Applied Deep Learning: Self-supervised pre-training for segmentation

23213994

UCL

1 Introduction

The challenge of learning effective visual representations without human supervision has been a persistent obstacle in the field of machine learning. Self-supervised learning offers a promising approach by utilizing pre-training models on vast quantities of unlabeled data, followed by finetuning in a supervised setting for a downstream task. However, the representations learned are generally useful for classification, but often exclude crucial information for object segmentation like transformations which don't affect classification (e.g. translations)[20].

This report delves into *contrastive methods* [4], which are pivotal in this domain. Notably, contrastive learning stands out as one of the most effective self-supervised techniques, achieving top-1 accuracy rates as high as 86.7% [12]. A remarkably simple self-supervised learning algorithm that has proven excellent results is SimCLR [4], which emphasize the similarities and differences between pairs of augmented views of the input data.

The aim of this report is to explore how well contrastive methods perform on segmentation challenges. In particular, for the MRP we investigate how the quality of segmentations produced are influenced by:

- the structure of the backbone architecture.
- the size of the datasets used for pre-training and finetuning.
- pre-training the backbone on different datasets and image resolutions.
- the content on the pre-training dataset.

Moreover (OEQ), vision data is an abundant and readily available resource. Our goal is to unlock the full potential of self-supervised learning by leveraging video data, treating the model as a visual learner akin to social learning theory [1]. We envision enabling observational learning through a simple low-resolution camera, serving as a source of knowledge acquisition. To this end, we will evaluate the performance of the Animalia Kingdom dataset [13], a collection of animal videos, as a pre-training dataset and compare it with other datasets studied in this report.

2 Methods

SimCLR Implementation. Our implementation of SimCLR comprises four principal components adapted to our compute constraints and dataset characteristics:

- Random Transformation Module: Each image example is transformed into two correlated views, $\tilde{x}_i$ and $\tilde{x}_j$, treated as positive pairs. For our experiments, we implemented transformations such as color distortions, cropping, and resizing. We selectively applied Gaussian blur based on the specific dataset; notably, for smaller images like those in CIFAR10, Gaussian blur was not applied, as recommended by the original SimCLR study [4].
- Backbone Encoder $f(\cdot)$: Due to our limited computational resources, we employed ResNet-34 [8] as the backbone encoder, in place of more computationally intensive transformers [19]. We adapted the ResNet architecture by replacing the initial 7x7 convolution with a stride of 2 with a 3x3 convolution with a stride of 1, and removed the max pooling layer. This modified architecture enhances the feature extraction capabilities essential for segmentation tasks [6]. For our purposes, we also excluded the final average pooling during finetuning and inference stages to optimize performance when integrating the encoder with a decoder structure.
- Projection Head $g(\cdot)$: We followed the simplicity suggested in the literature by implementing a projection head comprising a single hidden layer, ReLU activation, and batch normalization following the hidden layer. This setup effectively maps the representations to a feature space suitable for contrastive loss computation.
- Contrastive Loss Function: We implemented the NT-Xent loss (normalized temperature-scaled cross entropy loss) [17], defined as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(z_i, z_j)/\tau\right)} \tag{1}$$

  In our implementation, we use a temperature τ of 0.5, and sim represents the cosine similarity between the encoded vectors.

Datasets. Given our computing resource constraints, we were unable to utilize ImageNet [15] at its full resolution of 224x224, as recommended by previous research. Therefore, considering the importance of using larger batch sizes[4], we have decided to use CIFAR-10 [10] and iNaturalist [7] as datasets to address different aspects of our research objectives.

CIFAR-10, with its smaller image size (32 x 32), allows for an increased batch size, facilitating the computation of contrastive loss across a greater number of images simultaneously. This setup is crucial for efficient learning with limited computational power. Conversely, iNaturalist, which boasts approximately 2.7 million images across 10,000 species, offers a rich, diverse set of natural images. In particular, we only kept the subset of animal images since we considered it being advantageous for our subsequent finetuning on the Oxford-IIIT Pet Dataset [16], which includes 14,698 images across 37 pet categories. We hypothesize that pre-training on a broad dataset of animal images will enhance performance on pet-specific segmentation tasks. To optimize processing time and efficiency, we resized iNaturalist images to 64 x 64 pixels, reducing dataset preprocessing time by up to 30% by eliminating the need for real-time resizing during batch loading.

