DEPARA: Deep Attribution Graph for Deep Knowledge Transferability
– *Supplementary Material* –

Jie Song[1]*, Yixin Chen[1]*, Jingwen Ye[1], Xinchao Wang[2], Chengchao Shen[1], Feng Mao[3],
and Mingli Song[1]
[1]Zhejiang University, [2]Stevens Institute of Technology
[3]Alibaba Group

Here we provide the additional details and results that are left in the main text to this supplementary material. Firstly, we give detailed descriptions of the probe data used in this paper. Then more details about the proposed method and additional experimental results of both task transferability and layer transferability are provided.

1. Probe Data

Here we provide more details about the probe datasets used for task transferability on taskonomy models and layer selection in pre-trained VGG-19.

1.1. Task Transferability on Taskonomy Models

Following [4], we adopt three types of probe data to investigate the transferability of tasks involved in taskonomy [5].

1.1.1 Taskonomy Data

On taskonomy data [5], we construct the probe data by selecting images from the validation data of the TINY partition. In the validation set of TINY partition, images are randomly collected from 5 different buildings. We randomly select 200 images from each of these 5 buildings, constructing a probe dataset consisting of 1,000 images.

1.1.2 Indoor Scene

Indoor Scene [3] is a dataset used for indoor scene recognition. The original database contains 67 indoor categories, and a total of 15,620 images. We randomly select 15 images from each of these 67 categories, constructing a probe dataset consisting of 1,005 images.

1.1.3 COCO

The COCO [1] dataset is designed for multiple purposes including detection and captioning. On this dataset, we randomly select 1,000 images from the 2014 Val dataset to construct the probe dataset for evaluating the proposed method.

The styles of images in these three datasets vary a lot. The textures of images in taskonomy are in general simple, but those in Indoor Scene and COCO are relatively complex.

1.2. Layer Selection in Pre-trained VGG-19

The experiments of layer selection are conducted on the Syn2Real-C [2] data. For both the source and the target data in Syn2Real-C, the data is split into three groups used for training (70%), validation (10%) and test (20%). Both the two PR-DNNs, DNN-ImageNet and DNN-Source, will be transferred to the target data in Syn2Real-C. Thus we randomly sample 200 images from the validation set from the target data in Syn2Real-C as the probe data.

2. Task Transferability

In this section, to give a more comprehensive view of the proposed method, we provide more details and experimental results of the proposed method.

2.1. Algorithms Involved in DEPARA

Here we give the computation processes of $ascending\_rank$ in Eq. (1) and $descending\_rank$ in Eq. (5). The detailed computation processes are summarized in Algorithm 1 and 2, respectively.

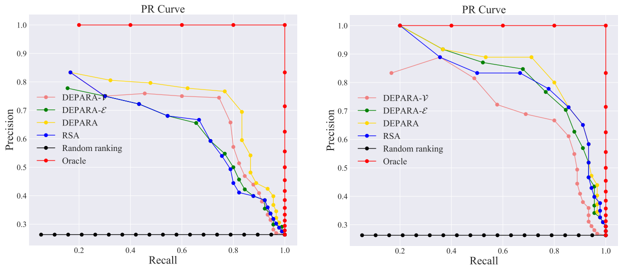
2.2. Precision-Recall Curves

In Figure S1, we depict the Precision-Recall Curves (PRCs) of the proposed method using probe data from Indoor Scene and COCO. We can see that using probe data from Indoor Scene and COCO, the proposed method still produces task transferability highly similar to that of taskonomy (the task similarity trees are depicted in Figure S2). Furthermore, utilizing both the nodes and the edges simultaneously (DEPARA) outperforms the utilizing only

