EFFICIENT PREDICTION OF MODEL TRANSFERABILITY IN SEMANTIC SEGMENTATION TASKS

Yang Tan[1] *Yicong Li*[1,2] *Yang Li*[1,⋆] *Xiao-Ping Zhang*[1]

[1] Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
[2] John A. Paulson School of Engineering and Applied Sciences, Harvard University

ABSTRACT

How to efficiently select highly transferable pretrained models remains a challenging problem in few-shot semantic segmentation tasks. Existing transferability metrics for classification tasks are difficult to compute on segmentation data due to the high-dimensional output of the segmentation model. In this work, we generalize existing transferability metrics to efficiently predict the transferability of semantic segmentation models, by calculating transferability scores over the sampled pixel-wise features. Then with the help of transferability, we propose a transferability-weighted finetuning method which puts more importance on those low-transferability regions to improve the overall transfer accuracy on the target task. Experiments on a challenging benchmark show that the transferability scores produced by our adaptation method are highly correlated with the ground-truth transfer accuracy, achieving 0.718 Spearman's correlation coefficient on average and at least $67\times$ gain on efficiency. In addition, our transferability-weighted finetuning method outperforms vanilla fine-tuning by 4% in transfer accuracy.

***Index Terms*—** Transferability estimation, Semantic segmentation, Transfer learning, Model finetuning

1. INTRODUCTION

Obtaining a well-performed semantic segmentation model usually requires massive labeled data for supervised training. However, annotating a semantic label mask is costly such that it is difficult to acquire sufficient training data for diverse practical scenarios. Therefore, transferring reusable knowledge from related source tasks (models) to the few-shot target task is an effective way to ease the scarcity of labeled training data. Given a set of source models and a target task, how to efficiently select the highly transferable models is an essential problem in transfer learning.

Early works [1, 2] empirically evaluate the task relationships using the transfer training loss or validation accuracy, which involves expensive computation in retraining neural networks. Recently, a series of works including H-score [3], LEEP [4], OTCE [5] and LogME [6] attempt to address this problem for classification tasks (or regression tasks [6]) in the context of *transferability estimation*. They work as a function of the source and target data that efficiently approximates the ground-truth transfer accuracy. Therefore, the predicted transferability scores can serve as the indicators of selecting source models.

However, currently there are few transferability studies on semantic segmentation tasks. And there are difficulties in applying existing transferability metrics to segmentation data. The inherent difference is that the classification or regression tasks considered in previous works are low-dimensional, i.e., a classification model only predicts the category of the input instance, while the segmentation model produces a high-dimensional output since it classifies each pixel on the input image. Existing transferability scores are computed over the *global* feature of the input instance. Applying the same strategy for segmentation tasks is problematic, since the computation resources and time cost are not acceptable in practice. Another solution is taking semantic segmentation as a regression task such that a regression transferability metric like LogME can be applied. However, the regression-based LogME score is computationally more expensive than the classification-based LogME, and is prone to memory issue on practical data.

Consequently, in this work, we propose an adaptation method generalizing existing transferability metrics to semantic segmentation data. We split the global feature map into pixel-wise feature representations according to pixel locations, such that transferability scores can be computed over the pixel-wise features treating each pixel as an instance of classifying. The advantages of our adaptation method are twofold. Firstly, it is compatible with all existing metrics proposed for the classification task. Secondly, the pixel-wise feature with less dimensions reduces the computation complexity, and we can further improve the efficiency using a sampled subset of pixels for transferability estimation.

In addition to using transferability scores for source model selection, how to enhance the transfer accuracy on the target task with the help of transferability is another important problem. As transferability reveals the hardness of transfer,

⋆ Corresponding author. This study is supported in part by the Tsinghua SIGS Scientific Research Start-up Fund (Grant QD2021012C) and Natural Science Foundation of China (Grant 62001266)

it inspires us to propose a transferability-weighted finetuning method. Specifically, it measures the transferability at different pixel locations, and then uses it as a weighting coefficient plugging into the cross-entropy loss function, to encourage the model to focus more on low-transferability regions. In summary, our contributions are threefold:

1) An efficient, flexible transferability estimation framework for semantic segmentation tasks. It achieves 0.718 Spearman's correlation with the ground-truth transfer accuracy, and at least $67\times$ gain on efficiency.

2) A transferability-weighted finetuning method which measures pixel-wise transferability and uses it to improve the transfer accuracy on the target task, with 4% gain.

