Are Masked Autoencoders Actually Scalable Spatiotemporal Learners?

Eric Cai
eycai@andrew.cmu.edu

Roshan Roy
roshanr@andrew.cmu.edu

Kyutae Sim
ktsim@andrew.cmu.edu

Michaela Tecson
mitecson@andrew.cmu.edu

1. Problem Setup & Motivation

Modern NLP models benefit from a unified self-supervised learning (SSL) based pre-training paradigm, via masked token prediction. In Computer Vision, DINO-like self-distillation methods have recently shown great promise for images, but there is no clear resolution for video sequences. The community remains fragmented for self-supervised video-based learning in contrastive, masked prediction, and self-distillation approaches. Masked Video Encoders (MVEs) have shown promise, but most papers demonstrate results on activity recognition benchmarks such as Kinetics-400 [11] and SomethingSomething [8]: their effectiveness beyond activity recognition remains unclear. In this project, we aim to investigate the hypothesis that masked video pre-training alone can serve as a universal backbone for diverse downstream video tasks such as tracking, optical flow estimation, etc. Additionally, we aim to investigate whether masking across videos can outperform traditional image-based SSL techniques, i.e. whether spatio-temporal masking of video patches capture spatial semantics as effectively as spatial masking of images. Since humans natively process all data as sequential frames rather than static images, we believe in the need for a unified pre-training strategy. Our findings will be experimentally supported by designing benchmarks, systematically comparing methods, and evaluating across diverse vision tasks.

2. Related Work

2.1. Self-supervised Learning for Images

Following the success of self-supervised learning (SSL) objectives (e.g. BERT [6], GPT [1]) in NLP, extensions of these methodologies to the image domain have been shown to be effective general-purpose feature-learning paradigms. These methods primarily fall under three categories - 1) masked autoencoding (e.g. MAE [7]), 2) contrastive learning (e.g. MoCO [4], CLIP [18]), and 3) teacher-student distillation (e.g. DINO [2], DoRA [16]) - and have collectively closed the gap between supervised pre-training and unsupervised pre-training for image-based vision foundation models, with broad applications for downstream vision tasks.

2.2. Self-supervised Learning for Videos

Inspired by the promising results in image domain, there has been increasing interest in adapting techniques mentioned previously to the video domain, motivated in particular by the infeasibility and inaccessibility of large-scale annotated video datasets.

Masking-based methods, originally proposed for image domains, have been extended to videos by leveraging temporal dependencies. Feichtenhofer et Al. [7] extends MAE to spatio-temporal learning by randomly masking out space-time patches in videos and learns an autoencoder to reconstruct videos. Contrastive learning has proven effective in capturing motion patterns and scene dynamics without explicit supervision. SimCLR [3], MoCo [4], and BYOL [9] have been extended to video data, leveraging temporal augmentation and spatiotemporal constraints to improve representation learning on methods such as CVRL [14], VideoMoco [13], and more. TCRL [5] in particular perform temporal contrastive learning on varying lengths of clips from the video. In self-supervised settings, teacher-student distillation is used to refine video representations by progressively enhancing the student's feature learning without requiring labeled data. Methods like DINO (Self-Distillation with No Labels) [2] employ this approach, where the student network learns by aligning its representations with those of the teacher, often using momentum-based updating or feature similarity objectives. This method has been particularly useful in learning hierarchical spatiotemporal features for videos.

3. Proposed Methodology

3.1. Self-Supervised Video Learning Backbone

While many SSL architectures have been investigated in the video learning literature, masked autoencoding remains the most popular approach. Moreover, given the high levels of

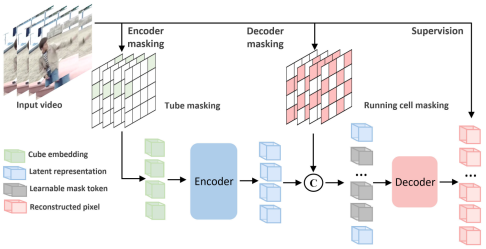

Figure 1. The VideoMAE V2 architecture, graciously borrowed from [17]. For our initial experiments, our plan is to turn the video encoder into an image encoder by repeating the image across the time dimension to obtain an image-based feature volume.

compute required to train video-based models from scratch, our intention is to begin with a pre-trained MAE-based model with open-source weights: namely, VideoMAE v2.