[*] Equal contribution.

```
Algorithm 1: Algorithm of $ascending\_rank$ in Eq. (1)
  Input: The knowledge pool $\Omega$, the source knowledge $\mathcal{F}^{(i)}$, the target task $t_j$, the data distribution $P_j$ of $t_j$, and the labeled target data $D$
  Output: The transferability of $\mathcal{F}^{(i)}$ to task $t_j$: $\mathcal{T}_{\mathcal{F}^{(i)} \to t_j}$
  Embedding $D$ into all the embedding spaces in $\Omega$, getting the set of embeddings $\mathcal{F}^{(k)}(D)$ for each $k$;
  For each $k$, producing the hypothesis $h_{\mathcal{F}^{(k)}(D)}$ on the embeddings $\mathcal{F}^{(k)}(D)$;
  Computing the standard expected risk $\mathcal{R}^k$ of $h_{\mathcal{F}^{(k)}(D)}$ on $P_j$ for each $k$;
  Sorting the standard expected risks $\{\mathcal{R}^1, \mathcal{R}^2, ..., \mathcal{R}^N\}$ in the ascending order;
  Setting $\mathcal{T}_{\mathcal{F}^{(i)} \to t_j}$ to be the order of $\mathcal{R}^i$;
```

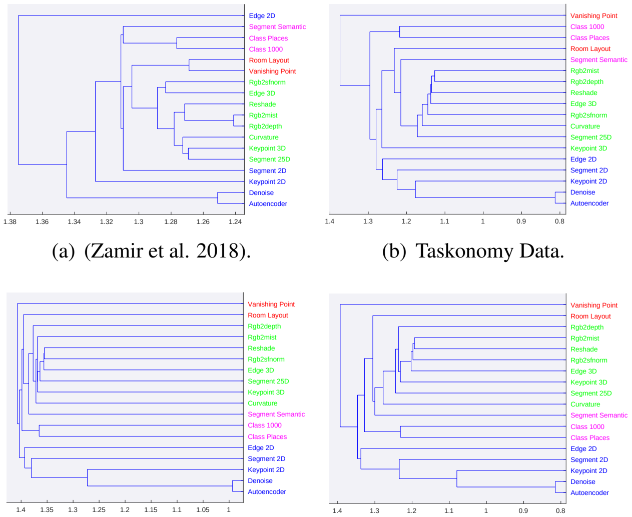
```
Algorithm 2: Algorithm of $descending\_rank$ in Eq. (5)
  Input: The knowledge pool $\Omega$, the source knowledge $\mathcal{F}_e^i$, the target knowledge $\mathcal{F}_e^j$, and the probe data $D_p$
  Output: The transferability of $\mathcal{F}_e^i$ to task $t_j$: $\mathcal{T}_{\mathcal{F}_e^i \to t_j}$
  For each $\mathcal{F}_e^k$ in $\Omega$ and the target knowledge $\mathcal{F}_e^j$, computing the DEPARA $\mathcal{G}^k$ on $D_p$ (Note that the source knowledge $\mathcal{F}_e^i$ is already in $\Omega$);
  For each $k$, computing the similarity $s_k$ between $\mathcal{G}^k$ and $\mathcal{G}^j$;
  Sorting the similarity $\{s_1, s_2, ..., s_N\}$ in the descending order;
  Setting $\mathcal{T}_{\mathcal{F}_e^i \to t_j}$ to be the order of $s_i$;
```



(a) COCO.

(b) Indoor Scene.

Figure S1. Precision-Recall Curves (PRC) on probe data from Indoor Scene and COCO.



(a) (Zamir et al. 2018).

(b) Taskonomy Data.

(c) COCO.

(d) Indoor Scene.

Figure S2. Task similarity trees of taskonomy [5] and the proposed method using taskonomy data, COCO and Indoor Scene as the probe data.

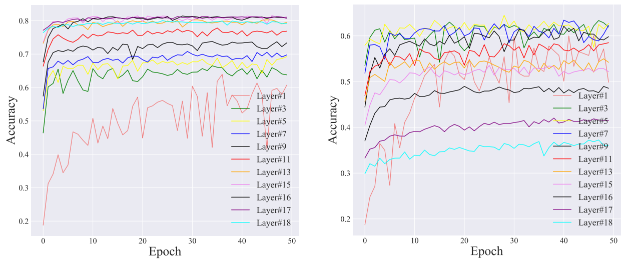
the nodes (DEPARA-$\mathcal{V}$) or edges (DEPARA-$\mathcal{E}$) by a considerable margin, which again verifies the essentiality of both the nodes and the edges for task transferability. These results are consistent with that of using taskonomy data as the probe data. Except for these findings, another interesting observation is that DEPARA-$\mathcal{V}$ outperforms DEPARA-$\mathcal{E}$ on COCO, but behaving worse on Indoor Scene. It indicates that for different probe data, the relative importance of the nodes and the edges is changing for quantifying the knowledge transferability. Thus the trade-off hyper-parameter λ in Eq. (5) needs to be tuned accordingly.

