Semantic Segmentation of Images using Self-supervised Learning

Anonymous Author

University College London

Abstract. This project explores the effectiveness of self-supervised learning using contrastive learning algorithm for semantic segmentation tasks on the Oxford-IIIT Pet Dataset. It compares the performance of self-supervised models with fully supervised baseline models, and investigates the impacts of fine-tuning dataset size and pre-training dataset similarity on segmentation model performance respectively.

Keywords: Self-supervised learning · Semantic segmentation · SimCLR

1 Introduction

1.1 Background

Semantic segmentation is a classical computer vision task that involves assigning a class label to each pixel in an image [14]. The label could be the name of the corresponding object or the category it belongs to. However, this task requires a large amount of labelled data for training, which is labour-intensive and time-consuming [2]. Self-supervised learning (SSL), on the other hand, has shown to be an effective approach to solve this obstacle without the need for expensive human-annotated labels [7]. SSL is referred to as a subset of unsupervised learning with focuses on learning general feature representations from unlabelled data that can be applied for various downstream tasks [4]. SSL mainly operates by self-generating its own supervision, allowing for pre-training on cheaper and unlabelled datasets [16]. It contains two phases: (i) pre-training - to construct a general representation on unlabelled data in a task-agnostic way and (ii) fine-tuning - to fine-tune on labelled data in a task-specific way [15]. Self-supervised learning is particularly beneficial for the semantic segmentation task, since it can not only reduce reliance on large labelled datasets, but also improve the generalization of the model [11]. One can therefore adopt SSL methods to pre-train a model on unlabelled data and then fine-tune it on a small labelled dataset for semantic segmentation tasks.

1.2 Aim and Objectives

The primary aim of this project is to explore the effectiveness of self-supervised learning for semantic segmentation task of differentiating foreground and background on the Oxford-IIIT Pet Dataset [12] and then compare it with fully supervised baseline methods. In addition, two different pre-training datasets (i.e.,

CIFAR-10 and CAT&DOG from HuggingFace repository [6]), are experimented respectively to explore the "open-ended question" of "How similar the pretraining and fine-tuning/test data need to be, for a better segmentation model?"

2 Methodology

2.1 Pre-training using SimCLR Architecture

Contrastive learning is a popular self-supervised learning method that has shown to be effective for learning general-purpose representations from unlabelled data [9]. SimCLR (Simple Framework for Contrastive Learning of Visual Representations) [3] is a state-of-the-art contrastive learning framework that can pull similar image samples together and push dissimilar image samples apart to achieve high-quality image representations on multiple downstream tasks [1].

Hence, this project adopts the SimCLR framework, which consists of four major components: (1) **Data Augmentation**, (2) **Encoder Network**, (3) **Projection Head**, and (4) **Contrastive Loss**.

Specifically, the data augmentation module is used to generate two augmented views of the same image. This project uses Random Cropping and Resizing, Random Colour Distortions and Random Gaussian Deblur as the data augmentation techniques.

Then, these augmented images are passed through the encoder network with a pre-trained ResNet-50 model as the backbone network to extract image features.

Afterwards, the projection head uses a two-layer MLP (Multi-Layer Perceptron) with a ReLU activation function to map the image representations to a higher-dimensional space.

Finally, the contrastive loss function is utilised to help the model learn similar representations for augmented views of the same image and dissimilar representations for different images. Because the model is self-supervised, positive pairs and negative pairs are defined as the augmented views of the same and different images respectively. The project therefore uses the NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss function to maximise the similarity between positive pairs and minimise the similarity between negative pairs. Hence, the total batch loss L is calculated by averaging the loss over all positive pairs and negative pairs in the batch [5].

$$L = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \tag{2.1}$$

where l is the contrastive loss for an image pair in a batch.

2.2 Fine-tuning on the Oxford-IIIT Pet Dataset

The fine-tuning method involves training the model on a subset of labelled Pet dataset to adapt the learned representations to the task of animal images segmentation. It consists of five main steps: (1) **Load** the pre-trained SimCLR

model, (2) **Modify** the model architecture for the segmentation task, (3) **Train** the model on the labelled dataset, (4) **Evaluate** the model on the validation set, and (5) **Test** the model's performance on the unseen test set.

