Redefining Temporal information in video classification problem

Woojoo Hahm[1], Hansol Kim[1], Jaewon Son[1] and Kwansu Kim[2]

[1] Sungkyunkwan University, Suwon, Korea, {hahmwj, skku.hansolkim, thswodnjs3}@ gmail.com
[2] Sungkyunkwan University, Suwon, Korea, {kim.kwangsu}@ skku.edu

Abstract

The main challenge in machine-learning for video classification is understanding 'Spatial Information' and 'Temporal Information.' While significant progress has been made in extracting spatial information by developing 2D image classification models, the 'Temporal Information' extraction has not advanced as much. One possible reason is that a comprehensive definition of temporal information has not yet been established. This paper proposes a novel definition of 'Temporal Information' in video classification consisting of 'Movement Information' and 'Temporal Ordering Information.' To demonstrate this, we conduct simple experiments using different timeline variations: Original, Reverse, and Stack. Furthermore, we evaluate how well-existing video classification models capture temporal information. To assess the meaningfulness of temporal ordering information in the feature vector obtained from video classification models, we modify the classifier to predict the Original and Reverse data. These experiments show that most existing video classification models struggle to recognize temporal ordering information. Our findings are validated using benchmark datasets such as UCF101 and Kinetics400, along with several well-established baseline models.

Keywords— *Video Classification, Temporal information*

I. Introduction

Most video classification models have been studied through methods of trying to find spatial and temporal information in video. 'Spatial information' is a feature from a single frame of video, meaning visual information, and 'Temporal information' is a feature from multi-frames of video. There are two types of models that perform video classification: a model based on a two-stream network that separately detects 'Spatial information' and 'Temporal information'[17, 14, 20, 5] and a model that detects 'Saptio-temporal information'[7, 3, 2, 8, 4] which implies the combination of spatial and temporal information jointly.

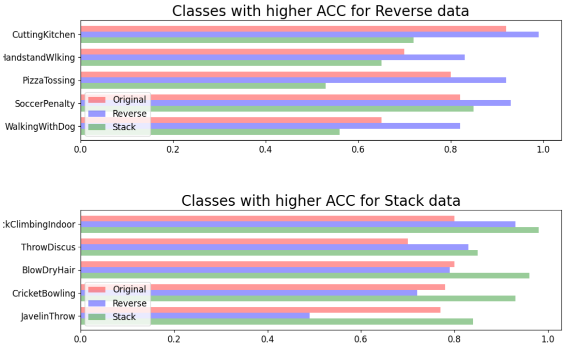

Fig. 1. This result is a class where Reverse or Stack showed higher accuracy than Original in a model trained with Original data. Original is forward video, Reverse is reverse video, and Stack is video created with only one image representing the video.

We propose that existing video classification models exhibit a bias towards spatial information. This bias arises from two main factors: the utilization of transformed large pre-trained image classification models to detect temporal information and the practice of increasing the number of input frames in the models. When employing a large pre-trained image classification model, the training process optimizes the model to detect temporal information, as spatial information is an inherent aspect of the model. Similarly, increasing the input frames may inadvertently introduce more spatial information, as the additional frames contribute to extracting spatial features. Consequently, the optimal approach to isolate and detect only temporal information in situations where existing models are biased toward spatial information remains a question. This paper proposes dividing temporal information into 'Movement information' and 'Temporal ordering information.' This distinction serves as our suggested definition for temporal information, allowing for a more comprehensive understanding of the temporal aspects present in the data.

A prominent approach for evaluating temporal information involves manipulating the order of frames in a video. To examine data without temporal ordering information, certain studies have categorized videos into three classes: normal videos, reversed videos, and videos with randomly shuffled frames[13, 16]. Shuffled data possesses a randomized timeline, although it still contains the inherent movement of objects within the video. We assert that these features also contribute to temporal information. To fully exclude such influences, we introduce Stack data, which comprises a single frame representing the entire video.

It is commonly believed that models trained on Original data exhibit the highest performance when tested with Original data, compared to Reverse and Stack data. This belief aligns with the general understanding; indeed, we can observe higher accuracy with Original data. The existing video classification model is suitable for detecting temporal ordering information. However, the results in Fig 1 show that certain classes exhibit higher accuracy for Reverse and Stack data. Previous studies[13, 16] have suggested that this discrepancy can be attributed to the dataset's inductive biases. Specifically, the dataset comprises classes where temporal order is not crucial, and instantaneous spatial information holds more importance for the models to make accurate judgments. Nonetheless, this explanation does not fully account for the classes that display a notably high accuracy for Reverse data. To thoroughly comprehend our results regarding the Reverse data, this paper primarily focuses on identifying potential issues within the models rather than the dataset itself. A detailed examination of this matter is provided in the **IV** section of the paper.

