Algorithmic Video Reduction by Estimating Chunk Significance Using Spatial-Temporal Masked Auto Encoders

***Reshab Chhabra**
Boston University
reshabc@bu.edu

***Mark D. Yang**
Boston University
mdyang@bu.edu

Abstract

This research paper proposes a new algorithmic video reduction technique called *CARV ST MAE* — Chunk Algorithmic Reduction for Videos (CARV) that utilizes spatial-temporal (ST) masked autoencoders (MAEs) [1, 2] to estimate chunk significance and preserve informational content while downsizing the video. Although transformer-based video compression techniques exist [3], they do not preserve ST properties. The proposed approach prioritizes the preservation of important ST chunks by sampling an ST MAE to estimate their "importance" and then applying an algorithmic compression approach. The resulting reduced videos can be used for various purposes, including training machine learning models. The results obtained were mostly inconclusive but seemed promising in terms of maintaining ST properties. Further investigation into the CARV method and its practicality are needed.

1 Introduction

Motivation Current video compression methods either lose the spatiotemporal aspect or are not designed for temporal reduction. Our approach utilizes ST MAEs to capture the connections between video chunks while approximating the relative significance of each chunk through sampling of the marginal mask losses.

By incorporating chunk prioritization, we can retain valuable information that is otherwise lost in conventional video reduction methods. This approach can then facilitate the reduced video creation, which would retain sufficient informational content. This video could then be used to train video-oriented models more quickly without compromising their performance.

Background and Related Work

Masked Autoencoders A masked autoencoder (MAE) is a type of autoencoder where some of the input units are randomly masked during training. This makes MAEs useful in tasks that involve missing data, including image or speech denoising. Over the years, MAE models have proven highly effective across various tasks. For instance, Kaiming He, Xinlei Che, et al. [5] demonstrated their remarkable performance as visual learners, contributing significantly to their widespread adoption. These models have their roots in denoising autoencoders [7], which are neural networks trained to reconstruct their input. In terms of video, several studies have explored the use of MAEs in video processing. For instance, Zhan Tong, Yibing Song, et al. developed *VideoMAE* [2], a video-specific variant of MAE. Similarly, Christopher Feichtenhofer, Haoqi Fan, et al. introduced *STMAE* [1] – a spatial-temporal variation of an autoencoder that can process temporal data e.g. video.

Chunk Chunk refers to the selected elements that are not hidden during the masking process of a video. In this paper, we investigate whether our proposed chunk sampling technique can significantly improve the performance of ST MAE videos compared to a random masking approach by viewing how both types of videos' comparison metrics. By selecting chunks of video frames, we aim to improve the spatiotemporal representation of the input data, which can enhance the network's ability to learn relevant features and improve its reconstruction accuracy.

Video Metrics Video metrics are quantitative measures used to evaluate the quality of video content. As there are many, determining the quality of a video is non-trivial due to the ambiguity of the task. This paper defines video quality through several video metric techniques that compare a reference video to a modified one. Commonly used metrics include Mean Square Error (MSE), Structural Similarity (SSIM) [9], Peak Signal-to-Noise Ratio (PNSR), and Spatio-Temporal Reduced Reference Entropic Differencing (ST-RRED) index [10]. A lower MSE signifies a higher-quality video. A higher mean SSIM signifies a higher overall similarity. A higher PNSR signifies a higher quality video. ST-RRED is split into full reference (ST-RRED) and reduced reference (ST-RRED SSN). A higher ST-RRED index signifies a higher quality video and correlation among adjacent frames.

It is worth noting that each video metric may not directly correlate with the desired results due to differing definitions of quality.

Problem This project aims to answer several research questions related to improving the performance of ST MAEs. Firstly, we investigate whether using the average reconstruction MSE is a sensible metric for evaluating the importance of video chunks. Secondly, we explore which chunks are deemed important and how they relate to features like motion. We then analyze what masking ratios give the best chunk importance in terms of reconstruction quality. Finding the optimal ratio is important as too little data can result in poor reconstruction accuracy, while too much data can lead to exploding variance. Finally, we assess whether the reconstructed information can be easily translated to useful information, which relates to the feasibility of preserving both critical and ST information concurrently.

