Supplementary Materials for LogME: Practical Assessment of Pre-trained Models for Transfer Learning

Kaichao You* [1] **Yong Liu*** [1] **Jianmin Wang** [1] **Mingsheng Long** [1]

A. Dataset description and statistics

Aircraft: The dataset contains fine-grained classification of 10,000 aircraft pictures which belongs to 100 classes, with 100 images per class.

Birdsnap: The dataset contains 49,829 images of 500 species of North American birds.

Caltech: The dataset contains 9,144 pictures of objects belonging to 101 categories. There are about 40 to 800 images per category. Most categories have about 50 images.

Cars: The dataset contains 16,185 images of 196 classes of cars. The data is split into 8,144 training images and 8,041 testing images.

CIFAR 10: The dataset consists of 60,000 32x32 colorful images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images.

CIFAR 100: The dataset is just like the CIFAR 10, except it has 100 classes containing 600 images each.

DTD: The dataset contains a collection of 5,640 textural images in the wild, annotated with a series of human-centric attributes. It has 47 classes and 120 images per class.

Pets: The dataset contains 7,049 images of cat and dog species which belongs to 47 classes, with around 200 images per class.

SUN: The dataset contains 39,700 scenery pictures with 397 classes and 100 samples per class.

For all the datasets we use, we respect the official train / val / test splits if they exist, otherwise we use 60% data for training, 20% data for validation (hyper-parameter tuning) and 20% data for testing.

B. Comparing LogME to re-training head

A naïve way to measure the relationship between features and labels is to train a classification / regression head for the downstream task, and to use the head's performance as an assessment (sometimes it is called "linear probing" or "linear protocol evaluation"). Actually we have considered this idea but find that it works not as well as expected.

The issues of re-training head are studied by researchers in visual representation learning, too. Kolesnikov et al. (2019) found that (1) re-training head by second-order optimization is impractical; (2) first-order optimization with gradients is sensitive to the learning rate schedule and takes a long time to converge.

Apart from issues discussed by Kolesnikov et al. (2019), Kornblith et al. (2019) also note that hyper-parameter of logistic regression (strength of L2 regularization) should be tuned extensively, making head re-training inefficient.

Our empirical experiments agree with the above concerns with re-training head, and also find that re-training head does not work as well as expected. In the Caltech dataset, we extract features from 10 pre-trained models, train softmax regressors with tuned hyper-parameters (the L2 regularization strength), and plot the correlation between the best head accuracy and the transfer performance *w.r.t.* the number of hyper-parameter trials in Figure 1. The correlation of LogME is plotted as a reference. Computing LogME requires 3× less time than re-training a head with one fixed hyper-parameter, and re-training head with exhaustive hyper-parameter search is still much inferior to LogME.

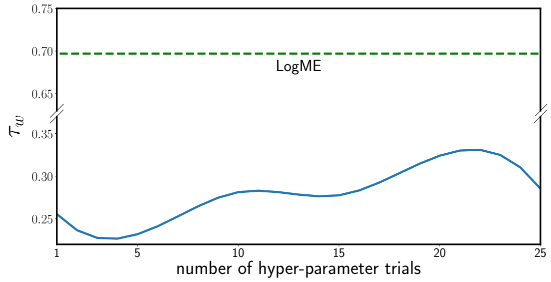

Figure 1. The correlation of re-training head *w.r.t.* the number of hyper-parameter trials. It is clear that re-training head is much worse than LogME.

As a side issue, even if we re-train a head for the downstream

*Equal contribution [1]School of Software, BNRist, Tsinghua University, Beijing 100084, China.
Kaichao You <youkaichao@gmail.com>. Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

task, it is unclear what quantity of the head should be used to measure pre-trained models. Since the performance of downstream tasks are evaluated by accuracy and MSE in transfer learning, it may somewhat cause over-fitting if we use the accuracy and MSE of the re-trained head. Indeed, in Figure 1, when the number of hyper-parameter trials increases, the correlation can even go down, showing the effect of somewhat over-fitting.

