An Information-Geometric Distance on the Space of Tasks

Yansong Gao[1] and Pratik Chaudhari[2]

[1]Applied Mathematics and Computational Science, University of Pennsylvania.
[2]Department of Electrical and Systems Engineering, University of Pennsylvania.
Email: gaoyans@sas.upenn.edu, pratikac@seas.upenn.edu

June 19, 2021

When is transfer learning between two tasks easy?

1. Deep networks pre-trained on a particular task typically perform well on many tasks.

When is transfer learning between two tasks easy?

1. Deep networks pre-trained on a particular task typically perform well on many tasks.
2. But there are also situations when transfer learning does not work well.
   - e.g., a pre-trained model on ImageNet is a poor representation to transfer to classification of medical images.

When is transfer learning between two tasks easy?

1. Deep networks pre-trained on a particular task typically perform well on many tasks.
2. But there are also situations when transfer learning does not work well.
   - e.g., a pre-trained model on ImageNet is a poor representation to transfer to classification of medical images.

Motivation. We would like to theoretically characterize the distance between two learning tasks.

Desiderata for a task distance

1. A learning task is defined to be a joint distribution $p(x; y)$ between input x and label y.

Desiderata for a task distance

1. A learning task is defined to be a joint distribution $p(x; y)$ between input x and label y.
2. There are many distances between task distributions.
   - e.g., KL divergence, Wasserstein distance....

Desiderata for a task distance

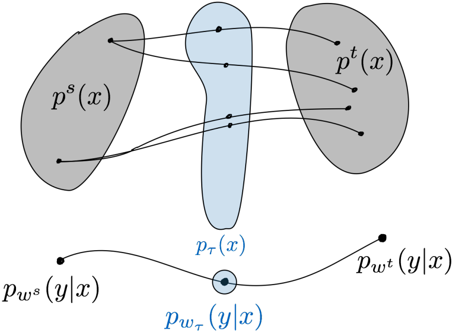
1. A learning task is defined to be a joint distribution $p(x; y)$ between input x and label y.
2. There are many distances between task distributions.
   - e.g., KL divergence, Wasserstein distance....
3. Distance between "learning" tasks is NOT the distance between two probability distributions.
   - Learning a new task depends on the capacity of hypothesis class that is used to transfer.
   - It is observed that transferring larger models is easier. A proper task distance needs to capture this fact.

Desiderata for a task distance

1. A learning task is defined to be a joint distribution $p(x; y)$ between input x and label y.
2. There are many distances between task distributions.
   - e.g., KL divergence, Wasserstein distance....
3. Distance between "learning" tasks is NOT the distance between two probability distributions.
   - Learning a new task depends on the capacity of hypothesis class that is used to transfer.
   - It is observed that transferring larger models is easier. A proper task distance needs to capture this fact.
4. Task distance should be comparable across different network architectures.

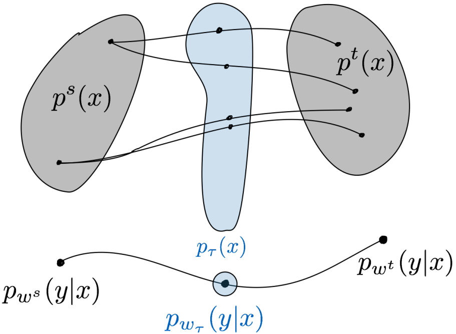
Coupled transfer distance: Modifying the task and classifier synchronously during transfer

$$p^s(x;y) = p^s{}_{w_s}(yjx)\; p^s(x) \quad ! \quad p^t(x;y) = p^t{}_{w_t}(yjx)\; p^t(x): \qquad (1)$$



Coupled transfer distance: Modifying the task and classifier synchronously during transfer

$$p^s(x;y) = p^s{}_{w_s}(yjx)\, p^s(x) \quad ! \quad p^t(x;y) = p^t{}_{w_t}(yjx)\, p^t(x): \quad (1)$$



Roughly speaking, the length of shortest trajectory connecting $p^s(x;y)$ and $p^t(x;y)$ on statistical manifold parametrized by $w\ 2\ W$ is our transfer distance.

Technical description of the coupled transfer distance

1. Manifold $M := \{p_w(Z) : w \in \mathbb{R}^p\}$ of positive measures on space Z specified by a vector parameter w.

Technical description of the coupled transfer distance

1. Manifold $M := \{p_w(Z) : w \in \mathbb{R}^p\}$ of positive measures on space Z specified by a vector parameter w.
2. Use $\mathrm{KL}[p_w, p_{w'}] = \int \mathrm{d}p_w(z) \log p_w(z)/p_{w'}(z)$ to obtain a Riemannian structure,

$$\mathrm{d}s^2 = 2\mathrm{KL}[p_w, p_{w+\mathrm{d}w}] = \sum_{i,j=1}^{p} g_{ij}\, \mathrm{d}w_i \mathrm{d}w_j, \qquad (2)$$

where (g_{ij}) is the Fisher Information Matrix (FIM).

Technical description of the coupled transfer distance

1. Manifold $M := \{p_w(Z) : w \in \mathbb{R}^p\}$ of positive measures on space Z specified by a vector parameter w.
2. Use $\mathrm{KL}[p_w, p_{w'}] = \int \mathrm{d}p_w(z) \log p_w(z)/p_{w'}(z)$ to obtain a Riemannian structure,

$$\mathrm{d}s^2 = 2\mathrm{KL}[p_w, p_{w+\mathrm{d}w}] = \sum_{i,j=1}^{p} g_{ij}\, \mathrm{d}w_i \mathrm{d}w_j, \qquad (2)$$

where (g_{ij}) is the Fisher Information Matrix (FIM).