2 Methods

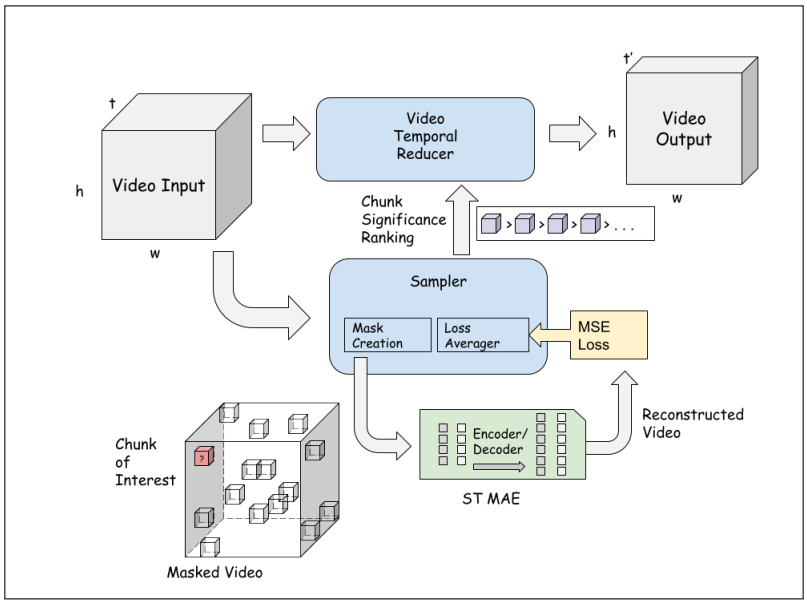

Figure 1: Model diagram of the overall *CARV ST MAE* algorithm

The study employs a pre-trained video MAE model to process video inputs. The videos are preprocessed into normalized arrays of 16 frames by 224 pixels by 224 pixels by 3 channels ($16 \times 224 \times 224 \times 3$) and split into $2 \times 16 \times 16 \times 3$ chunks. This approach maintains consistency and uniformity by applying this standard to all observed videos. These dimensions are additionally enforced upon using the pre-trained model [1]; hence, in the future, we hope to expand the model's usability by developing our own flexible ST MAE model.

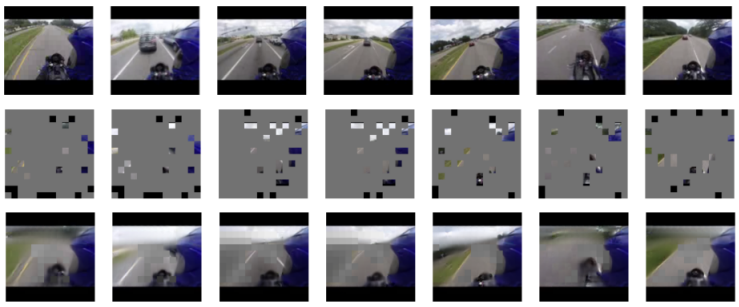

Figure 2: Example of a reconstructed ST MAE video (bottom) using a mask ratio of 90% (middle), compared with the original video (top)

The MAE model is utilized to create a reconstruction of all masked pixels by passing the mask and video (Figure 2) to the ST MAE model. To obtain an averaged significance for each chunk involved, this process is repeated multiple times using different randomized masks, sampled without spatial preference. Each chunk would then be evaluated by the Mean Squared Error (MSE) metric. This metric is responsible for developing a chunk significance value — how well the average image fragment containing that chunk performs compared to the original video.

In consideration of the influence of additional hyperparameters on the metric in question, a rigorous analysis of the MSE value distribution was performed on an ST MAE-masked video, tuning the hyperparameters (Figure 3).

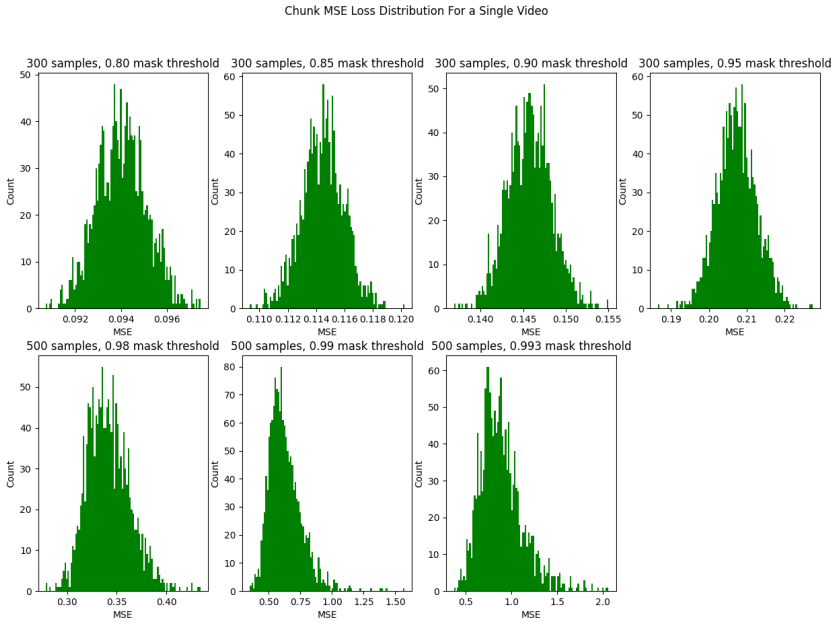

Figure 3: Chunk reconstruction MSE per masking ratio on a single video. Notice how both the shape and values change as masking increases.

