

A Large-scale Study of Representation Learning and the Benchmarking in Video Action Recognition

Md Ashik Khan

Dedicated to my dearest parents

A Large-scale Study of Representation Learning and the Benchmarking in Video Action Recognition

*Thesis submitted to
Indian Institute of Technology Kharagpur
for the award of the degree of*

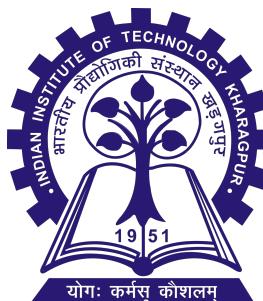
**Master of Technology
in
Computer Science and Engineering**

by

**Md Ashik Khan
21CS60A02**

Under the supervision of

Dr. Abir Das



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
WEST BENGAL - INDIA, 721302**

APRIL 2023

©2023 Md Ashik Khan. All Rights Reserved.



Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
West Bengal-India, 721302

CERTIFICATE

This is to certify that the thesis entitled **A Large-scale Study of Representation Learning and the Benchmarking in Video Action Recognition**, submitted by **Md Ashik Khan**(Roll Number: **21CS60A02**) a postgraduate student of **Department of Computer Science and Engineering** for the award of the degree of Master of Technology is a record of bona fide work under my supervision. I hereby accord my approval of it as a study carried out and presented in a manner required for its acceptance in the fulfillment for the Post Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements per the Institute's regulations and reached the submission standard.

Dr. Abir Das

Assistant Professor

Department of Computer Science and Engineering,
Indian Institute of Technology, Kharagpur

DECLARATION

I certify that

- a. The work contained in the thesis is original and has been done by me under the guidance of my supervisor;
- b. The work has not been submitted to any other institute for any other degree ;
- c. I have followed the guidelines provided by the Institute in preparing the thesis;
- d. I have conformed to ethical norms and guidelines while writing the thesis;
- e. Whenever I have used materials (data, models, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis, sharing their details in the references, and taken permission from the copyright owners of the sources, whenever necessary.

Md Ashik Khan

ACKNOWLEDGEMENTS

First and foremost, I take this opportunity to express my heartfelt gratitude and sincere thanks to my supervisor, **Dr. Abir Das**, Assistant Professor, Department of Computer Science & Engineering, Indian Institute of Technology, Kharagpur, for his constant guidance, suggestion, and gracious encouragement at every stage of this dissertation work.

I would also like to acknowledge my mentor **Owais Iqbal**, Research Scholar, Indian Institute of Technology, Kharagpur, for his valuable suggestion, constant guidance, and kind co-operation throughout the journey. I shall forever remain indebted to my parents, teachers, and friends for supporting me at every stage. Their constant encouragement and support have been helping me throughout my academic career, especially during the entire project work.

Md Ashik Khan

Department of Computer Science and Engineering,
IIT Kharagpur,
April 2023.

ABSTRACT

One of the major challenges for visual scene understanding is video action recognition. Due to the introduction of deep learning, we have seen significant improvements in this field in the last decade. Representation learning for action recognition learns a higher dimensional representation of videos generally expressed in RGB space so that these representations help classify the videos performing different actions. Notwithstanding the remarkable progress in video action recognition, without an unified evaluation of different approaches, it is always hard to gauge the true progress as different variabilities including sensitive hyperparameters, pretraining datasets, amount of data in down-the-line datasets, can affect the performance to a large extent. Additionally, diverse long-range temporal information in clips, the high processing cost of different action recognition architectures as well as assessment protocol differences have made it challenging for the researchers to bring different action recognition models and datasets to a standard benchmark. The study develops a Video Action Recognition Benchmark (**VARB**) and provides a complete examination of representation learning for video action recognition. The lack of a standardized assessment for all broad visual representations in action recognition is filled by **VARB**. With **VARB**, a thorough analysis of numerous well-liked, openly accessible video action recognition approaches has been conducted. Confounding factors like architecture and budget are carefully managed. The 13 publicly available action recognition datasets ranging from six domains are chosen for the purpose. Then, six popular video activity recognition models are chosen, starting with 2-D convolutional neural networks, moving on to 3-D convolutional neural networks, and ending with more modern video vision transformer architectures(both in 2-D and 3-D). In the process, we also curated a new dataset from the noisy and untrimmed Youtube-8M clips by manually trimming and annotating videos with industrial actions and included it in our study. The last step examines unresolved challenges and highlights possible applications for video action recognition to promote novel research ideas.

Keywords: video action recognition, transfer learning, full-finetuning, linear evaluation, limited evaluation, few-shot, resnet50, 2D,3D-CNN,kinetics-400, MiT

Contents

1	Introduction	1
2	Related Work	4
3	The Benchmarking in Video Action Recognition	9
3.1	Tasks	10
3.2	Transfer Strategy	10
4	Large-Scale Study	12
4.1	Dataset	12
4.1.1	Upstream Data	12
4.1.1.1	Kinetics - 400	12
4.1.1.2	Moments In Time	13
4.1.2	Downstream Data	13
4.1.2.1	HMDB	13
4.1.2.2	UCF101	13
4.1.2.3	Activity Net	13
4.1.2.4	FineGym - Sports	14
4.1.2.5	Diving	14
4.1.2.6	InfAR	14
4.1.2.7	ARID	15
4.1.2.8	Tiny Virat	15
4.1.2.9	UAV-Human Dataset	15
4.1.2.10	Mini Something Something V2	15

4.1.2.11	Construction	16
4.1.2.12	Ikea Furniture Assembly	16
4.1.2.13	InHARD - Industrial Human Action Recognition	16
4.1.2.14	”NTU RGB+D” -Masked Depth Maps	17
5	Experimental Results & Analysis	18
5.1	Evaluation Scheme	18
5.1.1	Hyperparameter Sweeps	19
5.1.2	Full Fine-tuning	20
5.1.3	Linear Evaluation	22
5.1.4	Limited Evaluation	24
5.2	Analysis	26
6	Conclusion & Future Scopes	36
6.1	Conclusion	36
6.2	Future Scopes	36
	References	38

List of Figures

1.1	Dataset vs Size	2
1.2	Dataset vs Category	2
2.1	large Scale Datasets vs Deep Learning Models	5
2.2	Workflow of the TSM model	6
2.3	Workflow of the TSN model	6
2.4	Workflow of the SlowOnly model	6
2.5	Workflow of the I3D model	7
2.6	Workflow of the TimeSFormer model	7
2.7	Workflow of the SIFAR model	8
5.1	The Validation Top-1 Accuracy on all six models by applying all combinations of the Hyperparameters on Construction Dataset	19
5.2	Full-Finetuning Scheme	20
5.3	Full Fine-tuning on Construction Dataset	21
5.4	Full fine-tuning Accuracy vs Dataset pertained on Kinetics-400	21
5.5	Full fine-tuning Accuracy vs Dataset pertained on MiT	22
5.6	Linear Evaluation Scheme	22
5.7	Linear Evaluation on ARID Dataset	23
5.8	Linear Evaluaiion Accuracy vs Dataset pretained on Kinetics-400	23
5.9	Linear Evaluaiion Accuracy vs Dataset pretained on Moments in Time	24
5.10	Limited Evaluation Scheme	24
5.11	Limited Evaluation on HMDB51 Dataset	25
5.12	Limited Evaluaiion Accuracy vs Dataset pretained on Kinetics-400	25
5.13	Limited Evaluation On TimeSFormer	26

5.14 Full Fine-tuning VS Linear Evaluation VS Limited Evaluation Using Kinetics-400 Pretraining	27
5.15 Full Fine-tuning using Kinetics-400 vs MiT Pretraining	31
5.16 Linear Evaluation over TimeSFormer using Kinetics-400 vs MiT Pre-training	33
5.17 Limited Evaluation over TimeSFormer using Kinetics-400 vs MiT Pre-training	34

List of Tables

5.1	Performance of different models on the downstream datasets in terms of top-1 accuracy with full fine-tuning sorted by domains pretrained on Kinetics-400	28
5.2	Performance of different models on the downstream datasets in terms of top-1 accuracy with Linear Evaluation sorted by domains pretrained on Kinetics-400	29
5.3	Performance of different models on the downstream datasets regarding top-1 accuracy with Limited Evaluation using Kinetics-400 pretraining	30
5.4	Performance of different models on the downstream datasets regarding top-1 accuracy with full fine-tuning pre-trained on Moments in time.	32
5.5	Performance of different models on the downstream datasets regarding top-1 accuracy with Linear Evaluation pretained on Moments in time.	33
5.6	Comparison between Performance of TimeSFormer on the downstream datasets in terms of top-1 accuracy with Limited Evaluation using Kinetics-400 and MiT Pretraining	34
5.7	Comparison between Performance of TimeSFormer on the downstream datasets in terms of top-1 accuracy with Limited Evaluation using Kinetics-400 and MiT Pretraining	35

Chapter 1

Introduction

In Computer Vision, video action recognition identifies when a person in a video is performing a given action. Video action recognition, an essential task in visual scene understanding, has been an active area of research for a long time.

Over the last decade, there has been growing research interest in video action recognition with the emergence of high-quality, large-scale action recognition datasets [1, 2, 3, 4, 5]. We compare the accuracy and efficiency of frequently used approaches on the same set of datasets. Additionally, we provide our implementations for complete repeatability. Towards this end, we conduct a large-scale study regarding deep learning models on video action recognition, linear evaluation, and limited evaluation. We precisely benchmark six deep learning models with fourteen target datasets pre-trained on two large-scale source datasets. To accelerate future research, we detail this area’s difficulties, unsolved issues, and potential.

The accuracy of deep learning models often increases as the training data amount grows. This indicates that we require large-scale annotated datasets to develop efficient models for video action recognition. It has also been observed that after pre-training with the same large-scale dataset, the performance varies significantly with downstream datasets when they are from diverse domains. For this reason, we explored datasets from six different domains. The domains are Human Action, Sports, IR Videos, Low- Resolution, Industrial Action Related Videos, and Masked Videos. For training and evaluation, we have chosen three types of deep learning models, e.g., the 2D CNN models, the 3D CNN models, and the transformer models. The size and the categories of these datasets are compared in (Figure 1.1) and (Figure 1.2).

