An Exploration of Charades: action video recognition

1. Introduction

Investigating action recognition in complex real world video settings using the Charades dataset came about due to my aim of building a model who could understand sign language, specifically Swedish Sign Language (TSP). However, due to me having difficulties finding a dataset built for machine learning on TSP I decided to widen my search – resulting in finding the Charades dataset, and interpreting understanding every day actions as a first step in models learning sign language.
Action recognition in video presents fundamental challenges beyond static image classification – in contrast to images videos contain: temporal dynamics, multiple spontaneous actions, and compositional structures where actions emerge from verb-object interactions. While there have been advances in video recognition achieving strong performance on specific benchmarks (such as Charades), recognising everyday activities in naturalistic environments remains challenging.
I aim to address the challenges of the inherent complexity, overlapping of actions and long temporal dependencies in video recognition using a multi-head SlowFast network architecture. My architecture explicitly decomposes actions into verbs, objects and compositional actions. The SlowFast architecture's dual-pathway design processes spatial semantics and temporal motion at different resolutions, aligning well with Charades'object-centric activities. I compare two model variants: a from-scratch implementation trained solely on Charades, and a transfer learning approach leveraging Kinetics-400 pretraining (Hara et al. 2018; Montalvo-Lezama & Escalante 2023).
My results demonstrate that transfer learning substantial improvements on mAP (the benchmark metric used for Charades) compared to the from-scratch baseline. However, both models reveal systematic challenges: a task hierarchy where compositional actions prove harder than individual verbs or objects, precision-recall imbalances indicating conservative strategies, and significant performance gaps between frequent and rare classes. The findings illustrate that while transfer learning is effective, fundamental challenges remain in handling compositional reasoning, class imbalance, and multi-label prediction in realistic video understanding.

2. Background

2.1. Images to video

The fundamental difference between images and video is their number of dimensions. While an image is 2D, (w, h), a video has an additional temporal dimension, (w, h, t) – an image is only spatial in nature, while a video is spatio-temporal (Feichtenhofer et al. 2019).
In image recognition the two spatial dimensions are usually treated symmetrically. As the temporal dimension of videos is not symmetrical to their spatial dimensions, due to slow motions being more likely than fast ones, the dimensions must be treated asymmetrically in model architectures (Feichtenhofer et al. 2019). The temporal information increases the complexity and amount of data needed for the model to process and subsequently produce a video recognition model, making the field of computer vision more time-consuming than image recognition (Carreira & Zisserman 2017). However, temporal information also offers an opportunity to disambiguate actions where the difference primarily lies in its speed and trajectory, such as the difference between 'throwing' and 'handing' an item (Ting-Long 2024).

2.2. Model architectures

Video recognition architectures must handle asymmetric spatiotemporal dimensions – slow motions are more common than faster ones, requiring specialised temporal modeling (Feichtenhofer et al. 2019). Modern approaches include 3D CNNs (C3D, I3D), two stream networks separating spatial and temporal processing, and biologically-inspired dual-pathway architectures like SlowFast (Karpathy et al. 2014: Tran et al. 2015; Carreira & Zisserman 2017; Feichtenhofer et al. 2019; Feichtenhofer 2020; Piergiovanni et al. 2021; Ting-Long 2024).

I chose SlowFast for its state-of-the-art performance on Charades (Feichtenhofer et al. 2019), balanced spatial-temporal modeling through parallel pathways, and compatibility with transfer learning from Kinetics-400 pretraining (Hara et al. 2018; Montalvo-Lezama & Escalante 2023). See Table 1 for a more detailed justification.