In deciding an appropriate masking ratio to perform the compression, it is desirable to have a mix of low values but a high standard deviation of average chunk MSE values. Intuitively, most information is extracted from a mixture of bad and good reconstruction losses, which can capture chunk dependencies. Hence, the masking ratio 98% was selected with a sample size of 500. Figures 4 and 5 demonstrate the outcome of this result.

Heatmap of the 150 most (green) and least (red) important chunks constructed from randomly sampling the ST MAE model chunks with 500 samples and a 98% masking ratio

Figure 4: Person squeegeeing a window (semi-static movement)

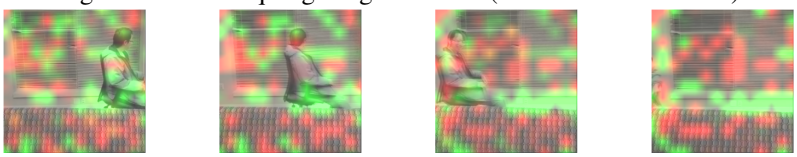

Figure 5: Person moving in a chair (dynamic movement)

Examining from a qualitative level, the colors on the heat map potentially are directly correlated with the level of movement and noise in the video. To exemplify its relation, the following classification label scheme was defined (in order from least to greatest movement respectively): static, semi-static, semi-dynamic, and dynamic. Our results will also be stratified in this manner.

In order to create output videos, the significance of each chunk is then utilized to determine the speed of frame changes. Specifically, lower chunk significance values for a chunk result in faster frame changes, meaning higher average chunk MSE values are given slower frame changes. The motivating idea behind this choice is that high-impact regions of video do not need to be refreshed as often, while the high-detailed low-impact regions of the video require a higher refresh rate to preserve information.

Moreover, as the videos in the ST MAE model are inputted as $16 \times 224 \times 224 \times 3$, CARV scales the chunks wider to accommodate for the video's actual dimension: the original video's quality is used, but the chunks have the described chunk-importance-dependent delays.

3 Results

Figure 6: Brief illustration of CARV video output. As the video continues, each chunk is updating at a different rate. For more detail, a *gif* file is displayed in the paper's GitHub repository *README*.

Reduction Method	MSE	$\overline{\text{PSNR}}$	$\overline{\text{SSIM}}$	ST-RRED	ST-RRED SSN
Video of a person playing piano (very static)					
CARV	11.1892	33.7121	0.9609	143.9922	13.3402
Every Nth Frame	9.2403	38.1359	0.9744	27.6665	1.7782
Nth Random Chunk	11.1678	34.3914	0.9620	127.6347	11.7886
Video of cheerleaders moving in a gym (static)					
CARV	22.8551	17.8534	0.8480	146.9948	8.0939
Every Nth Frame	23.0226	17.8522	0.8478	154.5442	10.0306
Nth Random Chunk	22.9208	17.8534	0.8479	135.8052	7.5192
Video of a man squeegeeing an window (semi-static)					
CARV	45.7002	12.7673	0.6982	672.5347	101.7814
Every Nth Frame	52.0060	12.6816	0.6500	675.3055	135.8617
Nth Random Chunk	46.3292	12.7686	0.6945	703.3399	106.1572
Video of a person rolling back and forth on a rolling chair (semi-dynamic)					
CARV	66.7845	17.0843	0.5761	1955.4026	72.3051
Every Nth Frame	66.6660	21.1114	0.5834	1273.5481	73.3168
Nth Random Chunk	66.2591	17.4845	0.5852	1671.6145	46.6337
Video of children riding toy bikes in a playground (dynamic)					
CARV	38.0973	10.9738	0.6270	3304.5288	670.4664
Every Nth Frame	33.4257	11.0673	0.6672	1752.6041	303.8985
Nth Random Chunk	37.7482	10.9840	0.6312	3359.6110	722.0799

The table above compares several compression techniques through several video metrics. The video metrics used are mean square error (MSE), structural similarity (SSIM), peak signal-to-noise ratio (PNSR), and spatiotemporal reduced reference entropic differencing index (ST-RRED, ST-RRED SSN). The reduced videos are compared with the original video as the reference. The compression techniques are CARV (this paper), every Nth frame (updating every Nth frame), and Nth Random Chunk (updating every Nth frame with a random offset for each chunk).

