Video-Based Activity Recognition for Infants

Keyuan Wu
SCPD, Stanford University

Introduction

OBJECTIVE

- Infants under 12 months of age are unable to speak and may be at risk if they are not within the caregiver's line of sight. The objective of this study is to develop a network that can accurately recognize different baby actions.

CHALLENGE

- Most of the available open source or commercial action unit detectors are trained with adults. Kid-specific dataset is more complex to generalize to than the adult-specific dataset (Olaere, 2021).

RELATED WORK

- The work of Sujitha Balasathiya et al. encompasses a diverse range of activities, but it lacks modern neural network techniques.
- Olalere's research also includes diverse activities, focusing on children aged 0-12 years old, specifically studying their involvement in sports activities.
- Studies by Hammal et al., Khan, Dechemi et al., and Adewopo et al, on the other hand, concentrate on specific and narrow conditions, such as analyzing sleeping gestures, crib-related sleeping and awakening, facial analysis, and reaching actions.
- Differing from these, my study included a diverse set of activities, yet specifically for infants under one year old. By employing transformed-based models, my paper serves the general purpose of recognizing various infant actions.

Video-Based Activity Recognition for Infants

Keyuan Wu
SCPD, Stanford University

Method

1

Model I. Transfer Learning Vision Transformer (ViT)

- Pre-trained VIT - “vit_base_patch16_224”
- For predicting the video labels-
  - Majority Vote: Selected the label that most frequently appeared across the frames
  - Probability Based: Chose the label from the frame with the highest predicted probability as the video label

2

Model II. Transfer Learning Swin transformer + LSTM (late fusion)

- A feature extractor, an LSTM layer and a fully connected layer
- Pre-trained Swin Transformer - “swin_tiny_patch4_window7_224”

3

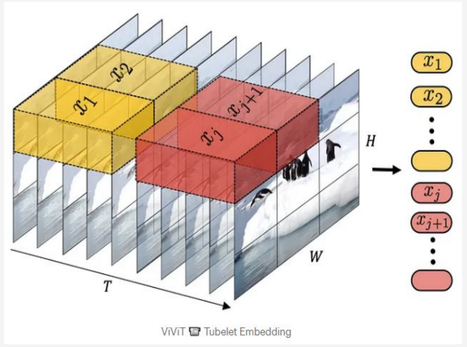
Model III. ConvNet (early fusion)

- ConvNet (including convolutional layers, ReLU activation functions, and max pooling layers)

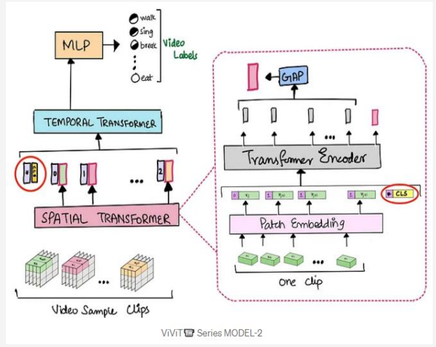
4

Model IV. Transfer Learning Video Vision Transformer (ViViT) Video Vision Transformer

- Pre-trained ViViT: model = ViViT(224, 16, 100, 16) *https://github.com/rishikksh20/ViViT-pytorch*
  - Tubelet Embedding
  - Factorized Encoder-Decoder



ViViT ☎ Tubelet Embedding

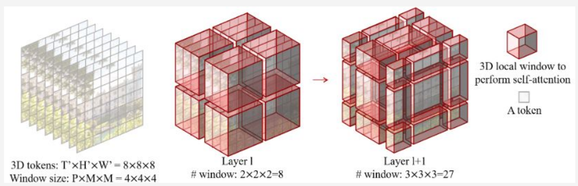


ViViT ☎ Series MODEL-2

5

Model V. Transfer Learning Video Swin Transformer

- Pre-trained Swin-T model *https://github.com/SwinTransformer/Video-Swin-Transformer*
  - Shifted Window-based Multi-head Self-Attention (SW-MSA)

