

Table 6: Comparison of different MTL methods on CIFAR100 and VOC07+12 datasets. The 1st/2nd best results are indicated in red/blue.

Method	Before evolution (%) \uparrow			After evolution (%) \uparrow		
	CLS	DET	AVG	CLS	DET	AVG
Expert(CLS)+MAD	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	70.52 \pm 0.17	82.58 \pm 0.13	76.55\pm0.15
Expert(CLS)+Pix2SeqV2	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	70.24 \pm 0.15	80.14 \pm 0.09	75.19 \pm 0.12
Expert(CLS)+Frozen	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	70.68\pm0.00	59.92 \pm 0.06	65.30 \pm 0.03
Expert(DET)+MAD	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	70.52 \pm 0.17	82.58 \pm 0.13	76.55 \pm 0.15
Expert(DET)+Pix2SeqV2	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	70.24 \pm 0.15	80.14 \pm 0.09	75.19 \pm 0.12
Expert(DET)+Frozen	0.47 \pm 0.00	83.89 \pm 0.00	42.18 \pm 0.00	43.05 \pm 0.08	83.89\pm0.00	63.47 \pm 0.04
DISC	70.68 \pm 0.00	0.00 \pm 0.00	35.34 \pm 0.00	72.28\pm0.08 _(+1.60)	84.92\pm0.04 _(+1.03)	78.60\pm0.06 _(+2.05)

Table 7: The Relationship from human society to machine society to method design

Human Society	Machine Society	Method Design
Social hierarchy	Hierarchical organizational structures	Layer-wise hierarchical structures
Sequential progression	Progressive interaction modes	Dynamic hierarchical collaboration
Cultural learning	Strong-guided communication mechanisms	Dynamic selective collaboration

Table 8: Comparison of different backbones, training strategies and tasks on CIFAR100 and VOC07+12 datasets. The 1st/2nd best results are indicated in red/blue.

Method	CLS (%) \uparrow	DET (%) \uparrow	SEG (%) \uparrow	AVG (%) \uparrow
ResNet-50 (Scratch)	61.15 \pm 0.08	40.11 \pm 0.06	50.83 \pm 0.12	50.69 \pm 0.09
ResNet-50 (ImageNet)	64.02 \pm 0.03	67.74 \pm 0.09	55.63 \pm 0.04	62.46 \pm 0.05
ResNet-101 (ImageNet)	64.73\pm0.09	69.66\pm0.05	56.29\pm0.06	63.56\pm0.07
ResNet-152 (ImageNet)	64.97\pm0.06 _(+0.24)	71.18\pm0.07 _(+1.52)	57.68\pm0.03 _(+1.39)	64.61\pm0.05 _(+1.05)

Table 9: Comparison of different knowledge acquisition methods on CIFAR100 and VOC07+12 datasets. The 1st/2nd best results are indicated in red/blue.

Method	CLS (%) \uparrow	DET (%) \uparrow	AVG (%) \uparrow
Data (Direct Experience)	69.84 \pm 0.07	82.96 \pm 0.11	76.40 \pm 0.09
Data-augmentation (Direct Experience)	70.68 \pm 0.09	83.89 \pm 0.07	77.29 \pm 0.08
General-Model (Indirect Experience)	70.95 \pm 0.07	83.91 \pm 0.08	77.43 \pm 0.08
Specialist-Model (Indirect Experience)	71.13\pm0.15	84.01\pm0.12	77.57\pm0.14
DISC (Direct and Indirect Experience)	72.28\pm0.08 _(+1.15)	84.92\pm0.04 _(+0.91)	78.60\pm0.06 _(+1.03)

Table 10: Comparison of performance across different tasks on CIFAR10, Food-101 and WIDER FACE(E, M, H) datasets. The 1st/2nd best results are indicated in red/blue.

