Do outliers matter? Assessing the Impact of Outliers on Video Classifications

Sayanton Dibbo and Cameron Keith
Department of Computer Science Dartmouth College
Hanover, NH 03755
`{sayanton.v.dibbo.gr, cameron.s.keith.26}@dartmouth.edu`

Abstract

Training sample quality impacts on deep learning model performances. While studies in the literature explored the association of outlier samples to model performance in modalities like text or images in the NLP and computer vision domains, it is relatively underexplored in the domain of video classifications. Researchers focused on anomaly detection or theoretical bounding of outliers towards video classifications. However, explicit, systematic empirical studies of the impacts of these outliers on video classification modeling are still yet to be explored. Therefore, to bridge this gap, in this work, we systematically analyze the impacts of outliers, specifically in-distribution outliers, on video classification performances and show that reducing the outliers from training can improve video classification performances. Our codes and models can be found at `https://github.com/ckeith26/video-outlier-optimization`

1 Introduction

With the ever-increasing data volume, deep learning models have become more popular, including applications on different domains and different data modalities Assran et al. [2023], Jin et al. [2018], Tong et al. [2024], Matthew et al. [2023], Dibbo et al. [2024], Ramachandram and Taylor [2017], Gilakjani et al. [2011], Vhaduri and Poellabauer [2019], Carlini and Wagner [2018], Wang et al. [2018], Xiao et al. [2024], Mullapudi et al. [2019], Fedorov et al. [2024]. For example, natural language processing (NLP) and computer vision (CV) have become more popular. However, video applications, commonly called video understanding, are relatively underexplored. Although the NLP and video both deal with sequential data, i.e., having time dependencies, video is comparatively more complicated since it deals with spatial and temporal dimensions.

Among the different properties of deep neural network-based models, *generalization* is an important property that assesses the model's real-life applicability. Researchers have focused on the generalization of the neural network models for NLP and CV applications. However, the generalization of video modeling has not been explored adequately. It is also clear from state-of-the-art (SOTA) research that the training data distribution significantly impacts model performance and, hence, the model's generalization capability. Motivated by SOTA research in other modalities like NLP, CV, or Audio domains, in this project, we explore the generalization of video modeling and the associated impact of the training data distribution on model performance. In particular, we want to systematically explore how the outliers impact model performance in video understanding and contribute to the generalizability of performance. Outliers can be (1) in-distribution outliers, which are necessarily training samples with distinct patterns, and (2) out-of-distribution (OOD) samples generated through augmentation or from other similar datasets than the training set, e.g., for the first category, an outlier can be playing basketball video in blurred that resumes playing soccer. We want to explore the first category, i.e., in-distribution outliers since those are necessarily training samples and can impact the performance of the model at a large scale. Specifically, we aim to explore the following research

Preprint. Under review.

question **RQ:** Do outliers (in-distribution) impact video classification performance? If so, how much, and whether the impact shows a monotonically increasing pattern, i.e., constantly impacting in a linear fashion or whether there exists an optimal threshold beyond which the impact drops or changes the pattern?

We make the following contributions in this paper:

- We provide a systematic study of the impacts of outliers on video classification through Isolation Forest (IF)-based outlier detection and thresholding-based outlier removal approaches
- We perform an exhaustive analysis on the impacts of in-distribution outliers towards the performance of video classification task
- We compare performances of video classification tasks among the baseline model (with all outliers) and outlier-removed models utilizing our proposed approaches

2 Related Work

Action recognition and optimization have been widely researched over the past few years. We build on prior works from two main categories: (1) outlier and out-of-distribution (OOD) data detection, and (2) video classification and anomaly detection.

2.1 Outlier and OOD Detection

Many works have focused on detecting outliers and optimizing model performance. Schubert et al. [2014] provided a generalized view on outlier detection across different domains, including video, and a theoretical foundation for understanding outliers. Ren et al. [2019] proposed a method for detecting OOD samples using likelihood ratios. Liu et al. [2020] introduced an energy-based method for detecting OOD samples in image classification tasks, while Hsu et al. [2020] proposed methods for detecting OOD images without the need for OOD data during training. Zhang et al. [2021] investigated the generalization of probabilistic image models to OOD data, focusing on image data and probabilistic modeling approaches.

However, these papers do not specifically address their impact on deep learning models for video classification. Ren et al. and Liu et al. focused on image datasets rather than video datasets, emphasizing OOD detection rather than the impact of outliers on model generalization. While Schubert et al. developed a framework for detecting outliers in various modalities, they did not specifically address the performance impact on deep learning models for video classification. Our approach differs as we analyze the effect of detecting and removing outliers on training video classification models.

