EfficientViT for Video Action Recognition

Derek Jin
Noble and Greenough School
Dedham, MA 02026, USA
djin25@nobles.edu

Shengyou Zeng
Tradelegs, Inc.
Boston, MA 02135, USA
shengyou@tradelegs.com

***Abstract*—Video action recognition is a critical challenge in a wide range of practical applications. Most of the video recognition models are computationally expensive and energy intensive. EfficientViT is a new family of vision transformers that offer high efficiency while maintaining state of the art accuracy on vision tasks such as image classification and semantic segmentation including on daily devices and cloud computing. However, the effectiveness has only been validated on images so far. In this project, we extend EfficientViT to video processing by replacing 2D convolution with 3D convolution. The updated EfficientViT model B1-r288 was successfully trained with the Epic-Kitchens-100 video dataset, achieving top 1% verb accuracy of 27.7%, a 5.4% improvement than the original model.**

Keywords—EfficientViT, Video action recognition, Epic-Kitchens-100

I. Introduction

Video action recognition promises innovative solutions across many industries. It can streamline manufacturing, optimize agriculture, enhance education, improve driving safety, strengthen security, transform entertainments, and revolutionize healthcare [1, 2]. Most existing action recognition models use 3D convolution or attention mechanisms to model the spatial-temporal relation [3,4,5] and are computationally heavy during both training and inference, prohibiting them from a wider range of real-world applications. Another line of approaches uses more efficient operations such as running 2D convolutions [6,7] on separate frames. However, the efficacy is mostly validated on video datasets biased towards static scenes [3]. Differentiating subtle differences between human actions, such as "open refrigerator" vs. "close refrigerator", requires more complex temporal modeling.

In this paper, we describe research on adapting EfficientViT models [8] for video action recognition. EfficientViT is a new family of vision transformers that replace the softmax attention [9] in quadratic complexity with a lighter-weighted ReLU linear attention [10]. Although EfficientViT has been tested on image processing tasks, its efficacy on videos is unexplored. We studied two ways of adaptation. The first was to apply EfficientViT on individual frames and average the predictions similar to [6]. The second was to upgrade EfficientViT to 3D convolution to model the spatial-temporal relationship. To further enhance the pre-training capability on large-scale images, we adopted the inflation technique [3] so that the 2D convolution kernels learned on ImageNet can be used as a starting point for learning 3D convolution kernels.

We conducted experiments on the Epic-Kitchens 100 (EK100) video action classification dataset. We observe that the performance gain of applying EfficientViT on individual frames quickly saturates when the number of frames was increased, indicating that the original image-level modeling fails to capture the rich temporal information. In contrast, adapting EfficientViT to 3D convolutions achieves better performance.

While artificial intelligence is transforming the world at an unprecedented pace, it is also fueling a boom in data centers and challenging the global energy demand [11]. EfficientViT is fast and accurate enough to be deployed on daily devices and cloud computing, thus its efficiency has a great potential to significantly reduce energy usage and minimize the climate impact of AI applications.

II. Related Work

Video recognition is the process of training a computer to identify human actions in a video. It classifies the actions in the video into predefined sets of action classes [12,13]. Video recognition encompasses a wide range of high-impact societal applications, such as virtual and augmented reality [1] and senior health care [2].

Many video recognition models have been proposed in the last decade, including two-stream [14], 3d convolution [3,15] and transformer-based approaches [4,5]. However, most of them are computationally expensive and energy intensive, thus are not suitable for real world deployment.

Cai et al introduced a new family of vision transformers[8]. In the paper, models of this new family were applied to image classification and trained on ImageNet. The model EfficientViT-L2-r384 obtains 86.0% top1 accuracy, a highly competitive performance against the state of the art (SOTA): +0.3 accuracy gain over EfficientNetV2-L and 2.6x speedup on A100 GPU [8].