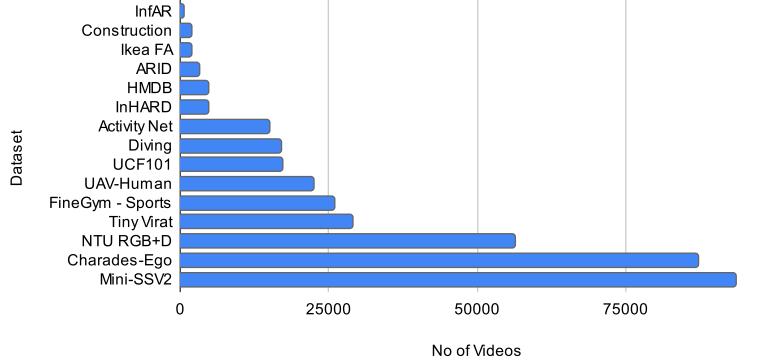


Figure 1.1: Dataset vs Size

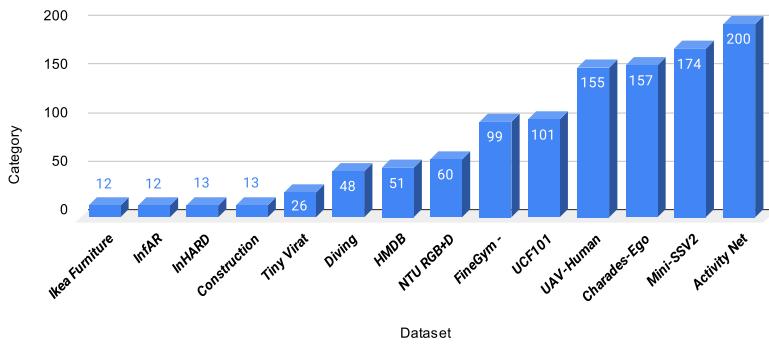


Figure 1.2: Dataset vs Category

Despite the existence of such deep learning-based models for video action recognition, there needs to be a standard benchmarking for these models. Proposed Video Action Recognition Benchmark (VARB) would define good representations as adapting to diverse tasks with/without few examples. We compare the performance and accuracy of these models using three distinct kinds of evaluation schemes:

- Full fine-tuning: All the parameters of the model are learned using the full training set.
- Linear evaluation: Only the last layer parameters of the model is learned using the full training set.
- Limited evaluation: All the parameters of the model are learned using 5 percentile data for each class of the training set.

We perform the experiments on 14 datasets. Before performing these operations, the

models were pretrained on upstream datasets such as **Kinetics-400** and **Moments in Time**.

Chapter 2

Related Work

This section reviews available action recognition benchmark datasets and deep learning models in this domain. Bibliophiles are referred to these survey papers for a more extensive list of action recognition datasets and deep learning models ([6, 7, 8, 9, 10, 5]).

Due to the availability of massive datasets and the rapid development of deep learning models, there has been a significant rise in deep learning-based models for video action recognition. Considering Moments in Time [2] and Kinetics [3] as case studies, large video models are trained on these massive datasets before being fine-tuned on smaller action detection datasets like UCF101 [6], HMDB [11], FineGym - Sports [12], ActivityNet [5], UAV-Human Dataset [7], ARID [13], Diving [10], Ikea Furniture Assembly [2], InfAR [14], InHARD - Industrial Human Action Recognition Dataset [15], Tiny Virat [10], Charades-Ego, "NTU RGB+D" -Masked Depth Maps [16], and "Mini Something Something V2" [17]. The models can transfer their learned features to smaller sets and develop condensed representations from the wide range of videos in massive datasets through fine-tuning.

The Kinetics 400 [3] dataset is described on the deep mind website as follows: "A large-scale, high-quality dataset of URL links to approximately 650,000 video clips that cover 400 human action classes, including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. Each action class has at least 400 video clips. Each clip is human annotated with a single action class and lasts around 10s."

Moments in Time [2], an extensive human-annotated collection of one million brief videos representing dynamic events occurring within three seconds, was released in

2018. One million 3-second video snippets are labeled with a dictionary of 339 classes.

Here is a timeline of recent significant contributions in (Figure 2.1).

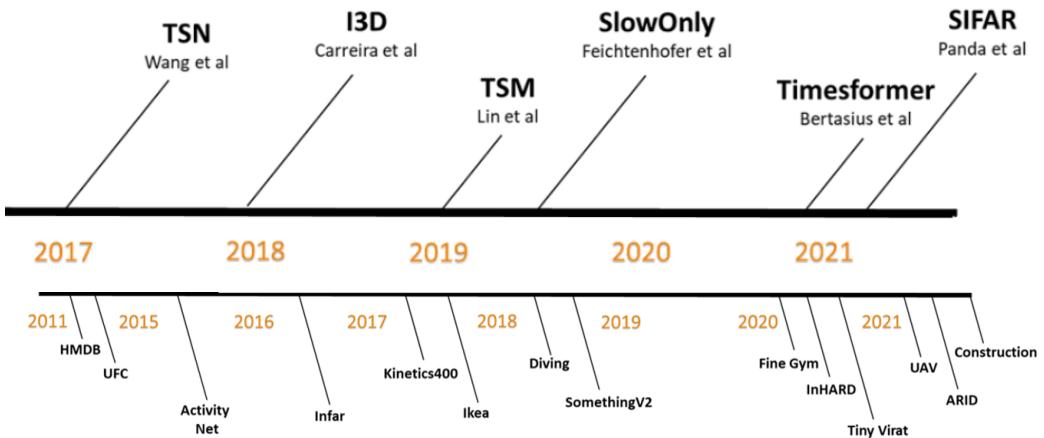


Figure 2.1: large Scale Datasets vs Deep Learning Models

One of the early attempts to use convolutional neural networks with videos was DeepVideo [18]. In the past, it has been popular to model temporal visual information using 3D convolutional kernels, such as I3D, R3D, S3D, Non-local, SlowFast, SlowOnly, etc. This field’s rapid development focused on increasing computing efficiency so that it could expand to increasingly more enormous datasets and be employed in practical applications. Examples are TSM [9],, Hidden TSN [8],, and others. The next big thing is migrating from CNN, categorizing 0–9 numbers for better understanding using Transformers. Finally, transformer-based models like TimeSFormer and SIFAR became the movement’s main focus.

2D MODELS

The Temporal Shift Module (TSM) [9] (Fig. 2.2) is a versatile deep-learning model with excellent performance and efficiency. In particular, it can perform on par with 3D CNN while keeping the complexity of 2D CNN. By moving some channels along the temporal axis, TSM helps information flow between neighboring frames. It may be integrated into 2D CNNs for temporal modeling without computation and parameters. The application of TSM to an online context is enhanced by offering real-time low-latency online video identification and object recognition.

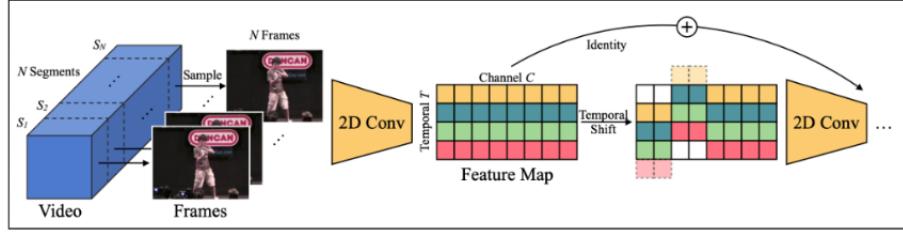


Figure 2.2: Workflow of the TSM model

Temporal Segment Network (TSN) [8] (Fig. 2.3) is a cutting-edge system for recognizing actions in videos. It is founded on the notion of modeling long-range temporal structure. It combines a sparse temporal sampling strategy with video-level supervision to make learning from the entire action video efficient and effective.

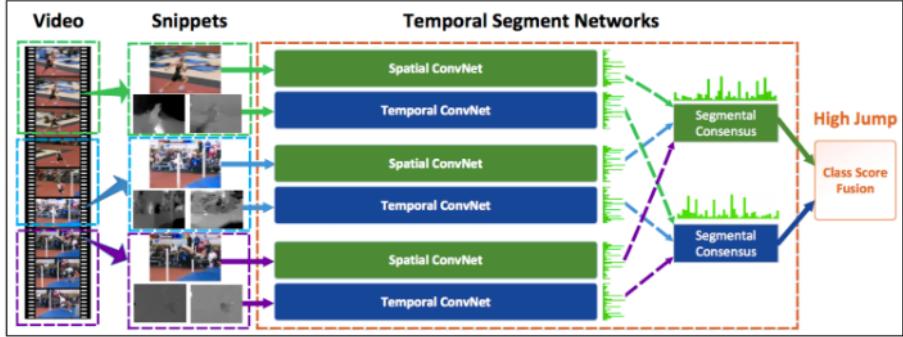


Figure 2.3: Workflow of the TSN model

3D MODELS

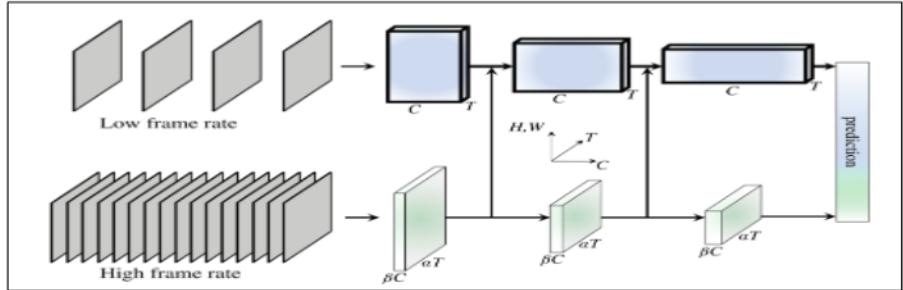


Figure 2.4: Workflow of the SlowOnly model

SlowOnly (Fig. 2.4) is a SlowFast network for video identification. In this approach, there is a Fast pathway that operates at a high frame rate to record motion

with adequate temporal precision and a Slow pathway that uses a low frame rate to capture spatial semantics. By lowering its channel capacity, the Fast route may be very light while learning relevant temporal information for video identification. The models perform well for both classification and detecting actions in videos. We report cutting-edge accuracy on the most important video recognition benchmarks, Kinetics.

I3D (Fig. 2.5) is a Two-Stream Inflated 3D ConvNet (I3D), which is based on 2D ConvNet inflation. By expanding very deep image classification ConvNets' filters and pooling kernels into 3D, I3D makes it possible to learn seamless spatio-temporal feature extractors from video while utilizing successful ImageNet architecture designs and their parameters.