2.2 Video Classification and Anomaly Detection

Researchers have attempted to detect anomalies in datasets to boost performance for downstream tasks, including video classification. Wang et al. [2018] concentrated on video anomaly detection and localization using a joint video representation and an OCELM (One-Class Extreme Learning Machine) model. Chang et al. [2022] explored a method for video anomaly detection using spatio-temporal dissociation. Georgescu et al. [2021] presented a self-supervised and multi-task learning approach for anomaly detection in video data. In an older paper, /citeay2001 present new techniques for outlier detection in high-dimensional data by examining the behavior of data projections, addressing the challenges of data sparsity and proximity in such spaces. Their method identifies outliers through the analysis of density distributions in lower-dimensional projections, rather than relying on full-dimensional proximity measures which are less effective in high-dimensional contexts.

While these studies focus on anomaly detection, they do not directly address the broader impact of outliers and OOD data on video classification models. Mahmood et al. [2021] proposed a method for anomaly event detection and localization in video clips using global and local outliers. Pang et al. [2020] discussed a self-trained deep ordinal regression approach for end-to-end video anomaly detection. Zhu et al. examined context-aware activity recognition and anomaly detection in video data Zhu et al. [2013]. Deb et al. [2018] attempted to detect violence in videos using a novel outlier-resistant feature encoding method. Mullapudi et al. [2019] argued for online model distillation to

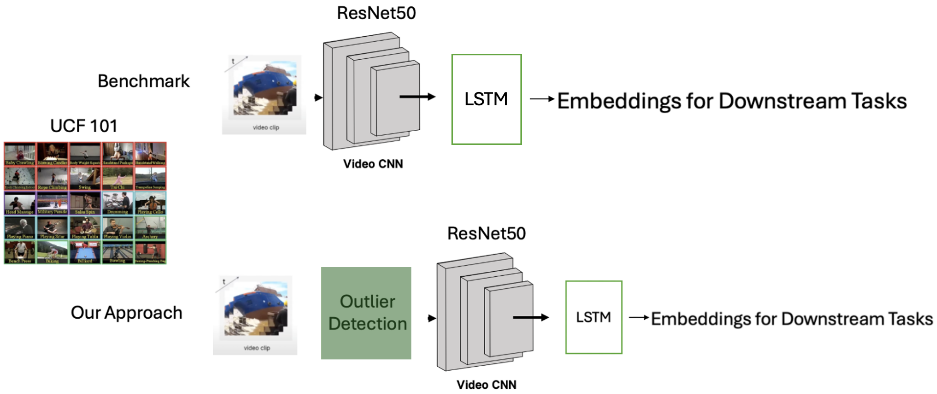

Figure 1: Overview of our proposed approach. First, we train the baseline benchmark m_{vbase} model with all training samples using the ResNet50 backbone. In addition, we apply our outlier removal approach to train our outlier removed models $m_{vdistil}$ with different outlier removal ψ values. We finally compare all model performances.

improve the efficiency of video inference, addressing efficiency but not directly tackling the questions of generalization and the impact of outliers or OOD data on model performance Wang et al. [2020] present a novel approach called Cluster Attention Contrast for video anomaly detection, which aims to establish reliable subcategories of normality by employing a contrastive representation learning task

Our research is distinct in that we methodically detect outliers, compare the results between base models and various outlier detection models, and aim to improve model generalization in video classification.

3 Our Proposed Approach

In this work, we first design a video classification model, then we work with the training data and consider different ways to generate training samples with different outliers to observe their impact while including them in the model during the training phase.

In Figure 1, we present the overview of our approach. We first train a video classification model as described below. In the next step, we apply the outlier detection technique and consider different thresholds for removing different numbers of outlier samples from the pool of all training samples. At each time, we then retrain models with the outlier removed sets of samples and compare the performances of those new models with the baseline model (i.e., trained including all outliers). We describe our approach in detail below.

3.1 Baseline Model Training

Our first step in this work is to train a target model for video classification. We train the model with all samples, including outliers. We consider such a model as our baseline model m_{vbase} for video classification. Our goal is to compare this model's accuracy on test sets with the outlier-removed model performances and draw a conclusion based on findings about the impact of outliers on video classification performances.