Method	CLS (CIFAR10) \uparrow	CLS (Food-101) \uparrow	DET (WIDER FACE (E)) \uparrow	DET (WIDER FACE (M)) \uparrow	DET (WIDER FACE (H)) \uparrow	AVG (%) \uparrow
LSKD(CLS)+MTL	89.03 \pm 0.05	77.16 \pm 0.17	66.28 \pm 0.11	66.53 \pm 0.16	50.22 \pm 0.14	69.84 \pm 0.13
LSKD(DET)+MTL	68.26 \pm 0.09	56.02 \pm 0.15	88.04 \pm 0.16	87.15 \pm 0.12	71.18 \pm 0.11	74.13 \pm 0.13
CrossKD(CLS)+MTL	89.72\pm0.08	78.01\pm0.17	68.15 \pm 0.12	67.31 \pm 0.17	51.25 \pm 0.14	70.89 \pm 0.14
CrossKD(DET)+MTL	69.04 \pm 0.05	57.28 \pm 0.12	88.92\pm0.16	87.99\pm0.12	71.78\pm0.11	75.00\pm0.11
PPAL(CLS)+MTL	89.15 \pm 0.09	77.35 \pm 0.13	67.24 \pm 0.16	67.05 \pm 0.11	50.72 \pm 0.17	70.30 \pm 0.13
PPAL(DET)+MTL	68.62 \pm 0.07	56.63 \pm 0.17	88.31 \pm 0.12	87.63 \pm 0.15	71.53 \pm 0.16	74.54 \pm 0.13
DISC	90.98\pm0.05 _(+1.26)	79.25\pm0.12 _(+1.24)	90.82\pm0.15 _(+1.90)	89.35\pm0.13 _(+1.36)	73.02\pm0.16 _(+1.24)	84.68\pm0.12 _(+9.68)

Table 11: Comparison of computational cost.

Method	FLOPs (G)	Params (M)
LSKD(CLS)	0.14	0.47
LSKD(DET)	52.96	22.59
CrossKD(CLS)	0.16	26.91
CrossKD(DET)	83.32	32.22
PPAL(CLS)	0.17	29.76
PPAL(DET)	85.05	36.45
MAD(CLS)	0.26	90.62
MAD(DET)	194.39	107.85
Pix2SeqV2(CLS)	0.29	116.89
Pix2SeqV2(DET)	204.81	132.62
DISC(CLS)	0.24	67.69
DISC(DET)	152.47	85.42

Proof. (Dynamic over static) Let $D_{\text{train}} = \{x_i, y_i\}_{i=1}^N$ be a training dataset of N samples, and $\hat{\mathbb{E}}(f^m)$ be the empirical error of the m -th model f^m on D_{train} . For any hypothesis $f \in \mathcal{H}$ (i.e., $\mathcal{H} : \mathcal{X} \rightarrow \{-1, 1\}$), with probability at least $1 - \delta$, the generalization error is bounded by:

$$\text{GError}(f) \leq \underbrace{\sum_{m=1}^M \mathbb{E}(w^m) \hat{\mathbb{E}}(f^m)}_{\text{Term-L (empirical loss)}} + \underbrace{\sum_{m=1}^M \mathbb{E}(w^m) \mathfrak{R}_m(f^m)}_{\text{Term-C (complexity)}} + \underbrace{\sum_{m=1}^M \text{Cov}(w^m, \ell^m)}_{\text{Term-Cov (covariance)}} + M \sqrt{\frac{\ln(1/\delta)}{2N}}, \quad (1)$$

where $\mathbb{E}(w^m)$ is the expected collaboration weight, $\mathfrak{R}_m(f^m)$ is the Rademacher complexity of model f^m , and $\text{Cov}(w^m, \ell^m)$ is the covariance between the weight and the loss.

In *static collaboration*, the weights w_{static}^m are constant, hence

$$\text{Cov}(w_{\text{static}}^m, \ell^m) = 0. \quad (2)$$

In *dynamic collaboration*, the collaboration weight w_{dynamic}^m increases as the model loss ℓ^m decreases. Thus,

$$\text{Cov}(w_{\text{dynamic}}^m, \ell^m) < 0, \quad (3)$$

which effectively reduces the Term-Cov and thereby lowers the generalization bound.

Based on the principle of convexity, it can be concluded that:

$$\sum \mathbb{E}(w_{\text{dynamic}}^m) \hat{\mathbb{E}}(f^m) = \sum w_{\text{static}}^m \hat{\mathbb{E}}(f^m). \quad (4)$$

$$\sum \mathbb{E}(w_{\text{dynamic}}^m) \mathfrak{R}_m(f^m) \leq \sum w_{\text{static}}^m \mathfrak{R}_m(f^m). \quad (5)$$

Since the confidence term $M\sqrt{\frac{\ln(1/\delta)}{2N}}$ is independent of the collaboration strategy, it remains the same. Therefore, suppose the hypothesis space is $\mathcal{H} : \mathcal{X} \rightarrow \{-1, 1\}$. Then for any $f_{\text{dynamic}}, f_{\text{static}} \in \mathcal{H}$, and for $1 > \delta > 0$, it holds that

$$\mathcal{O}(\text{GError}_{\text{dynamic}}) \leq \mathcal{O}(\text{GError}_{\text{static}}), \quad (6)$$

proving that dynamic collaboration yields a tighter generalization error bound than static collaboration. \square

