LFME: A Simple Framework for Learning from Multiple Experts in Domain Generalization

Liang Chen[1], Yong Zhang[2], Yibing Song, Zhiqiang Shen[1], Lingqiao Liu[3]
[1] MBZUAI [2] Meituan Inc [3] The University of Adelaide

NEURAL INFORMATION PROCESSING SYSTEMS

Objective and Motivation

Data in test would share similarity with that in training, more or less. If we could use the corresponding expert that specialized in the domain that the test data partially lies at, it might be beneficial for generalization.

There are problems for the motivation:

- We don't know which domain the test data lies at;
- We cannot utilize all domain expert during test, as it requires too much resources if the training domain number is very large.

A Simple Solution

We decide to obtain an expert that can specialize in all source domains, the two problems aforementioned can thus be alivaliated.

For each domain, we train an expert for each i-th domain, and we use their knowledge to refine the predictions of the target model through a logit regularization term:

$$\mathcal{L}_i = \mathcal{H}(q^{E_i}, y^i), \text{ s.t. } q_c^{E_i} = \text{softmax}(z_c^{E_i}) = \frac{exp(z_c^{E_i})}{\sum_j^K exp(z_j^{E_i})}$$

$$\mathcal{L}_{guid} = \|z - q^E\|^2$$

$$\mathcal{L}_{all} = \mathcal{L}_{cla} + \frac{\alpha}{2}\mathcal{L}_{guid}$$

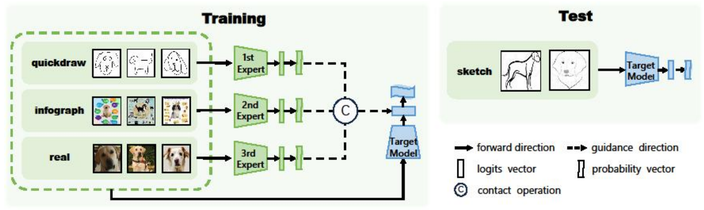
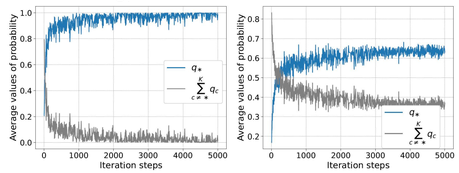


Figure 1: Pipeline of LFME. Experts and the target model are trained simultaneously. To obtain a target model that is an expert on all source domains, we learn multiple experts specialized in corresponding domains to help guide the target model during training. For each sample, the guidance is implemented with a logit regularization term that enforces similarity between the logit of the target model and probability from the corresponding expert. Only the target model is utilized in inference.

Why Logit Regularization Is Helpful

- Enable the target model to utilize more information for prediction (similar as that in label smoothing)



(a) q from ERM (b) q from LFME

Figure 2: Values of probabilities from the ERM model and LFME

- Enable the target model to mine hard samples from the expert, i.e. the dark knowledge in knowledge distillation. The larger the magtitude of the rescaling factor (postively correlated w.r.t to the probability from the expert), the more the training affected by the corresponding expert:

$$\mathcal{F} = \frac{\partial \mathcal{L}_{all}}{\partial z_*} / \frac{\partial \mathcal{L}_{cla}}{\partial z_*} = 1 - \alpha \frac{z_* - q_*^E}{1 - q_*}$$

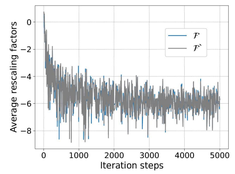
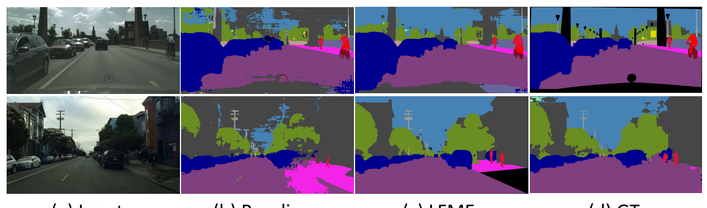


Figure 3: Rescaling factors of LFME from the gt and non-gt predctions.

(e) $\mathcal{F}$ and $\mathcal{F}'$

Results in Semantic Segmentation



(a) Input (b) Baseline (c) LFME (d) GT

	Cityscapes (%)		BDD100K (%)		Mapillary (%)		Avg. (%)	
	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc	mIOU	mAcc
Baseline†	35.46	-	25.09	-	31.94	-	30.83	-
IBN-Net [56]†	35.55	-	32.18	-	38.09	-	35.27	-
RobustNet [19]†	37.69	-	34.09	-	38.49	-	36.76	-
Baseline	37.19	48.75	27.95	39.04	32.01	48.88	32.38	45.56
PinMem [40]	41.86	48.30	34.94	44.11	39.41	49.87	38.74	47.43
SD [60]	34.77	46.63	28.00	40.33	31.41	48.18	31.39	45.05
Ours	38.38	48.99	35.70	46.16	41.04	53.71	38.37	49.62

Table 1: Evaluations on the generalizable semantic segmentation task. Gray results are from RobustNet.