Zhao and Krähenbühl [16] identified IO, CPU, and GPU computation to be the three bottlenecks in the video training pipeline. Through optimizing these three bottlenecks, they established a highly efficient video training pipeline that achieves higher accuracies with 1/8 of the computation compared to prior SOTA. The improved video training pipeline makes it possible to train an SOTA video model on a single machine with eight consumer-grade GPUs in a day. However, this still requires a GPU workstation which is prohibitive to train for common individuals.

III. METHODS

A. Preliminaries: EfficientViT

The core of EfficientViT is a new multi-scale linear attention module that enables the global receptive field and multi-scale learning with hardware-efficient operations. EfficientViT replaces the inefficient heavyweight softmax attention with a lightweight ReLU linear attention. This replacement reduces computational complexity from quadratic to linear. ReLU linear attention also reduces the memory footprint. The efficient ReLU linear attention is augmented by introducing a multi-scale learning module. Multi-scale learning is accomplished by aggregating nearby tokens (or patches) with small-scale convolutions to generate multi-scale tokens. Applying ReLU linear attention to multi-scale tokens combines the global receptive field with multi-scale learning. ReLU linear attention cannot generate concentrated attention maps, making it weak at capturing local information. This is addressed by enhancing ReLU linear attention with convolution by inserting a depthwise convolution layer (DWConv) in each FFN layer. The ReLU linear attention captures context information and the FFN+DWConv captures local information. With the use of ReLU linear attention to reduce computational complexity and improve efficiency, token aggregation to achieve multi-scale learning, and ReLU linear attention on multi-scale tokens to accomplish global receptive field, EfficientViT strikes a balance between performance and efficiency, making it practical to deploy high resolution dense prediction on edge devices with limited resources, like mobile and IoT devices.

B. Adapting EfficientViT for Video Inputs

We explored two ways to adapt EfficientViT for video inputs.

1) EfficientViT on individual frames. Similar to [6], we applied EfficientViT on individual frames and average the predictions. Let's denote the input frames I_1, I_2, ..., I_T , and the EfficientViT backbone F. The final activation is denoted by $\frac{1}{T}\sum_{t=1}^{T} F(I_t)$, which is later projected by a classification head to give the prediction.

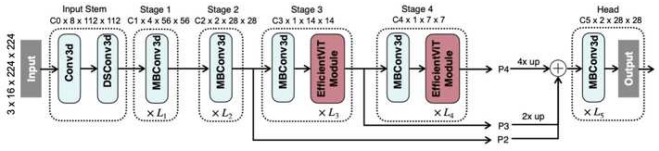

Fig 1: Macro Architecture of EfficientViT Updated for Video. We adapted the macro architecture of EfficientViT (Cai *et al* [8]). The input shape is 3x16x224x224. All 2D convolutions in the original architecture are updated to 3D convolutions. Other aspects of the architecture are preserved.

2) Adapting EfficientViT with 3D convolutions. As will be illustrated in the experiments, applying EfficientViT on individual frames failed at complex temporal modeling. To fix this, we incorporated 3D convolutions to model the nonlinear spatial-temporal relation between neighboring pixels. We replaced 2D convolutions with 3D by inflating the convolution kernel through the time axis. We also followed I3D [3] to copy the convolution weights by T times along the time axis and divide each value by a constant factor T, so that the weights trained from ImageNet can be reused when training on videos. Because convolution is a local operator, the added complexity is much lower compared to the global attention operation proposed in [16].

IV. EXPERIMENTS

A. Datasets

A resized version of the EK100 video dataset was downloaded from the AVION GitHub repository [17].

EK100 consists of 100 hours of egocentric videos, recorded by 37 participants in 45 kitchens across Europe and North America. The dataset contains 20 million video frames, 90 thousand action segments, 97 verb classes and 300 noun classes. Image files were extracted from the action video segments in the dataset by using the information in the train and validation annotation files (downloaded from [18]).

