Siamese Masked Autoencoders

Rasmus Hammer, Sebastien Roig, Shiuan-Ting Lin

December 15, 2023

Abstract

This study aims to replicate the result from [1], focused on learning visual correspondence from videos. We employ UCF101 [2] as the training dataset, deviating from the original Kinetics-400 [3] due to resource limitations. The SiamMAE model, while conceptually straightforward, demonstrates superior performance on tasks such as video object segmentation, pose keypoint propagation, and semantic part propagation. Due to our constrained training resources, the learned features are specifically applied to the Video Object Segmentation task on DAVIS [4]. The results of the reproduction show worse performance of the model, but indicate similar behavior when comparing the ablation studies to the main method. Our code is available at `https://github.com/Jeremylin0904/DeepLearning_final`.

1 Introduction

The goal of [1] is to find a method that would learn visual correspondence from videos. To do that the authors try to extend Masked Autoencoders [5] to videos, with the hope that it could improve temporal correspondence understanding. The main idea is to train the model to reconstruct future frames given a past frame and a masked version of the concerned future frame. Thus, they combine the ideas of a siamese encoder that take the two frames as their respective input, and asymmetric masking.

In this work we tried to reproduce the experiments of [1]. The goal was to do a complete pretraining on UCF-101 [2] and then to evaluate the learned representation on three downstream tasks. However, due to limited amount of time, we focused on DAVIS [4] dataset object segmentation for our downstream task.

We also tried to reproduce some tables of the ablation study to see if we could get similar results and thus confirm the architectural choices in [1].

2 Related work

In this report we have attempted a reproduction of [1], which combines several previously known concepts to build the model, and achieve their results.

2.1 Self-supervised representation learning

Self-supervised learning, increasingly popular for leveraging vast amounts of unlabelled data, focuses on learning versatile representations for various tasks. In computer vision, imaginative pretext tasks, like contrastive learning exemplified by SimCLR [6], are essential for effective learning. A key challenge addressed by techniques like SiamMae is the reliance on data augmentation.

2.2 Masked Autoencoders

Masked Autoencoders (MAE) [5] are a scalable self-supervised learning method in computer vision, designed to prevent overfitting in modern models exposed to large image datasets. They operate by masking a substantial portion of an image, then training the encoder on the visible patches. The decoder processes these encoded patches with mask tokens to reconstruct the original image.

2.3 Vision Transformers

Transformers [7], introduced in 2017 by Vaswani et al. for machine translation, have since dominated natural language processing. Their adaptation to images was challenging due to the quadratic computational cost of applying self-attention to each pixel. Vision Transformers [8] (ViT), developed in 2020 by Dosovitskiy et al., addressed this by dividing images into patches and using a learnable embedding vector for the patch sequence. This vector serves as the image representation for tasks like classification. Position embeddings are also added to the patch embeddings in order to retain positional information.

2.4 Siamese Masked Autoencoders

In [1], an extension of Masked Autoencoders is introduced, forming a siamese network where inputs are processed independently by each encoder and then combined in the decoder using cross-self-attention. This network, based on a Vision Transformer, employs asymmetric masking: the first encoder receives unmasked images, while the second has a 95% mask ratio. This design aims to enhance the model's ability to learn object motion, improving performance in object segmentation, pose keypoint propagation, and semantic part propagation.

3 Methods

We use the same ViT base model as the one used in the original paper, but with a different size due to limitations in computational resources.

Encoder: While the encoder blocks used are similar in architecture to the original MAE design [5], the model was modified to use two different encoders which independently encode the initial frame and the target frame respectively. The outputs of these is then fed into the decoder. Aside from the siamese encoder, joint encoder was also used for the ablation study.

