A Study of Transfer Learning Performance for Semantic Segmentation from Various Source Domains

Kefan Wang Choi Lam Wong Jingze Xu Lishen Chen Beiyu Xu
Jiajie Shi Siyuan Ren

University College London

Abstract. Multi-task learning (MTL) is a learning paradigm which can potentially improve generalisation. In this study, we investigated the effectiveness of transfer learning, an approach of MTL, on the semantic segmentation task on Oxford-IIIT pets dataset. By performing transfer learning with 5 datasets and comparing the resulting models, we explored the relationship between properties of source domain and performance of the model. An ablation study is also conducted to contrast effects of fixing and not fixing the backbone during transfer learning.

Keywords: Transfer Learning · Sequential Multi-Task Learning · Semantic Segmentation

1 Introduction

Image segmentation is the task of partitioning the images into different areas based on their belonged classes or instances, it could be categorised into semantic segmentation and instance segmentation. The former labels pixels with classes, e.g. person, car. The latter labels pixels with instances of objects, e.g. different individuals [1]. Some representative work in the field of image segmentation are ConvNets [2], SegNet [3], FCN [4] and DeepLab [5]. Multi-task learning (MTL), as a machine learning paradigm aiming to enhance generalisation [6], has been incorporated in image segmentation model in the literature and resulting in a better performance [7, 8, 9, 10, 11, 12, 13]. The essence of MTL is knowledge sharing between different tasks, which are typically categorised into target tasks and auxiliary tasks. Two different approaches of conducting knowledge sharing are simultaneous learning and sequential learning. With simultaneous learning, the models are trained to perform target tasks and auxiliary tasks simultaneously. Whereas, with sequential learning, or transfer learning, models are first trained on the auxiliary tasks in the source domain and subsequently trained on the target task in the target domain with the aids of learnt knowledge from the previous tasks. Both approaches of MTL have been utilised in image segmentation tasks in studies. [7, 8, 9, 10] utilized transfer learning in image segmentation tasks and reported varying amount of improvements on performance. Simultaneous learning has been leveraged in [11, 12, 13] and proven to be successful in

boosting performance.

A well-performing deep learning model usually contains a large number of parameters, and training requires a large amount of data. Due to overfitting, the network can't generalise well when the training set is small [14, 15]. In this report, we utilise a provided subset of the pre-processed Oxford-IIIT Pet Dataset [16], which contains 2210 training samples, 738 validation samples and 738 test samples. We aim to investigate the effectiveness of transfer learning on semantic segmentation from different source domain. Transfer Learning can be considered to be very adaptive if it can perform well when pretraining with a Less/Un-related dataset. Related-domain transfer learning for classifying The Oxford-IIIT Pet Dataset has been done previously. Kornblith et al. [17] investigated the transfer learning from ImageNet [18] to classify The Oxford-IIIT Pet Dataset. Ngiam et al. [19] investigated the transfer learning from entire JFT dataset [20]/JFT subsets/entire ImageNet/ImageNet subsets to classify The Oxford-IIIT Pet Dataset.

2 Methodology

A typical semantic segmentation network involves a feature extraction network and a classifier. The feature extraction network takes images as its input and outputs high level features. The classifier takes the features as its input and classifies each pixels. In our study, we perform transfer learning by training the feature extraction network on the source domain, afterwards, the classifier is trained on the target domain with the feature extraction network fixed. In cases where the tasks of source domain and target domain differ, the structure of the classifier used during pretraining and the subsequent step might be different as well, the feature extraction network is resued regardless.

3 Experiments

In order to investigate the effectiveness of transfer learning on different datasets, transfer learning with different source domains are performed and compared. We used ResNet50 [21] backbone as our feature extraction network, DeepLabv3 [22] as our classifier and semantic segmentation on various datasets as our source domains. For each source domain, a DeepLabv3 network with ResNet50 backbone is pretrained by the source domain. The transfer learning is conducted subsequently by replacing the classifier with an untrained DeepLabv3 network and training with the training set of the provided OXFORD-IIIT Pet dataset, while the pretrained ResNet50 backbone is fixed. Finally, the network is evaluated on the test split by metrics including accuracy, precision, recall, F1 and IoU.

The adopted network, DeepLabv3, employs atrous convolution and achieves outstanding mIoU on PASCAL VOC 2012 [23]. Furthermore, DeepLabv3+ [24],

DeepLabv3's extension, is the current state-of-the-art solution for semantic segmentation task.

