Impact of the pre-training data distribution on the fine-tuned performance of MAEs

Student Number: 23077650 (Group 13)

1 Introduction

Self-supervised learning approaches have gained significant popularity in recent years, specifically when used for the pre-training of very large foundation models, both in NLP and Computer Vision [5, 7, 11, 15]. These methods allow for pre-training on very large amounts of data (for which collection of labels would be impractical), resulting in models which are capable of learning richer embedded representations. These models can then be fine-tuned to specific tasks using significantly less data than if trained from scratch.

In computer vision, these methods are mostly divided in invariance-based methods [1, 2, 6] and generative methods [4, 16]. The former are based on training an encoder-like network to produce similar embeddings for images of the same scene, but with different views. This is the idea behind contrastive learning [8]. The embeddings will thus provide a representation with high semantic value. Generative methods (in which masked autoencoders (MAES) [16] are included) are based on corrupting portions of the input images, and learning to predict these corrupted portions. In doing so, the model learns meaningful representations, albeit of a lower level than contrastive methods. Recent work has proposed a third approach, named IJEPA [3], which also does not rely on data augmentation, but differs from MAEs by computing the loss in embedding space, thus promoting meaningful embedded representations.

A natural question that arises from the pre-training of large models is how the distribution of the pre-training data affects the model's performance on downstream tasks [12]. In this paper, we hypothesise that pre-training on data from a similar distribution to the fine-tuning data should result in better model performance. We implement a Masked Autoencoder [16], and pre-train it on the MS COCO [18] dataset. This pre-trained model is then fine-tuned to perform image segmentation on the Oxford Pet dataset [19]. Different splits of the MS COCO pre-training dataset are attempted so as to study the impact of the pre-training data distribution on the down-stream segmentation results.

2 Methods

2.1 Masked Autoencoder pre-training and fine-tuning

As described in the MAE paper [16], we implement a Masked Autoencoder model with Vision Transformers (ViTs) to pre-train on a large dataset. The ViT parameters are shown in Table 1, and correspond to a base ViT. This choice of

ViT size allows for computation on our hardware, while being suitable for the relatively low data setting (larger ViTs demand much larger amounts of data).

Table 1. Settings of the ViT used in the MAE.

patch_size	embed_dim	depth	num_heads
16	768	12	12
decoder_embed_dim	decoder_depth	decoder_num_heads	mlp_ratio
512	8	16	4

We first pre-trained the model on 25739 images of 10 different animals (including cats and dogs) from the MS COCO dataset [18]. It is possible for an image to have multiple labels in this dataset. Thus, we filtered it so that any data used corresponded to images with only a specific animal's label. This dataset was chosen due to its medium scale, allowing for a large enough amount of data to pre-train, while remaining suitable for our available technology. Furthermore, the dataset has a large variety of realistic scenes, thus providing a good choice to use in pre-training. In particular, MS COCO has fewer categories but more instances per category than ImageNet [10], which is hypothesised to allow for more detailed object recognition [18]. However, note that against my recommendation, this diversity of data was not actually used in the basic pre-training described in Section 3.1(we only consider animals here), and that the number of images used for pre-training was very small for a typical pre-training setting. We validate our pre-training results on 1090 animal images.

The same model is used for end-to-end fine-tuning. Importantly, the masking of the MAE is now set to 0%. This is because the role of masking (and in particular of a high masking percentage) during pre-training is to remove redundancy and thus require the model to do more than simple extrapolation, and instead learn representations at a semantic level. However, during fine-tuning, the model uses these representations to perform a simpler task. Thus, masking simply makes the segmentation task much harder (if not impossible at some patches), and does not improve the generalisation abilities of the model. However, this means our encoder is not sparse, and thus the model takes longer to train for a fixed dataset size.