Consequently, this paper is motivated by the observation that certain classes for existing video classification models ineffectively handling temporal information. We conducted an investigation with the hypothesis that these challenges predominantly arise from the models themselves rather than being inherent to the dataset. In summary, our contributions are as follows: (i) Introducing a fresh perspective on defining temporal information within the realm of video classification;(ii) Proposing the potential for enhanced performance of existing models by integrating increased temporal information during the training process.

II. Related work

A. *Video classification model*

Video classification models can be categorized into two main types: the two-stream model[14, 17, 20, 9] and the one-stream model[3, 7, 6, 11, 2, 8, 4]. In the case of the two-stream model, spatial information is extracted by applying a common image classification model to video frames. To capture temporal information, this model employs a technique called 'Optical flow.' Optical flow represents the displacement vectors between consecutive frames, generated based on the assumptions of 'Brightness Consistency,' 'Temporal Persistence,' and 'Spatial Coherence.' These assumptions ensure that corresponding positions in consecutive frames have similar brightness, minimal displacement, and consistent directional changes in neighboring pixels. However, computing optical flow is time-consuming and does not allow for end-to-end learning. More recently, researchers have explored using large pre-trained models in the two-stream model. They utilize a pre-trained model for capturing spatial information and incorporate the concept of NLP's Adapter module to handle temporal information. This approach enhances the efficiency and effectiveness of training the two-stream model[10, 9].

The one-stream model can be divided into two main approaches. The first approach is feeding video frames into a 2D image classification model to generate a feature vector. This feature vector is then passed through an RNN-based or Transformer model to extract spatio-temporal information. The second approach is to extend existing 2D image classification models to 3D. This approach uses a 3D Convolution base model, where the filter size becomes (time × 3 × width × height) to account for temporal information. Recently, with the successful application of Vision Transformer(VIT)[1] in image tasks, researchers have also started applying VIT to video analysis. They divide the video into patches, incorporating time information, and utilize VIT-based models for video classification[2, 8, 5].

B. *Temporal information*

Many existing machine-learning based video classification models often overlook the initial step of defining temporal information and directly analyzing it. However, it is crucial to establish a clear definition of temporal information at the outset. Recent studies [13, 16] try to define temporal information by conducting various experiments, including categorizing videos into Original, Reverse, and Shuffle sequences. These experiments demonstrate the significance of temporal order in video classification tasks. Original refers to videos with frames in the correct chronological order, Reverse indicates videos with frames in reverse order, and Shuffle represents videos with randomly shuffled frames. The mentioned studies highlight a specific issue observed in the **I** section, where certain classes demonstrate higher accuracy when videos are presented in a Reverse or Stack order. These studies attribute this phenomenon to the dataset's inductive bias. They argue that factors such as camera angles, labeling methods, and actor movements within the datasets tend to de-emphasize the importance of temporal information. Additionally, the relatively short duration of the data contributes to this issue, with UCF101 having an average length of 5.8 seconds and Kinetics400 having an average length of 9.7 seconds. The authors suggest that these durations are insufficient for capturing and leveraging the full temporal dynamics of the videos. However, this paper contends that the root cause of this problem lies in the model's approach to capturing temporal information.

A recent study[15] examined the impact of temporal ordering information in Video-to-Language models. The researchers established a baseline using a 'before/after' setup and systematically flipped video clips to assess the models' ability to predict temporal order. The study demonstrated that existing Video-to-Language models were unable to determine temporal ordering information accurately. Moreover, the paper proposed fine-tuning the models using a loss function that specifically incorporates temporal ordering, leading to overall performance improvements. While this approach shares similarities with our paper, the critical difference lies in the task itself. The mentioned study focused on Video-to-Language models, where altering the temporal order of input video clips would result in different output text. In contrast, a different experimental setup was required for the Video Classification task addressed in our paper to investigate the influence of temporal ordering, as detailed in the **III** section.

III. Method

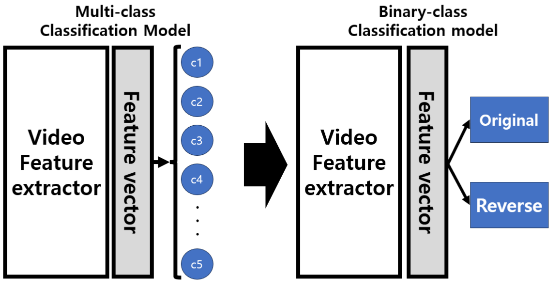

Fig. 2. The overall method of testing temporal ordering information in video classification problem.