Observing general trends, we can see that the the chunk-based reduction methods – CARV and Nth Random Chunk — have sufficiently higher ST-RRED values than the Every N^{th} Frame. This indicates a much higher correlation between adjacent frames and the higher preservation of spatiotemporal properties.

In the video metrics MSE, mean PSNR, and mean SSIM, CARV sees slightly worse performance on dynamic videos with little movement, with slightly better performance on static videos. This shows that CARV is worse at replicating pixel-by-pixel differences than other naive methods, like compression by every Nth frame. However, for the spatial-temporal metrics ST-RRED and ST-RRED SSN the opposite trends appear. Compared to the other methods, CARV performs better on dynamic videos and worse on static videos. This could imply that CARV is better at capturing spatial-temporal data in dynamic videos. To summarize, CARV performs good on dynamic videos on ST metrics, but does poorly on preserving pixel by pixel values. In contrast, on dynamic videos CARV performs better at preserving pixel by pixel values but scores worse on spatial-temporal metrics.

In total considering the video metrics, CARV sees some to no improvements to the other methods. CARV generally has similar video metrics to Nth random chunk, suggesting that the estimated importance has little to no effect choosing each chunk, but the ST property appears to hold well.

4 Limitations

Fixed size The current CARV model implementation uses an ST MAE, which is created by fine tuning an pretrained MAE model. The MAE model takes in images of a fixed size, so implementing CARV for high definition or long videos requires training an ST MAE from scratch.

Temporal restriction The base ST MAE model tunes a pretrained image MAE model. Thus, the ST MAE may not have as strong of a temporal persistence as a ST MAE model trained from scratch on videos instead of images.

Computational Cost CARV is a very computationally expensive method. As a result, it is not a very practical choice for video compression, especially considering the marginal benefits it provides.

5 Next Steps

Training custom MAE models Moving forward, it would be better to train a custom ST MAE from scratch instead of fine-tuning from a pre-trained checkpoint. This would allow us to have more control over the model architecture and training process, leading to potentially better performance and customization for this specific use case. However, this option would require a large investment in resources out of the scope of this project.

Modifying video chunk structure In order to optimize the video compression approach, further inquiry should explore modifying the structure of video chunks. This could involve experimenting with different chunk sizes or incorporating adaptive chunking techniques. Fine-tuning the chunk structure could potentially lead to improved compression efficiency and video quality.

Shrinkability maps for context-aware video scaling The videos could be compressed using another approach: context-aware video scaling [4]. A base assumption is that chunk importance relates to information of each chunk, meaning they can be compressed more aggressively without significant loss of visual quality. Context-aware video scaling provides just this: chunk importance could be used to define shrinkability maps in order to create scaling smooth in both an image and a spatial-temporal context.

Training data performance evaluation An avenue of exploration is to use videos compressed by the method as training data for other machine learning models in order to assess the transferability of information in the videos. This would theoretically explore the generalization capability of models and the potential of having this ST property maintained.

References

[1] Christoph Feichtenhofer & Haoqi Fan et al. (2021) Masked Autoencoders As Spatiotemporal Learners. arXiv:2205.09113v2

[2] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv:2203.12602, 2022.

[3] Fabian Mentzer, George Toderici, David Minnen, Sung-Jin Hwang, Sergi Caelles, Mario Lucic, Eirikur Agustsson. (2022) VCT: A Video Compression Transformer. arXiv:2206.07307

[4] Yi-Fei Zhang, Shi-Min Hu, Ralph R. Martin. (2008) Shrinkability Maps for Content-Aware Video Resizing. Computer Graphics Forum, 27: 1797-1804. https://doi.org/10.1111/j.1467-8659.2008.01325.x

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. (2022) Masked autoencoders are scalable vision learners. arXiv:2111.06377

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, Andrew Zisserman. (2017) The Kinetics Human Action Video Dataset. arXiv:1705.06950

[7] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. (2008) Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, pages 1096–1103.

[8] Shuhao Cao, Peng Xu, David A. Clifton. How to Understand Masked Autoencoders. arXiv:2202.03670

[9] Wang, Zhou, et al. "Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13.4 (2004): 600-612.

[10] Soundararajan, Rajiv, and Alan C. Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Transactions on Circuits and Systems for Video Technology 23.4 (2012): 684-694.