3) A challenging benchmark evaluating the performance of transferability estimation and model finetuning. It contains diverse cross-domain cross-task transfer configurations, using four datasets and six model architectures.

2. MEASURING TASK TRANSFERABILITY

2.1. Transferability Definition

Suppose there are a source task dataset $D_s = \{(\boldsymbol{x}_s^i, \boldsymbol{y}_s^i)\}_{i=1}^M \sim P_s(\boldsymbol{x}, \boldsymbol{y})$ and a target task dataset $D_t = \{(\boldsymbol{x}_t^i, \boldsymbol{y}_t^i)\}_{i=1}^N \sim P_t(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ represents the input image and $\boldsymbol{y}$ denotes the label mask. We have $\boldsymbol{x}_s^i, \boldsymbol{x}_t^i$ from the input space $\mathcal{X} = \mathbb{R}^{W\times H\times 3}$, and $\boldsymbol{y}_s^i, \boldsymbol{y}_t^i$ from the source label space $\mathcal{Y}_s = \{0,1\}^{W\times H\times C_s}$ and the target label space $\mathcal{Y}_t = \{0,1\}^{W\times H\times C_t}$ respectively. Here W, H denote the *width* and *height* of image. C_s, C_t represent the number of categories of the source and target tasks respectively. $P(\boldsymbol{x}_s) \neq P(\boldsymbol{x}_t)$ indicates that there exists domain shift, and $\mathcal{Y}_s \neq \mathcal{Y}_t$ indicates that the semantic contents of two tasks are different. For neural network based transfer learning, we usually transfer a source model θ_s pretrained on the source data to the target task, in which $\theta_s : \mathcal{X} \to \mathcal{P}(\mathcal{Y}_s)$ and $\mathcal{P}(\mathcal{Y}_s)$ is the space of all probability distributions over $\mathcal{Y}_s$. Here we also use θ_s to represent model parameters.

During the transfer training phase, we first use the source model parameters θ_s to initialize the target model θ_t. Then we finetune θ_t on the target training data via supervised learning. Formally, once the optimized target model is obtained, we define the *empirical transferability* as,

Definition 1 *The empirical transferability from the source task S to the target task T is measured by the expected log-likelihood of the θ_t on the testing set of target task:*

$$\begin{aligned}\mathrm{Trf}(S\to T) &= \mathop{\mathbb{E}}_{(\boldsymbol{x}_t,\boldsymbol{y}_t)\in\mathcal{X}\times\mathcal{Y}_t}\left[\log P(\boldsymbol{y}_t|\boldsymbol{x}_t;\theta_t)\right] \\ &= \mathop{\mathbb{E}}_{(\boldsymbol{x}_t,\boldsymbol{y}_t)\in\mathcal{X}\times\mathcal{Y}_t}\left[\log \prod_{j=1}^{W}\prod_{k=1}^{H} P(\boldsymbol{y}_t^{j,k}|\boldsymbol{x}_t^{j,k};\theta_t)\right].\end{aligned} \tag{1}$$

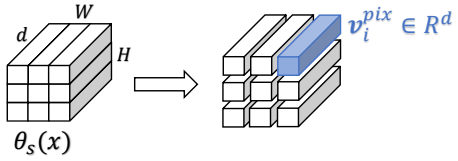

Fig. 1. Illustration of converting an embedded global feature map $\theta_s(\boldsymbol{x})$ to be pixel-wise feature representations $\{\mathbf{v}_i^{pix}\}$.

We follow the convention in deep learning based segmentation paradigm using the pixel-wise cross-entropy loss for training. In practice, we take the commonly used criteria MIoU (Mean Intersection over Union) as an approximation of the log-likelihood. Although the empirical transferability describes the ground-truth transfer performance, it is computationally expensive to obtain.

2.2. Adaptation Method

We propose a flexible adaptation method to generalize existing transferability metrics [3, 4, 5, 6] to work on segmentation data. It splits the global feature map into pixel-wise features according to pixel locations, as illustrated in Fig. 1. Formally, the pixel-wise feature set is defined as,

$$D^{pix} = \{(\boldsymbol{v}_i^{pix}, y_i^{pix})\}_{i=1}^{N\times W\times H}, \tag{2}$$

where $\boldsymbol{v}_i^{pix} \in \mathbb{R}^d$ represents the d-dimensional pixel-wise feature vector from the predicted feature maps $\{\theta_s(\boldsymbol{x}_j)\}_{j=1}^N$, and y_i^{pix} is the pixel-wise label from the ground-truth label masks $\{\boldsymbol{y}_j\}_{j=1}^N$. Here N denotes the number of image samples, and W, H represent the *width* and *height* of the image, respectively. In practice, a common segmentation dataset with 100 images of size 1024×512 contains more than 10^7 pixels. To ensure the computation efficiency, we randomly sample a subset of the source data D_s^{pix} and target data D_t^{pix} for computing transferability scores.