Our decision to employ datasets with different lengths and image resolution serves two purposes. Firstly, it accommodates our computational constraints. Secondly, it enables us to investigate the influence of batch size on model performance and the effects of image resolution on classification accuracy. Smaller image sizes frequently present significant challenges for class discrimination, offering valuable insights into the limitations and capabilities of contrastive learning frameworks.

Furthermore, we are exploring the potential of self-pre-training on video through the Animal Kingdom dataset, also known as *Animalia* [13], which features videos depicting animal behavior. This inclusion aims to test the hypothesis that datasets focused on animal imagery yield better segmentation results due to the thematic alignment with our target application in pet image segmentation.

Decoder. Our decoder architecture employs a two-block upsampling method designed to minimize artifacts, such as checkerboard patterns commonly associated with deconvolution layers [14]. Each block in our decoder consists of the following sequence: a convolution layer, batch normalization, ReLU activation, a dropout layer, and a bilinear upsampling. The first block prepares the features for upsampling, while the second block includes another convolution followed by a bilinear upsampling step that adjusts the output to the desired resolution.

This architecture was specifically chosen to mitigate issues arising from upsampling. We discovered that traditional decoder configurations often struggled with accurate reconstruction of smaller images (e.g., 32 x 32 pixels). To address this, we experimented with increasing the input image size to both 128x128 and 256x256 pixels, observing immediate improvements in accuracy with no discernible difference in performance between these two resolutions.

Optimiser and Learning rate schedulers. The original paper uses LARS optimiser[21] during the optimization procedure. LARS' main strength resides in computing data-parallel synchronous SGD, since they are training their model on several TPUs. In our case, we opted to use ADAM [9], a renowned stochastic optimization algorithm popular in the deep learning literature. In addition, we used two learning rate schedulers, a linear warm up for the first 10 epochs, similar to the paper, and a cosine annealing schedule with warm restarts every 50 epochs [11].

Finetuning. After doing pre-training we finetune both pre-trained and non-pre-trained models on the pet dataset to observe how they respond to downstream task adaptation. For this, we simply load the pre-trained weights on the encoder part of the network and train to minimise the cross entropy loss. We decided to split the given segmentation in a One Hot Encoded tensor with three layers each corresponding to one of the possible values given by the masks of the PET dataset. Thus, we transformed the given masks from a (batch_size, 1, 128, 128) to (batch_size, 3, 128, 128).

3 Experiment study.

In this section we cover the details of the experiments that were carried out. All our experiments are performed with a single Nvidia 3090 Ti GPU.

1. Pixel-wise accuracy comparison. We study pixel-wise accuracy with respect to ground truth images of segmentations produced by our baseline model (100 epochs) trained in a supervised manner on PET-III dataset; and SimCLR pre-trained on Cifar10 (300 epochs), iNat (300 epochs) and Animal Kingdom (300 epochs) and later finetuned on PET-III dataset (100 epochs).

2. Segmentation plots comparison. Pixel-wise accuracy alone may not capture all the important information about segmentation quality. While it helps quantify the overall accuracy of segmentation, it may miss crucial details such as precision and finer segmentation boundaries. For that reason, we also also analyze quality of segmentations generated by visually plotting and inspecting the segmentation outputs.

3. Pixel-wise accuracy for different PET dataset sizes. The first experiment aims to investigate the impact of fine-tuning data quantity on segmentation quality. We conduct this experiment using varying proportions of the PET-III dataset, specifically 1%, 10%, and 100% of the dataset's total size. By evaluating the segmentation performance across these different data quantities, we can study how the amount of fine-tuning data influences the quality of segmentation results.