The following datasets of different sizes and fields are adopted for the pretraining:

PASCAL VOC 2012. It contains 20 categories of objects where cats and dogs are included [23]. A subset of size 2,913 is used for pretraining.
COCO train2017. A large scale dataset containing 118,287 images with 80 types of common objects including all those from the **PASCAL VOC 2012**. The majority of the images are non-iconic [25, 26].
ISIC 2018 Task 1 dataset. A dataset containing 2,594 images of skin with segmentation masks marking locations of lesion [27, 28].
Cityscapes dataset. It has various high-quality pixel-level stereo video sequences recorded in street scenes of 50 different cities, which contains 34 classes of objects including commonly seen objects in city such as pedestrians, vehicles and constructions [29]. A subset of size 2,975 is used for pretraining.
MAS3K dataset. A dataset containing images shot underwater of 37 types of marine animals such as crab and starfish [30]. A subset of size 2,910 is used for pretraining.

In terms of data preprocessing and data augmentation, data samples of **ISIC-2018** and **MAS3K** are resized to 256×256 where the nearest pixels are adopted; data samples of **Cityscapes** and **PASCAL-2012** are center cropped to 256×256. Horizontal flip of 50% probability is also performed for better generalisation.

10 training epochs are ran for pretraining and transfer learning. The model with the lowest validation loss during transfer learning is considered as the best model and is saved for evaluation. Furthermore, cross-entropy loss and Adam optimiser with learning rate 0.001 are employed for training. Cross-entropy loss is used because the essence of semantic segmentation is pixel-wise classification. Adam is adopted because of its computational efficiency and it outperforms other stochastic gradient descent algorithms in empirical studies [31].

A baseline is obtained by training a DeepLabv3 with ResNet50 backbone on the provided training set of OXFORD-IIIT pets dataset with no prior training, so as to test whether transfer learning can enhance performance. An ablation study is also conducted to compare the performance with and without fixing the ResNet backbone from the COCO pretrained model.

4 Results

	Accuracy	Recall	Precision	F1 Score	IoU
Baseline	0.931	**0.927**	0.907	0.916	0.846
COCO	**0.943**	0.912	**0.954**	**0.932**	**0.873**
COCO-Ablation	0.927	0.900	0.927	0.912	0.839
ISIC2018	0.833	0.824	0.763	0.789	0.653
Cityscapes	0.808	0.788	0.753	0.761	0.616
MAS3K	0.841	0.816	0.796	0.803	0.672
VOC2012	0.899	0.873	0.888	0.879	0.786

Table 1: Performance on Oxford-IIIT pets test set

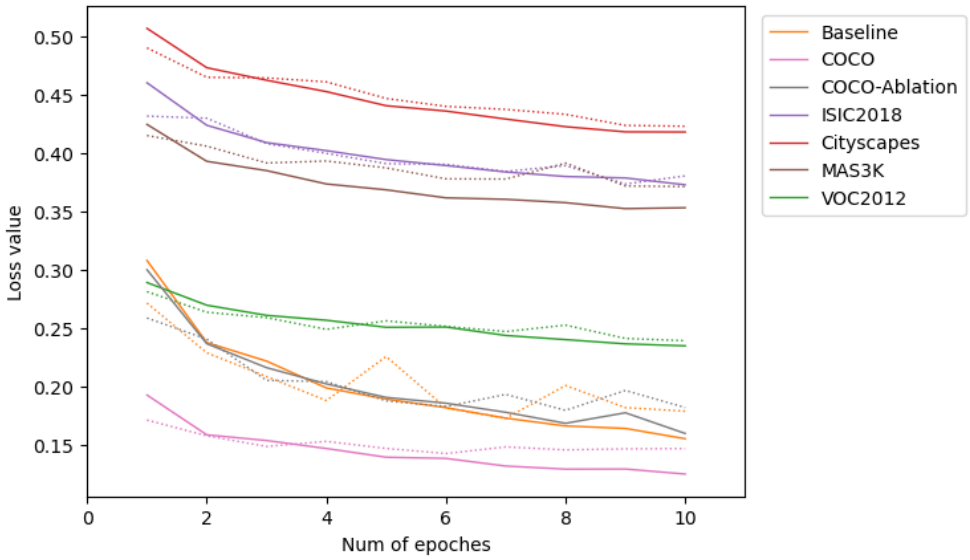

Fig. 1: Training and validation losses over tranfer learning epoches. Dotted lines and solid lines represent validation loss and training loss respectively.

Table 1 shows the metrics of performance of different models on the provided Oxford-IIIT pets test set. As described in the previous section, **Baseline** is not pretrained, the feature extraction network of **COCO-Ablation** is pretrained on COCO dataset and further trained on the Oxford-IIIT pets train set, while the other models have pretrained and fixed feature extration networks. The model with the best overall performance is **COCO** and the second is **Baseline**, followed by **VOC2012**, **MAS3K**, **ISIC2018** and **Cityscapes**.