3. TRANSFERABILITY-WEIGHTED FINETUNING

Task Transferability scores are very useful in selecting highly transferable source models. A further question is how to utilize transferability to help downstream transfer learning. We are inspired to propose a transferability-weighted finetuning method to enhance the transfer accuracy on the target task, as shown in Fig. 2.

Specifically, we define a pixel-wise transferability map $\boldsymbol{t} \in \mathbb{R}^{W\times H}$, and $\boldsymbol{t}^{j,k}$ represents the transferability score at a pixel coordinate (j,k), where $j \in [1, W], k \in [1, H]$. Formally,

$$\boldsymbol{t}^{j,k} = \begin{cases} Trf(\{\theta_s(\boldsymbol{x}_t^i)^{j,k}\}_{i=1}^N) & \text{(for [4, 6])} \\ Trf(\{\theta_s(\boldsymbol{x}_s^i)^{j,k}\}_{i=1}^M, \{\theta_s(\boldsymbol{x}_t^i)^{j,k}\}_{i=1}^N) & \text{(for [5])}, \end{cases} \tag{3}$$

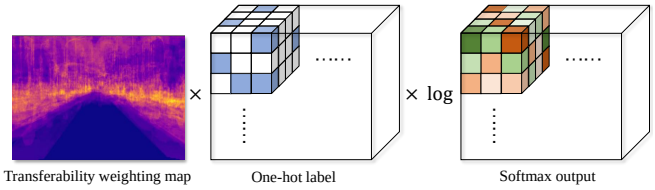

Fig. 2. Illustration of transferability-weighted cross-entropy loss function.

where $Trf()$ is the transferability metric from [4, 5, 6]. Note that H-score [3] is inapplicable here since it cannot be used to compare the transferability of a same source model with respect to different target tasks.

Pixel-wise transferability reveals the hardness of transfer at a local area, which inspires us to use it as a weighting coefficient to encourage the neural network to focus on the low-transferability regions during the finetuning phase. Formally, we define a weighting map $\boldsymbol{w} \in \mathbb{R}^{W \times H}$, where

$$\boldsymbol{w}^{j,k} = \exp\left(\frac{-\boldsymbol{t}^{j,k} - \min(-\boldsymbol{t})}{\max(-\boldsymbol{t}) - \min(-\boldsymbol{t})}\right). \tag{4}$$

Then the transferability-weighted cross-entropy loss is defined as,

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{W}\sum_{k=1}^{H} \boldsymbol{w}^{j,k} \sum_{l=1}^{C_t} \boldsymbol{y}_i^{j,k,l} \log \boldsymbol{o}_i^{j,k,l}, \tag{5}$$

where $\boldsymbol{o}_i^{j,k,l}$ is the softmax output of the target model. In practice, we can use patch-wise transferability to improve the computation efficiency.

4. EXPERIMENT

4.1. Evaluation Benchmark

We propose a challenging benchmark to evaluate the effectiveness of transferability metrics in *source model selection* and *transferability-weighted finetuning* for semantic segmentation tasks. It contains 18 source models pretrained on three datasets: BDD100K [7], GTA5 [8], ADE20K [9], with increasing domain shifts and task differences with respect to the target dataset Cityscapes [10]. More details are described in Fig. 3 and Table 1. For each source dataset, we pretrain six models with different architectures including Fcn8s [11], UNet [12], SegNet [13], PspNet [14], FrrnA and FrrnB [15]. We randomly select four cities including *aachen, cologne, jena, strasbourg* from Cityscapes as target tasks. To simulate the few-shot transfer learning scenarios, each target task has 20 labeled images for transfer training.

4.2. Evaluation on Source Model Selection

We adopt the commonly used Spearman's rank correlation coefficient (Spearman's ρ coefficient) and the Kendall rank correlation coefficient (Kendall's τ coefficient) to evaluate the correlation between the ground-truth transfer accuracy (MIoU) and predicted transferability scores. A higher correlation result indicates that the transferability metric is more accurate in prioritizing transferable source models.