Table 1. SlowFast Model Architecture: Advantages and Disadvantages

	Reason	Motivation
Advantages	State-of-the-art accuracy on Charades	As Charades is a multi-label dataset which requires the model to both capture spatial semantics and motions, a model is needed which can handle both – such as SlowFast's two-stream design (Feichtenhofer et al. 2019)
	Balanced spatial and temporal modeling	The structure of the SlowFast model matches Charades object-centric actions. Having a slow and fast pathway allows the model to both focus on the spatial and temporal changes in the video. Actions in Charades combining the two pathways, such as "open fridge", require detection of change in both the spatial and temporal dimension (Feichtenhofer et al. 2019).
	Ecosystem support	There is good documentation of the SlowFast model available (Fan et al. 2020; Feitchenhofer et al. 2019), in addition to the model being integrated into libraries such as PyTorchVideo.
	Scalable backbone	The SlowFast model family is easy to scale due to the scalability of its backbone (Feichtenhofer 2020).
Disadvantages	High computational and memory cost	To run and train a SlowFast model access to a GPU is required. It requires much more compute and memory compared to other models such as Tiny Video Networks – about 30 GFLOPs compared to 13 GFLOPs (Feichtenhofer 2020).
	Training cost	In addition to having different dropout rates and separate optimisers for each pathway,the parameter count doubles – prolonging training compared to other models (Almushyti & Li 2022).
	Less specialised for	In contrast to models such as GLIDIN,

	human-object-interaction (HOI)	graph-based models who explicitly model HOI, the SlowFast architecture is not adapted for fine-grained HOI actions – making it a suboptimal choice for those specific actions (which can be found in Charades) (Almushyti & Li 2022).
	No built-in intent or long-range temporal reasoning	As Charades contains casual chains, i.e. actions such as 'walking to the fridge and opening the fridge door to grab a drink', it requires the model to be intent-aware or able to handle Asynchronous Temporal Fields (ATF) to capture these chains – which SlowFast lacks (Sigurdsson et al. 2017a).

2.3. Datasets for video recognition

Three datasets dominate video recognition benchmarks: Kinetics (400/600/700) for pretraining and scale, Something-Something for temporal reasoning, and Charades for multilabel long-term activities (Carreira & Zisserman 2017; Feichtenhofer et al. 2019). Charades uniquely combines multilabel classification (6.8 actions per video on average) with 30-second clips of everyday indoor activities, making it ideal for evaluating models on realistic, complex action recognition (Sigurdsson et al. 2016A).

3. Dataset

The charades dataset was developed as a response to the bias at the time present in video recognition datasets. At the time the datasets for video recognition mainly consisted of short clips of actions from sports and scripted movies – lacking complexity, contexts, and noise needed for realworld application of videorecognition models (Sigurdsson et al. 2017b).
In total the dataset consists of 9 848 videos, each on average around 30 seconds, all recorded by 267 M'Turk workers. The scrips given to the workers to act out were built from a vocabulary of 40 objects and 30 verbs – creating 157 action classes, an action class consists of an action and an object, 46 object classes, and 66 500 temporally annotated action intervals (Sigurdsson et al. 2016A). Although there in total are 15 different types of scenes in the dataset, this is not taken into consideration (i.e. no class is created for the scenes), possibly resulting in the model trained on the dataset acquiring a scene bias (Choi et al. 2019).
The strengths of the Charades dataset lies in the ecological validity of the videos – them being recorded in 'real' homes, not a set or artificial environment – the actions being everyday actions and the richness of the dataset annotations enabling research in multiple fields connected to computer vision. Due to its strengths it is a dataset used as a benchmark within the field of computer vision (Zhu et al. 2020; Wu, C. Y. et al. 2020; Ting-Long 2023; Almushyti & Li 2022; Piergiovanni et al. 2021). However, the dataset is restricted to indoors activities and limited to few objects and actions (out of all possible actions and objects in the world) which subsequently limits generalisability – models trained on it might not perform well on actions not in the dataset nor on actions in scenes not in the dataset.

4. Model architecture

To address the multi-label action recognition in Charades I developed a multi-head SlowFast architecture – extending the SlowFast paradigm to have task-specific output heads. The architecture consists of a shared SlowFast backbone for feature extraction and three separate classification heads for verb, object, and action predictions. I implemented two variants of this architecture – one 'from-scratch' implementation with no pre-training, and one with a Kinetics-400 pre-trained SlowFast backbone.