Firstly, this project starts with loading the pre-trained SimCLR model with the ResNet-50 backbone and the projection head. The pre-trained model is loaded with the weights learned during the self-supervised pre-training phase as mentioned above.

Then, it replaces the final layers of the pre-trained SimCLR model with a U-Net architecture for the segmentation task (i.e., this project uses the decoder of the U-Net architecture to generate the segmentation mask for each input image). The U-Net architecture is a popular architecture for image segmentation tasks, which contains an encoder-decoder structure with skip connections to capture both local and global features in the image [13].

Subsequently, this modified model is trained on the labelled dataset using the Adam optimizer with a learning rate of 10^{-4} and the binary cross-entropy loss function (BCE) (i.e., measure the difference between the predicted mask and the ground truth mask).

Next, the project trains the model for 60 epochs on the labelled dataset and evaluates the model on the validation set with the use of Intersection over Union (IoU) to prevent overfitting. Therefore, the model with the lowest validation loss is saved during all training epochs.

Finally, the project tests the model's performance on the unseen test set. Performance metrics, including IoU, F1-score, Dice loss, Focal loss, BCE loss, recall, and accuracy, are used to assess its generalisation ability.

3 Experiments

3.1 Experiment of Minimum Required Project

This project designs two comparison experiments for meeting the "Minimum Required Project": (1) Compare the SimCLR framework with a baseline model trained on the same fine-tuning data using fully supervised methods and (2) Compare the benefit of the pre-trained segmentation model using different fine-tuning dataset sizes.

Comparison (1): The project trains a baseline model through fully supervised learning methods to compare it with the proposed SimCLR framework. This baseline model utilises a U-Net architecture with ResNet-50 as the encoder. As for the SimCLR framework, it pre-trains two models with 20 epochs (limited by the usage of GPU) on CIFAR-10 and CAT&DOG datasets respectively using the same pre-training methodology as described above for this comparison. During the fine-tuning, both SimCLR and baseline models split the Oxford-IIIT Pet dataset into development (**80%**) and test (**20%**) sets for training and testing respectively. Within the development dataset, they further split the data into train (**90%**) and validation (**10%**) sets for training and validations respectively. Also, binary cross entropy (BCE) loss function are kept consistent for both models to

ensure a fair comparison. Afterwards, this project uses the same performance metrics, as mentioned above, to measure the performance of two models. Finally, based on these metrics, one can therefore determine the effectiveness of self-supervised learning for semantic segmentation tasks.

Comparison (2): This project proposes a method of gradually reducing the size of development set in the fine-tuning dataset (i.e., development set decreases from 80%, 50% to 20%)to explore how it affects the performance of the pre-trained model. Namely, three different dataset distributions: (1) **80% dev + 0% unused + 20% test**, (2) **50% dev + 30% unused + 20% test**, and (3) **20% dev + 60% unused + 20% test**, are configured for this experiment. It is noteworthy that the 20% holdout test set remains unchanged under all conditions (including those instances) for a fair comparison. Besides, the train/validation split ratio is still 90:10 within each development set. Therefore, if the pre-training model shows effectiveness, comparable results between models using different fine-tuning dataset sizes should be obtained.

3.2 Experiment of Open Ended Question

To investigate how similar the pretraining and fine-tuning/test data need to be, for a better segmentation model, two different datasets, CIFAR-10 and CAT&DOG, are utilised to generate two pre-trained models respectively.

CIFAR-10 [10] comprises 60,000 images divided into 10 classes, offering a diverse range of visual contexts, from animals to inanimate objects. Its diversity in images provides a broad understanding of visual features in the model.

CAT&DOG dataset [6] contains 25,000 images of cats and dogs, which is a domain-specific animal dataset. It is particularly beneficial for learning animal-specific features that are expected to be captured in the Oxford-IIIT Pet Dataset.