The model is fine-tuned to perform trimap image segmentation on the Oxford Pet dataset [19], which has 7349 images of cats and dogs. This dataset was split into 70% training, 15% validation and 15% test data. We modify the labels so that each pixel can have a label of 0 (animal), 1 (background), or 2 (edges/boundaries/anything else that isn't directly the animal or the background). We do not aim to perform semantic segmentation, that is, distinguish specifically between cats and dogs.

2.2 Supervised Baselines: Non Pre-trained MAE and ResNet

To provide baselines for the method based on pre-training followed by fine-tuning, we implement two baselines based on typical supervised learning. Firstly,

we simply train the MAE to perform image segmentation on the Oxford Pet dataset, without any pre-training. Secondly, we implement ResNet50 [17]. The choice of ResNet avoids the issues of vanishing gradients typically present in CNNs [20]. We choose ResNet50, as opposed to a more complex model such as ResNet101, to allow for relatively fast training of our baseline. Furthermore, this is a relatively simple segmentation task (especially due to not performing semantic segmentation), and thus we deem ResNet50 to allow for enough complexity for such a task (note ResNet50 has roughly half the amount of parameters of the ResNet101 model). Note we removed its fully connected layer, and instead upsampled the features to match $(3, 224, 224)$, the output shape corresponding to the probability of each class for each pixel.

3 Experiments

3.1 Self-Supervised Pre-Training + Fine-tuning

As explained in Section 2.1, we pre-train the MAE on approximately 25 thousand images of animals of the MS COCO dataset for 150 epochs. Randomised cropping, randomised horizontal flipping and image normalisation are applied to the input, so as to prevent overfitting. The learning curves were monitored, and we did not observe overfitting, with the final validation loss being 0.547. Note these had not yet converged. An example output is shown in Figure 1. We did not perform an analysis on the test set since the MS COCO test set lacked the necessary labels to retrieve only animal images (as our model is only pre-trained on animal images, we would require a similar distribution on the test set).

Two fine-tuning processes are then experimented: in the first one, we fine-tune (end-to-end) the pre-trained MAE on the training set of the Oxford Pet dataset. In the second process, we halve the fine-tuning training set (random split), and repeat the procedure. The optimizer settings and parameters were set as in the original implementation of the MAE [14].

Finally, we compare the loss functions, IOU (binary) and accuracy (not plotted in this report) of the two self-supervised pre-training followed by fine-tuning settings, with the two baselines described in Section 2.2 at each epoch. Randomised cropping and randomised horizontal/vertical flipping are applied to the input images in all methods, so as to prevent overfitting. The results for loss function and IOU can be seen in Figure 2, and Table 2.

3.2 Pre-training data distribution study

This paper aims to study the impact of the pre-training data distribution on the downstream performance of MAEs. As seen in Figure 1, we study three different settings for the pre-training data, all from the MS COCO dataset: in the first one, the model is pre-trained only on animal data (25739 images) as in the previous part; in the second one, it is trained exclusively on data without animal labels (25739 images) in the MS COCO dataset; and in the third one it

selects 50% animals and 50% non-animals (25918 images). Each of these three pre-trained models was then fine-tuned on the full-sized Oxford Pet dataset for the same 100 epochs. The results are shown in Figure 3, and Table 2.

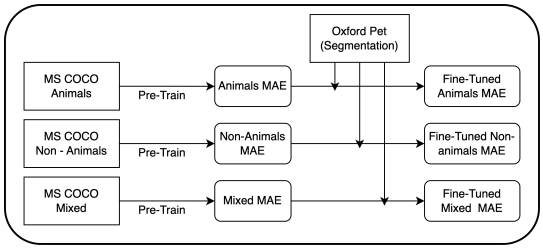

Fig. 1. Schematic of pre-training data distribution study.

4 Results

4.1 Self-Supervised Pre-Training + Fine-tuning

Figure 2 shows the pre-training on the MAE achieved a realistic reconstruction. However, note these results are worse than those presented in the original MAE paper [16], suggesting a larger pre-training dataset should have been used.