To train the baseline model m_{vbase}, we consider the model backbone as the hybrid LSTM CNN model. In particular, in the first part of the model m_{vbase}, we consider incorporating the ResNet-50 pretrained model, which will generate the embeddings from all training samples D_{tr}. We choose ResNet-50 pretrained model as opposed to other pretrained models since ResNet-50 is trained on an image dataset, i.e., an ImageNet dataset, instead of action video datasets that are used to train other pretrained models. This is because any pretrained models that are trained with action datasets (video) can impact model performance highly, and that would interfere with the model performance due to sample quality since we aim to identify the impact of sample quality on model training for video classification tasks. In the next layer after the ResNet-50, we consider adding the LSTM layers and, finally, different linear layers up to the classification layer. This is the overall architecture of the video classification model m_{vbase}. For our models trained with removed outliers, i.e., distilled models $m_{vdistil}$, we also use the same model architecture but with different training sets of samples so that the model architecture does not impact the performances.

3.2 Outlier Detection and Outlier Removal

We apply the Isolation Forest (IF) technique (unsupervised) to identify the outliers in the training data (in-distribution). We first create an IF model with the outlier detection variable set to a specific threshold, i.e., ψ. For experimental purposes, we consider $\psi = 5\%, 10\%, 15\%$. After creating the IF, all training samples are fit to the forest and passed to get prediction scores. Based on the returned scores from the IF, we identify two categories of samples, i.e., non-outliers $D_{tr_{nout}}$, and outliers $D_{tr_{out}}$. For example, if the threshold $\psi = 5\%$, 5% of the total training samples (outliers) will be in the $D_{tr_{out}}$ and the remaining non-outlier samples set will belong to $D_{tr_{nout}}$. In summary, $D_{tr_{nout}}$ is the set of non-outlier training samples. We train our distilled models using the $D_{tr_{nout}}$ sets.

3.3 Training Distilled Models

We train the distilled models $m_{vdistil}$ with the non-outlier sample sets $D_{tr_{nout}}$ (Section 3.2) obtained through applying our outlier removal technique utilizing the IFs. We obtained different $D_{tr_{nout}}$ sets using different threshold ψ. For each of the $D_{tr_{nout}}$ sets, we train the $m_{vdistil}$ models, where these models have the same architecture as the baseline classification model m_{vbase} (Section 3.1). We finally tested each of the $m_{vdistil}$ models and baseline classification model m_{vbase} on the same test set to compare the performances of different models trained with different training sets, highlighting the impact of outliers inclusion in the training samples.

4 Experiments and Results

In this section, we present the experimental setup for our proposed technique and detailed analysis, along with discussions of our findings. We illustrate our analysis step by step.

4.1 Experimental Setup

We trained both our baseline model m_{vbase} and distilled models $m_{vdistil}$ for 15 epochs on the UCF-101 dataset. We consider training m_{vbase} with 100k training samples, i.e., D_{tr}. We then apply steps illustrated in Section 3.2 to obtain the outlier removed sets, i.e., $D_{tr_{nout}}$ used to train $m_{vdistil}$ models. We also tested all models with 20k test samples from the dataset.

We trained all models with 128 batch sizes with cross-entropy loss optimized with the SGD optimizer. Specifically, we consider training with learning rate $lr = 0.01$ and momentum $m = 0.9$. In Table 1, we present our hardware details, including GPU and RAM for the experiments.

For comparisons, we consider two major performance metrics, i.e., TOP-1 ACCURACY and TOP-5 ACCURACY. We compare (in terms of these performance metrics) both m_{vbase} and $m_{vdistil}$ with different numbers of outliers removed.

Table 1: Hardware Details of our Experiments.

Parameter	Value
CPU	INTEL i5-13600k
Core	14
RAM	16GB
GPU	NVIDIA 3090
VRAM	24GB
Space	1TB

Table 2: Performance comparisons of our model with and without different outliers (Note that ψ indicates the percentage of outliers removed from training sets).

Accuracy	ψ	Top-1 Acc	Top-5 Acc
m_{vbase}	0%	**0.590**	0.827
$m_{vdistil}$	5%	0.583	**0.836**
$m_{vdistil}$	10%	0.583	0.816
$m_{vdistil}$	15%	0.564	0.806

Table 3: Workload for Our Project.

Category	Sayanton	Cameron
1. Determine Datasets and Models	✓	✓
2. Choose Outlier Algorithms	✓	✓
3. Develop Optimization Framework	✓	✓
4. Present Results	✓	✓

4.2 Performance Comparisons and Analysis

In Table 2, we present the performances of our m_{vbase} and $m_{vdistil}$ models, tested on the same test set. Observe that, as expected for each model, the test accuracy improves in the top-5 accuracy metric compared to the top-1 accuracy. Note that the clean test top-1 accuracy is lower since we trained models only for 15 epochs. A higher number of epochs can improve the top-1 accuracy further.