Therefore, *re-training head is neither efficient nor effective as LogME*.

C. Original Results in Figures

Original results in figures are shown in the Table 1, Table 2, and Table 3.

References

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting Self-Supervised Visual Representation Learning. In *CVPR*, 2019.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *CVPR*, 2019.

Table 1. Original results in Figure 4.

task		ResNet-50	ResNet-101	ResNet-152	DenseNet-121	DenseNet-169	DenseNet-201	Inception v1	Inception v3	MobileNet v2	NASNet-A Mobile	τ_w
Aircraft	Accuracy	86.6	85.6	85.3	85.4	84.5	84.6	82.7	88.8	82.8	72.8	-
	LEEP	-0.412	-0.349	-0.308	-0.431	-0.340	-0.462	-0.795	-0.492	-0.515	-0.506	0.11
	NCE	-0.297	-0.244	-0.214	-0.296	-0.259	-0.322	-0.348	-0.250	-0.411	-0.444	0.40
	LogME	0.946	0.948	0.950	0.938	0.943	0.942	0.934	0.953	0.941	0.948	**0.54**
Birdsnap	Accuracy	74.7	73.8	74.3	73.2	71.4	72.6	73.0	77.2	69.3	68.3	-
	LEEP	-1.647	-1.553	-1.481	-1.729	-1.756	-1.645	-2.483	-1.776	-1.951	-1.835	0.27
	NCE	-1.538	-1.479	-1.417	-1.566	-1.644	-1.493	-1.807	-1.354	-1.815	-1.778	**0.74**
	LogME	0.829	0.836	0.839	0.810	0.815	0.822	0.806	0.848	0.808	0.824	0.67
Caltech	Accuracy	91.8	93.1	93.2	91.9	92.5	93.4	91.7	94.3	89.1	91.5	-
	LEEP	-2.195	-2.067	-1.984	-2.159	-2.039	-2.122	-2.718	-2.286	-2.373	-2.263	0.27
	NCE	-1.820	-1.777	-1.721	-1.807	-1.774	-1.808	-1.849	-1.722	-2.009	-1.966	0.65
	LogME	1.509	1.548	1.567	1.365	1.417	1.428	1.440	1.605	1.365	1.389	**0.70**
Cars	Accuracy	91.7	91.7	92.0	91.5	91.5	91.0	91.0	92.3	91.0	88.5	-
	LEEP	-1.570	-1.370	-1.334	-1.562	-1.505	-1.687	-2.149	-1.637	-1.695	-1.588	0.41
	NCE	-1.181	-1.142	-1.128	-1.111	-1.192	-1.319	-1.201	-1.195	-1.312	-1.334	0.35
	LogME	1.253	1.255	1.260	1.249	1.252	1.251	1.246	1.259	1.250	1.254	**0.65**
CIFAR10	Accuracy	96.8	97.7	97.9	97.2	97.4	97.4	96.2	97.5	95.7	96.8	-
	LEEP	-3.407	-3.184	-3.020	-3.651	-3.345	-3.458	-4.074	-3.976	-3.624	-3.467	0.64
	NCE	-3.395	-3.232	-3.084	-3.541	-3.427	-3.467	-3.338	-3.625	-3.511	-3.436	0.43
	LogME	0.388	0.463	0.469	0.302	0.343	0.369	0.293	0.349	0.291	0.304	**0.78**
CIFAR100	Accuracy	84.5	87.0	87.6	84.8	85.0	86.0	83.2	86.6	80.8	83.9	-
	LEEP	-3.520	-3.330	-3.167	-3.715	-3.525	-3.643	-4.279	-4.100	-3.733	-3.560	0.61
	NCE	-3.241	-3.112	-2.980	-3.304	-3.313	-3.323	-3.253	-3.447	-3.336	-3.254	0.44
	LogME	1.099	1.130	1.133	1.029	1.051	1.061	1.037	1.070	1.039	1.051	**0.79**
DTD	Accuracy	75.2	76.2	75.4	74.9	74.8	74.5	73.6	77.2	72.9	72.8	-
	LEEP	-3.663	-3.718	-3.653	-3.847	-3.646	-3.757	-4.124	-4.096	-3.805	-3.691	-0.08
	NCE	-3.119	-3.199	-3.138	-3.198	-3.218	-3.203	-3.082	-3.261	-3.176	-3.149	-0.38
	LogME	0.761	0.757	0.766	0.710	0.730	0.730	0.727	0.746	0.712	0.724	**0.48**
Pets	Accuracy	92.5	94.0	94.5	92.9	93.1	92.8	91.9	93.5	90.5	89.4	-
	LEEP	-1.031	-0.915	-0.892	-1.100	-1.111	-1.108	-1.520	-1.129	-1.228	-1.150	0.65
	NCE	-0.956	-0.885	-0.862	-0.987	-1.072	-1.026	-1.076	-0.893	-1.156	-1.146	**0.84**
	LogME	1.029	1.061	1.084	0.839	0.874	0.908	0.913	1.191	0.821	0.833	0.58
SUN	Accuracy	64.7	64.8	66.0	62.3	63.0	64.7	62.0	65.7	60.5	60.7	-
	LEEP	-2.611	-2.531	-2.513	-2.713	-2.570	-2.618	-3.153	-2.943	-2.764	-2.687	0.58
	NCE	-2.469	-2.455	-2.444	-2.500	-2.480	-2.465	-2.534	-2.529	-2.590	-2.586	0.77
	LogME	1.744	1.749	1.755	1.704	1.716	1.718	1.715	1.753	1.713	1.721	**0.86**

