Algorithmic Video Reduction by Estimating Chunk Significance Using Spatial-Temporal Masked Auto Encoders (Milestone Meeting)

***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 — 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 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 effectiveness of the proposed approach is demonstrated through experiments and comparisons among existing techniques. The preliminary results demonstrate that the proposed chunk selection method reveals some patterns among different masking thresholds.

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 potentially facilitate the creation of reduced videos that retain sufficient informational content and could be used to train video-oriented models more quickly without compromising their performance.

Background and Related Work

Masked Autoencoders Over the years, MAE models have proven to be highly effective across various tasks. For instance, Kaiming He, Xinlei Che, et al. [5] demonstrated their remarkable performance as visual learners, which has contributed significantly to their widespread adoption. These models have their roots in denoising autoencoders [7], which are neural networks trained to reconstruct their input. 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, such as image or speech denoising. The network learns to capture the most salient features of the input data and can reconstruct the missing parts of the input based on this information. Several studies have explored the use of MAEs in video processing. For instance, Zhan Tong, Yibing Song, et al. developed VideoMAE, a video-specific variant of MAE. Similarly, Christopher Feichtenhofer, Haoqi Fan, et al. introduced a ST MAE – a variation of an autoencoder that processes spatial temporal data including video.

Chunk The term "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 selecting chunks of video frames, we aim to improve the spatio-temporal representation of the input data, which can enhance the network's ability to learn relevant features and improve its reconstruction accuracy.

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. Thirdly, we 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

The study employs a pre-trained video MAE model to process videos, with the possibility of custom-trained MAE models in the future. The videos are preprocessed into normalized arrays of 16 frames, 224 pixels by 224 pixels by 3 channels, and split into 2-frame chunks of 16 pixels by 16 pixels to maintain regularity among the videos observed.

The MAE model is utilized to create a reconstruction of all masked pixels by passing the mask and video (figure 1). This is sampled many times with different randomized masks (sampled spatially agnostic) in order to obtain an averaged significance for each chunk involved. Each chunk would then be evaluated by the Mean Squared Error (MSE) metric, which we would label chunks' importance directly related to its MSE. Framing it another way, chunk significance is equivalent to how well the average image fragment containing that chunk performs. Figures 3 and 4 show the results of this process, with the most important chunks highlighted in green and the least important or harmful chunks highlighted in red.

For further steps, the chunk importance can be used to create shrink-ability maps used in context-aware scaling. This will hopefully provide a shrunken video whilst preserving spatiotemporal video properties.

3 Preliminary Results

Chunk MSE 1-Variate Statistics, Characterized by Various Masking Ratios and Sample Size

Masking Ratio (%)	# Sampled	MSE Mean	MSE Std.	MSE min	MSE max
85	300	0.0940	0.0011	0.0907	0.0974
90	300	0.1457	0.0026	0.1368	0.1550
95	300	0.2074	0.0052	0.1865	0.2273
98	500	0.3414	0.0216	0.2783	0.4345
99	500	0.6293	0.1312	0.3504	1.5722
99.3	500	0.8885	0.2372	0.3685	2.0584

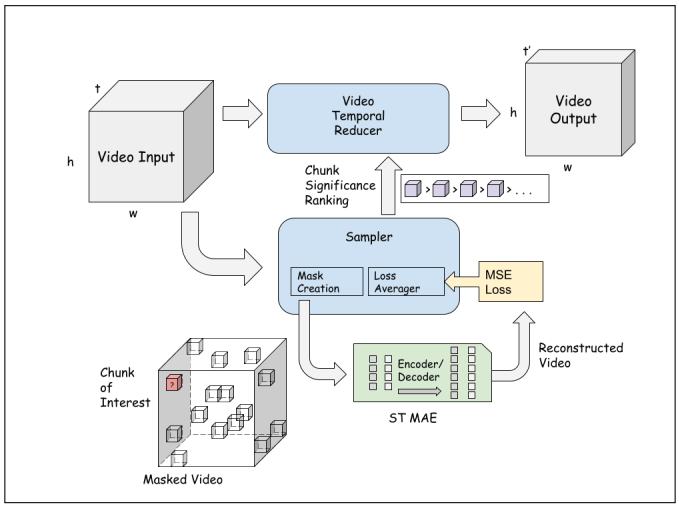

Figure 1: Proposed Model Layout. The project has currently reached the chunk significance rating stage of the model.

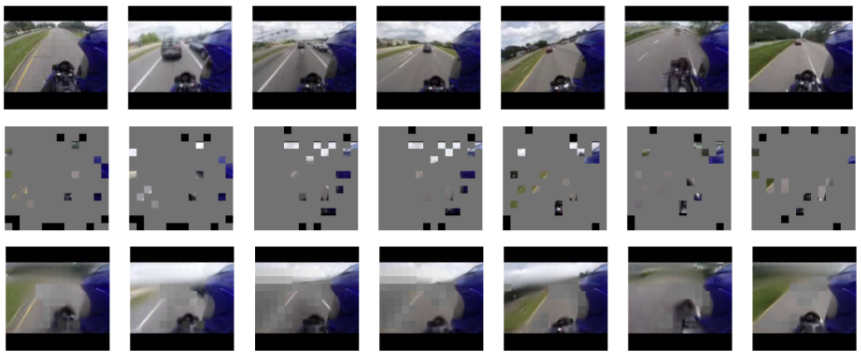

Figure 2: Top Row: Original Video, Middle Row: Masked Video, Bottom Row: Reconstructed Video. This highlights the capabilities of spatiotemporal MAEs.