4.1. Base architecture

As previously discussed , the SlowFast architecture is inspired by biological visual structures and consists of two parallel pathways who process video at different temporal resolutions (Feichtenhofer et al. 2019). This design explicitly addresses the asymmetry between spatial and temporal dimensions in video data (where slow motions are more likely than fast ones). The slow pathway operates at a low frame rate (temporal stride of 16) to capture spatial semantics and appearance, while the fast pathway operates at a high frame rate (temporal stride of 2) to detect motion and temporal changes (Feichtenhofer et al. 2019).
The dual pathway design is well-suited for Charades, as it contains object-centric actions that require both strong spatial understanding (e.g. ability to identify objects) and temporal reasoning (e.g. detecting the change in state of the object). Actions such as 'open fridge' or 'throw pillow' exemplify this, as they combine static object recognition with dynamic motion patterns.

4.2. Implementation

The slow pathway (temporal stride 16, full channel capacity) captures spatial semantics while the fast pathway (temporal stride 2, 0.125× channels) detects motion. Both use a 3D ResNet backbone with lateral connections (5×1×1 convolutions) fusing fast-pathway motion cues into the slow pathway. After the final stage (res5), pathway features are concatenated, pooled, and passed through dropout (p=0.5) to produce a 512-dimensional representation.
To enable information flow from the fast pathway to the slow pathway, lateral connections are inserted after each residual stage. These connections use 5×1×1 convolutions with temporal stride 8, effectively fusing the high-frequency information from the fast pathway to the slow pathway. The temporal dimension is then interpolated to match the slow pathway's temporal resolution using trilinear interpolation. This fusion mechanism allows the slow pathway to incorporate motion cues while maintaining its focus on spatial semantics.
After the final residual stage (res5), the features from both pathways are fused using concatenation along the channel dimension. The fast pathway features are first temporally interpolated to match the slow pathway's temporal resolution. The concatenated features pass through an adaptive average pooling layer (reducing to 1×1×1), followed by dropout (p=0.5) and a fully connected layer that produces a 512-dimensional feature representation.
To address Charade's multi-label nature, I employ a multi-head architecture with three separate linear classification heads:

- Verb head: Predicts verb classes (e.g. 'open', 'throw')
- Object head: Predicts object classes (e.g. 'fridge', 'pillow')
- Action head: Predicts the full action classes (e.g. 'open fridge', 'throw pillow')

This decomposition explicitly models the compositional structure of actions, enabling the model to learn shared representations across verbs and objects while maintaining the flexibility to predict complete action labels. Each head receives the same 512-dimensional feature vector from the backbone, ensuring consistent semantic understanding across all prediction tasks.

4.3. Model variants

4.3.1. From-scratch model implementation

The from scratch variant initializes all network parameters randomly and trains the entire architecture end-to-end on Charades. This approach allows for the model to learn representations specifically tailored to the Charades distributions and task requirements, without any inductive bias from external datasets. However, it requires more training data and computational resources to converge, and may be more prone to overfitting given Charade's relatively modest size of ~9,88 videos.

4.3.2. Pre-trained model implementation

The transfer learning variant initialises the SlowFast backbone with weights pre-trained on Kinetics-400, a large-scale action recognition dataset containing 400 action classes across hundreds of thousands of video clips. As previously mentioned, Kinetics serves as a standard benchmark for pre-training and evaluation in video recognition. Pre-training on Kinetics provides the model with general spatiotemporal features that transfer well to other action recognition tasks.
The multi-head classification layer are randomly initialised and trained from scratch, as they are task-specific to Charades' label space. During training, I employ different learning rates for the pre-trained backbone (lower learning rate) and the classification heads (higher learning rate), following standard practice in transfer learning. This approach leverages the powerful representations learned from large-scale data while adapting them to Charade's specific characteristics, including its multi-label nature, long-term temporal structure, and everyday activity setting.