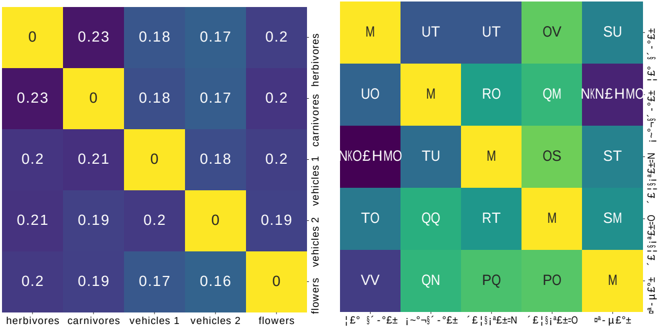
3. **Coupled transfer distance** between learning tasks is the solution of the following optimization problem.

$$\min_{\in} \int_0^1 \mathop{\mathbb{E}}_{(x,y) \sim p\ (x,y)} \sqrt{2\mathrm{KL}\, p_{w(\ )}(\cdot|x),\ p_{w(\ +\mathrm{d}\ )}(\cdot|x)}$$

$$\text{subject to } \frac{\mathrm{d}w(\ )}{\mathrm{d}} = \nabla_w \mathop{\mathbb{E}}_{(x,y) \sim p} \log p_{w(\ )}(y|x) \qquad (3)$$

$$\text{and } p\ (x) = \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} {}_{ij}\ (1\ \ )x_i + x'_j(x)$$

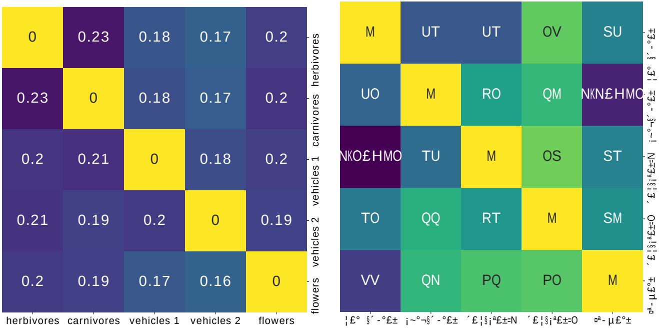
Experiments: transferring across super-classes of CIFAR-100



(a) Coupled Transfer Distance ($r = 0{:}14; p = 0{:}05$ with fine-tuning)

(b) Fine-Tuning ($r = 0{:}36; p = 0{:}03$ with itself)

Experiments: transferring across super-classes of CIFAR-100

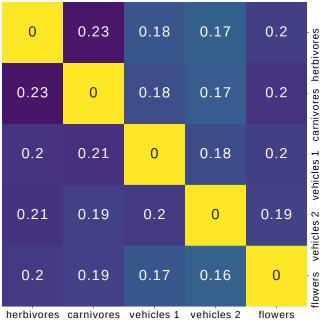


(a) Coupled Transfer Distance ($r = 0.14$, $p = 0.05$ with fine-tuning)

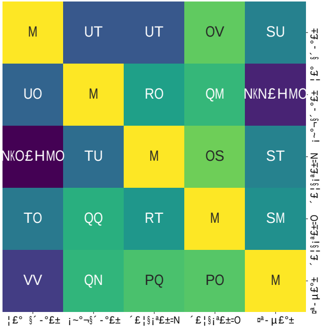
(b) Fine-Tuning ($r = 0.36$, $p = 0.03$ with itself)

1. We use the Mantel test to accept/reject the null hypothesis that variations in two distance matrices are correlated. Large r with small p indicates better correlation.

Experiments: transferring across super-classes of CIFAR-100



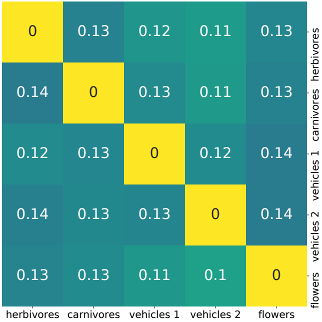
(a) Coupled Transfer Distance (r = 0:14; p = 0:05 with fine-tuning)



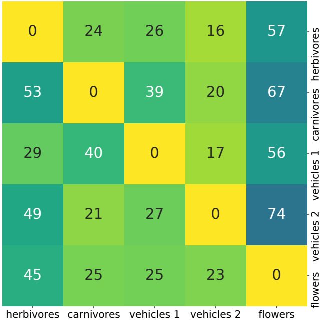
(b) Fine-Tuning (r = 0:36; p = 0:03 with itself)

1. We use the Mantel test to accept/reject the null hypothesis that variations in two distance matrices are correlated. Large r with small p indicates better correlation.
2. Task2Vec(Achille et al., 2019) does NOT correlate with the difficulty of fine-tuning well.

Larger model capacity results in smaller task distance



(a) Couple Transfer Distance(WideRes), (r = 0:15; p = 0:01 with fine-tuning)



(b) Fine-Tuning(WideRes), (r = 0:39; p = 0:01 with itself)

The larger WRN-16-4 model has a smaller task distance for all pairs compared to the smaller convolutional network on the previous slide.

Discussion

1. Our work is an attempt to theoretically understand when transfer is easy and when it is not.

Discussion

1. Our work is an attempt to theoretically understand when transfer is easy and when it is not.
2. Coupled transfer distance accurately reflects the difficulty of transfer/fine-tuning.

Discussion

1. Our work is an attempt to theoretically understand when transfer is easy and when it is not.
2. Coupled transfer distance accurately reflects the difficulty of transfer/fine-tuning.
3. Future work: Both task and weights are modified synchronously here, we would like to use the tools developed here for practical applications, e.g., to design methods that can select the best source task or the best architecture to transfer.