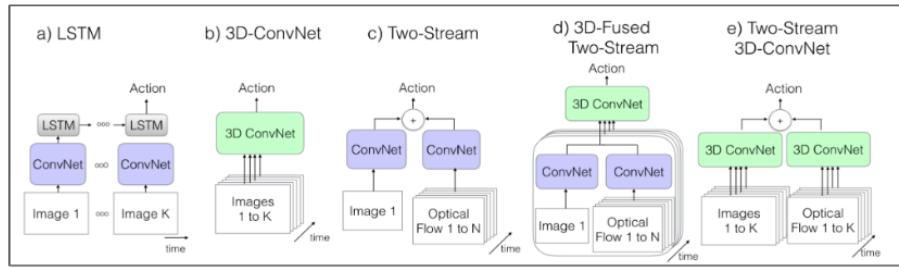


Figure 2.5: Workflow of the I3D model

TRANSFORMER BASED MODELS

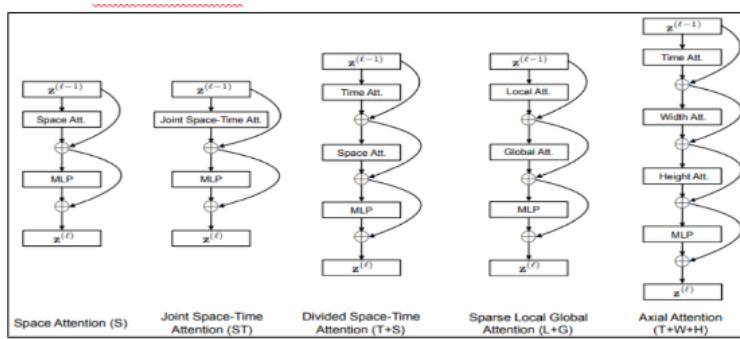


Figure 2.6: Workflow of the TimeSFormer model

TimeSformer (Fig. 2.6) expands the Transformer architecture's capabilities to video

by allowing spatiotemporal feature learning straight from a series of frame-level patches. The model based on split attention, where both temporal and spatial attention are independently applied within each block, yields the most incredible video classification accuracy among the design choices considered.

SIFAR (Fig. 2.7) recasts the video recognition issue as an image recognition challenge; they investigate a different angle on video action recognition. The method rearranges the video frames used as input into super pictures, enabling the training of an image classifier to perform the action identification task in the same manner as image classification. They demonstrate that using such a straightforward notion, a transformer-based image classifier alone can be sufficient for action recognition.

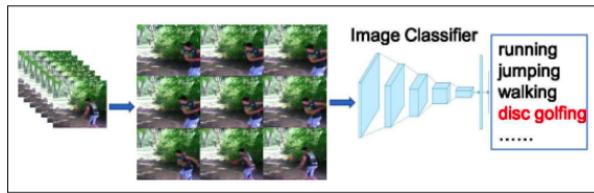


Figure 2.7: Workflow of the SIFAR model

Particularly, their strategy outperforms SOTA techniques on a number of open datasets, including Kinetics400, Moments in Time, Something-Something V2 (SSV2), Jester, and Diving48.

Chapter 3

The Benchmarking in Video Action Recognition

The transferability of learned representations from majorly three types of models(2D,3D and Transformer) is extensively examined in this work using 14 downstream datasets covering six different domains.

With three training protocols per task, the search for algorithms that excel on a wide range of video action recognition challenges. We establish this objective and specify a workable benchmarking process to gauge advancement.

A dataset D is a set of n instances with observations(i.e., clips) $x_i \in X$ and labels $y_i \in Y$. A prediction function is a classifier $F : X \rightarrow Y$, which maps x_i to y_i . An algorithm, A , takes as input a dataset and outputs a prediction function. For instance, A might be a pre-trained network attached to a training procedure. We use a technique for evaluation(E), which inputs F and produces a scalar measuring F 's performance (e.g., top1 accuracy, F1 score, etc.). A task T is a tuple with a task-specific dataset distribution D and assessment method. We are looking for a technique that optimizes predicted performance over the distribution of tasks P .

Given a source dataset D_s with n instances $\{(x_{si}, y_{si})\}$ where $1 \leq i \leq n$ with a marginal distribution M_s and a target D_t n instances $\{(x_{ti}, y_{ti})\}$ where $1 \leq i \leq n$ with a marginal distribution M_t , where (x_i, y_i) is the video and action pair, and generally $M_s \neq M_t$. Learning a target prediction function f using knowledge of D_s is the goal of transfer learning. We research several target prediction tasks, such as full-network fine-tuning, linear evaluation, and limited evaluation for video action recognition.

3.1 Tasks

The tasks in the VARB span a variety of domains and semantics. HUMAN ACTION, SPORTS, IR/DARK VIDEOS, LOW-RESOLUTION, INDUSTRIAL ACTION, and DEPTH VIDEOS are the six categories into which these are divided.

- The HUMAN ACTION group represents the classical vision challenge. This task contains a number of human actions either taken from movie clips or captured. The group comprises UCF101, HMDB, and Activity Net.
- The SPORTS group has videos of sports action (mainly gymnasium and diving) performed by trained professionals.
- The IR/DARK VIDEOS contain videos taken in a dark setting where recognizing actions by models becomes difficult.
- The LOW-RESOLUTION group has clips taken from above the ground. We include UAV-Human Dataset and Tiny Virat here.
- The INDUSTRIAL ACTION Related Videos group contains clips related to industry and construction. We include Construction(homegrown dataset), In-HARD, and Ikea.
- The DEPTH VIDEOS group, where the depth/masked videos are considered.NTU Depth Masked Videos includes this group.

3.2 Transfer Strategy

The representations must be modified to complete the new tasks because they were not pretrained on the evaluation tasks. In deep representation learning, pre-training the network is followed by,

- Full-finetuning - where the whole model is tuned with the new task data,
- Linear Evaluation - where the model’s weights are frozen, allowing only the last linear layer to train and
- Limited Evaluation - where the full models are tuned but using limited in-domain data.

The original weights can be adjusted more effectively when large differences exist between the upstream and downstream datasets [19, 20]. VARB does not constrain the transfer approach; in this case, we choose fine-tuning because it typically works the best.

Chapter 4

Large-Scale Study

4.1 Dataset

The accuracy of deep learning techniques often increases as training data volume grows. This indicates that we require large-scale annotated datasets to develop efficient models for video action recognition.

Datasets for video action recognition are frequently created using the method described below:

- (1) Depending on the use case, create an action list by merging labels from earlier action recognition datasets and adding additional categories.
- (2) Obtain videos by matching the video title and subtitle to the action list from various websites, including YouTube and motion pictures.
- (3) Finally, de-duplicate and filter the dataset after manually adding temporal annotations to show where the activity begins and ends.

4.1.1 Upstream Data

4.1.1.1 Kinetics - 400

In 2017, Kinetics400[3] —a collection of 240k training and 20k validation videos with 400 categories of human action was released. The Kinetics family has continued to grow, with the introduction of the 480K video Kinetics-600 in 2018 and the 650K video Kinetics-700 in 2019.

4.1.1.2 Moments In Time

Moments in Time[2], a large dataset created for event analysis, was launched in 2018. One million 3-second video clips are labeled with a dictionary of 339 classes. The Moments in Time collection includes data on people, animals, objects, and natural events, contrasting previous datasets designed to explain human action. By extending the number of videos to 1.02 million, removing ambiguous classes, and increasing the number of labels for each video, the dataset was expanded to Multi-Moments in Time (M-MIT) in 2019.

4.1.2 Downstream Data

HUMAN ACTION

4.1.2.1 HMDB

In 2011, HMDB51 [11] was released. It was mainly compiled from videos, with a tiny portion coming from open-access sources like YouTube, Google Videos, and the Prelinger Archives. The dataset consists of 6,849 clips, grouped into 51 action categories, each with at least 101 clips. Three documented splits exist for the dataset. Most previous studies provide the top-1 classification accuracy on split one or the average accuracy across three splits.

4.1.2.2 UCF101

An expansion of the earlier UCF50 dataset, UCF101[6], was released in 2012. 13,320 YouTube videos from 101 different human activity categories are included. The dataset was analyzed similarly and has three formal divides like HMDB51 [11].

4.1.2.3 Activity Net

ActivityNet [5] was first presented in 2015, and since then, the ActivityNet family has gone through repeated editions. Two hundred actions from a person's everyday life are included in the most current version of ActivityNet 200 (V1.3). Videos for training, validation, and testing total 10,024; 4,926; and 5,044, respectively. Each class typically contains 137 untrimmed videos and 1.41 activity instances.

SPORTS

4.1.2.4 FineGym - Sports

A dataset with an emphasis on gymnastics videos named FineGym [12] was released in 2020. Due to its high-quality and action-centric data, consistent annotations at various temporal and semantic granularities, and a variety of interesting action instances, FineGym stands apart from other action recognition datasets in a number of ways.

4.1.2.5 Diving

Diving48 is a fine-grained video dataset comprising 48 distinct dive sequences from competitive diving. It contains around 18,000 trimmed video clips. Modern action recognition algorithms find this a difficult challenge since dives might vary in three phases (takeoff, flight, and entry), necessitating the modeling of long-term temporal dynamics.

IR AND DARK VIDEOS

4.1.2.6 InfAR

The construction of the Infrared dataset [14], involved gathering videos of 12 typical human actions. Handclapping, walking, running, jumping, skipping, handshakes, hugs, pushing, punching, and fighting are among the actions kinds. Wave1 and Wave2 are one-hand waves. 30 videos are available for each movement type. Each action is carried out by 25 distinct people. The movies are taken using a handheld infrared camera, model IR300A. The average duration of each clip is 4 seconds. With a resolution of 293x256, the frame rate is 25 frames per second. In every video, a single individual or a group of people is seen carrying out one or more acts. Interactions between different people make up some of them.

4.1.2.7 ARID

Over 3,780 video clips with 11 different action classes make up the Action Recognition in the Dark (ARID) [13] dataset. The first dataset is devoted to analyzing human action in low-light videos. The video datasets' randomly recorded sequences may not feature identifiable human activities, making them unsuitable for action recognition. Contrarily, the ARID dataset classifies various human actions in dark videos.