4.4. Architecture justification

My multi-head SlowFast architecture addresses several key challenges (which I previously mentioned):

1. Asymmetric spatiotemporal processing: The dual-path design explicitly handles the asymmetry between spatial and temporal dimensions, aligning with the theoretical motivation for SlowFast networks.
2. Multi-label prediction: The multi-head design accommodates Charade's multi-label nature (average of 6.8 actions per video), enabling the model to predict multiple simultaneous actions without architectural constraints.
3. Compositional action understanding: By decomposing actions into verbs and objects, the architecture can potentially learn more generalisable representations and handle compositional novel actions.
4. Computational efficiency: While SlowFast networks are computationally demanding, they achieve a favourable balance between accuracy and efficiency compared to other state-of-the-art video models. The reduced channel capacity of the fast pathway reduces the computational overhead of processing high-frame-rate video.
5. Transfer learning capacity: The architecture's compatibility with Kinetics-400 pre-training enables leveraging large-scale external data, partially addressing the limited size of the Charade dataset.

The primary limitation of this architecture, as previously noted, is its lack of explicit modeling of long-range temporal reasoning and causal chains. Charades contains sequences of actions, which require intent-awareness or asynchronous temporal field modeling – which SlowFast does not provide. Future work could address this limitation by incorporating temporal attention mechanisms or hierarchical temporal modeling.

5. Results

Overall mAP (the benchmark metric used for Charades) showed an increased performance of 88% (~0.293 compared to ~0.155) for the pretrained model compared to the 'from-scratch' baseline. The results clearly demonstrate that the pretrained model outperforms the 'from-scratch' baseline on all metrics measured – indicating a strong transferability from Kinetics-400 to Charades. However, the pretrained model still has room for improvement as it is well below the current state-of-the-art models who typically rest at or above 0.4 mAP (Feichtenhofer et al. 2019).

Table 2. mAP metrics for pretrained and 'from scratch' models.

mAP scores				
model	overall mAP	verb mAP	object mAP	action mAP
pretrained	**0.293374**	**0.385151**	**0.318120**	**0.176851**
from scratch	0.155430	0.223310	0.166162	0.076817

Table 3. F1 scores for the macro average for both the pretrained and 'from scratch' model.

F1 scores (Macro Averages)			
model	verb F1 (macro)	object F1 (macro)	action F1 (macro)
pretrained	**0.228410**	**0.155038**	**0.048693**
from scratch	0.044352	0.014891	0.000000

Table. 4 F1 scores for the micro average weighted by class frequency for both the pretrained and 'from scratch' model.

F1 score (micro average weighted by class frequency)			
model	verb F1 (micro)	object F1 (micro)	action F1 (micro)
pretrained	**0.507683**	**0.283766**	**0.104451**
from scratch	0.294891	0.114693	0.000000

Table 5. Precision and recall for the macro average for both the pretrained and 'from scratch' model.

Precision and Recall (macro average)						
model	verb precision	verb recall	object precision	object recall	action precision	action recall

pretrained	**0.485955**	**0.185081**	**0.377649**	**0.115075**	**0.169138**	**0.033382**
from scratch	0.073202	0.045691	0.014620	0.015172	0.000000	0.000000

Both models display a hierarchy in difficulty between the tasks – verb recognition being the easiest, followed by objects, and action recognition as the hardest – suggesting that compositional actions (actions consisting of both a verb and an object) are inherently more challenging than identifying individual verbs/objects. As the hierarchy is affecting both models, it suggests that it is an inherent part of the Charades dataset instead of a model limitation.

6. Discussion

The pretrained SlowFast model successfully transfers learning from Kinetics-400, indicating that despite domain differences between the datasets (daily activities in Charades and internet videos in Kinetics), low-level motion patterns and object recognition capabilities transfer effectively. Consequently, the dual-pathway architecture of SlowFast has shown to be beneficial for compositional action understanding.