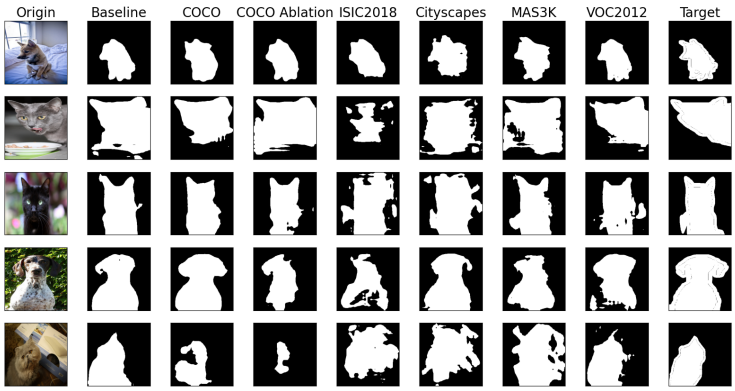

Fig. 2: Visualisation of outputs from the provided test set

COCO has higher Accuracy, Precision, F1 score and IoU than **Baseline**, proving that pretraining can indeed improves performance on target domain if applied properly. However, in other instances of transfer learning, pretrainig has shown to be harming the performance as their metrics are lower than those of **Baseline**. The causes of differences in efficacy of transfer learning could be the dismilarities of datasets sizes, task relatedness between the source and target domain. Regarding sizes, the COCO train2017 dataset has a superior size of 118,287 while the others has less than 3,000. As for task relatedness, COCO test2017 and VOC 2012 have higher relatedness than others since both of them contain object categories of cats and dogs, moreover, other categories of similar looking animals are also included such as cows, horses and sheeps (four legs and tail). Additionally, images of **MAS3K** are shot in underwater environments, images of **Cityscapes** are shot in cities and training samples of **ISIC2018** are skin images. The above properties of source domains explain part of the results, the outperforming **COCO** could have been benefited from its superior size and having better task relatedness than others. Among other datasets with similar size ranging from 2,594 to 2,975, **VOC 2012** toke advantage of its task relatedness and resulted in a better performance.

The ablation study revealed the effect of parameter fixing. **COCO** with parameter fixing outperforms **COCO-Ablation**. The reason could be that parameter fixing allows the model to benefit from the larger datasets of the source domain and therefore generalises better on unseen training set, whereas, **COCO-Ablation** can be only viewed as training on the same dataset as **Baseline** with a special weight initialisation procedure. This initialisation procedure does not seem to have significant impact on the optimisation given the training and validation loss curves depicted by Fig. 1, also, the final metrics of **COCO-Ablation**

and **Baseline** are only differed by a small amount. However, based on Fig. 2, **Baseline** outperforms **COCO-Ablation** on the outputs.

5 Discussions

Surprisingly, **Baseline** has similar loss but better output compared with **COCO-Ablation**, as mentioned in Section 4, **COCO-Ablation** can be considered as **Baseline** but with an initialisation strategy, which reduces the performance of the model. Another reason of effecting the models performance is the categories of the datasets, as a result, model trained by datasets which contains the class of images (cats and dogs) as the target task has a better and intuitive performance.

By considering the total time consumed for training models with different datasets, a limited number of epochs is applied to each training process and each image is resized/cropped to [256, 256], enable to produce models in a efficient way. However, this may cause lower quality compared with models trained with larger number of epochs and higher resolution images. The size of dataset is another factor to the quality of the models, during experiments, dataset used are slightly large compared with the Oxford-IIIT Pet Dataset. Another limitation is that only COCO dataset is used for the ablation version, rather than an ablation version for each dataset used.

Models trained by different datasets produces different results, however, the relationship between models' performance and the properties of corresponding datasets is still unknown. To be specific, properties such as, size of dataset, the similarity of content from different dataset, videos frame based dataset or independent images based dataset will requires future study and experiments. On the other side, how different model structure influences the performance of trained model is another direction of further study.

6 Conclusion

We demonstrated a pipeline to do transfer learning. In the experiments, we used DeepLabV3 ResNet50 model, and 5 datasets to pretrain the model. Among those datasets, some are more relevant to the target Oxford-IIIT Pet dataset, while others are the opposite. Through the experiments, we found that transfer learning can indeed improve the performance of the baseline model if applied properly. And the ablation study showed that fixing the backbone has better performance than not fixing the backbone. In our experiments, the performance of models transferred from other datasets except COCO was worse than the baseline. However, we observed that transferring knowledge from more related datasets would perform better than transferring knowledge from less related datasets. There were limitations in the experiments, such as different dataset sizes, different preprocessing methods. Future work can be solving the limitations and experimenting more diverse datasets.