The current video classification models demonstrate excellent performance in classification tasks but often fall short of fully capturing and utilizing temporal information. This deficiency arises due to the need for a clear definition of temporal information within the context of video classification using machine-learning. In this paper, we aim to address this issue by defining temporal information and evaluating the existing video classification models' capability to detect and utilize such information. To show that, we conduct two simple experiments.

First, we experiment by categorizing videos into three classes: Original, Reverse, and Stack. For the Stack data, we extract a single image from the video's midpoint and concatenate it with the remaining video frames to match the data shape of the other videos. We adopt this approach assuming that since UCF101 is a dataset comprising short-trimmed and single-action videos, the image from the video's midpoint represents the video's overall behavior.

Second, we modify the classifier component of the existing video classification model to create a binary model that distinguishes between Original and Reverse classes, as depicted in Fig 2. The Original class represents videos played in the forward direction, while the Reverse class denotes videos played in reverse. We hypothesize that if the feature vector immediately before the classifier in the existing video classification model adequately incorporates and represents temporal information, a simple fine-tuning process would effectively enable the model to address the temporal classification problem.

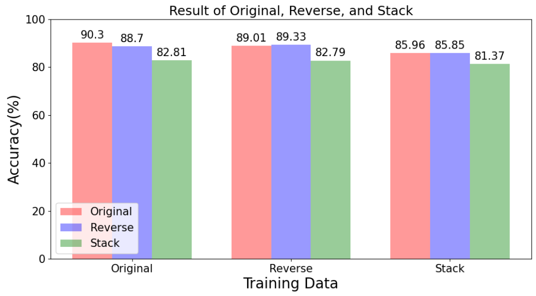

Fig. 3. The graph summarizes the performance of a model trained and tested on datasets with different temporal orders. The x-axis represents the training data, including Original, Reverse, and Stack sequences. The y-axis indicates the accuracy achieved when testing the model on different datasets.

IV. Experiments

A. *Datasets*

To evaluate the effectiveness of our method, we conducted experiments on two datasets: UCF101[12] and a subset of Kinetics400[19]. UCF101 is a motion recognition dataset encompassing various challenges such as camera motion, object appearance, and pose variations. It consists of 13,321 short-trimmed videos across 101 action categories. Kinetics400, on the other hand, is an action recognition dataset comprising realistic action videos collected from YouTube. It includes 306,254 short-trimmed videos spanning 400 action categories. We utilized different base models for our experiments: CNN-RNN-based, CNN-Transformer, and X3D models were trained using a base model pre-trained on UCF101, while X3D, I3D, R3D, and SlowFast models were trained using a model pre-trained on Kinetics400.

B. *Rethinking temporal information*

Fig 3 shows the average accuracy results for different training and testing configurations. The highest average accuracy is observed when training and testing are conducted on the Original data, followed by the second highest accuracy when training and testing are performed on the Reverse data. Also, when the Original and Reverse data are switched during training and testing (i.e., Train: Original, Test: Reverse, and vice versa), a drop in accuracy is observed. These findings indicate that existing video classi-

Model	LSTM	GRU	Transformer	X3D(U)	X3D(K)	I3D	R3D	Slowfast
Base model Train Acc	53.17%	48.18%	97.37%	93.12%	91.79%	86.60%	91.63%	92.74%
Base model Test Acc	60.31%	60.71%	85.95%	83.53%	72.12%	71.65%	74.58%	76.86%
1HL model with SD	50.10%	50.12%	50.10%	64.73%	59.36%	43.22%	67.66%	62.21%
5HL model with SD	50.00%	50.05%	50.00%	61.98%	53.56%	45.81%	57.85%	48.23%
1HL model with LD	50.06%	50.06%	50.14%	61.52%	32.70%	41.78%	39.91%	52.08%
5HL model with LD	50.08%	50.10%	50.04%	68.61%	39.83%	46.23%	41.66%	66.11%

Table 1. In this experiment, we evaluated different models to determine their ability to distinguish between Original and Reverse classes. HL represents hidden layers applied to the classifier, and the following number indicates the count of hidden layers. SD and LD represent small dataset and large dataset, respectively, while U and K represent UCF101 and Kinetics400 datasets. The results presented are the average values obtained from multiple experiments. The models CNN-LSTM, CNN-GRU, CNN-Transformer, and X3D(U) exhibit a deviation of ±5%. X3D(K), I3D, R3D have a deviation of ±20%, and SlowFast has a deviation of ±10%.

fication models effectively capture temporal information. However, a different pattern emerges when examining the results at the class level. When testing the model trained on the Original data using Original, Reverse, and Stack data, the accuracy for Reverse data surpasses that of other classes in 25 out of 101 categories, and the accuracy for Stack data is highest in 15 out of 101 categories. These results indicate that there is remaining temporal information that existing models need to learn.