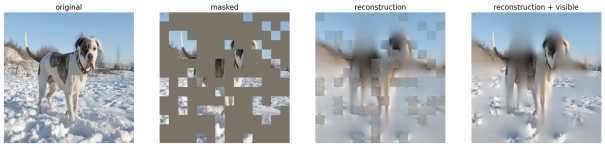

Fig. 2. Example MAE output after 150 pre-training epochs on 25739 animal images.

Figure 3 shows the training and validation losses, as well as the IOU for the two different fine-tuning procedures, as well as the two baselines. Figure 4 shows example segmentation masks on the test set. Test set quantitative results are shown in the first four rows of Table 2 ("fully supervised" stands for non pre-trained MAE). Note the non fine-tuned MAE performs the worse, as it is a data-hungry model and thus requires pre-training (or a larger dataset than OxfordPet) to perform well. The ResNet follows, and in particular we clearly see that overfitting happens from about epoch 40 onwards. This is because the ResNet is a relatively small model, and thus it overfits to our relatively small dataset if ran for too many epochs. However, note the ResNet achieves the best accuracy on the test set (89.3%). DEPENDS ON OUTPUT MASKS. Finally, we see that the MAE fine-tuned after pre-training achieves the best performance, as expected. Furthermore, halving the fine-tuning data worsens its performance significantly.

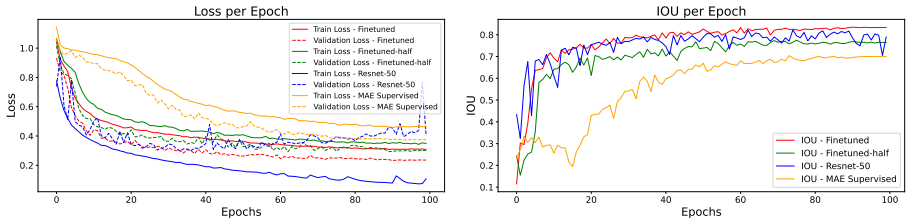

Fig. 3. Loss functions (on training and validation sets) and validation set IOU for the different fine-tuned MAE models and baselines ("MAE supervised" is the not pre-trained MAE).

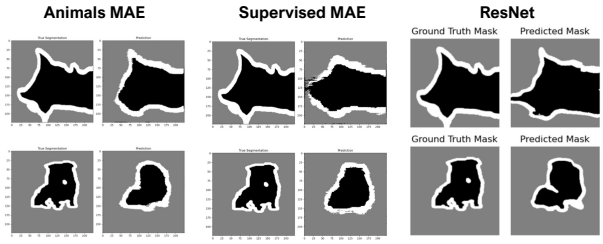

Fig. 4. Two example segmentation masks on test set for three of the models. Ground truth on the left, predicted segmentation on the right.

4.2 Pre-training data distribution study

Figure 5, as well as the third, fifth and sixth rows of Table 2, show the results. Firstly, we see that the pre-training on non-animal data performs significantly worse across all metrics. This supports our hypothesis that the data the model is fine-tuned on should at least be present in the pre-training set. However, the results for the mixed and full animal settings on the test set are very similar, and too close to conclude any advantage of either. This is a very interesting result, and one that supports a realistic pre-training followed by fine-tuning pipeline. In particular, it seems that the pre-training and fine-tuning data do not ought to have the same distribution, but that instead the fine-tuning data ought to be at least similar to a significant section of the pre-training data. This supports the typical use case of pre-training on very vast amounts of data, and then fine-tuning to specific tasks for which parts of the pre-training data are relevant. Any stronger conclusions would benefit from studying more percentages of animal vs non-animal data split, between 0% and 50%, to understand how the segmentation performance degrades as the pre-training data has less and less animal images.