LOW-RESOLUTION

4.1.2.8 Tiny Virat

The TinyVIRAT dataset [10], based on the VIRAT dataset, was created by clipping small action clips from the VIRAT movies to solve real-world tiny action detection issues. With 26 action labels, TinyVIRAT has 7,663 training and 5,166 test videos. Since the activities in TinyVIRAT videos have multiple labels and were taken from surveillance footage, they are more challenging and realistic.

4.1.2.9 UAV-Human Dataset

A large dataset known as UAV-Human [7] includes 67,428 annotated video sequences of 119 subjects for action recognition, 22,476 annotated frames for position estimation, 41,290 annotated frames of 1,144 identities for person re-identification, and 22,263 annotated frames for attribute recognition. Using a DJI Matrice 100 platform, the dataset was gathered. UAV-Human is the most critical, complex, and complete UAV dataset for comprehending human activity, position, and behavior.

4.1.2.10 Mini Something Something V2

In 2018, Something-Something V2 [21] was released. Another well-liked benchmark, this family has 174 action classes illustrating how people use ordinary objects to accomplish simple tasks. V2 has 220,847 videos. Since most actions cannot be predicted based just on physical parameters, it should be noted that the Something-Something dataset requires strong temporal modeling. Mini Something Something V2 is a part

of the Something-Something V2 dataset(basically 50% of the original dataset).

INDUSTRIAL ACTION RELATED VIDEOS

4.1.2.11 Construction

There are many challenging benchmark datasets for action recognition, as earlier mentioned, but not in the construction area. But in our perspective, these datasets are too demanding for a challenge like action recognition. This is because most of the video benchmarks currently in use comprise construction-related actions and feature sequences with construct equipment-related activities. Because of this, it could be challenging to recognize or distinguish actions among various classes of the Construction domain(which needs to be explored more).

To achieve this, we propose a new dataset named Construction, which includes all the fundamental tasks associated with the construction industry, including bulldozing, excavating, lifting, loading, moving, digging, lowering, paving, carrying, assembling, concreting, carpentry, and unloading. With this dataset, we expect to anticipate a subject’s equipment usage once a few frames have been viewed that show a construction activity. The collection contains roughly 1815 films, each lasting between 2 and 30 minutes and being captured at 30 frames per second. There are 13 distinct classes assigned to this dataset. The YouTube8M dataset, released in 2016 and by far the largest-scale video collection, comprises 8 million YouTube videos and is annotated with 3,862 action classes. This dataset covers particular construction-based activities that were gathered from that dataset.

4.1.2.12 Ikea Furniture Assembly

The Ikea Furniture Assembly (Ikea FA) dataset captures the fundamental steps people take to put together a modest Ikea piece of furniture. After a few observed frames that feature an assembly motion, this dataset forecasts the subject’s stance. The dataset contains 101 films, each taken at 30 frames per second and lasted between 2-4 minutes. Twelve labels make up the dataset.

4.1.2.13 InHARD - Industrial Human Action Recognition

“Industrial Human Action Recognition Dataset” (InHARD) [15] from a real-world setting for industrial human action recognition with over 2 million frames collected

from 16 distinct subjects. This dataset has more than 4800 action samples, including 13 distinct industrial action classifications. The dataset is based on a real-world use case from an industrial setting. In this use case, the robotic arm UR10 is used to assemble various parts and components at various stages. This latter device is a Collaborative Industrial Robotic Arm designed for large-scale jobs, automating processes including packing, palletizing, assembling, and pick-and-place.

DEPTH VIDEOS

4.1.2.14 "NTU RGB+D" -Masked Depth Maps

A large-scale RGB+D human action recognition dataset [16] with more than 56 thousand video samples and 4 million frames was collected from 40 distinct subjects. 60 activity classifications are included in this dataset, encompassing everyday, interpersonal, and medical acts. This large-scale dataset can be used as data-driven learning techniques, such as Long Short-Term Memory networks, to address this issue and obtain performance accuracy that is superior to hand-crafted features because of the size of the acquired data.

Chapter 5

Experimental Results & Analysis

In this section, we compare popular approaches on selected benchmark datasets. To be specific, we first introduce standard evaluation schemes. Then we divide common benchmarks into different domains such as human action, sports, IR videos, low-resolution, dark videos, industrial action-related videos, and depth videos. Ultimately, we present a fair comparison of recognition accuracy among 2D,3D, and transformer base deep learning models.

5.1 Evaluation Scheme

The first stage is running six models on 14 datasets using the hyperparameter-sweep to pick the appropriate learning rate and weight decay for performing three operation types: full-finetuning, Linear Evaluation and Limited Evaluation on the pre-trained models created from large-scale upstream dataset kinetics-400 [3] & Moments in Time [2]. (Fig. 5.1) illustrates how hyperparameter sweep plays a vital role in achieving these three types of operations.

For setting the standard benchmark, the selected dataset is full-finetuned with all six models with the achieved learning rate and weight decay from hyperparameter-sweep to check which model performs better in accuracy. After completing the full-finetuning, the following experiment is Linear Evaluation, whether all the layers are frozen except the last layer and perform training and evaluation.Finally, we take on Limited Evaluation where we will take only 5% of training data across all classes of a target dataset while training with a model and evaluate the complete testing dataset for the standard benchmark judgment.

5.1.1 Hyperparameter Sweeps

One must search through high dimensional hyperparameter spaces to rapidly locate the best effective model, which might become expensive. The most effective technique to run a model competition and select the most accurate model is to use hyperparameter sweeps. They make this possible by automatically aggregating hyperparameter values (learning rate, weight decay) to determine the best values.

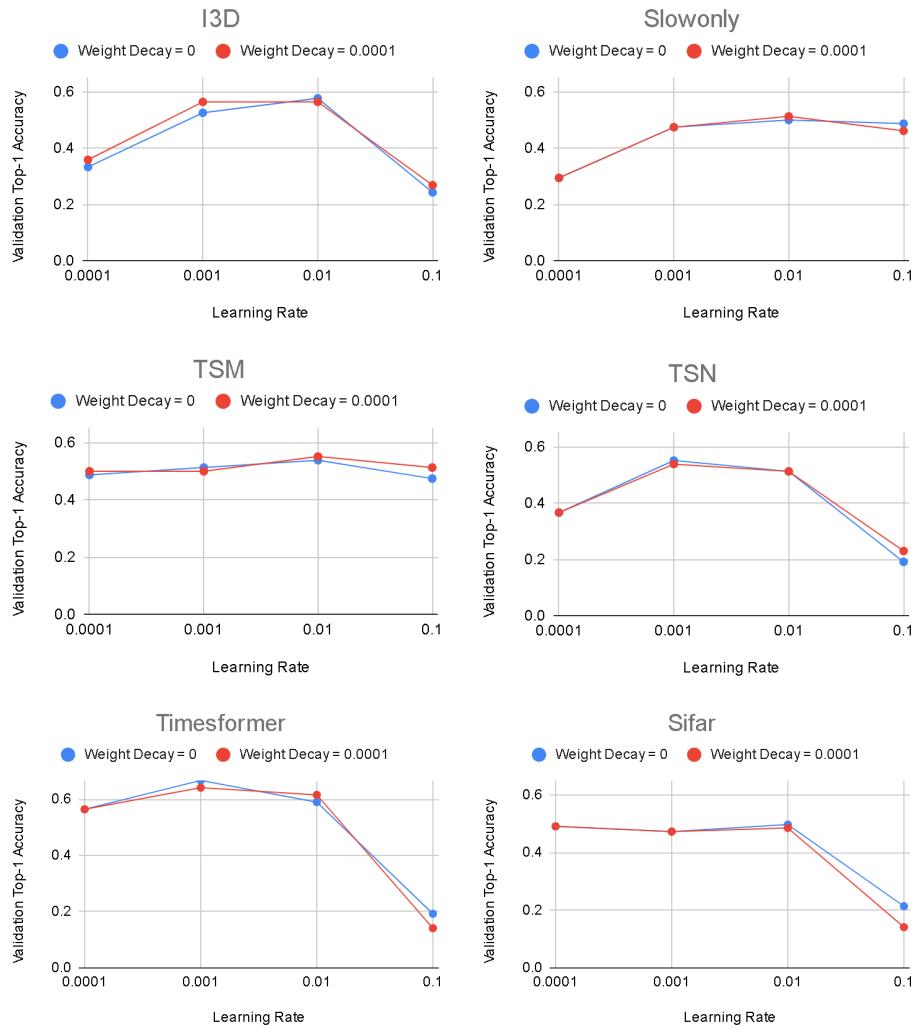


Figure 5.1: The Validation Top-1 Accuracy on all six models by applying all combinations of the Hyperparameters on Construction Dataset

Weight decay is a regularization technique by adding a minor penalty, usually the L2 norm of the weights (all the weights of the model), to the loss function. Hyperparameter sweeps make it feasible to identify the optimal combinations of hyperparameter values for models on a given dataset.

We separated the validation set from the training set (using 95% as training and 5% as validation from the original training set) and swept the hyperparameters (learning rate, batch size, and weight decay) for each dataset before applying transfer learning to the downstream datasets. The whole training set (train+val) is then trained using the best hyperparameters, and the test set is used for evaluation.

The variations of learning rate as [0.1, 0.001, 0.001, 0.0001] and two values for weight decay [0, 0.0001] are considered for the experiments with fixed batch size. A model and a dataset will have $4 \times 2 = 8$ experiments in a hyper-parameter sweep to find the best parameters.

5.1.2 Full Fine-tuning

To get the desired output or performance, full fine-tuning (Fig. 5.2) typically involves minor adjustments to a process. By full fine-tuning a model that has already been trained for a specific activity, it may now execute a second task that is comparable to the first. Full fine-tuning consists of unfreezing and re-training the entire model on new data.

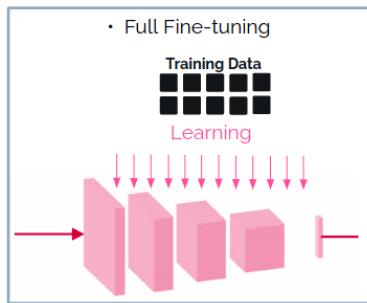


Figure 5.2: Full-Finetuning Scheme

The complete training and the testing dataset are used for full fine-tuning, which uses the parameters obtained from the hyper-parameter sweep with seed values of

$[0, 50, 100]$ for training and validation. Then record the model’s final benchmark value over the dataset as the average accuracy value from these three experiments.

The experiments of full fine-tuning for the Construction Dataset over all models pre-trained on Kinetics-400 are featured in (Figure 5.3).