This dual-dataset for pretraining ensures that it can learn both general and domain-specific representations for the model, making it insightful in investigating the impact of dataset similarity between pre-training and fine-tuning on image segmentation performance.

Therefore, this experiment can be divided into two parts: (1) Pre-training on CIFAR-10 and fine-tuning on Oxford-IIIT Pet Dataset and (2) Pre-training on CAT&DOG and fine-tuning on Oxford-IIIT Pet Dataset. Besides, this experiment uses the same fine-tuning dataset distribution and same performance metrics as described in the "Minimum Required Project" to ensure a fair comparison. Given that the experiment in section 3.1 integrates the implementation of this experiment, there is no necessity to repeat explaining the implementation details in this section. The report will directly present the results and analysis in the next section.

4 Results

After successfully implementing the above experiments using the corresponding methodologies, one can find a summary of all model performance metrics in the

table 4.1 below. This table demonstrates all the testing results on the Oxford-IIIT Pet dataset for three models: (1) baseline model, (2) CIFAR-10 pre-trained model and (3) CAT&DOG (abbreviated as Pet in the table 4.1) pre-trained model respectively. One should notice that development splits ranging from 20% to 80% are also included in the table, providing insights on the correlation between model performance and dataset size.

Table 4.1. Model Performance Metrics

ID	Model	Loss Func	Dev Split	IOU	F1	Dice	Focal	BCE	Recall	Accuracy
0	baseline	BCE	20	0.758	0.862	0.150	0.292	0.369	0.867	0.893
1	baseline	BCE	50	0.828	0.906	0.110	0.179	0.243	0.904	0.925
2	baseline	BCE	80	0.906	0.951	0.060	0.108	0.144	0.942	0.961
3	CIFAR	BCE	20	0.754	0.860	0.150	0.306	0.383	0.871	0.890
4	CIFAR	BCE	50	0.852	0.920	0.091	0.214	0.261	0.903	0.935
5	CIFAR	BCE	80	0.900	0.947	0.066	0.109	0.149	0.937	0.958
6	Pet	BCE	20	0.759	0.863	0.148	0.295	0.371	0.867	0.892
7	Pet	BCE	50	0.843	0.915	0.099	0.184	0.239	0.909	0.932
8	Pet	BCE	80	0.902	0.948	0.064	0.130	0.167	0.929	0.958

4.1 Results of Minimum Required Project

Comparison (1): As illustrated in the figure 4.1 (a) below, it shows SimCLR framework and baseline model performance under 80% development splits. It contains three sub-plots for baseline model, CIFAR-10 pre-trained model and CAT&DOG pre-trained model, all under 80% development split conditions. Also, original image, ground-truth mask and predicted mask are displayed from LHS to RHS in each sub-plot, to reveal its performance of differentiating foreground and background. The IoU score associated with each model is around 0.90, which indicates that there is no significant impact on the model performance when applying the proposed self-supervised pre-training methods. **Comparison (2)**: As shown in the figure 4.1 (b) below, it demonstrates CIFAR-10 pre-trained model results with different development set sizes from 80%, 50% to 20%. The model's IoU score has explicitly degraded from 0.9000, 0.8524 to 0.7543, which indicates that the pre-trained segmentation model can significantly benefit from larger fine-tuning dataset sizes.

4.2 Results of Open Ended Question

To answer how similar the pretraining and fine-tuning/test data need to be, for a better segmentation model, one can refer to table 4.1 above. It can be noticed that CIFAR-10 pre-trained model with 20% development split achieves 0.754 in IoU score, while CAT&DOG pre-trained model with the same development split obtains 0.759 in IoU. This indicates that the model pre-trained on

a domain-specific dataset will result in a slightly higher performance in image segmentation. However, this improvement is not significant, suggesting that the pretraining dataset similarity does not have a substantial impact on the segmentation model. Therefore, the model can still achieve a good performance on the segmentation task even if the pretraining and fine-tuning datasets are not similar. This result is consistent with the findings in the literature [8], which show that the model can learn general features from a diverse dataset and still effectively adapt to a specific task with a small amount of labelled data.