Another fascinating finding is that the model trained on Stack data, which lacks explicit time information, exhibited higher accuracy when tested with Original and Reverse data. This suggests that object motion plays a significant role in video classification, leading to the introduction of the term 'Movement information' in this paper. In summary, we can recognize the two results in the experiments above: First, there is still temporal information that existing models can learn to improve their performance. Secondly, temporal information can be further categorized into 'Movement information' and 'Temporal ordering information' to provide a more refined understanding of its features.

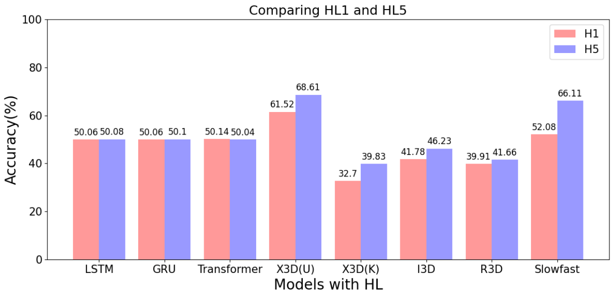

Fig. 4. Change in test accuracy when increasing the number of hidden layers in Binary classification model with Large data.

C. *Evaluate temporal ordering information*

The experiments were conducted from two perspectives: the fine-tuning layer and the size of training data. We investigated whether it is feasible to discern temporal order information in the feature vector of a video generated using an existing model. To accomplish this, we divided the classifier, distinguishing between the original and reversed videos, into 'Light model' and 'Deep model.'

The reason for distinguishing between the 'Light model' and the 'Deep model' was that the existing feature vectors were optimized for video classification, and we hypothesized that the classification task might be challenging due to insufficient training parameters. The 'Light model' only modified the output layer to binary within the existing video classification model. On the other hand, the 'Deep model' gradually increased the depth of the 'Light model' from 1 to 5 in the existing model to determine if the training parameters were sufficient. Additionally, to assess the detectability of temporal order from the feature vector, we divided the classifier training data into small data, utilizing only 10% of the total data, and large data, incorporating all available data. We conducted experiments with 7 base models, including 2 CNN-RNN-based models, a CNN-Transformer model[20, 17], an R3D, I3D, X3D, and a SlowFast model[7, 3, 11, 4]. For the CNN-RNN-based models and CNN-Transformer model, we employed EfficientNet[18] as the backbone to detect spatial information.

In Table 1, 'Base model training' and 'Base model test' refer to the base model's training accuracy and test accuracy, respectively. The remaining values indicate the test accuracy after transforming the base model into a binary classification model that detects Original and Reverse videos, where the expected accuracy is 50%. The table demonstrates that the values are randomly distributed, with the majority centered around 50%. The values in Table 1 are averaged from multiple experiments. The LSTM, GRU, Transformer, and X3D(U) models exhibited an error margin of ±5%, while the SlowFast model showed an error margin of ±10%. The X3D(K), I3D, and R3D models displayed an error margin of ±20% from the values. The relatively lower variation in the results obtained from the model trained on the UCF101 dataset is noteworthy. This observation could be attributed to the larger size of the UCF101 dataset compared to the Kinetics400 dataset used in the experiments.

In Fig 4, most values exhibit a slight increase in accuracy, around 50%, while X3D(U) and Slowfast demonstrate a notable improvement in accuracy. The Slowfast model employs two pathways: a slow pathway for detect-

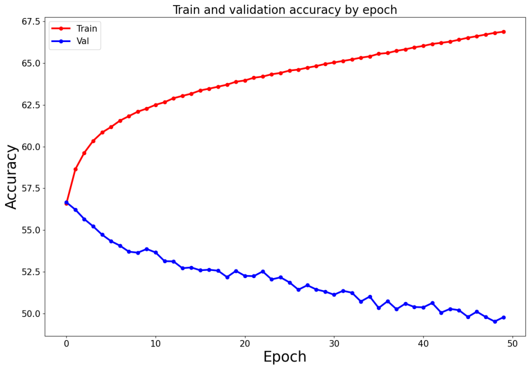

Fig. 5. The mean training and validation accuracies per epoch during the training of a binary classification model designed to discern between the original and reverse videos.

ing semantic information and a fast pathway for capturing motion. This signifies that two distinct timelines are analyzed within the same model. Consequently, the Slowfast model, compared to other models, is expected to incorporate a substantial amount of temporal information in the feature vector, resulting in the previous outcomes. It is anticipated that X3D(U) yielded similar results due to the complexity of the model compared to the data.