Figure 5.3: Full Fine-tuning on Construction Dataset

We perform the $6 \times 14 \times 3$ experiments for six deep learning models over 14 datasets with different seed values of $[0, 50, 100]$ for standardizing the benchmark whether the models were pre-trained twice from Kinetics-400 [3] and Moments in Time [2]. The results of full-finetuning pre-trained on Kinetics-400 now are figured in (figure 5.4), where in terms of datasets UCF and in terms of model, SlowOnly is performing better.

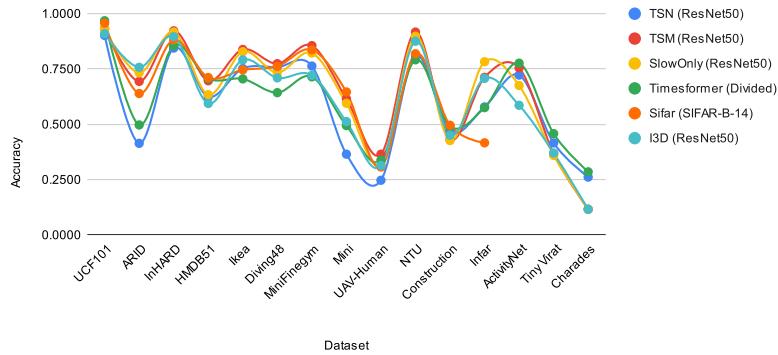


Figure 5.4: Full fine-tuning Accuracy vs Dataset pertained on Kinetics-400

After pre-trained in the Moments in Time dataset, the results of full-finetuning showed some interesting facts. The graphical view is figured in (figure 5.5), where

in terms of datasets UCF & NTU and terms of model, TimeSFormer is performing better.

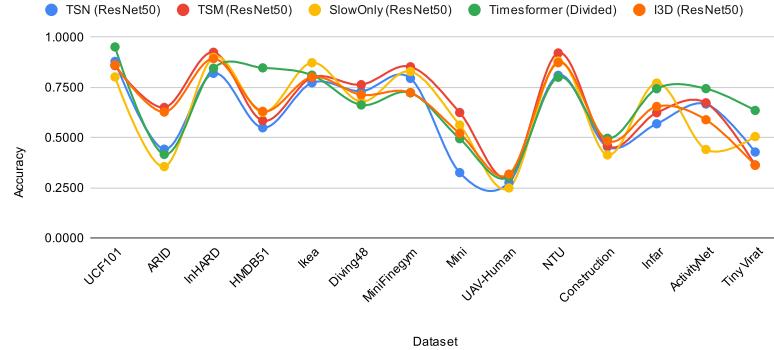


Figure 5.5: Full fine-tuning Accuracy vs Dataset pertained on MiT

5.1.3 Linear Evaluation

Linear Evaluation (Fig. 5.6) is a process where we perform both hyper-parameter sweep and the same as like full-fine-tuning task but only by taking the model’s last layer and freezing all layers except that layer.

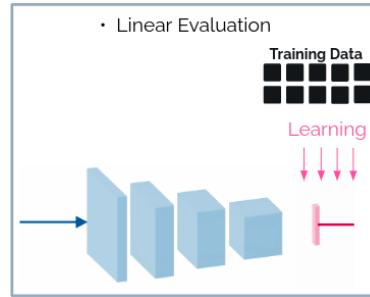


Figure 5.6: Linear Evaluation Scheme

In terms of cost efficiency, Linear Evaluation reduces the model’s sample and performs better cost efficiently than full finetuning. We split the training dataset into 60:40 to obtain the training and validation dataset for the hyper-parameter sweep. Finally, we serve 4×2 (learning rate \times weight decay) experiments of hyper-parameter sweep with frozen layers for a model over a dataset. With the best parameters resulting from the hyper-parameter sweep, we go through the Linear Evaluation over the complete

training and validation dataset. The Linear Evaluation of the ARID dataset is stated in (Figure 5.7).

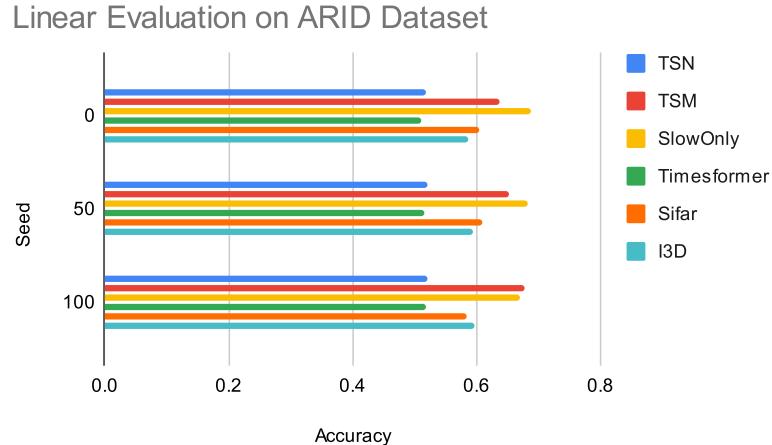


Figure 5.7: Linear Evaluation on ARID Dataset

We perform the experiments for six deep learning models over 14 datasets pertained in Kinetics & Moments in Time. The overall results of Linear Evaluation pertained in Kinetics-400 are figured in (figure 5.8).

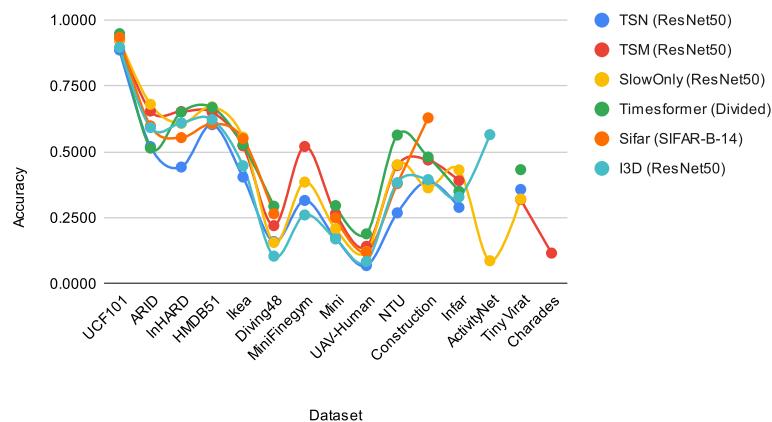


Figure 5.8: Linear Evaluation Accuracy vs Dataset pretrained on Kinetics-400

The overall results of Linear Evaluation pertained in Moments in Time till now are figured in (figure 5.9).

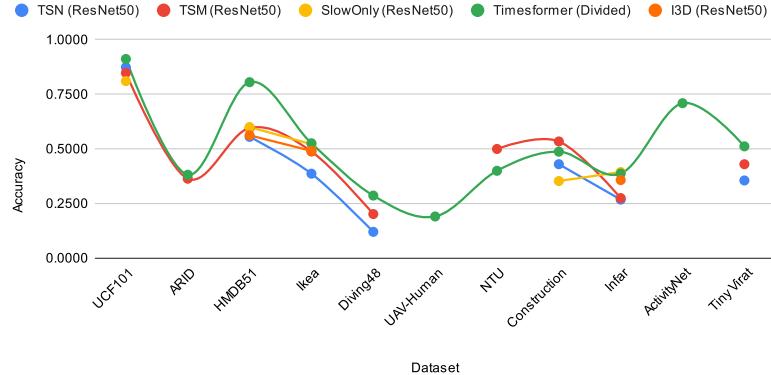


Figure 5.9: Linear Evaluation Accuracy vs Dataset pretrained on Moments in Time

5.1.4 Limited Evaluation

Limited Evaluation (Fig. 5.10) is a type of evaluation in which we use just 5% of the training dataset's data from all classes and the best-performing hyper-parameter provided by full-fine-tuning for training to evaluate the complete validation dataset to determine how well-trained our model is to handle it. Low execution costs result from Limited Evaluation. Therefore, the method will be highly cost-effective for the dataset if Limited Evaluation is successful.

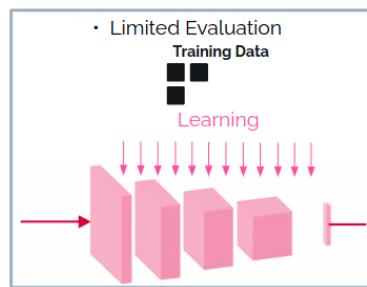


Figure 5.10: Limited Evaluation Scheme

We take three different 5% splits from the training dataset and experiment with seed values of [0,50,100] alike full fine-tuning for a model over a dataset. Linear Eval-

uation for the HMDB dataset is stated in (Figure 5.11).

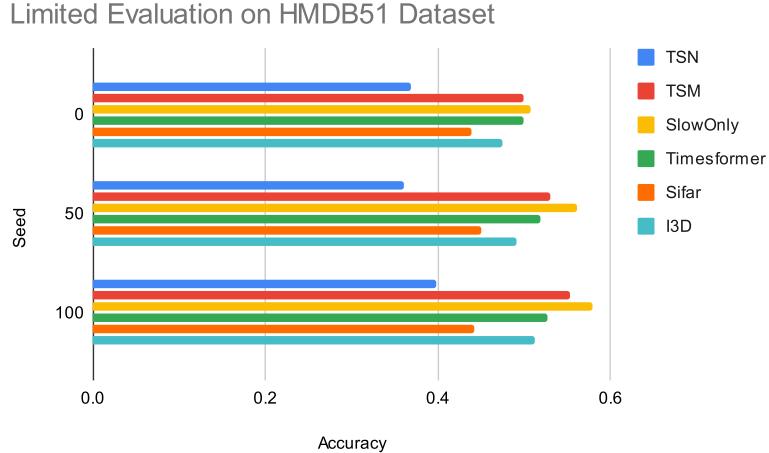


Figure 5.11: Limited Evaluation on HMDB51 Dataset

Limited Evaluation is very cost-efficient compared to Linear Evaluation and full fine-tuning as it's trained in a minimal-size dataset. We perform the $6 \times 14 \times 3$ experiments for six deep learning models over 14 datasets. The overall Limited Evaluation pretrained on Kinetics-400 results are figured in (figure 5.12).