2.3. Task Similar Trees

Figure S2 depicts the task similarity trees produced by taskonomy [5] and the proposed method using taskonomy data, COCO and Indoor Scene as the probe data. The task similarity tree is acquired from agglomerative clustering of the tasks based on their transferring-out behavior [5]. The tree shows how tasks would be hierarchically positioned with respect to each other when measured based on providing information for solving other tasks; the closer two tasks, the more similar their role in transferring to other tasks. 3D, 2D, geometric, and semantic tasks clustered together in taskonomy. It can be seen that with different probe data, the proposed method produces task similarity trees alike that of taskonomy. These results demonstrate that the proposed method is insensitive the probe data to some degree, which relieves us of our burden for collecting the probe data.

3. Layer Transferability

Here we provide more results and analyses of the proposed method for tackling the layer selection problem in transfer learning.



(a) DNN-ImageNet. (b) DNN-Source.

Figure S3. Test accuracy curves of different layers when transferred to the target data in 0.1-T mode.

3.1. Test Accuracy Curves

In Figure S3, we depict the test accuracy curves of different layers when transferred to the target data. Here the experiments are conducted in 0.1-T mode. The results shown in Figure S3 further demonstrate the layers selected by the proposed method are more suitable for being transferred to the target than other layers. For example, for the PR-DNN DNN-ImageNet, the proposed method picks out the #15, #16, #17, #18 layers for being transferred. In Figure S3, it can be seen that these layers converge much faster than other layers when re-trained for the target task. The final accuracy also tends to be higher than that of other layers. Both these two characteristics are desirable for being transferred to the target task. Furthermore, layers in DNN-ImageNet produce more smooth test accuracy curves than DNN-Source, which indicates that the embedding space learned by DNN-ImageNet are more easily adapted to the target task. The embedding space learned by DNN-Source, however, is quite different in topological structure (as indicated by the low similarity of edges in DEPARA) from that learned on the target data. When adapted to the target data, it will be largely destroyed and rebuilt for the target, thus the test accuracy curves oscillate and the transferring performance is poor.

3.2. More Observations from Table 2

In the main body of this paper, we provide four main observations from the results shown in Table 2. Here we give some other interesting discoveries from Table 2 for better understanding the results: (1) Intuitively, shallow layers potentially encode richer information then deep layers. It can be observed from the results of DNN-Source, where the similarity of $\mathcal{V}$ decreases as the layers go deeper. However, for DNN-ImageNet, as the layers go deeper, the similarity of $\mathcal{V}$ firstly increases then decreases. The reason underling this phenomenon may be that DNN-Source learns from the source data decision patterns which are unsuitable for handling the target data. Thus as the layers in DNN-Source go deeper, the learned embedding space becomes less suitable. However, DNN-ImageNet learns the embedding space that is very easily adapted to the target data. As the layers in DNN-ImageNet go deeper, the learned knowledge is more suitable to the target task, thus the similarity of $\mathcal{V}$ firstly increases. However, as the layers go excessively deeper, the learned knowledge becomes too specific to the classification task of ImageNet. Thus the similarity of $\mathcal{V}$ starts to decrease after the #12 layer. (2) The main assumption in the proposed method is that higher similarity between DEPARAS indicates higher transferability between the learned deep knowledge. This assumption can be verified from Table 2 where the similarity values are monotonically correlated with the transferring performance. Specifically, the Spearman's rand-order correlation between the similarity and the transferring performance is 0.875 (in 0.1-T mode) and 0.832 (in 0.01-T mode) for DNN-ImageNet and 0.954 (in 0.1-T mode) and 0.989 (in 0.01-T mode) for DNN-Source, respectively. These high correlation coefficients imply that the similarity derived from the proposed DEPARA is a good indicator for layer selection in transfer learning.

References

[1] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[2] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *CoRR*, abs/1806.09755, 2018. 1

[3] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 1

[4] Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. Deep model transferability from attribution maps. In *NeurIPS*, pages 6179–6189. Curran Associates, Inc., 2019. 1

[5] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR 2018*, June 2018. 1, 2