Fig 5 illustrates that as the number of epochs increases during classifier training, the training accuracy shows a consistent improvement, whereas the validation accuracy stabilizes around 50%. These findings imply that although the temporal ordering feature has the potential to be learned, the existing video classification model's feature extractor fails to extract meaningful and comprehensive temporal ordering features. In other words, enhancing the feature extractor's structure to capture temporal ordering information effectively would result in more informative feature vectors.

V. Conclusion

This paper emphasizes the need for more understanding of temporal information by video classification models and attributes this issue to the inadequate definition of temporal information within machine-learning based video classification problems. To address this limitation, the study proposes a detailed approach that divides temporal information into 'Movement information' and 'Temporal ordering information' and emphasizes the need for such distinction. To demonstrate this, we conducted experiments using a novel approach, employing a Stack that eliminates temporal information instead of the traditional method. Furthermore, the study evaluated whether existing video classification models can capture 'Temporal ordering information' through simple experiments. The findings revealed that most models failed to effectively incorporate 'Temporal ordering information,' except in cases where the models were significantly more complex than the data or specific tricks were employed, such as increasing the temporal order. The authors propose this novel perspective on defining 'Temporal information' in the context of video analysis to facilitate the development of more robust video classification models in future research endeavors.

VI. Acknowledgement

This work was supported by Korea Internet & Security Agency(KISA) grant funded by the Korea government(PIPC) (No.RS-2023-00231200, Development of personal video information privacy protection technology capable of AI learning in an autonomous driving environment)

References

[1] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Alexey Dosovitskiy, Lucas Beyer and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Georg Heigold Chen Sun Mario Lučić Anurag Arnab, Mostafa Dehghani and Cordelia Schmid. Vivit: A video vision transformer. *In: Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Jitendra Malik Christoph Feichtenhofer, Haoqi Fan and Kaiming He. Slowfast networks for video recognition. *In: Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[5] Maya Zohar Daniel Neimark, Omri Bar and Dotan Asselmann. Video transformer network. *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021.

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

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

[8] Karttikeya Mangalam Yanghao Li Zhicheng Yan Jitendra Malik Haoqi Fan, Bo Xiong and Christoph Feichtenhofer. Multiscale vision transformers. *In: Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6284–6835, 2021.

[9] Jiyoung Lee Jungin Park and Kwanghoon Sohn. Dual-path adaptation from image to video transformers. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2203–2213, 2023.

[10] Xiatian Zhu Jing Shao Junting Pan, Ziyi Lin and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. *Advances in Neural Information Processing Systems*, pages 26462–26477, 2022.

[11] Hirokatsu Kataoka Kensho Hara and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. *In: Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

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

[13] Zhicheng Yan Vedanuj Goswami Matt Feiszli Laura Sevilla-Lara, Shengxin Zha and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. *In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 535–544, 2021.

[14] Zhe Wang Yu Qiao Dahua Lin Xiaoou Tang Limin Wang, Yuanjun Xiong and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. *In European conference on computer vision. Springer, Cham*, pages 325–329, 2017.

[15] Makarand Tapaswi Piyush Bagad and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. *arXiv preprint arXiv:2301.02074*, 2023.

[16] Akash Kumar Praveen Tirupattur Shruti Vyas Yogesh Singh Rawat Rajat Modi, Aayush Jung Rana and Mubarak Shah. Video action detection: Analysing limitations and challenges. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[17] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 2017.

[18] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *In: International conference on machine learning. PMLR*, pages 6105–6114, 2019.

[19] Karen Simonyan Brian Zhang Chloe Hillier Sudheendra Vijayanarasimhan Fabio Viola Tim Green Trevor Back Paul Natsev Mustafa Suleyman Will Kay, Joao Carreira and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[20] Shawn Newsam Yi Zhu, Zhenzhong Lan and Alexander G. Hauptmann. Hidden two-stream convolutional networks for action recognition. *In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, Revised Selected Papers, Part III 14. Springer International Publishing December 2–6*, pages 363–378, 2018.