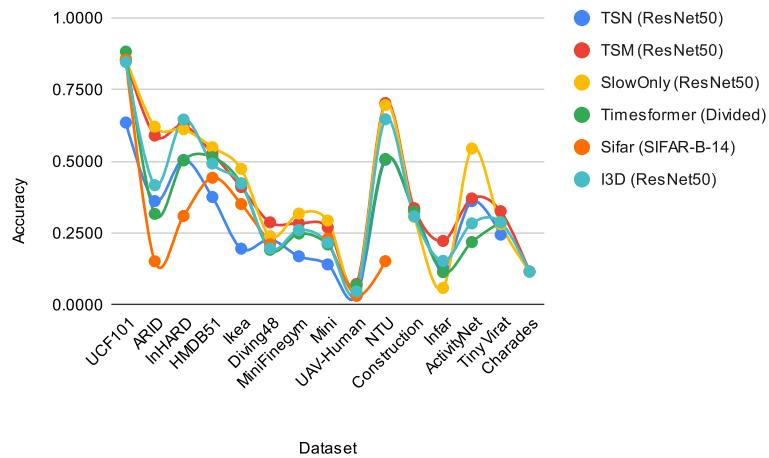


Figure 5.12: Limited Evaluation Accuracy vs Dataset pretrained on Kinetics-400

From the full fine-tuning and Linear Evaluation over models pre-trained on the MiT dataset, it's noticed that TimeSFormer is the best-performing model across six models. So we conducted a Limited Evaluation on TimeSFormer pre-trained on MiT. Here is the comparison between TimeSFormer pre-trained in Kinetics-400 and Moments in Time figured in (figure 5.13).

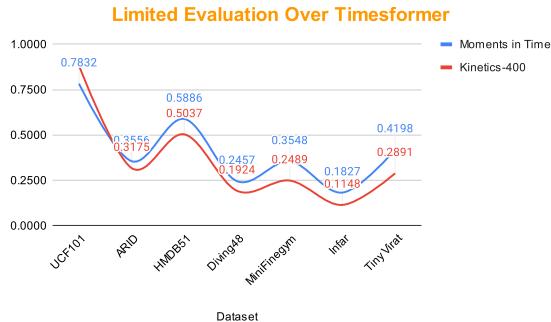


Figure 5.13: Limited Evaluation On TimeSFormer

5.2 Analysis

Let's look at (Figure 5.14) where the comparison between Full-finetuning, Linear Evaluation, and Limited Evaluation has been sketched.

From Figure 5.14 , we closely look at the accuracy of different datasets performed on the TSN model. We will notice a few facts, such as in terms of the ARID dataset, Linear Evaluation outperforms Full Fine-tuning, and Limited Evaluation is in the same pitch as Full Fine-tuning. The margin of accuracy for UCF, HMDB, Construction, and Tiny Virat datasets is very low for Linear and Full-finetuning. These datasets, such as UCF-101 and HMDB-51, share high similarities with Kinetics-400 regarding actions. Thus the performance of these datasets may only partially reflect the effectiveness of the evaluated model.

The experiments with TSN also experience the same kind of facts as for UCF, ARID, InHard, Construction, and TinyVirat. Full fine-tuning and Linear Evalu-

tion are getting more than near accuracy, and Limited Evaluation is touching the benchmark a few times.

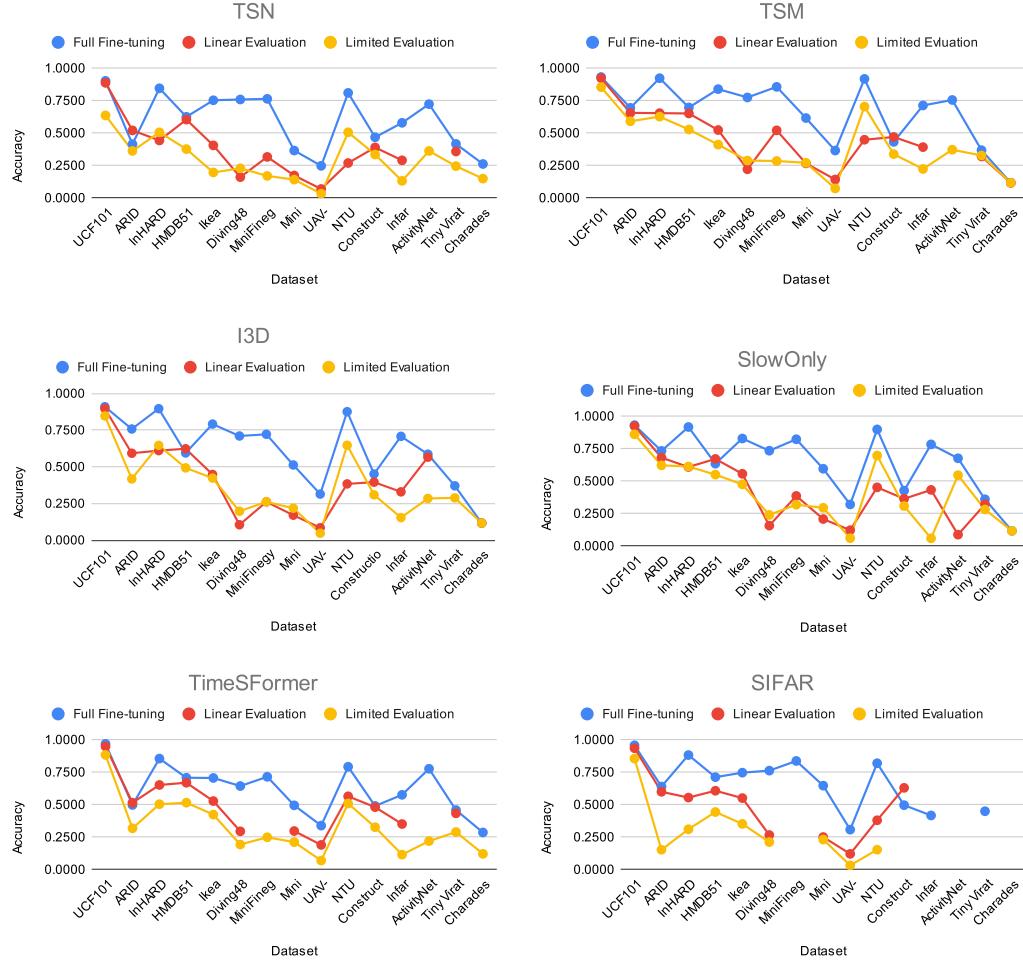


Figure 5.14: Full Fine-tuning VS Linear Evaluation VS Limited Evaluation Using Kinetics-400 Pretraining

The same scenario is repeated with I3D, SlowOnly, TimeSFormer, and SIFAR. For I3D, full fine-tuning and Linear Evaluation accuracy is quite similar for UCF, HMDB, Construction, and TinyVirat Dataset. For SlowOnly, the accuracy of full fine-tuning and Linear Evaluation is similar for UCF, ARID, HMDB, Construction, ActivityNet, and TinyVirat Dataset.

For TimeSFormer, the accuracy of full fine-tuning and Linear Evaluation is similar for UCF, ARID, HMDB, Construction, and TinyVirat Dataset. For SIFAR, the

Full Fine-tuning Using Kinetics-400 Preatining						
Dataset	TSN	TSM	SlowOnly	Timesformer	Sifar	I3D
Human Action						
UCF101	0.9019	0.9316	0.9308	0.9682	0.9572	0.9098
HMDB51	0.6250	0.6970	0.6342	0.7072	0.7118	0.5945
ActivityNet	0.7226	0.7549	0.6758	0.7765	0.4881	0.5861
Mini SS V2	0.3653	0.6156	0.5952	0.4946	0.6465	0.5125
Sports						
Diving48	0.7585	0.7743	0.7342	0.6433	0.7621	0.7101
MiniFinegym	0.7636	0.8558	0.8223	0.7148	0.8366	0.7217
IR and Dark Videos						
Infar	0.5789	0.7124	0.7829	0.5757	0.4167	0.7078
ARID	0.4143	0.6942	0.7327	0.4974	0.6395	0.7578
Low-Resolution Videos						
UAV-Human	0.2474	0.3652	0.3201	0.3385	0.3075	0.3141
Tiny Virat	0.4161	0.3680	0.3600	0.4582	0.4492	0.3704
Industrial Action Related Videos						
InHARD	0.8448	0.9227	0.9170	0.8550	0.8821	0.8963
Ikea	0.7526	0.8385	0.8282	0.7055	0.7464	0.7914
Construction Videos						
Construction	0.47	0.43	0.4268	0.4898	0.4959	0.4512
Masked Depth Videos						
NTU	0.8094	0.9172	0.8984	0.7919	0.8191	0.8758

Table 5.1: Performance of different models on the downstream datasets in terms of top-1 accuracy with full fine-tuning sorted by domains pretrained on Kinetics-400.

accuracy of full fine-tuning and Linear Evaluation is identical for UCF, ARID, and, Construction.

If we shift from Full Fine-tuning to Linear Evaluation, we see a significant variation in batch size while maintaining the same space for all Deep Learning Models. When using linear evaluation, the batch size significantly grows. On average, we could schedule five times more batches in Linear Evaluation than in full fine-tuning models.

Linear Evaluation Using Kinetics-400						
Human Action						
Dataset	TSN	TSM	SlowOnly	Timesformer	Sifar	I3D
UCF101	0.8868	0.9236	0.9240	0.9487	0.9347	0.8972
HMDB51	0.6035	0.6509	0.6708	0.6690	0.6072	0.6231
ActivityNet	0.5586	0.6995	0.6114	0.7185	0.5384	0.5657
Mini SS V2	0.1728	0.2663	0.2078	0.2967	0.2504	0.1702
Sports						
Diving48	0.1603	0.2207	0.1568	0.2944	0.2658	0.1049
MiniFinegym	0.3162	0.5209	0.3858	0.5957	0.4706	0.2920
IR and Dark Videos						
Infar	0.2898	0.3915	0.4320	0.3507	0.4167	0.3294
ARID	0.5205	0.6553	0.6808	0.5148	0.5990	0.5926
Low-Resolution Videos						
UAV-Human	0.0692	0.1421	0.1217	0.1896	0.1207	0.0846
Tiny Virat	0.3579	0.3195	0.3215	0.4329	0.4056	0.3089
Industrial Action Related Videos						
InHARD	0.4428	0.6533	0.6085	0.6516	0.5545	0.6105
Ikea	0.4049	0.5233	0.5562	0.5276	0.5505	0.4479
Construction Videos						
Construction	0.3892	0.4695	0.3638	0.4807	0.6294	0.3953
Masked Depth Videos						
NTU	0.2692	0.4484	0.4517	0.5642	0.3798	0.3836

Table 5.2: Performance of different models on the downstream datasets in terms of top-1 accuracy with Linear Evaluation sorted by domains pretrained on Kinetics-400

Kinetics-400 [3] is a large-scale, high-quality dataset covering a diverse range of human actions. Its commonly used for pretraining models. For constructing **VARB**, we tried to see whether the number of classes matters or the number of data points. That's why we have chosen Moments in Time [2]. Compared with Kinetics-400, Moments in Time has a 1,000,000 number of videos which is greater than Kinetics-400 (contains 234,584 videos), and has less number of classes than Kinetics-400. After performing Full-fine tuning, Linear Evaluation & Limited Evaluation over models pre-trained on Kinetics-400, We perform all types of experiments using MiT pretraining. We analyze the performance of models in terms of pre-training the models.