Decoder: The decoder blocks was modified to receive two inputs, one from each encoder. The output from the first encoder is used to get the key and value for the attention layer, while the second encoders output is used for the query. The attention function, the Scaled Dot-Product Attention, is given by the function:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Each decoder block for the main model consist of a $cross - attention$ layer, and a $self - attention$ layer. In the cross attention layer, the pixels from the first frame (f1) attend to the pixels in the second frame (f2). The output of this layer is then fed into the self-attention layer. Finally the output of the self-attention layer is fed into an MLP, and is then normalized and fed into a final linear layer used for prediction of the pixels in the reconstructed image.

Masking: For the main model 95% masking of the patches in the second frame, and 0% masking in the first frame, was used together with a siamese encoder and cross-self-decoder. Other configurations that was examined was siamese encoder with 95% masking and cross-decoder, as well as joint-decoder, and siamese encoder with cross-self-decoder and 50%, 75%, and 90% masking.

Label propagation: The label propagation algorithm consists of computing similarities between the patch embeddings and using those to propagate the labels. Given the parameters m (queue length), and k (top-k neighbors), the different steps of the algorithm are: 1. Resize the labels from first frame into the patchified image size and one-hot encode them. 2. Encode each frame of the video with the encoder. 3. Compute similarities of each patch of target frame with patches from the past m frames. 4. Select top-k similarities. 5. Propagate the labels by taking the weighted sum of the labels of the selected patches. 6. Apply softmax to these distributions and upsample the result to image size.

4 Data

We use the UCF101 dataset, which is similar to the kinetics dataset in that it consists of videos of human actions, but is much smaller in size. The dataset consists of 13,320 video clips, classified into 101 categories, covering Body motion, Human-human interactions, Human-object interactions, Playing musical instruments and Sports. All videos are collected from Youtube and is introduced in [2]. We did normalization of all images as well as minor data augmentation. During training two frames were randomly sampled at different intervals.

5 Experiments and results

Setup of the training process: The model was trained for 25 epochs on the main task, and 5 epochs for ablation studies, with Adam optimizer and a learning rate of 1e-4. The ViT-T/16 (Tiny ViT with patch-size 16) was used, and a batch size of 64, with minimal augmentations to the data consisting of random cropping and horizontal flipping. More precisely, we followed the setup of the paper using repeated sampling factor of 2 [9].

Evaluation methodology : We evaluate the quality of learned representation with label propagation on one downstream task: video object segmentation (Davis-2017 [4]). For the parameters, we used a queue length of 20 frames, 7 neighbors, and a radius for local attention of 7 patches. We use a smaller neighborhood size than the original paper because we used down-sampled image for inference.

5.1 Model performance

As can be observed in table 1 the model is able to learn a representation even with a small dataset and thus have better performance than more basic techniques. However, the visual results show sub-optimal performance. We can see in table 1 that the model performs best with fewer epochs. We theorize that this is due to over-fitting and perhaps the lack of inductive biases in vision transformers [8]. Indeed, even though the l_2 continues to decrease, the model performance on the downstream task get worse beyond 5 epochs.

5.2 Object segmentation

In figure 1 we observe that the network is able to model the object in the image, particularly for easier tasks. On harder tasks however the representation of the object is worse. Further the model has a harder time modeling motion of the object in the image, especially if the movement is large and relatively fast. Figure 2 does however show that the network is able to capture some amount of movement, although the representation becomes worse the farther into the future the frame is taken from.

Table 1: Results by number of epochs for pre-training measured on Davis

Number of epochs	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
3	44.4	40.7	48.0
5	44.1	40.4	47.8
10	42.8	39.2	46.3
25	39.5	36.1	42.9

The difference in the results between this report and the original one can be attributed to the different datasets used for the models. The dataset used in this

Figure 1: Video object segmentation on DAVIS for an "easy" and a "hard" task.

Figure 2: Visualization of the model trying to capture movement of the object.

report is significantly smaller than the original one, and consequently the model is much worse at generalizing, and is prone to overfit the training data. To avoid overfitting, a smaller version of the model used in the original paper was chosen (ViT-T/16), and it was also trained for fewer epochs. However the issue still remains of the model not being able to learn representations good enough to perform well on the downstream task, and the results indicate that the data is not large enough in quantity for the model to be able to do this consistently. In particular the model is able to quite well propagate the labels in simpler videos, with less movement, while being unable to accurately segment the main object in harder videos, as shown in figure 1. That being said, for objects which shape to not change much although the movement might be large, the model is able to somewhat accurately model the motion, as shown in 2.