As a general outline for our approach, our aim is to repurpose the learned video encoder (Figure 1) for downstream vision tasks through supervised fine-tuning (and a minor additional exception for optical flow). Most of the task-specific architectures that we will use for downstream tasks utilize image encoders - since VideoMAE v2 (and most video SSL methods in general) relies on spatiotemporal embeddings, we plan to use an approach similar to [12], where we repeat image across the time dimension of the embedding cube size.

3.2. Video Tasks: Optical Flow and Tracking

The vast majority of investigations into downstream task performance from the video-based SSL literature only investigate tasks such as action classification and recognition. For our project, we would like to see if the features learned from the MAE paradigm transfer well to a broader variety of domain-specific tasks, as has been shown in the image domain.

For optical flow and tracking, our initial goal is to perform supervised fine-tuning respectively on the RAFT [15] and CoTracker [10] architectures, with their corresponding image encoders replaced by our pre-trained video encoder using the image-repeating method described above. To the best of our knowledge, there are no video-based SSL papers that evaluate performance on these downstream tasks, so our goal is to benchmark performance against the reported results from the original RAFT [15] and CoTracker [10] architectures that are trained from scratch.

3.2.1. Analytical Gradients for Optical Flow

Similar to the methodology described in GradCAM, we would also like to explore analytical gradients as a fully unsupervised mechanism for flow estimation. As roughly visualized in Figure 2, if Frame 1 (unmasked) and Frame 2 (masked) are well-reconstructed with a video-based MAE, then we would expect the image gradient of Frame 2 at point B to be quite high with respect to the pixel values of Frame 1 at point A. Through a some thresholding metric, we would then be able to obtain an approximate, dense flow estimate. While this almost certainly will not obtain state-of-the-art performance, we are interested in using it as a simple baseline that does not require any supervised training.

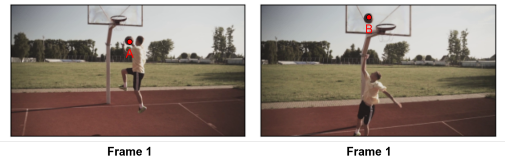

Figure 2. A very rough visualization of the proposed analytical reconstruction gradient approach for flow estimation.

3.3. Image Tasks: Depth, Segmentation, Classification, etc.

While it naturally follows that well-learned video features can aid in video-based tasks, we also argue that such features should facilitate object-centric reasoning and 3D understanding towards many image-based tasks (e.g. depth estimation, segmentation, and classification) as well. Some works in video-based SSL have shown minimal results along these lines for classification[7, 12], but this is a largely unexplored question. While we do not expect to set the state-of-the-art in these domains, our hope is that we can show spatiotemporal supervised learning to be a data-efficient feature learner for a broad variety of vision tasks beyond the video domain.

4. Goals

1. **Evaluating Masked Video Pretraining on Diverse Tasks:** To check if masked video encoders can serve as a universal backbone for downstream video tasks. Specifically, can they work on tasks that require fine-grained motion understanding, such as optical flow estimation and object tracking.
2. **Evaluating Video-Based SSL on Image-Centric Tasks:** To test the possibility of a unified image and video learning pipeline, we systematically compare video-based pretraining against traditional image-based SSL on standard image tasks.

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen,

Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 1

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1

[4] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[5] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, page 103406, 2022. 1

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv:2205.09113*, 2022. 1, 2

[8] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The ”something something” video database for learning and evaluating visual common sense, 2017. 1

[9] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 1

[10] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together, 2024. 2

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 1

[12] A. Emin Orhan, Wentao Wang, Alex N. Wang, Mengye Ren, and Brenden M. Lake. Self-supervised learning of video representations from a child’s perspective, 2024. 2

[13] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples, 2021. 1

[14] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning, 2021. 1

[15] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. 2

[16] Shashanka Venkataramanan, Mamshad Nayeem Rizve, Joao Carreira, Yuki M Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[17] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking, 2023. 2

[18] Floris Weers, Vaishaal Shankar, Angelos Katharopoulos, Yinfei Yang, and Tom Gunter. Masked autoencoding does not help natural language supervision at scale, 2023. 1