Limited Evaluation Using Kinetics-400 Pretraining						
Dataset	TSN	TSM	SlowOnly	Timesformer	Sifar	I3D
Human Action						
UCF101	0.6358	0.8536	0.8601	0.8830	0.8548	0.8468
HMDB51	0.3764	0.5273	0.5498	0.5151	0.4433	0.4928
ActivityNet	0.3619	0.3715	0.5454	0.2194	0.5291	0.2844
Mini SS V2	0.1413	0.2706	0.2944	0.2112	0.2311	0.2177
Sports						
Diving48	0.1697	0.2847	0.3186	0.2489	0.2117	0.2612
MiniFinegym	0.1697	0.2847	0.3186	0.2489	0.4006	0.2612
IR and Dark Videos						
Infar	0.1315	0.2232	0.0593	0.1148	0.2409	0.1538
ARID	0.3621	0.5905	0.6213	0.3175	0.1523	0.4176
Low-Resolution Videos						
UAV-Human	0.0326	0.0732	0.0602	0.0693	0.0320	0.0473
Tiny Virat	0.2450	0.3270	0.2814	0.2891	0.3412	0.2884
Industrial Action Related Videos						
InHARD	0.5054	0.6265	0.6125	0.5037	0.3102	0.6462
Ikea	0.1963	0.4110	0.4745	0.4233	0.3517	0.4233
Construction Videos						
Construction	0.3344	0.3374	0.3079	0.3262	0.3627	0.3089
Depth Videos						
NTU	0.5059	0.7038	0.6965	0.5089	0.1527	0.6471

Table 5.3: Performance of different models on the downstream datasets regarding top-1 accuracy with Limited Evaluation using Kinetics-400 pretraining

From Figure 5.15, it's noticeable for Full-fine tuning that, while using MiT pre-training, transformer-based video models have been more effective than CNNs if we compare them with Kinetics-400 pretraining. TimeSFormer pre-trained on Moments in Time overperforms using Kinetics-400 pretraining for 9 out of 14 datasets. With respect to datasets, the top1 accuracy of HMDB, Ikea, Infar & Tiny Virat on TimeSFormer pre-trained on MiT was far better than that of TimeSFormer pre-trained on Kinetics-400. TSN and I3D pre-trained on MiT performed better for 6 out of 14 datasets. So, the performance of TSN and I3D pre-trained on Kinetics-400 and Moments in Time is too close to call.

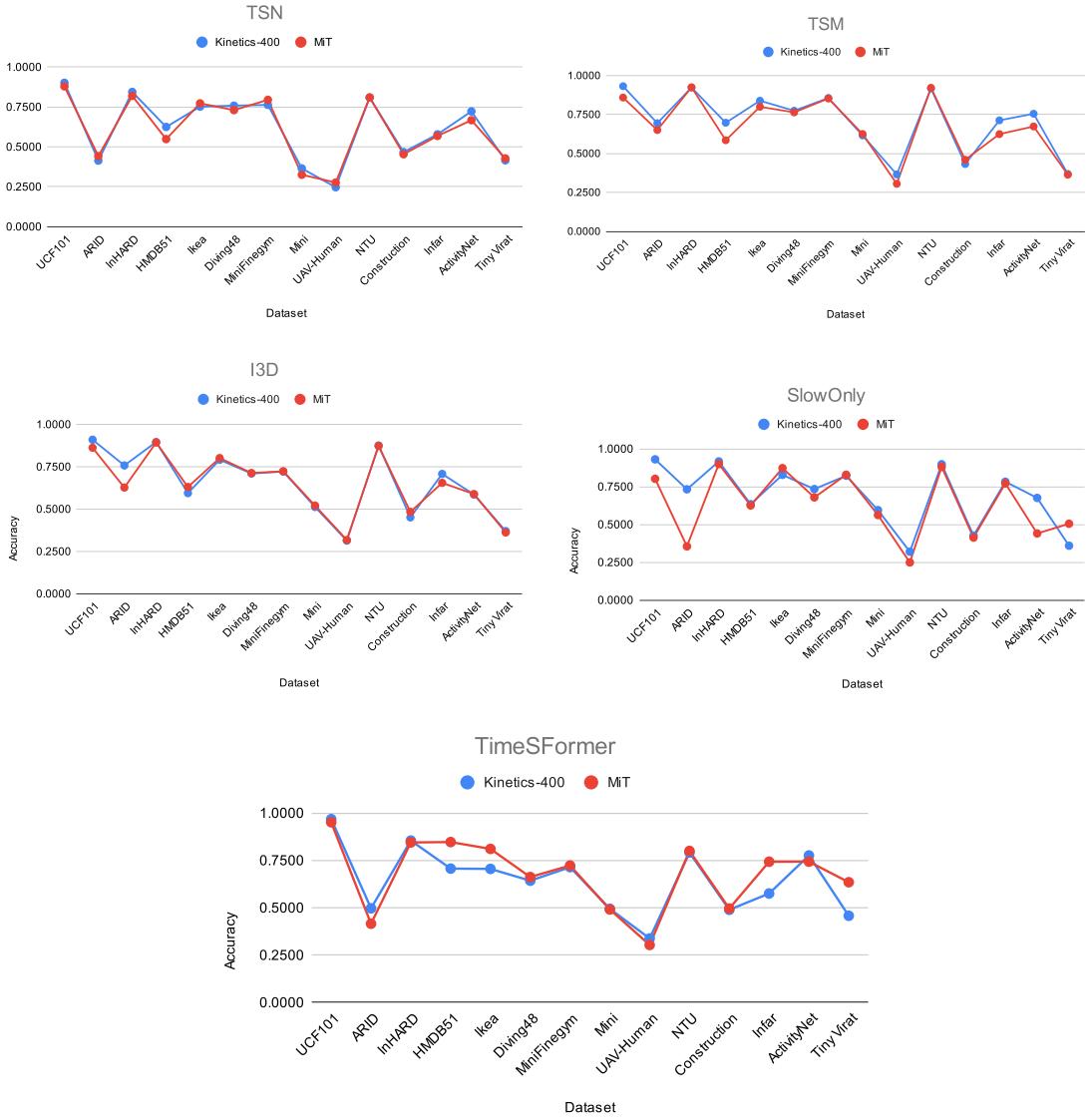


Figure 5.15: Full Fine-tuning using Kinetics-400 vs MiT Pretraining

The Top-1 accuracy of each model on the downstream datasets, sorted by domains pre-trained on the Moments in Time dataset, is presented in Table 5.4. Regarding Domains, Sports, Industrial Action Related Videos, Construction Videos, and Depth Videos, MiT pretraining performs better than Kinetics-400 pretraining.

Using MiT pre-trained models, 4 out of 5 models perform better on MiniFinegym and Ikea datasets. While 3 out of 5 performed better on Construction, Tiny Virat & NTU datasets.

Full Fine-tuning Using MiT Preatining					
Dataset	TSN	TSM	SlowOnly	Timesformer	I3D
Human Action					
UCF101	0.8795	0.8587	0.8023	0.9516	0.8623
HMDB51	0.5491	0.5847	0.6259	0.8472	0.6308
ActivityNet	0.6673	0.6733	0.4407	0.7438	0.5889
Mini SS V2	0.3257	0.6248	0.5256	0.4802	0.5208
Sports					
Diving48	0.7306	0.7636	0.6792	0.6626	0.7132
MiniFinegym	0.7951	0.8529	0.8289	0.7233	0.7235
IR and Dark Videos					
Infar	0.5689	0.6240	0.7716	0.7435	0.6550
ARID	0.4423	0.6501	0.3556	0.4164	0.6274
Low-Resolution Videos					
UAV-Human	0.2767	0.3049	0.2494	0.3041	0.3178
Tiny Virat	0.4282	0.3636	0.5052	0.6532	0.3620
Industrial Action Related Videos					
InHARD	0.8201	0.9249	0.8996	0.8452	0.8934
Ikea	0.7730	0.7996	0.8729	0.8110	0.8016
Construction Videos					
Construction	0.4543	0.4594	0.4136	0.4965	0.4837
Depth Videos					
NTU	0.8100	0.9215	0.8809	0.8004	0.8731

Table 5.4: Performance of different models on the downstream datasets regarding top-1 accuracy with full fine-tuning pre-trained on Moments in time.

If we look at the Model comparison, TimeSFormer performs better for Human Action Domain, reaching three out of four best top1 accuracy. TSM outperforms for Sports domain, achieving the best top1 accuracy for all datasets in this domain. TimeSFormer and TSM jointly achieved the best top1 accuracy for 10 out of 14 datasets.

Inspecting further, we can see that TimeSFormer excels in Linear evaluation for Low-Resolution and Construction Videos. Indeed, as shown in Table 5.5, SlowOnly is only comparable or inferior to TimeSFormer in the other domains.

The performance of TimeSFormer between pre-trained in Kinetics-400 and Moments in Time is featured in Figure 5.16. It shows that TimeSFormer pre-trained on MiT performs better for HMDB, Infar, Tiny Virat & Construction. But for most of the datasets, pre-trained on Kinetics-400 models performed better. So, in conclu-

Linear Evaluation Using MiT Preatining				
Dataset	TSN	TSM	SlowOnly	Timesformer
Human Action				
UCF101	0.8741	0.8479	0.8107	0.9118
HMDB51	0.5551	0.5956	0.6	0.8056
ActivityNet				0.7094
Sports				
Diving48	0.1200	0.2015		0.2862
IR and Dark Videos				
Infar	0.2678	0.2746	0.3940	0.3870
ARID	0.3633		0.3815	
Low-Resolution Videos				
UAV-Human				0.1904
Tiny Virat	0.3554	0.4298		0.5116
Industrial Action Related Videos				
Ikea	0.3865	0.4879	0.5215	0.5258
Construction Videos				
Construction	0.4299	0.5342	0.3527	0.4872
Depth Videos				
NTU		0.4997		0.3997

Table 5.5: Performance of different models on the downstream datasets regarding top-1 accuracy with Linear Evaluation pretrained on Moments in time.

sion, we may reach to the point that, Transfer learning with Kinetics-400 excels over Moments in Time.

