fei. Neural graph matching networks for few-shot 3d action recognition. In The European Conference on Computer Vision (ECCV), September 2018.
[9] Berthold KP Horn and Brian G Schunck. Determining optical flow. Artificial intelligence, 17(1-3):185–203, 1981.
[10] Mihir Jain, Jan C. van Gemert, Thomas Mensink, and Cees G. M. Snoek. Objects2action: Classifying and localizing actions without any video example. In The IEEE International Conference on Computer Vision (ICCV), December 2015.
[11] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In Advances in neural information processing systems, pages 3464–3472, 2014.
[12] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In The IEEE International Conference on Computer Vision (ICCV), Oct 2017.
[13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheerja Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.
[14] Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In 2011 International Conference on Computer Vision, pages 2556–2563. IEEE, 2011.
[15] Ji Lin, Chuang Gan, and Song Han. Tsn: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision, pages 7083–7093, 2019.
[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008.
[17] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shababz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
[18] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian
[19] Isabelle Rapin. Autistic children: Diagnosis and clinical features. *Pediatrics*, 87(5):751–760, 1991.

[20] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children’s social behavior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3414–3421, 2013.

[21] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] Shyam Sundar Rajagopalan, Abhinav Dhall, and Roland Goecke. Self-stimulatory behaviours in the wild for autism diagnosis. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

[24] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[26] Yali Wang, Lei Zhou, and Yu Qiao. Temporal hallucinating for action recognition with few still images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Hongtao Yang, Xuming He, and Fatih Porikli. One-shot action localization by learning sequence matching network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[29] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–712, 2018.