In the EK100 train and validation annotation file, a video segment is identified by narration_id, and the start and end of a video segment is specified by start_frame and end_frame. The combination of verb_class and noun_class value from the annotation files is used to form class names. Sixteen frames are extracted from each video segment. Shortest video segment has 10 frames, longest 14857 frames, with an average of 172.4 frames per video segment. A total of 1,261,882 images were extracted from 67,217 video segments in 9667 videos.

B. Model Specification

Two pretrained EfficientViT models (b1-r288 and l1-r244) and the EK100 video dataset were used in the experiments. Experiments were carried out in two stages. First, the pretrained EfficientViT models were trained on datasets of images extracted from video segments of specific actions. This stage assesses the performance of the models on series of images from the same action in a video. In the second stage, EfficientViT models were updated for video processing, and then the pretrained models were trained on the EK100 video dataset.

The train set was used to train the pretrained b1-r288 and l1-r244 model, downloaded from EfficientViT GitHub repository [19]. Training was performed using an Amazon Web Services (AWS) g5.8xlarge instance, which is equipped with a single NVIDIA A10G GPU (24GB) and a 16-core AMD EPYC CPU. We trained the model for 100 epochs (105,000 total iterations).

C. Results

EfficientViT on Individual Frames. Models were trained with the train dataset consisting of 8 frames extracted from video segment for each action in the videos. Model b1-r288 and l1-r244 were trained until accuracy saturated at 22.35% and 22.55% of top 1% verb accuracy, respectively.

Class prediction accuracy of the trained models were measured using validation sets of 1, 2, 4, 8, and 16 frames per video segment extracted from the validation set of the EK100 dataset. The action class prediction accuracies were presented in Fig. 2. The number of frames per video segment makes a slight

difference on prediction accuracy. Highest prediction accuracy of model b1-r288's is 15.6%, and that of model l1-r244 17.3%.

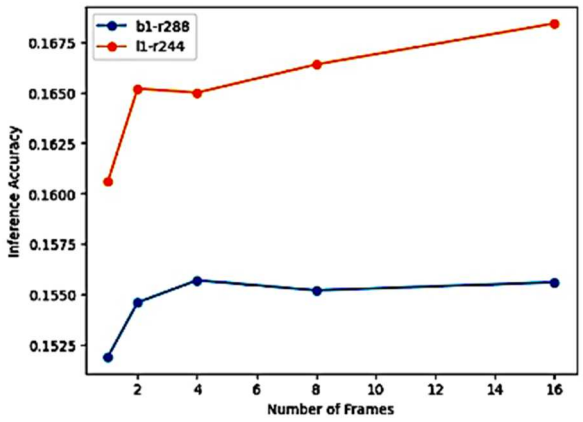

Fig. 2. Class prediction accuracy increases slightly with the number of frames per video segment. Experiments run on AWS 5.8xlarge instance.

Fig 3 shows per-image inference time vs. the number of frames extracted from a video segment. In both models, average inference time increased with the number of frames per video segment, from ~6 milliseconds for 1-frame to ~11 milliseconds for 16-frame validation set.

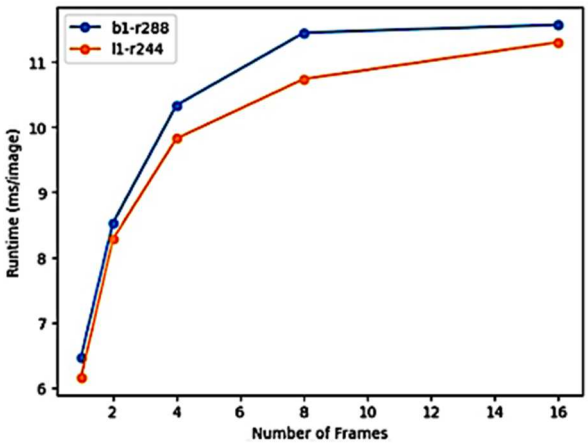

Fig. 3. Inference runtime increases with the number of frames per video segment. Experiments run on AWS 5.8xlarge instance.