For both models the action class shows the lowest performance as it requires understanding of both the verb in motion and the object being manipulated. However, unlike the pretrained model the 'from scratch' model fails to recognize any actions (see Table 3 and 4) suggesting: it being trained on insufficient data for compositional reasoning (Materzynska et al. 2020), no understanding of hierarchical modeling which leverages verb-object relationships (Hou et al. 2020), and class imbalance (Zhang et al 2023) – resulting in rare actions not being predicted. This suggests hierarchical compositional modeling or explicit verb-object factorization may improve action recognition (Hou et al. 2020; Materzynska et al. 2020).

As seen for both models in Table 5, but more prominently for the pretrained model, is an asymmetry between precision and recall – with high precision and low recall. This indicates that the model has learnt to be conservative in its predictions, only predicting the correct class a minority of the time. This could be due to the decision boundary being set to 0.5, a threshold the model rarely seems to exceed, which could stem from the class imbalance in Charades or the Binary Cross-Entropy (BCE) loss being a suboptimal loss function for this task and dataset (Zhang et al. 2023; Wu, T. et al. 2020). Lowering the threshold could be a first step in exploring whether F1 score, recall, and precision metrics could improve, especially for the action class. However, the observed trend could also be a consequence of the nature of the dataset (Zhang et al. 2023), as the multi-label nature of it could result in the model learning that only a handful of the 157 possible action classes are positive for each video. Hence, since a majority of the labels should be predicted as 'incorrect' for each video there might be a systematic tendency in the model towards predicting labels to be 'incorrect' – maximizing False Negatives (Durarte et al. 2021). Addressing this requires per-class threshold optimization (Zou et al. 2016) using validation data (e.g., finding thresholds that maximize F1 per class rather than using uniform 0.5) or asymmetric loss functions (Wu, T. et al. 2020). Additionally, probability calibration techniques could better align the model's confidence scores with True Positive rates, enabling more reliable threshold-based decision making (Zou et al. 2016).

The gap between macro and micro F1 scores (see Table 3 and 4) is a signature of the long-tail distribution problem – indicating that model performance highly varies across classes. While the micro F1 score aggregates all predictions across all classes, giving more weight to frequent classes, macro F1 treats each class equally. Consequently, a difference between the micro F1 score and macro

F1 score indicates that a model performs well on the most frequent classes, and poorly on any rare ones. The difference between the two metrics indicates that the class imbalance in the dataset is a serious issue which needs to be dealt with to improve the performance of the model (Durarte et al. 2021). Improving the long-tail distribution would require class-balancing training strategies, such as: re-weighting loss by inverse class frequency, class-balanced sampling that oversamples rare classes, or few-shot learning techniques that can generalize from limited examples of rare actions by leveraging shared verb and object representations (Wang et al. 2023).

7. Conclusion

My work demonstrates both the promise and limitations of transfer learning for multilabel action recognition in everyday activities. The pretrained SlowFast model achieves ~0.29 overall mAP, representing an 88% improvement over the from-scratch baseline, validating that spatiotemporal representations learned on large-scale internet videos transfer effectively to indoor activity recognition despite domain differences.

As previously discussed compositional action understanding remains difficult, especially for the from-scratch model, which demonstrated that learning compositional structures from limited data is challenging. Future work should explore explicit compositional modeling, such as factorizing action predictions through learned verb-object relationships, to possibly improve the model's compositional understanding. In addition, the substantial gap between macro F1 and micro F1 reveal severe performance disparities between frequent and rare classes – depicting a long-tail problem inherent to realistic action datasets, where some classes occur a lot more than others. To address this the model architecture needs to be adapted to limit the impact of class frequencies on the model's performance. Furthermore, the model needs to lower the classification threshold and potentially change the loss function to be able to optimize the precision-recall tradeoff (see Table 5).