5.3 Ablation studies

In our ablation studies, we endeavored to replicate various encoder and decoder designs mentioned in the paper under the default masking ratio of 95%. Additionally, we investigated the impact of varying masking ratios while maintaining the default encoder-decoder design as Siamese encoder and cross-self decoder.

5.3.1 Encoder-Decoder Designs

According to the results in table 2a, Siamese encoder with cross decoder performs best in regards to video object segmentation on the DAVIS dataset, however there is very little difference in performance between the cross-self-decoder, which performed the best in the reproduced paper, and the cross-decoder. The reason these are so similar might be because the model is only consistently able to learn object representations and not motion. Siamese encoder with joint decoder

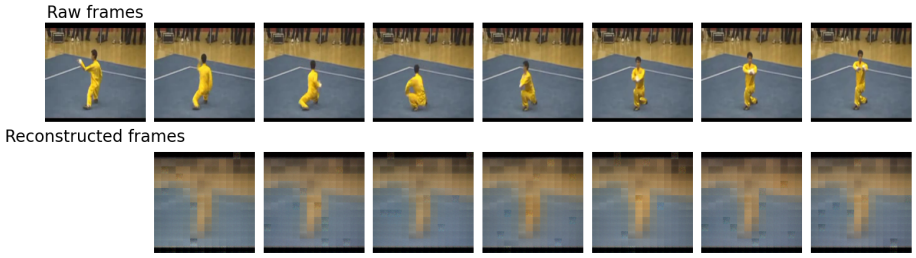

Figure 3: Visualization of 8 frames of videos, and reconstruction of the frames on UCF101 dataset with SiamMAE model trained with 95% masking ratio.

perform significantly worse than in the reproduced paper, which might be due to architectural differences or the network not being able to learn the representation with the limited amount of data. In regards to all versions using joint encoder the performance is very bad, and the results indicate representational collapse. This is consistent with the reproduced paper, in which the joint encoder also perform the worst for all configurations [1].

5.3.2 Masking ratio

We examined the impact of different masking ratios applied to the target frame when using Siamese encoder with cross-self decoder in video object segmentation on the DAVIS dataset. From table 2b, it is evident that a masking ratio of 95% yields the optimal performance. However the difference in performance for the different masking ratios is quite small, and so it is not obvious from these results if the masking has any tangible effect on the end performance. Most notably the model with a masking ratio of 50% performs on par with the one with 95%. This could be because they have different trade-offs, and as a consequence perform similar on the dataset for the downstream task as a whole. The model with 50% masking-ratio might be able to learn the features of the data better than the model with 95% masking, but is fast to overfit the training data. The model with 95% masking on the other hand might not be able to learn the feature representation as well as the one with 50%, but avoid overfitting.

6 Challenges

Computational resources : The paper utilizes the Kinetics dataset [3], around 450 GB of videos, for model pretraining. Our limited resources made using this dataset unfeasible, so we opted for the smaller UCF-101, 64 times smaller yet similar to Kinetics. Challenges included the video format, which prolonged training due to slow reading and storage constraints preventing us from saving videos as frames. Consequently, training was longer, and epoch count was limited. We used a better CPU to offset slow loading, increasing GCP VM costs, and with our budget constraints it limited our ability to explore the two other planned

Table 2: SiamMAE Ablation Experiments on DAVIS with different configurations

(a) Encoder-Decoder designs

Encoder	Decoder	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Joint	Joint	29.4	26.1	32.7
Joint	Cross	28.1	24.8	31.4
Joint	Cross-self	26.0	22.3	29.8
Siam	Joint	32.6	30.5	34.7
Siam	Cross	45.4	41.6	49.2
Siam	Cross-self	44.4	40.7	48.0

(b) Asymmetric masking

Mask ratio	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.50	44.2	40.7	47.7
0.75	42.4	39.2	45.6
0.90	41.3	37.8	44.8
0.95	44.4	40.7	48.0

downstream tasks. Additionally, the lack of higher memory GPUs restricted us from experimenting with smaller patch sizes, which we believe could have improved our reconstruction task results.