4. Open Ended Question. For the OEQ two approaches where considered. Initially, we simply extracted images sampled every 100 frames from the video and resized them to 64x64 resolution. However, aiming to improve the results, we propose modifying the loss function by selecting an anchor image x_0 with its embedded representation denoted by z_0, and a successive set of images $x_1, \ldots, x_p$ with corresponding embedded representations $z_1, \ldots, z_p$. Then, we compute a weighted loss:

$$\text{VideoSimCLRLoss} = \sum_{i=1}^{p} \lambda_i \text{NT-Xent}(z_0, z_i)$$

where $\lambda_i \in [0, 1]$ is the weight assigned to each loss term. The intuition behind this approach is to discount the contribution of successive frames while giving more importance to frames that are temporally further away from the anchor frame.

5. ResNet's first convolutional layer. The results of this experiment will not be included in the Results section due to brevity; however, all preceding experiments were conducted twice, using default ResNet-34 configuration (1st layer 7×7 kernel size and stride 2) and our modified version's 1st layer with kernel size 3×3 and stride 1. Since our image resolution is small, we noticed that having a bigger kernel diminishes performance.

4 Results

Within a few training epochs, we observe the formation of distinct clusters in the latent representation space (Figure 1). Although we could not replicate the image for iNaturalist, due to the large number of classes, or Animal Kingdom dataset, we can observe that SimCLR can capture latent representations correctly regardless of the image resolution. This observation aligns with the findings reported in the literature [18].

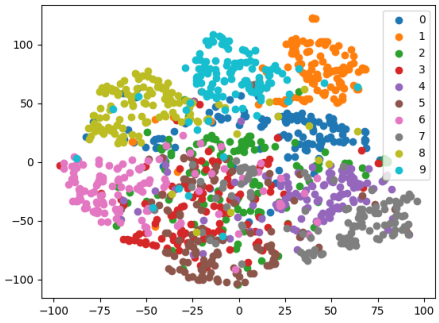

Fig. 1. t-SNE visualizations of representations from 10 classes in CIFAR-10 dataset.

Regarding image accuracy, Figure 2 shows that our models pre-trained using SimCLR outperform the baseline model. Notably, SimCLR models are capable of learning to segment the dataset with just a few-shots (1% of images). This pattern can be explained by the efficacy of the contrastive learning component during pre-training (Figure 1). However, as the finetuning dataset sizes for training the SimCLR and the baseline model increases, the performance gap between both decreases.

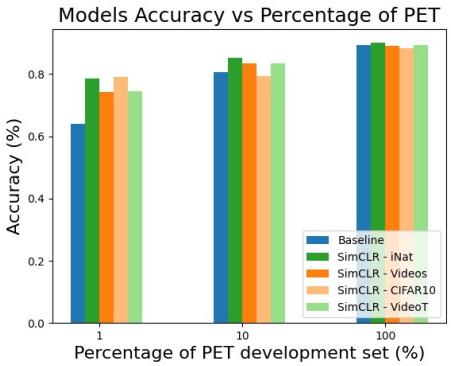

Fig. 2. Accuracy comparison between the baseline model and the SimCLR model pre-trained on *iNAT*, CIFAR-10 and the two video approaches (Videos is the first video approach whereas *VideoT* is the anchor approach)

Note that there is only a slight disparity between the performance of the SimCLR pre-trained methods. We observed that when training on an extremely small fraction of images, the video methods aren't able to adapt as well as others to the segmentation task. This might be explained by the lack of variety in the Animalia Dataset, as we extracted multiple frames from the same video. Conversely, as soon as we have more data, we can see the benefits of pretraining on an animal dataset (i.e., iNaturalist or Animal Kingdom) compared to a broader dataset (i.e., CIFAR-10). Additionally, we observe that the model is no longer biased towards the representations learned from CIFAR-10 since the baseline achieves the same accuracy. Finally, note that when training on the full dataset, the accuracies are almost the same.

Fig. 3. Segmentation comparison for 1% (left) and 100% (right) of the finetune data. Here, *Animalia* is the first video approach whereas *VideoT* is the anchor approach both pre-trained on the Animal Kingdom dataset.

In Figure 3, we observe a comparison of the segmentation performance among different models. Notice that the baseline results are the poorest. This aligns with the previous results (Figure 2).