5 Discussion

As discussed in Section 4.2, the results regarding the impact of pre-training data distribution provide a starting point. However, the similarities between the mixed

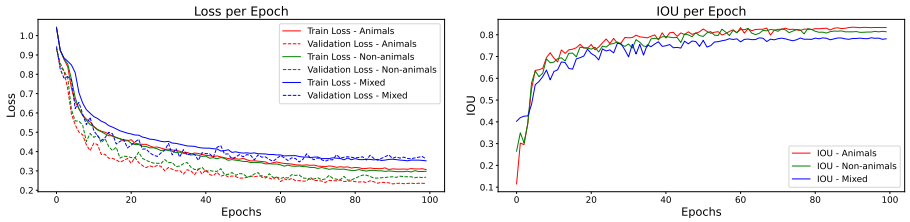

Fig. 5. Loss functions (on training and validation sets) and validation set IOU of the fine-tuned MAE for the different pre-training data distributions.

Table 2. Loss, IOU and accuracy on test set of all models described in this report.

Method	Test loss	Test IOU	Test Accuracy
ResNet-50	0.40	0.765	**89.3**
Supervised MAE	0.45	0.699	82.8
Finetuned-Animals	0.34	**0.785**	86.7
Finetuned-Half	0.35	0.757	86.5
Finetuned-Non-Animals	0.37	0.761	85.4
Finetuned-Mixed	**0.32**	0.781	86.8

and animals splits raise the question of exactly how similar the pre-training data has to be. A more thorough study, as performed in [13, 9], would allow us to better understand the benefits of self-supervised learning in pre-training large models [12], and how to select the pre-training data to reap these benefits.

Furthermore, the IOU metric used in this report is partly flawed, as it simply considers the "border" label as background. A better analysis would instead use mIOU, for multiple classes, to better understand model behaviour (this is relevant because the borders are exactly the hardest part for the model to learn). Furthermore, note once again the small pre-training dataset sizes, which should be much larger so as to mimic a realistic pre-training setting. As a matter of fact, the training curves of pre-training had not converged yet after 150 epochs.

Further work should investigate how these results extend to other methods, in particular IJEPA [3]. Note we initially attempted this method, but the loss on embedded space made it challenging to obtain interpretable results. Furthermore, here we focus on a very narrow task (segmentation of animals), and a more solid analysis would evaluate how the model performs across a variety of tasks.

6 Conclusion

In this paper, we see that the downstream performance of a pre-trained MAE model worsens if its fine-tuning images are from a completely different distribution from the pre-training images. However, we also observe the pre-training images do not need to come from the same distribution as the fine-tuning ones. Further work should study to what extent these distributions have to be similar before degradation of downstream MAE performance is observed.

References

1. Asano, Y., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning (2020)
2. Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., Ballas, N.: Masked siamese networks for label-efficient learning (04 2022)
3. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture (04 2023)
4. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: data2vec: A general framework for self-supervised learning in speech, vision and language (02 2022)
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. arxiv.org (05 2020)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments (01 2021)
7. Chen, R., Krishnan, R.: Self-supervised vision transformers learn visual concepts in histopathology (2022)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (06 2020)
9. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers (08 2021). https://doi.org/10.48550/arXiv.2104.02057
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition (06 2009)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (05 2019)
12. Emam, Z., Kondrich, A., Harrison, S., Lau, F., Wang, Y., Kim, A., Branson, E.: On the state of data in computer vision: Human annotations remain indispensable for developing deep learning models (07 2021). https://doi.org/10.48550/arXiv.2108.00114
13. Entezari, R., Wortsman, M., Saukh, O., Shariatnia, M.M., Sedghi, H., Schmidt, L.: The role of pre-training data in transfer learning (03 2023). https://doi.org/10.48550/arXiv.2302.13602
14. FAIR: Original mae implementation (04 2024), https://github.com/facebookresearch/mae
15. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations (2018)
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (12 2015)
18. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C., Dollár, P.: Microsoft coco: Common objects in context (2014)
19. Parkhi, O., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs (2012)

20. Shah, A., Shinde, S., Kadam, E., Shah, H., Shingade, S.: Deep residual networks with exponential linear unit (2016), https://arxiv.org/pdf/1604.04112.pdf