References

[1] Yuhan Yang Xiaolong Liu Zhidong Deng. "Recent progress in semantic image segmentation". In: *Artificial Intelligence Review (2018): 1-18, arXiv:1809.10198* (2018).

[2] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[5] Liang-Chieh Chen et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[6] Yu Zhang and Qiang Yang. "A survey on multi-task learning". In: *IEEE Transactions on Knowledge and Data Engineering* (2021).

[7] Mohsen Ghafoorian et al. "Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation". In: *CoRR* abs/1702.07841 (2017). arXiv: 1702.07841. URL: http://arxiv.org/abs/1702.07841.

[8] Annegreet van Opbroek et al. "Transfer Learning Improves Supervised Image Segmentation Across Imaging Protocols". In: *IEEE Transactions on Medical Imaging* 34.5 (2015), pp. 1018–1030. DOI: 10.1109/TMI.2014.2366792.

[9] Zhexin Jiang et al. "Retinal blood vessel segmentation using fully convolutional network with transfer learning". In: *Computerized Medical Imaging and Graphics* 68 (2018), pp. 1–15. ISSN: 0895-6111. DOI: https://doi.org/10.1016/j.compmedimag.2018.04.005. URL: https://www.sciencedirect.com/science/article/pii/S0895611118302313.

[10] Michael Wurm et al. "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150 (2019), pp. 59–69. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2019.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0924271619300383.

[11] Benjamin Bischke et al. "Multi-Task Learning for Segmentation of Building Footprints with Deep Neural Networks". In: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, pp. 1480–1484. DOI: 10.1109/ICIP.2019.8803050.

[12] Amine Amyar et al. "Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation". In: *Computers in Biology and Medicine* 126 (2020), p. 104037. ISSN: 0010-4825. DOI:

https://doi.org/10.1016/j.compbiomed.2020.104037. URL: https://www.sciencedirect.com/science/article/pii/S0010482520303681.

[13] Jelena Novosel, Prashanth Viswanath, and Bruno Arsenali. "Boosting semantic segmentation with multi-task self-supervised learning for autonomous driving applications". In: *Proc. of NeurIPS-Workshops*. Vol. 3. 2019.

[14] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[15] Tahir Mehmood et al. "Combining multi-task learning with transfer learning for biomedical named entity recognition". In: *Procedia Computer Science* 176 (2020), pp. 848–857.

[16] Yipeng Hu et al. "Pre-processed Oxpet Dataset". In: (2021). URL: https://weisslab.cs.ucl.ac.uk/WEISSTeaching/datasets/-/tree/oxpet/data_new.

[17] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. *Do Better ImageNet Models Transfer Better?* 2019. arXiv: 1805.08974 [cs.CV].

[18] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[19] Jiquan Ngiam et al. "Domain Adaptive Transfer Learning with Specialist Models". In: *CoRR* abs/1811.07056 (2018). arXiv: 1811.07056. URL: http://arxiv.org/abs/1811.07056.

[20] Chen Sun et al. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*. 2017. arXiv: 1707.02968 [cs.CV].

[21] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[22] Liang-Chieh Chen et al. "Rethinking Atrous Convolution for Semantic Image Segmentation". In: *CoRR* abs/1706.05587 (2017). arXiv: 1706.05587. URL: http://arxiv.org/abs/1706.05587.

[23] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. URL: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[24] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: (Feb. 2018).

[25] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[26] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. "Coco-stuff: Thing and stuff classes in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1209–1218.

[27] Noel C. F. Codella et al. "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)". In: *CoRR* abs/1902.03368 (2019). arXiv: 1902.03368. URL: http://arxiv.org/abs/1902.03368.

[28] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions". In: *CoRR* abs/1803.10417 (2018). arXiv: 1803.10417. URL: http://arxiv.org/abs/1803.10417.

[29] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *CoRR* abs/1604.01685 (2016). arXiv: 1604.01685. URL: http://arxiv.org/abs/1604.01685.

[30] Lin Li et al. "MAS3K: An Open Dataset for Marine Animal Segmentation". In: *Benchmarking, Measuring, and Optimizing*. Ed. by Felix Wolf and Wanling Gao. Cham: Springer International Publishing, 2021, pp. 194–212. ISBN: 978-3-030-71058-3.

[31] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980.