5 Discussion

These results provide insights into our research questions. The consistent outperformance of the SimCLR models over the baseline across all fine-tuning dataset sizes supports our hypothesis that self-supervised contrastive learning methods help achieve a higher performance even for segmentation tasks. On the other hand, we see that the dataset focused on animal images yield better results suggesting that the pre-training dataset has to be selected according to the target application.

Moreover, the modification to ResNet34's first convolutional layer has a major impact on performance. When using the original approach we had to use a bigger decoder structure to account for all the modifications and yet didn't achieve the same segmentation quality.

Regarding the OEQ, we attribute the performance loss in the video approaches to the small variance across images. Although we induce variance in the data with the augmentations, we still require to clearly differentiate between frames. Hence, we hypothesise that one could expect an increase in performance just by changing the pose of the camera while recording video. Note that the approach used is only useful to image segmentation since we do not require any temporal understanding of the scene compared to other approaches [5,2].

One limitation of our study is the inability to pre-train on the ImageNet dataset, which prevented us from directly comparing our segmentation results with the classification accuracy reported in the original paper. For future work, exploring other self-supervised algorithms or investigating end-to-end segmentation techniques like DETR [3] could be promising routes.

6 Conclusions

Our experiments underscore the effectiveness of SimCLR for segmentation tasks, outperforming fully supervised baselines, especially in data-scarce scenarios. Pre-training on datasets thematically aligned with the target domain yields better segmentation performance. While promising, observational learning from videos faces some challenges. Overall, it is a auspicious way of training models without extensive labelled data.

References

1. Bandura, A. *Social Learning Theory*. Prentice-Hall, Oxford, England, 1977.
2. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. V-JEPA: Latent video prediction for visual representation learning, 2024.
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers, 2020.
4. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations, 2020.
5. Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, S., Stiefelhagen, R., and Van Gool, L. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 1502–1512.
6. Ghesu FC, Georgescu B, M. A. Y. Y. N. D. P. P. V. R. B. J. C. Y. G. S. C. D. Contrastive self-supervised learning from 100 million medical images with optional supervision. *J Med Imaging (Bellingham). 2022 Nov;9(6):064503. doi: 10.1117/1.JMI.9.6.064503. Epub 2022 Nov 30. PMID: 36466078; PMCID: PMC9710476.* (2022).
7. Grant Van Horn et al., Oisin Mac Aodha, Y. S. Y. C. C. S. A. S. H. A. P. P. S. B. The inaturalist species classification and detection dataset. *arXiv(Cornell University)* (2018).
8. Kaiming He et al., Xiangyu Zhang, S. R. J. S. Deep residual learning for image recognition. *arXiv(Cornell University)* (2015).
9. Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization, 2017.
10. Krizhevsky, A. Learning multiple layers of features from tiny images. *Toronto University)* (2009).
11. Loshchilov, I., and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts, 2017.
12. Maxime Oquab et al., Timothée Darcet, T. M. H. V. M. S. V. K. P. F. D. H. F. M. A. E.-N. M. A. N. B.-W. G. R. H. P.-Y. H. S.-W. L. I. M. M. R. V. S. G. S. H. X. H. J. J. M. P. L. A. J. P. B. Dinov2: Learning robust visual features without supervision. *arXiv(Cornell University)* (2023).
13. Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., and Liu, J. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 19023–19034.
14. Odena, A., Dumoulin, V., and Olah, C. Deconvolution and checkerboard artifacts. *Distill* (2016).
15. Olga Russakovsky et al., Jia Deng, H. S. J. K. S. S. S. M. Z. H. A. K. A. K. M. B.-A. C. B. L. F.-F. Imagenet large scale visual recognition challenge. *arXiv(Cornell University)* (2014).
16. Omkar M Parkhi et al, A. V., and Andrew Zisserman, C. V. J. Cats and dogs. *EEE Conference on Computer Vision and Pattern Recognition* (2012).
17. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems* (2016).
18. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., and Gool, L. V. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems 34* (2021), 16238–16250.

19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2023.
20. Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. Dense contrastive learning for self-supervised visual pre-training, 2021.
21. You, Y., Gitman, I., and Ginsburg, B. Large batch training of convolutional networks, 2017.