Proof. (Dynamic weights depend on both itself and the collaborators) Given the dynamic collaboration formula delineated in $f(x) = \sum_{m=1}^{|\mathcal{M}|} \omega^m \cdot f^m(x^m)$, consider ℓ to be the convex logistic loss function. The softmax function is utilized to normalize $\omega^m = \frac{e^{w^m}}{\sum_{j=1}^{|\mathcal{M}|} e^{w^j}}$.

Considering the property of convex function, we have:

$$\ell(f(x), y) = \ell\left(\sum_{m=1}^{|\mathcal{M}|} \omega^m f^m(x^m), y\right) \leq \sum_{m=1}^{|\mathcal{M}|} \omega^m \ell(f^m(x^m), y). \quad (7)$$

When computing the expectation of Equation (7) and leveraging the properties of expectation, the subsequent equation is satisfied. To simplify notation, $\ell(f^m), y$ can be denoted as ℓ^m and D is an unknown dataset:

$$\begin{aligned} GE(f) &= \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell(f(x), y) \\ &\leq \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m \ell^m] \\ &= \frac{1}{|\mathcal{M}|} \left(|\mathcal{M}| \cdot \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m \ell^m] \right) \\ &= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m \ell^m] + (|\mathcal{M}| - 1) \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\left(1 - \sum_{j \neq m} \omega^j \right) \ell^m \right] \right) \\ &= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m] \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] + \text{Cov}(\omega^m, \ell^m) \right. \\ &\quad \left. - (|\mathcal{M}| - 1) \cdot \sum_{j \neq m} (\mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^j] \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] + \text{Cov}(\omega^j, \ell^m) - \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m]) \right) \\ &= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \left[\mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m] + (|\mathcal{M}| - 1) \cdot \left[1 - \sum_{j \neq m} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^j] \right] \right) \right. \\ &\quad \left. + \text{Cov}(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j \neq m} \text{Cov}(\omega^j, \ell^m) \right] \\ &= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \left[|\mathcal{M}| \cdot \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] \mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m] + \text{Cov}(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j \neq m} \text{Cov}(\omega^j, \ell^m) \right] \\ &= \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] \mathbb{E}_{(x,y) \sim \mathcal{D}} [\omega^m] + \frac{1}{|\mathcal{M}|} \cdot \left[\text{Cov}(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j \neq m} \text{Cov}(\omega^m, \ell^j) \right] \right) \\ &\leq \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell^m] + \frac{1}{|\mathcal{M}|} \text{Cov}(\omega^m, \ell^m) - \frac{|\mathcal{M}| - 1}{|\mathcal{M}|} \sum_{j \neq m} \text{Cov}(\omega^m, \ell^j) \right) \end{aligned} \quad (8)$$

To simplify Equation (8), we invoke Rademacher complexity theory, which establishes that with a confidence level of $1 - \Delta$ where $0 < \Delta < 1$, the following holds:

$$\mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell^m] \leq \hat{err}[f^m] + R_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}}. \quad (9)$$

In this context, $\hat{err}(f^m)$ represents the empirical error of the unimodal function f^m , and \mathcal{H} denotes the hypothesis set, defined as $\mathcal{H} : X \rightarrow \{-1, +1\}$, which includes f as a member. The Rademacher complexity is denoted by $R_N(\mathcal{H})$. Consequently, we assert that with a confidence level of $1 - \Delta$, where $0 < \Delta < 1$, the following relationship is upheld:

$$\begin{aligned} GE(f) \leq & |\mathcal{M}| \left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}} \right) + \sum_{m=1}^{|\mathcal{M}|} \hat{err}(f^m) \\ & + \sum_{m=1}^{|\mathcal{M}|} \left[\frac{1}{|\mathcal{M}|} \underbrace{Cov(\omega^m, \ell^m)}_{\text{Negative self-correlation in DSC}} - \frac{|\mathcal{M}| - 1}{|\mathcal{M}|} \sum_{j \neq m} \underbrace{Cov(\omega^m, \ell^j)}_{\text{Positive collaborative correlation in DSC}} \right], \end{aligned} \quad (10)$$

proving that dynamic weights depend on both itself and the collaborator, and that dynamic collaboration leads to a tighter generalization error bound.

□