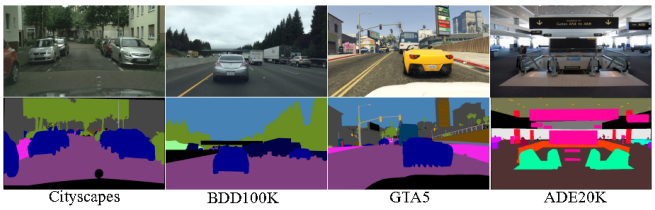

Fig. 3. Examples from semantic segmentation datasets.

Table 1. Datasets for semantic segmentation.

Dataset	Type	Categories	Training samples	Scene
Cityscapes	real captured	34	3,478	street
BDD100K	real captured	19	8,000	street
GTA5	computer game	19	24,966	street
ADE20K	real captured	150	20,210	diverse

Correlation results shown in Table 2 and Fig. 4 (target task is *aachen*) demonstrate that our adaptation method is effective to generalize existing transferability metrics to semantic segmentation data, where OTCE, LEEP, H-score and LogME are accurate in predicting the highly transferable source models, achieving an average of 0.718, 0.705, 0.676 and 0.513 on Spearman's correlation coefficient, respectively. In addition, we also notice that with the increasing of domain gaps and task differences (BDD100K $<$ GTA5 $<$ ADE20K) with respect to the target task, the transfer accuracy drops as expected. Moreover, without the requirement on GPU, the adapted transferability metrics achieve at least $67\times$ gain on efficiency compared to the *empirical transferability* (5,840s ($\sim$ 1.62h) using GPU(NVIDIA TITAN V)).

4.3. Study on Number of Sampled Pixels

To ensure the computation efficiency of transferability estimation, we propose to compute transferability scores over a sampled subset of pixel-wise features. Fig. 6 presents the effects of different numbers of sampled pixels on the accuracy of transferability estimation. It demonstrates that using a small set of pixels ($< 0.3\%$) achieves comparable performance as using the full pixels ($1,024 \times 512 \times 10 = 5,242,880$). We also notice that the accuracy of LogME drops with the increasing of pixel number, suggesting that LogME cannot converge well on a large dataset. Meanwhile, OTCE is unable to compute over the full pixels due to the memory limit. So in Table 2, we present the correlation results of OTCE, LEEP, H-score using 15,000 pixels, and LogME using 1,000 pixels.

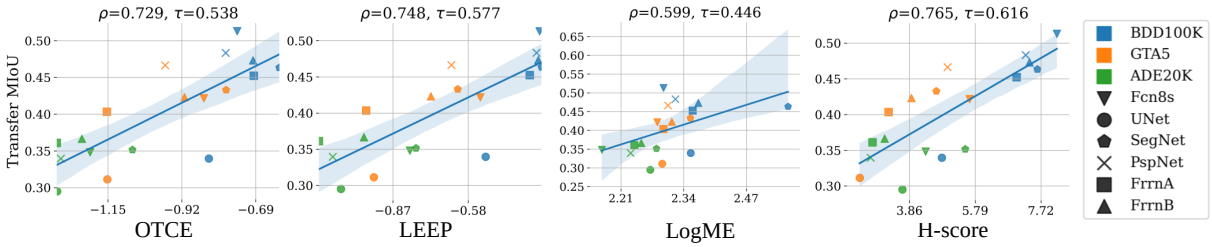

Fig. 4. Visual comparisons on the correlation between transfer accuracy (MIoU) and predicted transferability scores.

Table 2. Correlation results between transferability metrics and the ground-truth transfer accuracy (MIoU). For each target task, the upper row represents Spearman's ρ coefficient and the lower row represents Kendall's τ coefficient. The bottom row presents the computation time of transferability metrics for a single source-target pair, which is at least $67\times$ faster than the *empirical transferability* (~ 1.62h).