Table 2. Original results in Figure 5.

task		ResNet-50	ResNet-101	ResNet-152	DenseNet-121	DenseNet-169	DenseNet-201	Inception v1	Inception v3	MobileNet v2	NASNet-A Mobile	τ_w
dSprites	MSE	0.031	0.028	0.028	0.039	0.035	0.036	0.045	0.044	0.037	0.035	-
	LogME	1.53	1.64	1.63	1.35	1.25	1.34	1.18	1.22	1.18	1.39	0.84

Table 3. Original results in Figure 6.

task		RoBERTa	RoBERTa-D	uncased BERT-D	cased BERT-D	ALBERT-v1	ALBERT-v2	ELECTRA-base	ELECTRA-small	τ_w
MNLI	Accuracy	87.6	84.0	82.2	81.5	81.6	84.6	79.7	85.8	-
	LogME	-0.568	-0.599	-0.603	-0.612	-0.614	-0.594	-0.666	-0.621	0.66
QQP	Accuracy	91.9	89.4	88.5	87.8	-	-	-	-	-
	LogME	91.9	89.4	88.5	87.8	-	-	-	-	0.73
QNLI	Accuracy	92.8	90.8	89.2	88.2	-	-	-	-	-
	LogME	-0.565	-0.603	-0.613	-0.618	-	-	-	-	1.00
SST-2	Accuracy	94.8	92.5	91.3	90.4	90.3	92.9	-	-	-
	LogME	-0.312	-0.330	-0.331	-0.353	-0.525	-0.447	-	-	0.68
CoLA	Accuracy	63.6	59.3	51.3	47.2	-	-	-	-	-
	LogME	-0.499	-0.536	-0.568	-0.572	-	-	-	-	1.00
MRPC	Accuracy	90.2	86.6	87.5	85.6	-	-	-	-	-
	LogME	-0.573	-0.586	-0.605	-0.604	-	-	-	-	0.53
RTE	Accuracy	78.7	67.9	59.9	60.6	-	-	-	-	-
	LogME	-0.709	-0.723	-0.725	-0.725	-	-	-	-	1.00