However, with the current architecture I fail to address the issue of longrange temporal reasoning present in the Charades dataset, resulting in poorer performance. To improve performance and the model's longrange temporal reasoning ability, adaptations of the architecture must explore how best to capture the temporal attention mechanism and the hierarchical temporal modeling.

8. Bibliography

Almushyti, M. and Li, F.W. (2022) 'Distillation of human–object interaction contexts for action recognition', *Computer Animation and Virtual Worlds*, 33(5). doi:10.1002/cav.2107.

Carreira, J. and Zisserman, A. (2017) 'Quo Vadis, action recognition? A new model and the kinetics dataset', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. doi:10.1109/cvpr.2017.502.

Durand, T., Mehrasa, N. and Mori, G. (2021) 'PLM: Partial Label Masking for Imbalanced Multi-label Classification', *arXiv preprint* arXiv:2105.10782. Available at: https://arxiv.org/abs/2105.10782 (Accessed: 28 November 2025).

Fan, H., Li, Y., Xiong, B., Lo, W.-Y. and Feichtenhofer, C. (2020) *PySlowFast*. Available at: https://github.com/facebookresearch/SlowFast (Accessed: 28 November 2025).

Feichtenhofer, C., Fan, H., Malik, J. and He, K. (2019) 'SlowFast networks for video recognition', in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Available at: https://arxiv.org/pdf/1812.03982.pdf (Accessed: 28 November 2025).

Feichtenhofer, C. (2020) 'X3D: Expanding architectures for efficient video recognition', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 200–210. doi:10.1109/cvpr42600.2020.00028.

Hou, Z. *et al.* (2020) 'Visual compositional learning for human-object interaction detection', *Lecture Notes in Computer Science*, pp. 584–600. doi:10.1007/978-3-030-58555-6_35.

Karpathy, A. *et al.* (2014) 'Large-scale video classification with Convolutional Neural Networks', *2014 IEEE Conference on Computer Vision and Pattern Recognition* [Preprint]. doi:10.1109/cvpr.2014.223.

Liang, A. (2024) 'SlowFast Model Explained with a PyTorchVideo Implementation', *Medium*, 24 December. Available at: https://medium.com/@mlshark/slowfast-model-explained-with-a-pytorchvideo-implementation-dc835e17a9df (Accessed: 28 November 2025).

Materzynska, J. *et al.* (2020) 'Something-else: Compositional action recognition with spatial-temporal interaction networks', *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. doi:10.1109/cvpr42600.2020.00113.

Piergiovanni, A.J., Angelova, A. and Ryoo, M.S. (2022) 'Tiny Video Networks', *Applied AI Letters*, 3(1). doi:10.1002/ail2.38.

Sigurdsson, G.A. *et al.* (2016) 'Hollywood in homes: Crowdsourcing data collection for Activity Understanding', *Lecture Notes in Computer Science*, pp. 510–526. doi:10.1007/978-3-319-46448-0_31.

Sigurdsson, G.A. *et al.* (2017a) ‘Asynchronous temporal fields for action recognition’, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. doi:10.1109/cvpr.2017.599.

Sigurdsson, G.A., Russakovsky, O. and Gupta, A. (2017b) ‘What actions are needed for understanding human actions in videos?’, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2156–2165. doi:10.1109/iccv.2017.235.

Ting-Long, L. (2024) ‘Short-term action learning for video action recognition’, *IEEE Access*, 12, pp. 30867–30875. doi:10.1109/access.2024.3364810.

Tran, D. *et al.* (2015) ‘Learning spatiotemporal features with 3D convolutional networks’, *2015 IEEE International Conference on Computer Vision (ICCV)* [Preprint]. doi:10.1109/iccv.2015.510.

Wang, H. *et al.* (2023) ‘Free-form composition networks for Egocentric Action Recognition’, *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10), pp. 9967–9978. doi:10.1109/tcsvt.2024.3406546.