EfficientViT with 3D Convolution. We first trained the b1-r288 model on the EK100 video dataset with random initialization. Model was trained until top 1% verb accuracy saturated at 23.7%. Next, we trained b1-r288 with 3D convolution on the EK100 video dataset with ImageNet initialization and achieved top 1% verb accuracy of 27.7%. The second experiment achieved higher accuracy than the first because in the second experiment the model was pretrained and initialized with more image data. Note that both experiments achieved higher accuracy than EfficientViT on individual frames (22.3%).

V. Discussions

Training of original EfficientViT models from Cai et al [8] on individual frames quickly saturates. This indicates that the original image-level modeling fails to capture the rich temporal information in videos. We successfully adapted original EfficientViT models for video processing by updating 2D convolutions to 3D. Updated EfficientViT was successfully trained on a resized, chunked version of the EK100 egocentric video dataset.

The model runs in ~11 milliseconds on a single GPU instance (AWS g5.8xLarge), which equals to 90 FPS. This bodes well for deployments of EfficientViT models to solve image processing problems in real time. Future work will include investigating accuracy and runtime for video class predictions.

As a final note, training of transformer models, especially with large video datasets like the project presented here, requires large amount of resources [16]. We have limited ability to access AWS resources for training the models. It's reasonable to expect better results if more resources are available.

Acknowledgment

The authors thank Mr. Yue Zhao at University of Texas at Austin and Prof. Nicu Sebe at University of Trento, Italy for their advice on this video action recognition research.

References

[1] I. Chamusca, I. Winkler, T. Pagano, R. Loureiro, A. Santos, and T. Murari. Machine learning for object and action recognition in augmented and mixed reality: A literature review., 2023. URL https://www.preprints.org/manuscript/202310.0200/v1.

[2] H. Sun and Y. Chen. Real-time elderly monitoring for senior safety by lightweight human action recognition. In 2022 IEEE 16th International Symposium on Medical Information and Communication Technology (ISMICT)., volume 2022, pages 1–6. IEEE, 2022. URL https://ieeexplore.ieee.org/document/9828343/.

[3] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, 2017.

[4] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Luciˇc, and C. Schmid. Vivit: A video vision ´ transformer. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6836–6846, 2021.

[5] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? In ICML, volume 2, page 4, 2021.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In European conference on computer vision, pages 20–36. Springer, 2016.

[7] L. Wang, Z. Tong, B. Ji, and G. Wu. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1895–1904, 2021.

[8] H. Cai, J. Li, M. Hu, C. Gan, and S. Han. Efficientvit: Multi-scale linear attention for high resolution dense prediction, 2023. URL https://doi.org/10.48550/arXiv.2205.14756.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[10] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In

International conference on machine learning, pages 5156–5165. PMLR, 2020.

[11] A. Chow. How ai is fueling a boom in data centers and energy demand, 2024. URL https: //time.com/6987773/ai-data-centers-energy-usage-climate-change/.

[12] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li. A comprehensive study of deep video action recognition., 2020. URL http: //arxiv.org/abs/2012.06567.

[13] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian. Vision transformers for action recognition: A survey, 2022. URL http://arxiv.org/abs/2209.05700.

[14] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27, 2014.

[15] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.

[16] Y. Zhao and P. Krahenbuhl. Training a large video model on a single machine in a day, 2023. URL https://arxiv.org/pdf/2309.16669.

[17] Y. Zhao. Epic-kitchens-100 (ek-100), resized version, 2023. URL https://github.com/ zhaoyue-zephyrus/AVION/tree/main/datasets#epic-kitchens-100-ek-100.

[18] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, and J. Ma. Epic kitchens-100 (ek-100) annotations, 2022. URL https://github.com/epic-kitchens/ epic-kitchens-100-annotations.

[19] H. Cai. Efficientvit github repository, 2023. URL https://github.com/mit-han-lab/ efficientvit/blob/master/applications/cls.md#pretrained-models.