Unveiling the Dynamics of Information Interplay in Supervised Learning

Kun Song, Zhiquan Tan*, Bochao Zou, Huimin Ma†, Weiran Huang†*

Preliminaries

Cross-entropy loss $\mathcal{H}(p,q) = -\sum_{i=0}^{n} p(x_i)\log q(x_i)$

Matrix Entropy $\mathrm{H}(\mathbf{K}) = -\operatorname{tr}\left(\frac{1}{d}\mathbf{K}\log\frac{1}{d}\mathbf{K}\right)$

Matrix Mutual Information $\mathrm{MI}(\mathbf{K}_1,\mathbf{K}_2) = \mathrm{H}(\mathbf{K}_1) + \mathrm{H}(\mathbf{K}_2) - \mathrm{H}(\mathbf{K}_1 \odot \mathbf{K}_2)$

Matrix Mutual Information Ratio $\mathrm{MIR}(\mathbf{K}_1,\mathbf{K}_2) = \frac{\mathrm{MI}(\mathbf{K}_1,\mathbf{K}_2)}{\min\{\mathrm{H}(\mathbf{K}_1),\mathrm{H}(\mathbf{K}_2)\}}$

Matrix Entropy Difference Ratio $\mathrm{HDR}(\mathbf{K}_1,\mathbf{K}_2) = \frac{|\mathrm{H}(\mathbf{K}_1)-\mathrm{H}(\mathbf{K}_2)|}{\max\{\mathrm{H}(\mathbf{K}_1),\mathrm{H}(\mathbf{K}_2)\}}$

Theoretic Insights in Supervised Learning

Neural Collapse 1 $h(\mathbf{x}_i) = \mu_{y_i}\ (i = 1, 2, \cdots, n)$

Neural Collapse 2 $\cos(\tilde{\mu}_i, \tilde{\mu}_j) = \frac{C}{C-1}\delta_j^i - \frac{1}{C-1}$

Neural Collapse 3 $\frac{\mathbf{W}^T}{\|\mathbf{W}\|_F} = \frac{\mathbf{M}}{\|\mathbf{M}\|_F}$, where $\mathbf{M} = [\tilde{\mu}_1 \cdots \tilde{\mu}_C]$

Gram Matrix $\mathbf{G}(\mathbf{Z}) = \hat{\mathbf{Z}}^T\hat{\mathbf{Z}}, where\ \hat{\mathbf{Z}} = \left[\frac{\mathbf{z}_1}{\|\mathbf{z}_1\|} \cdots \frac{\mathbf{z}_N}{\|\mathbf{z}_N\|}\right]$

When Nerual Collaspe happens

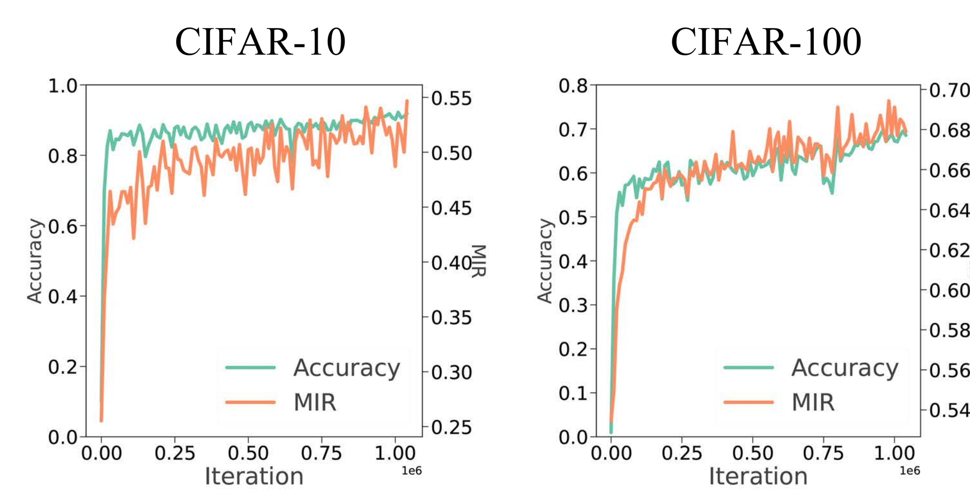
$$\mathrm{HDR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = 0$$