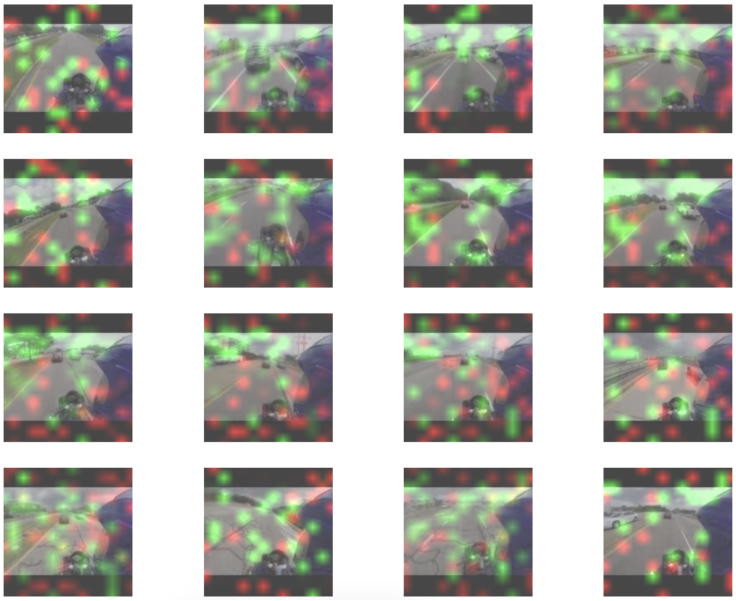

Figure 3: Video of a motorcyclist driving down a highway using a 0.90 Masking Ratio and 300 samples, colored with greens and reds: most beneficial and most detrimental respectively.

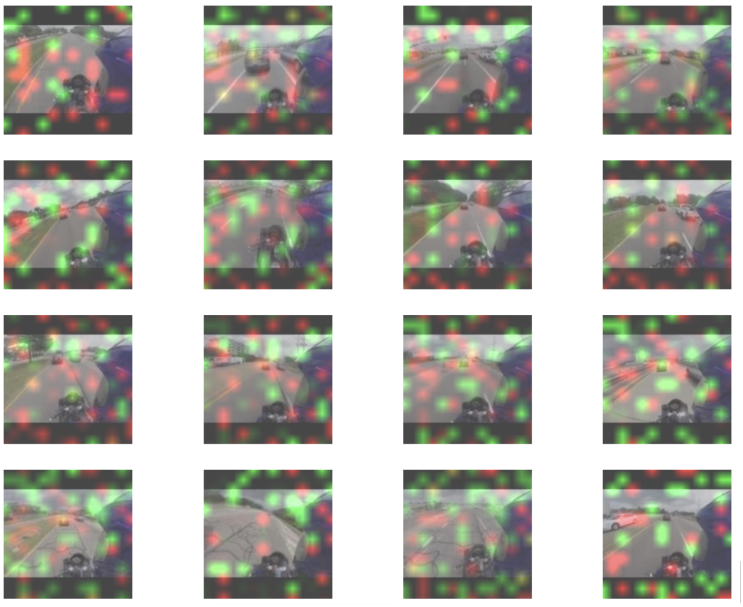

Figure 4: Video of a motorcyclist driving down a highway. Notice how the highlighted chunks are more randomly distributed. Used a 0.98 Masking Ratio and 500 samples, colored with greens and reds: most beneficial and most detrimental respectively.

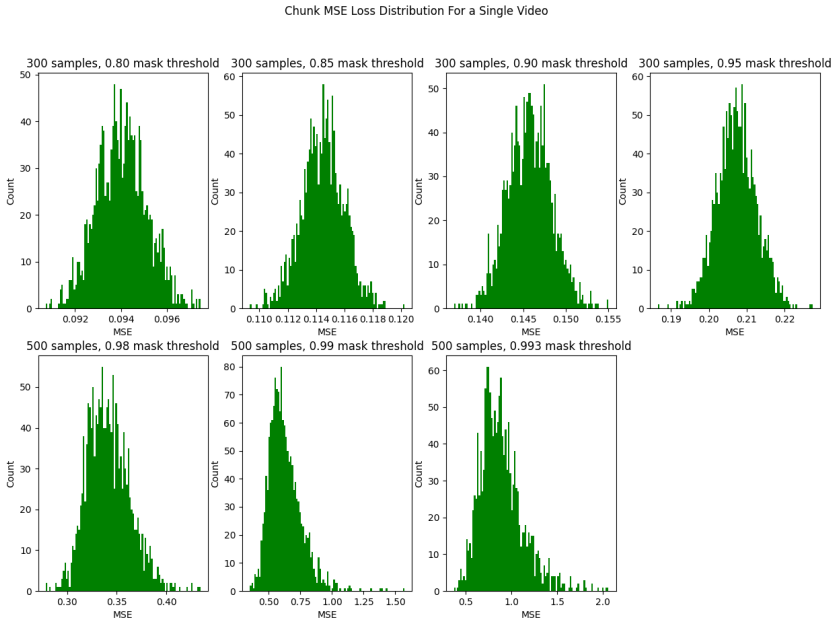

Figure 5: Chunk Reconstruction MSE per Masking Ratio. Notice how both the shape and values change as masking increases.

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