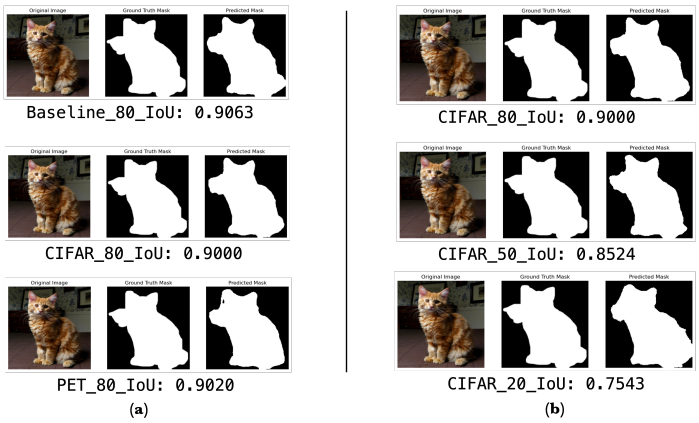

Fig. 4.1. (a) Comparison Result 1; (b) Comparison Result 2

5 Discussion and Conclusion

One of the limitations is that pre-training phase requires expensive and longtime computation (although A-100 GPU is used). Besides, the number of epochs used for pre-training is another limitation. This project uses only 20 epochs (to save money) for CIFAR-10 and CAT&DOG datasets pre-training, which may lead to a relatively low performance on the pre-trained models. Future works could involve the exploration of how different pre-training hyper-parameters (e.g., temperature parameters τ in NT-Xent loss function) can affect the model performance.

In conclusion, this project shows that the SimCLR framework can achieve a comparable performance of 0.9 in IoU score to the baseline model in the semantic segmentation task, which indicates that self-supervised learning is an effective approach for learning general-purpose representations from unlabelled data. Two comparison studies of using different fine-tuning dataset sizes and different pre-training dataset kinds further reveal that the model can benefit significantly from larger fine-tuning dataset size and that the pre-training dataset similarity does not have a substantial impact on the segmentation model's overall performance.

References

1. Albelwi, S.: Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. Entropy **24**(4) (2022). https://doi.org/10.3390/e24040551, https://www.mdpi.com/1099-4300/24/4/551
2. Aysel, H.I., Cai, X., Prügel-Bennett, A.: Semantic segmentation by semantic proportions (2023)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (13–18 Jul 2020), https://proceedings.mlr.press/v119/chen20j.html
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners (2020)
5. Chen, T., Luo, C., Li, L.: Intriguing properties of contrastive losses (2021)
6. Elson, J., Douceur, J.J., Howell, J., Saul, J.: Asirra: A captcha that exploits interest-aligned manual image categorization. In: Proceedings of 14th ACM Conference on Computer and Communications Security (CCS). Association for Computing Machinery, Inc. (October 2007), https://www.microsoft.com/en-us/research/publication/asirra-a-captcha-that-exploits-interest-aligned-manual-image-categorization/
7. Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., Tao, D.: A survey on self-supervised learning: Algorithms, applications, and future trends (2023)
8. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty (2019)
9. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1) (2021). https://doi.org/10.3390/technologies9010002, https://www.mdpi.com/2227-7080/9/1/2
10. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (05 2012)
11. Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., Tao, C.: Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022). https://doi.org/10.1109/tgrs.2022.3147513, http://dx.doi.org/10.1109/TGRS.2022.3147513
12. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: The oxford-iiit pet dataset, https://www.robots.ox.ac.uk/ vgg/data/pets/
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
14. Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., Herath, D.: Semantic segmentation using vision transformers: A survey (2023)
15. Xiong, Y., Ren, M., Zeng, W., Urtasun, R.: Self-supervised representation learning from flow equivariance (2021)
16. Zhang, C., Gu, Y.: Dive into self-supervised learning for medical image analysis: Data, models and tasks (2023)