Kungliga Tekniska högskolan

Deep Learning, Advanced Course
DD2412

# Project Proposal

*Authors:*
Friso de Kruiff
Magnus Tibbe
Casper Augustsson Savinov

October 31, 2023

# 1 Predicting the future with certainty: An extension of Siamese Masked Autoencoders

We decided to work on the paper Siamese Masked Autoencoder [1] because it combines the topics self-supervised learning and generative models and we see the opportunity to incorporate uncertainty estimation as well, which would cover all of our interests. The working title can be found above. The grade we are aiming for with our group is excellent (A). We would like to receive feedback on our proposal if our goals are in line with the aim of the grade.

First, we plan on replicating all the main results and figures in the paper. To that extend we want to replicate Figures 2, 3 and 5. We want to create our own version of Figure 1 to explain the method in one clear picture and we chose not to replicate Figure 4. Second, we aim to replicate the SiamMAE results of Table 1 for both ViT-S/8 and ViT-S/16, Table 2d and Table 3a. We chose these tables to make the project computationally feasible and because we believe these are the most interesting and important results. Thrid, to extend the research from the paper we have several directions we are considering.

1. We want to see if their method extends to more kinds of data augmentations, specifically rotation as this has been proven to work well in contrastive learning. The question would be how much rotation (in degrees) would be needed to get better performance than the baseline model from the paper.
2. Testing on more complex dataset (VSPW or UVO or KITTI) → more objects per frame.
3. Predicting multiple future frames.
4. Use uncertainty estimation to increase performance (maybe qualitative output could be an uncertainty heatmap of the predicted pixel values). We could use the uncertainty estimate to implement a "smart masking" method that masks the areas with high certainty first, in order to improve the prediction of the future frame or vise versa to make the task harder.

The deep learning packages that we will used to implement all the methods is JAX and PyTorch. In the original paper 4 datasets where used. One for training and testing and three for evaluation. The Kinetic-400 dataset [2] is

| Purpose | Name | Task | Statistics |
|---|---|---|---|
| Pre-training | Kinetics-400 [2] | Pre-Training the model | 400 human action classes, min 400 10s clips |
| Evaluation | DAVIS-2017 [3] | Evaluating Video Object Segmentation | 50 videos, 3455 frames, 480p resolution, 784 MB |
| Evaluation | JHMDB [4] | Evaluating For Pose Tracking | 31838 frames of 21 action classes. |
| Evaluation | VIP [5] | Evaluating Video Part Segmentation | 20 classes, 560x560 resolution, each video 120s |

Table 1: Scaled Datasets

between 400-500 GB in size, so we will require a significant amount of storage. In the original paper, 4 NVIDIA Titan RTX GPUs was used. We have access to 600 Google Cloud credits and one NVIDIA RTX 2060 Ti Super. We are unsure if we will have sufficient computational resources at this stage. Our plan is to work on free resources and deploy the code on Google Cloud once we confirm that it works. Below you can find for each of us a description of what we want to learn from this project:

- **Casper:** I want to gain a deeper understanding of auto encoders, the attention mechanism and self supervised learning. Additionally, I want to gain practical experience with popular python packages like PyTorch and JAX.
- **Magnus:** My main goal is to gain more experience in implementing deep learning pipelines using popular deep learning frameworks like JAX and PyTorch. Additionally, I want to learn more about how visual transformers work by implementing this paper.
- **Friso:** Theoretically, I want to understand representation learning and self-supervised learning through Auto-Encoders better. Second, I want to see how we can incorporate uncertainty estimation techniques and attention maps to enhance the performance of self-supervised models. Finally, I want to gain practical experience with JAX.

We will measure our success in two ways, learning and outcome. In learning we deem our project successfully if we were able to learn about the topics each of us described individually so that we can work with them, use them in code and explain them to others. In terms of outcome we deem our project successful if we are able to reproduce the results of the paper and we implement at least one major extension successfully. We are aware that the current project doesn't have it's own public GitHub repository, and therefore replicating the results will be a challenge. Additionally, the experiments done in the paper are computationally intensive and infeasible to do fully for our project. Therefore, we think that when we are able to replicate the most important results, and add on that through the described extensions, we meet the requirements for an excellent grade of the course. Finally, this paper combines multiple techniques from our course; generative modelling, undersupervised/self-supervised learning and we aim to incorporate uncertainty estimation. If we succeed in successfully combining multiple domains of the course we believe we have shown that we master the course material in an excellent way.

With this paper we have a few considerations we would like to get specific feedback on:

1. There is no public GitHub repository and therefore replication of the results itself is already a challenge. In terms of aims, grade and project contents we would like to receive feedback on if we decided to replicate the right things from the original paper, do we need to do more, are we to ambitious?
2. We have not been able to get a feeling for the feasibility of the pretraining on the Kinetrics-400 dataset and that is why it is hard to judge now how computationally feasibile our goals are at this point. Do you have any suggestions on this and could you advise us on how feasible pre-training would be?
3. At this stage we are unsure, how complex the proposed extensions will be with respect to time, computational resources and effort so we are not sure if implementing just one of them will be enough for the grade that we are aiming for.

# References

[1] Agrim Gupta et al. "Siamese Masked Autoencoders". In: *arXiv preprint arXiv:2305.14344* (2023).

[2] Will Kay et al. *The Kinetics Human Action Video Dataset*. 2017. arXiv: 1705.06950 [cs.CV].

[3] F. Perazzi et al. "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation". In: *Computer Vision and Pattern Recognition*. 2016.

[4] H. Jhuang et al. "Towards understanding action recognition". In: *International Conf. on Computer Vision (ICCV)*. Dec. 2013, pp. 3192–3199.

[5] Qixian Zhou et al. "Adaptive temporal encoding network for video instance-level human parsing". In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 1527–1535.