Target task	OTCE	LEEP	LogME	H-score
aachen	0.729	0.748	0.599	**0.765**
	0.538	0.577	0.446	**0.616**
cologne	0.787	**0.796**	0.475	0.752
	0.647	0.621	0.367	0.542
jena	**0.699**	0.686	0.587	0.583
	0.503	0.490	0.412	0.438
strasbourg	**0.657**	0.589	0.391	0.604
	0.477	0.425	0.255	0.425
Average	**0.718**	0.705	0.513	0.676
	0.541	0.528	0.370	0.506
Efficiency	87.39s	5.25s	4.33s	5.30s

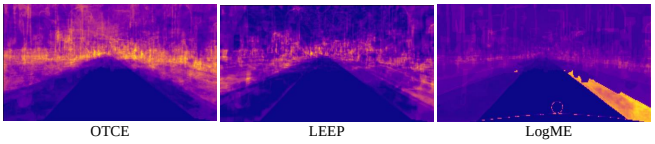

Fig. 5. An example of transferability weighting maps.

4.4. Evaluation on Transferability-Weighted Finetuning

We compare our transferability-weighted finetuning method with the commonly used vanilla finetuning under diverse transfer settings. We adopt an Adam optimizer with learning rate 0.0001 to finetune the source model on the target data for 20k iterations, and we preserve the checkpoint with best validation accuracy. Meanwhile, we compute transferability for each 4×4 patch instead of a single pixel.

Quantitative comparisons shown in Table 3 demonstrate that our transferability-weighted finetuning consistently outperforms the vanilla finetuning in the most of transfer experiments. It achieves 4% MIoU gain on average while taking OTCE as the transferability metric. As shown in Fig. 5, the weighting maps computed on street-scene segmentation datasets reveal that the areas surrounding the road exhibit higher uncertainties in transfer learning.

Table 3. MIoU of transferability-weighted finetuning (FT) and vanilla finetuning on the target task *aachen*.

Source	Model	Vanilla FT	Transferability-weighted FT		
			OTCE	LEEP	LogME
GTA5	UNet	0.3113	0.3610	**0.3640**	0.3632
	SegNet	**0.4330**	0.4324	0.4295	0.4199
	FrrnA	0.4033	**0.4634**	0.4170	0.4228
	FrrnB	0.4232	0.4351	**0.4369**	0.4288
	Fcn8s	0.4217	**0.4421**	0.4397	0.4304
	PspNet	0.4664	**0.4688**	0.4681	0.4660
	Average	0.4098	**0.4338**	0.4259	0.4219
BDD100K	Average	0.4541	**0.4643**	0.4599	0.4580
ADE20K	Average	0.3436	0.3553	0.3443	**0.3605**
Average all		0.4025	**0.4178**	0.4100	0.4135

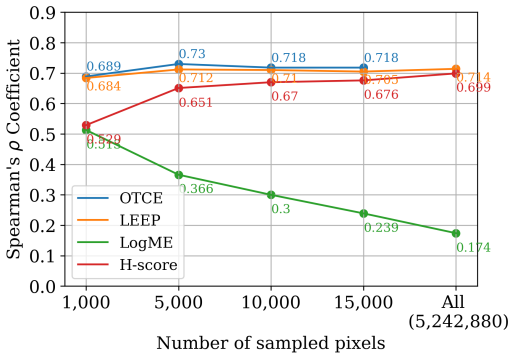

Fig. 6. Effects of the number of sampled pixels on computing transferability scores.

5. CONCLUSION

In this work, we propose a general framework compatible with existing transferability metrics to efficiently predict the transfer accuracy of semantic segmentation models, which is useful in source model (task) selection. Moreover, we propose a transferability-weighted finetuning method to improve the transfer accuracy for a given source-target task pair. Our method consistently outperforms the vanilla finetuning method in diverse transfer configurations. Future studies may further improve the accuracy of transferability estimation with involving more inherent characteristics of segmentation data like geometric relationships.

6. REFERENCES

[1] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese, "Taskonomy: Disentangling task transfer learning," 2018, pp. 3712–3722.

[2] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[3] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas, "An information-theoretic approach to transferability in task transfer learning," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2309–2313.

[4] Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and Matthias Seeger, "Leep: A new measure to evaluate transferability of learned representations," in *International Conference on Machine Learning*, 2020.

[5] Yang Tan, Yang Li, and Shao-Lun Huang, "Otce: A transferability metric for cross-domain cross-task representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, June 2021, pp. 15779–15788.

[6] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long, "Logme: Practical assessment of pre-trained models for transfer learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12133–12143.

[7] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[8] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun, "Playing for data: Ground truth from computer games," in *European conference on computer vision*. Springer, 2016, pp. 102–118.

[9] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] 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.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[15] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4151–4160.