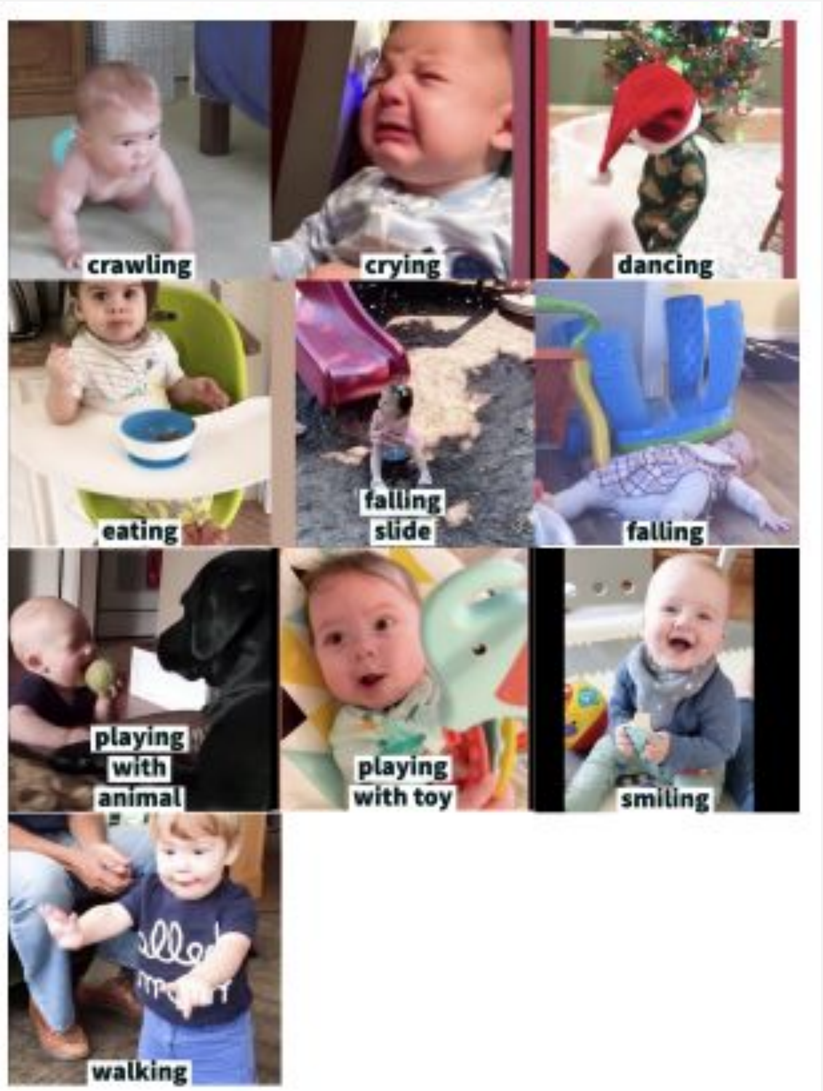


Video-Based Activity Recognition for Infants

Keyuan Wu
SCPD, Stanford University

Data

A collection of 305 videos with an average length of 6.33 seconds and a cumulative duration of 32 minutes was sourced from YouTube. These videos were manually labeled across 10 distinct categories.



Summary of Video Statistics

Category	Number of Videos	Average Length
crawling	29	6.36s
crying	35	6.87s
dancing	21	8.16s
eating	21	11.59s
falling slide	39	4.77s
falling	41	3.74s
playing with animals	25	7.77s
playing with toys	25	6.92s
smiling	26	7.93s
walking	43	4.17s

Video-Based Activity Recognition for Infants

Keyuan Wu
SCPD, Stanford University

Results

Experiments have been conducted using five transfer learning models, based on pre-trained Vision Transformer (ViT) and Swin Transformer models. The best performing model is the Swin Transformer with late fusion LSTM applied in transfer learning, which achieved a top-1 prediction accuracy of 50.00%.

Model Hyperparameters and Accuracies

Model	Learning Rate	Batch Size	Number of Epochs	Optimizer	Learning Rate Decay	Top1 Accuracy	Top5 Accuracy
Model I	1e-4	8	10	Adam	0.1	43.33%	93.33%
Model II	1e-4	8	10	Adam	-	50.00%	83.33%
Model III	1e-4	8	10	Adam	0.001	43.33%	86.67%
Model IV	1e-4	8	50	Adam	-	33.33%	80.00%
Model V	1e-4	8	100	Adam	-	36.67%	66.67%

Top-1 Prediction Accuracies by Category

Label	Videos	Model I	Model II	Model III	Model IV	Model V
crawling	3	0.67	0.33	0.67	0.67	0.33
crying	3	0.67	0.67	0.67	0.00	0.33
dancing	2	0.00	0.00	0.00	0.50	0.00
eating	2	0.50	0.00	0.50	0.50	0.00
falling slide	4	0.75	0.75	0.50	0.25	0.50
falling	4	0.50	0.75	0.75	0.75	0.75
playing w animals	3	0.67	0.67	0.33	0.33	0.00
playing w toy	2	0.00	0.00	0.00	0.00	0.00
smiling	3	0.00	0.33	0.33	0.00	0.33
walking	4	0.25	0.75	0.25	0.25	0.25

Video-Based Activity Recognition for Infants

Keyuan Wu
SCPD, Stanford University

Discussion

OBSERVATION

- Classifying certain infant actions such as crawling, crying, and falling proved straightforward, while categories like dancing and playing with toys posed challenges. These difficulties arose due to the complex task of capturing subtle movement nuances and the presence of various objects, which were only partially visible in the frames.

LIMITATION

- Data augmentation was not performed, which could have supplemented the data points and improved the model accuracy.
- The training process was limited to the last linear layers of the models, with other weights remaining frozen. Given more time for fine-tuning all the weights, the models might have achieved higher accuracy.
- The video-based models - ViViT and Video Swin Transformer, didn't perform up to expectations. Further investigation is needed to ascertain potential reasons, which could include exploring hyperparameter settings or optimizing the input data to better utilize the model architecture.

FUTURE WORK

- The fusion of audio and visual data can provide significant insights that enhance the safety and well-being of infants in real-world situations.