Wu, C.-Y. *et al.* (2020) ‘A multigrid method for efficiently training video models’, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. doi:10.1109/cvpr42600.2020.00023.

Wu, T. *et al.* (2020) ‘Distribution-balanced loss for multi-label classification in long-tailed datasets’, *Lecture Notes in Computer Science*, pp. 162–178. doi:10.1007/978-3-030-58548-8_10.

Zhang, W. *et al.* (2023) ‘Learning in Imperfect Environment: Multi-label classification with long-tailed distribution and partial labels’, *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1423–1432. doi:10.1109/iccv51070.2023.00137.

Zhu, S., Yang, T., Mendieta, M. and Chen, C. (2020) 'A3D: Adaptive 3D Networks for Video Action Recognition', *arXiv preprint* arXiv:2011.12384. Available at: https://arxiv.org/abs/2011.12384 (Accessed: 28 November 2025).

8.1. Reference to dataset

The Allen Institute for Artificial Intelligence (no date) *Charades*, *Perceptual Reasoning and Interaction Research*. Available at: https://prior.allenai.org/projects/charades (Accessed: 01 November 2025).

8.2. Bibliography for architecture design

Bleed AI Academy (2021) *Human Activity Recognition using TensorFlow (CNN + LSTM) | 2 Methods*. Available at: https://youtu.be/QmtSkq3DYko (Accessed: 28 November 2025).

Facebook Research (no date) *pytorchvideo.models.slowfast — PyTorchVideo documentation*. Available at: https://pytorchvideo.readthedocs.io/en/latest/api/models/slowfast.html (Accessed: 28 November 2025).

Fan, H., Li, Y., Xiong, B., Lo, W.-Y. and Feichtenhofer, C. (2020) *PySlowFast*. Available at: https://github.com/facebookresearch/SlowFast (Accessed: 28 November 2025).

Feichtenhofer, C., Fan, H., Malik, J. and He, K. (2019) 'SlowFast networks for video recognition', in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Available at: https://arxiv.org/pdf/1812.03982.pdf (Accessed: 28 November 2025).

GluonCV (no date) 'Getting Started with Pre-trained SlowFast Models on Kinetics400', *GluonCV Documentation*. Available at: https://cv.gluon.ai/build/examples_action_recognition/demo_slowfast_kinetics400.html (Accessed: 28 November 2025).

Kim, D.-K. (2025) 'SlowFast: Revolutionizing Video Recognition with Two-Speed Processing', *Medium*, 13 March. Available at: https://medium.com/@kdk199604/slowfast-revolutionizing-video-recognition-with-two-speed-processing-c9a5a68894f6 (Accessed: 28 November 2025).

Liang, A. (2024) 'SlowFast Model Explained with a PyTorchVideo Implementation', *Medium*, 24 December. Available at: https://medium.com/@mlshark/slowfast-model-explained-with-a-pytorchvideo-implementation-dc835e17a9df (Accessed: 28 November 2025).

Lindernoren, E. (no date) *Action-Recognition: Exploration of different solutions to action recognition in video, using neural networks implemented in PyTorch*. Available at: https://github.com/eriklindernoren/Action-Recognition (Accessed: 28 November 2025).

Mallikarjuna Infosys (2021) *Human Activity Recognition Project Execution*. Available at: https://youtu.be/eQqCpysGkIc (Accessed: 28 November 2025).

Qi, M., Qin, J., Zhen, X., Huang, D., Yang, Y. and Luo, J. (2020) 'Few-Shot Ensemble Learning for Video Classification with SlowFast Memory Networks', in *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. New York, NY: Association for Computing Machinery, pp. 3007–3015. doi: 10.1145/3394171.3416269.

Code With Aarohi (2022) *Video Classification with a CNN-RNN Architecture | Human Activity Recognition*. Available at: https://youtu.be/ezjnySXqdTo (Accessed: 28 November 2025).