Lack of details : The paper lacked a lot of details notably concerning the architecture of the decoder. More precisely, the description of the cross-self attention part is ambiguous and could lead to different interpretations of the architecture. We also found that some parts of the experimentation was unclear, and particularly the label propagation algorithm that was not explained at all. Even in the other papers that are mentioned, the algorithm is not totally described and never adapted to Vision transformers. We were finally able to reproduce it thanks to another paper which did a review of this method [10] and thanks to previous works such as [11]. Finally, there is no Github repository associated with the article at the moment, so trying to reproduce it proved quite difficult.

7 Conclusion

The results in section 4 is an attempt at reproducing the paper Siamese Masked Autoencoders [1]. Although the results are worse than the results from the original paper, they still show a somewhat similar trend. The masking ratio of 95% seem to produce the best result, although the performance for the different ratios is very similar.

Generally the model is this report seems to be able to learn a representation of objects, but is unable to consistently develop an understanding of movements.

We attribute this to there not being enough data for the model to generalize, and learn object motion, well.

If we had more computational resources at our disposal, we would have liked to extend our experimentation to the other downstream tasks, namely human pose propagation (JHMDB [12]), and semantic part propagation (VIP[13]), and we would also have liked to use the original dataset used, Kinetics, to see if this enhances performance of our model. Also, we would have liked to try to use smaller patch size to see the difference, particularly in the reconstruction results. Further it would have been interesting to do a grid search on the parameters for the downstream task depending on which configuration of the network that was trained.

8 Ethical consideration, societal impact, alignment with UN SDG targets

In the process of reproduction, we have encountered significant disparities, particularly for researchers lacking computational resources or companies with insufficient capital. The computational demands for training these models far exceed our initial expectations, leading us to acknowledge this as an ethical consideration. The societal impact of our innovation lies in its ability to address, to some extent, the challenges associated with image reconstruction. While this may assist historical projects requiring image reconstruction, the substantial consumption of computational resources raises concerns about energy consumption and contributes to increased CO2 emissions.

In the context of UN Sustainable Development Goals (SDGs), we believe our work has implications for Goal 4 (Quality Education) and Goal 9 (Industry, Innovation, and Infrastructure). In terms of Goal 4, we posit that well-trained models can enhance educational settings by offering improved tools and resources. Regarding Goal 9, the application of our image reconstruction technology can contribute to advancements in industry development, offering valuable assistance in the broader landscape of innovation and infrastructure.

9 Self assessment

In this project we reproduced a good part of [1], given our limited computational resources. This paper was really complex and required a lot of auxiliary implementation such as label propagation algorithm and the Vision Transformer architecture [8]. We had to make several modifications to make the experiment fit with our resources, such as the choice of an adapted dataset and the reduction of the model's size. Finally, we succeeded in producing quantitative as well as qualitative results. Thus, we believe the grade deserved for this project is an A.

References

[1] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders, 2023.

[2] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

[3] 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.

[4] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations, June 2020. arXiv:2002.05709 [cs, stat].

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[9] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment Your Batch: Improving Generalization Through Instance Repetition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8126–8135, Seattle, WA, USA, June 2020. IEEE.

[10] Daniel McKee, Zitong Zhan, Bing Shuai, Davide Modolo, Joseph Tighe, and Svetlana Lazebnik. Transfer of Representations to Video Label Propagation: Implementation Factors Matter, March 2022. arXiv:2203.05553 [cs].

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. arXiv:2104.14294 [cs].

[12] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[13] Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive temporal encoding network for video instance-level human parsing, 2018.