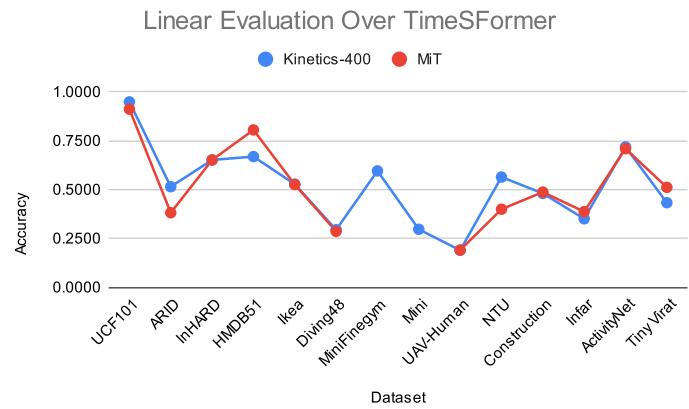


Figure 5.16: Linear Evaluation over TimeSFormer using Kinetics-400 vs MiT Pre-training

For both Full Fine-tuning and Linear Evaluation TimeSFormer is the best-performing model pre-trained in Moments in Time. So, we performed the Limited Evaluation over TimeSFormer pre-trained in Moments in Time which is stated in Table 5.6.

Limited Evaluation Over TimeSFormer		
Dwon-stream Dataset	Kinetics-400	MiT
UCF101	0.883	0.7832
HMDB51	0.5037	0.5886
Diving48	0.1924	0.2457
MiniFinegym	0.2489	0.3548
Infar	0.1148	0.1827
ARID	0.3175	0.3556
Tiny Virat	0.2891	0.4198

Table 5.6: Comparison between Performance of TimeSFormer on the downstream datasets in terms of top-1 accuracy with Limited Evaluation using Kinetics-400 and MiT Pretraining

Figure 5.17 demonstrates the impressive performance of TimeSFormer pre-trained on Moments in Time over Kinetics-400.

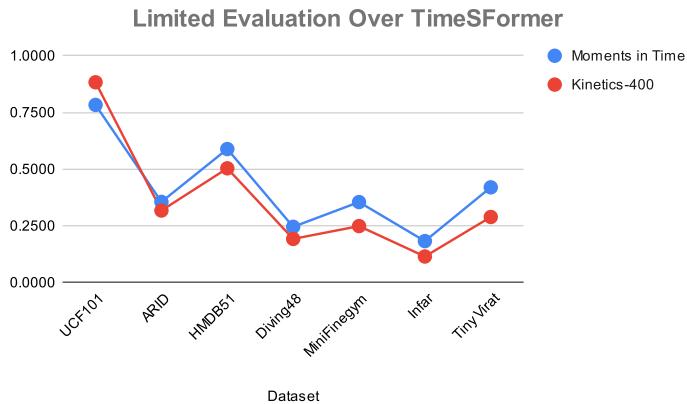


Figure 5.17: Limited Evaluation over TimeSFormer using Kinetics-400 vs MiT Pre-training

In our studies, transformer-based video models have been more effective than CNNs on several representative datasets. And TimeSFormer Model pre-trained in Moments in Time performs better than that of Kinetics-400. The main difference between transformer-based models and CNNs is the number of trainable parameters.

ters. Indeed, shown in Table 5.7, transformer-based models hold 3-4 times greater parameters than CNNs.

Model	No. of Trainable Params
TSM	23,612,531
TSN	23,606,384
SlowOnly	32,044,296
I3D	27,229,644
SIFAR	87,025,987
TimeSFormer	121,167,447

Table 5.7: Comparison between Performance of TimeSFormer on the downstream datasets in terms of top-1 accuracy with Limited Evaluation using Kinetics-400 and MiT Pretraining

So, The conclusion leads that the number of data points is more important than the classes in this post-transformer era.

Chapter 6

Conclusion & Future Scopes

6.1 Conclusion

We thoroughly evaluate the performance of deep learning models based on 2D, 3D, and Transformers on the downstream video action recognition datasets. According to our research, Transformer models consistently outperform other models in transfer learning. The accuracy of Human Action and Industrial Action datasets is higher than that of different areas. We discover a number of factors that make the tests cost-effective. A noticeable number of times, linear and limited evaluation may get the same results at a lesser cost when fine-tuning a model trained on a dataset as opposed to the conventional method. Full fine-tuning is optional if linear or limited evaluation meets the benchmark since such approaches are less expensive. In this post-transformer era, the number of data points is more important than the classes for pretraining a model.

6.2 Future Scopes

Benchmarking in video action recognition is an important research area that has the potential for several future advancements. Here are some potential future scopes:

- **Transfer learning and domain adaptation:** Transfer learning and domain adaptation can be leveraged to improve the performance of video action recog-

nition models across different domains and applications. The pretrained models on Moments in Time and Kinetics-400 can be fine-tuned on other smaller datasets for specific tasks or domains.

- **Multi-modal video analysis:** the integration of multi-modal information such as audio, text, and image data in addition to video is a promising area of research , which can improve the accuracy and robustness of video action recognition models.
- **Real-time video analysis:** The real-time analysis of videos has many applications in areas such as security, robotics, and sports. Future research can focus on developing real-time video action recognition models that can process large volumes of data in real-time while maintaining high accuracy.

References

- [1] Joonseok Lee Paul Natsev George Toderici Balakrishnan Varadarajan Sami Abu-El-Haija, Nisarg Kothari and Sudheendra Vijayanarasimhan. YouTube-8M: A Large-Scale Video Classification Benchmark. *Computer Vision and Pattern Recognition*, 2016.
- [2] Bolei Zhou Kandan Ramakrishnan Sarah Adel Bargal Tom Yan Lisa Brown Quanfu Fan Dan Gutfreund Carl Vondrick Aude Oliva Mathew Monfort, Alex Andonian. Moments in Time Dataset: one million videos for event understanding. *Computer Vision and Pattern Recognition*, 2018.
- [3] Karen Simonyan Brian Zhang Chloe Hillier Sudheendra Vijayanarasimhan Fabio Viola Tim Green Trevor Back Paul Natsev Mustafa Suleyman Andrew Zisserman Will Kay, Joao Carreira. The Kinetics Human Action Video Dataset. *Computer Vision and Pattern Recognition*, 2017.
- [4] Andrew Zisserman Joao Carreira. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Computer Vision and Pattern Recognition*, 2017.
- [5] Fabian Caba Heilbron; Victor Escorcia; Bernard Ghanem; Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. *IEEE*, 2015.
- [6] Mubarak Shah. Khurram Soomro, Amir Roshan Zamir. UCF101 - Action Recognition Data Set. *Center for Research in Computer Vision*, 2013.

- [7] Wei Zhang Yun Ni Wenqian Wang Zhiheng Li Tianjiao Li, Jun Liu. UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles. *Computer Vision and Pattern Recognition*, 2021.
- [8] Zhe Wang Yu Qiao Dahua Lin Xiaou Tang Luc Van Gool Limin Wang, Yuan-jun Xiong. Temporal Segment Networks for Action Recognition in Videos. *Computer Vision and Pattern Recognition*, 2017.
- [9] Song Han Ji Lin, Chuang Gan. TSM: Temporal Shift Module for Efficient Video Understanding. *Computer Vision and Pattern Recognition*, 2018.
- [10] Tushar Sangam Shruti Vyas Yogesh S Rawat Mubarak Shah Praveen Tirupattur, Aayush J Rana. TinyAction Challenge: Recognizing Real-world Low-resolution Activities in Videos. *Computer Vision and Pattern Recognition*, 2021.
- [11] T. Poggio T. Serre H. Kuehne; H. Jhuang, E. Garrote. HMDB: A Large Video Database for Human Motion Recognition. *IEEE*, 2011.
- [12] Bo Dai Dahua Lin Dian Shao, Yue Zhao. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. *Computer Vision and Pattern Recognition*, 2020.
- [13] Haozhi Cao Kezhi Mao Jianxiong Yin Simon See Yuecong Xu, Jianfei Yang. ARID: A New Dataset for Recognizing Action in the Dark. *Computer Vision and Pattern Recognition*, 2020.
- [14] Jiang Liu Jing Lv Luyu Yang Deyu Meng Alexander G. Hauptmann Chen-qiang Gao, Yinhe Du. InfAR dataset: Infrared action recognition at different times. *Neurocomputing* 212, 2016.
- [15] David BAUDRY Xavier SAVATIER Mejdi DALLET, Vincent HAVARD. In-HARD - Industrial Human Action Recognition Dataset in the Context of Industrial Collaborative Robotics. *IEEE*, 2020.
- [16] Tian-Tsong Ng Gang Wang Amir Shahroudy, Jun Liu. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *Computer Vision and Pattern Recognition*, 2016.

- [17] Vincent Michalski Joanna Materzyńska Susanne Westphal Heuna Kim Valentin Haenel Ingo Fruend Peter Yianilos Moritz Mueller-Freitag Florian Hoppe Christian Thurau Ingo Bax Roland Memisevic Raghav Goyal, Samira Ebrahimi Kachou. The "something something" video database for learning and evaluating visual common sense. *Computer Vision and Pattern Recognition*, 2017.
- [18] Lorenzo Torresani Matt Feiszli Du Tran, Heng Wang. Video Classification with Channel-Separated Convolutional Networks. *Computer Vision and Pattern Recognition*, 2019.
- [19] Clune J. Bengio Y. Yosinski, J. and Lipson. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems. *Computer Vision and Pattern Recognition*, 2014.
- [20] Shlens J. Kornblith, S. and Q. V. Le. Do better imagenet models transfer better? . *Computer Vision and Pattern Recognition*, 2019.
- [21] Lucas Beyer Alexander Kolesnikov, Xiaohua Zhai. Revisiting Self-Supervised Visual Representation Learning. *Computer Vision and Pattern Recognition*, 2019.