While we compare the performances among the m_{vbase} and $m_{vdistil}$ models, we observe that the model m_{vbase} has 0.01% worse top-5 accuracy than $m_{vdistil}$ model with $\psi = 5\%$. This indicates the model with 5% outliers removal boosts performances in video classification models. However, the baseline model has a slightly better top-1 accuracy, which is attributed to random noise in a single run. A more comprehensive experiment with multiple runs (mean ± std) could better illustrate the scenario.

We further run models by removing higher numbers of outliers, e.g., $\psi = 10\%, 15\%$. We observe that both top-1 and top-5 accuracies of $m_{vdistil}$ models are lower compared to the baseline m_{vbase} model, respectively. We believe the outlier removal can also negatively impact model performances. This is attributed to conclude that, there might exist an optimal value of ψ, which can positively impact video classification performances. Further increasing or reducing ψ can negatively impact model performances. One possible future research direction can be identifying a generalizable optimal threshold of outlier removal, i.e., ψ that can positively impact model performances, i.e., generalizability.

4.3 Project Workload

We identify the major workloads below. In Table 3, we present the details of the team workloads and breakdown. We follow this throughout our project during the term.

5 Conclusion

While analyzing the impacts of outliers on model performance is not relatively new in NLP or computer vision tasks, it is indeed an underexplored direction for the video classification domain. Although researchers have investigated the theoretical bounding of outliers and different outliers handling, i.e., anomaly detection techniques, there still remains a gap to systematically study the impacts of outliers on video classification task model performances. Our study bridges this gap by revealing outliers really matter in video classification. In particular, our analysis shows that, while outliers, in general, can reduce model performances, removing large-scale outliers can downplay overall performances. This calls for further investigations on the optimal threshold for outlier removal to boost video classification performances at scale.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*, pages 1–7. IEEE, 2018.

Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition*, 122:108213, 2022.

Tonmoay Deb, Aziz Arman, and Adnan Firoze. Machine cognition of violence in videos using novel outlier-resistant vlad. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 989–994. IEEE, 2018.

Sayanton V Dibbo, Adam Breuer, Juston Moore, and Michael Teti. Improving robustness to model inversion attacks via sparse coding architectures. *arXiv preprint arXiv:2403.14772*, 2024.

Alex Fedorov, Eloy Geenjaar, Lei Wu, Tristan Sylvain, Thomas P DeRamus, Margaux Luck, Maria Misiura, Girish Mittapalle, R Devon Hjelm, Sergey M Plis, et al. Self-supervised multimodal learning for group inferences from mri data: Discovering disorder-relevant brain regions and multimodal links. *NeuroImage*, 285:120485, 2024.

Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12742–12752, June 2021.

Abbas Pourhossein Gilakjani, Hairul Nizam Ismail, and Seyedeh Masoumeh Ahmadi. The effect of multimodal learning models on language teaching and learning. *Theory & Practice in Language Studies*, 1(10), 2011.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10951–10960, 2020.

SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 307–324, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf`.

Sawsen Abdulhadi Mahmood, Azal Monshed Abid, and Sadeq H Lafta. Anomaly event detection and localization of video clips using global and local outliers. *Indones. J. Electr. Eng. Comput. Sci*, 24:1063, 2021.

Jagielski Matthew, Nasr Milad, Choquette-Choo Christopher, Lee Katherine, and Carlini Nicholas. Students parrot their teachers: Membership inference on model distillation. *arXiv preprint arXiv: 2303.03446*, 2023.

Ravi Teja Mullapudi, Steven Chen, Keyi Zhang, Deva Ramanan, and Kayvon Fatahalian. Online model distillation for efficient video inference. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 3573–3582, 2019.

Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf`.

Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data mining and knowledge discovery*, 28:190–237, 2014.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

Sudip Vhaduri and Christian Poellabauer. Multi-modal biometric-based implicit authentication of wearable device users. *IEEE Transactions on Information Forensics and Security*, 14(12): 3116–3125, 2019.

Siqi Wang, En Zhu, Jianping Yin, and Fatih Porikli. Video anomaly detection and localization by local motion based joint video representation and ocelm. *Neurocomputing*, 277:161–175, 2018.

Ziming Wang, Yuexian Zou, and Zeming Zhang. Cluster attention contrast for video anomaly detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2463–2471, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413529. URL `https://doi.org/10.1145/3394171.3413529`.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. *arXiv preprint arXiv:2404.09403*, 2024.

Mingtian Zhang, Andi Zhang, and Steven McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34:3811–3823, 2021.

Yingying Zhu, Nandita M. Nayak, and Amit K. Roy-Chowdhury. Context-aware activity recognition and anomaly detection in video. *IEEE Journal of Selected Topics in Signal Processing*, 7(1): 91–101, 2013. doi: 10.1109/JSTSP.2012.2234722.