Results in DG Benchmark

	PACS	VLCS	OfficeHome	TerraInc	DomainNet	Avg.
MMD [46]	81.3 ± 0.8	74.9 ± 0.5	59.9 ± 0.4	42.0 ± 1.0	7.9 ± 6.2	53.2
RSC [36]	80.5 ± 0.2	75.4 ± 0.3	58.4 ± 0.6	39.4 ± 1.3	27.9 ± 2.0	56.3
IRM [1]	80.9 ± 0.5	75.1 ± 0.1	58.0 ± 0.1	38.4 ± 0.9	30.4 ± 1.0	56.6
DANN [24]	79.2 ± 0.3	76.3 ± 0.2	59.5 ± 0.5	37.9 ± 0.9	31.5 ± 0.1	56.9
GroupGRO [66]	80.7 ± 0.4	75.4 ± 1.0	60.6 ± 0.3	41.5 ± 2.0	27.5 ± 0.1	57.1
VREx [42]	80.2 ± 0.5	75.3 ± 0.6	59.5 ± 0.1	43.2 ± 0.3	28.1 ± 1.0	57.3
CAD [65]	81.9 ± 0.3	75.2 ± 0.6	60.5 ± 0.3	40.5 ± 0.4	31.0 ± 0.8	57.8
CondCAD [65]	80.8 ± 0.5	76.1 ± 0.3	61.0 ± 0.4	39.7 ± 0.4	31.9 ± 0.7	57.9
MTL [5]	80.1 ± 0.8	75.2 ± 0.3	59.9 ± 0.5	40.4 ± 1.0	35.0 ± 0.0	58.1
ERM [75]	79.8 ± 0.4	75.8 ± 0.2	60.6 ± 0.2	38.8 ± 1.0	35.3 ± 0.1	58.1
MixStyle [91]	82.6 ± 0.4	75.2 ± 0.7	59.6 ± 0.8	40.9 ± 1.1	33.9 ± 0.1	58.4
MLDG [44]	81.3 ± 0.2	75.2 ± 0.3	60.9 ± 0.2	40.1 ± 0.9	35.4 ± 0.0	58.6
Mixup [80]	79.2 ± 0.9	76.2 ± 0.3	61.7 ± 0.5	42.1 ± 0.7	34.0 ± 0.0	58.6
MIRO [9]	75.9 ± 1.4	76.4 ± 0.4	64.1 ± 0.4	41.3 ± 0.2	36.1 ± 0.1	58.8
Fishr [62]	81.3 ± 0.3	76.2 ± 0.3	60.9 ± 0.3	42.6 ± 1.0	34.2 ± 0.3	59.0
Meta-DMoE [89]	81.0 ± 0.3	76.0 ± 0.6	62.2 ± 0.1	40.0 ± 1.2	36.0 ± 0.2	59.0
SagNet [54]	81.7 ± 0.6	75.4 ± 0.8	62.5 ± 0.3	40.6 ± 1.5	35.3 ± 0.1	59.1
SelfReg [39]	81.8 ± 0.3	76.4 ± 0.7	62.4 ± 0.1	41.3 ± 0.3	34.7 ± 0.2	59.3
Fish [69]	82.0 ± 0.3	76.9 ± 0.2	62.0 ± 0.6	40.2 ± 0.6	35.5 ± 0.0	59.3
CORAL [70]	81.7 ± 0.0	75.5 ± 0.4	62.4 ± 0.4	41.4 ± 1.8	36.1 ± 0.2	59.4
SD [60]	81.9 ± 0.3	75.5 ± 0.4	62.9 ± 0.2	42.0 ± 1.0	36.3 ± 0.2	59.7
CausEB [17]	82.4 ± 0.4	76.5 ± 0.4	62.2 ± 0.1	43.2 ± 1.3	34.9 ± 0.1	59.8
ITTA [15]	83.8 ± 0.3	76.9 ± 0.6	62.0 ± 0.2	43.2 ± 0.5	34.9 ± 0.1	60.2
RIDG [14]	82.8 ± 0.3	75.9 ± 0.3	63.3 ± 0.1	43.7 ± 0.5	36.0 ± 0.2	60.3
Ours	82.4 ± 0.1	76.2 ± 0.1	63.2 ± 0.1	46.3 ± 0.5	36.1 ± 0.1	60.8
ERM† [75]	83.1 ± 0.9	77.7 ± 0.8	65.8 ± 0.3	46.5 ± 0.9	40.8 ± 0.2	62.8
Fish† [69]	84.0 ± 0.3	78.6 ± 0.1	67.9 ± 0.5	46.6 ± 0.4	40.6 ± 0.2	63.5
CORAL† [70]	85.0 ± 0.4	77.9 ± 0.2	68.8 ± 0.3	46.1 ± 1.2	41.4 ± 0.0	63.9
SD† [60]	84.4 ± 0.2	77.6 ± 0.4	68.9 ± 0.2	46.4 ± 2.0	42.0 ± 0.2	63.9
Ours†	85.0 ± 0.5	78.4 ± 0.2	69.1 ± 0.3	48.3 ± 0.9	42.1 ± 0.1	64.6

Table 1: Evaluations on the DomainBed benchmark [1].

Takeaway Notes

- A free lunch for generalizaton, using both cross-entropy and MSE losses for classification: $\mathcal{L}_{all} = \mathcal{H}(\text{softmax}(z), y) + \frac{\alpha}{2}\|z - y\|^2$. It's like label smoothing, but we don't need to worry about their balances, it can't hurt whatever value we select. Corresponding to Sec. 6.1
- A new knowledge distillation paradigm: using the probability of the teacher to refine the logit of the sutdent. It works better for DG than feature to feature or logit to logit.
- When mining hard samples for DG, focusing on those lie in the mixed region of different domains: those can either say with domain a or b. Corresponding to Sec. D.4.

References

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In ICLR, 2021