$$\mathrm{MIR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-2)}$$

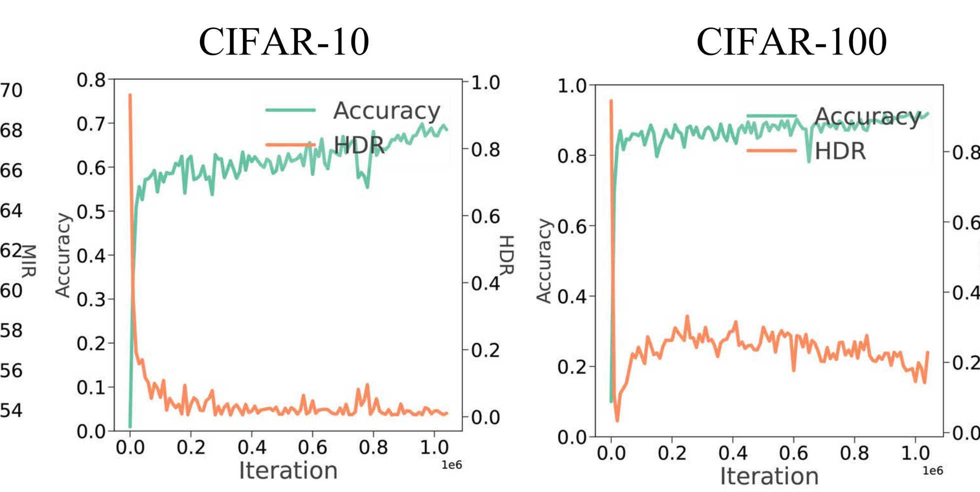
$$\frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)} \approx \frac{1}{C-1} + \frac{(C-2)\log(C-1)}{(C-1)\log(C-1)} = 1$$

Information Interplay in Supervised Learning

Information Interplay during Standard Supervised Learning



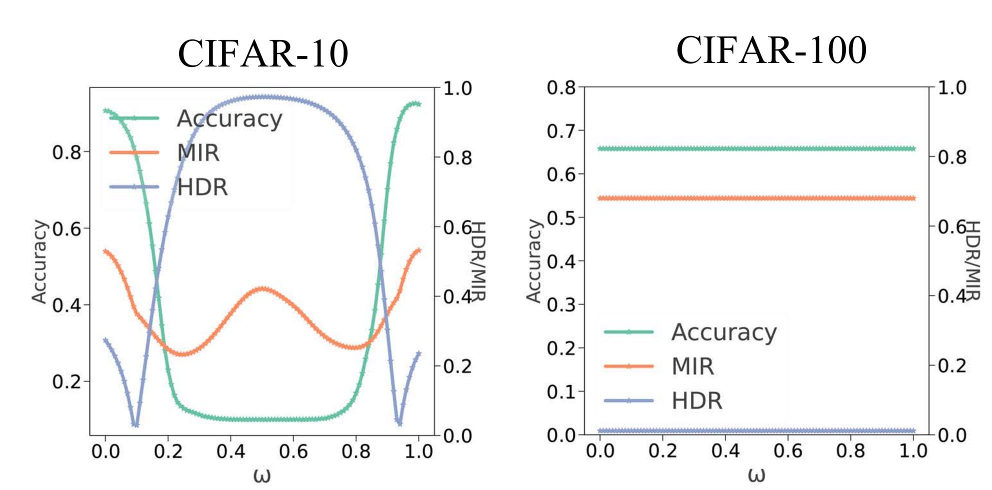
Accuracy and MIR on the test set during training



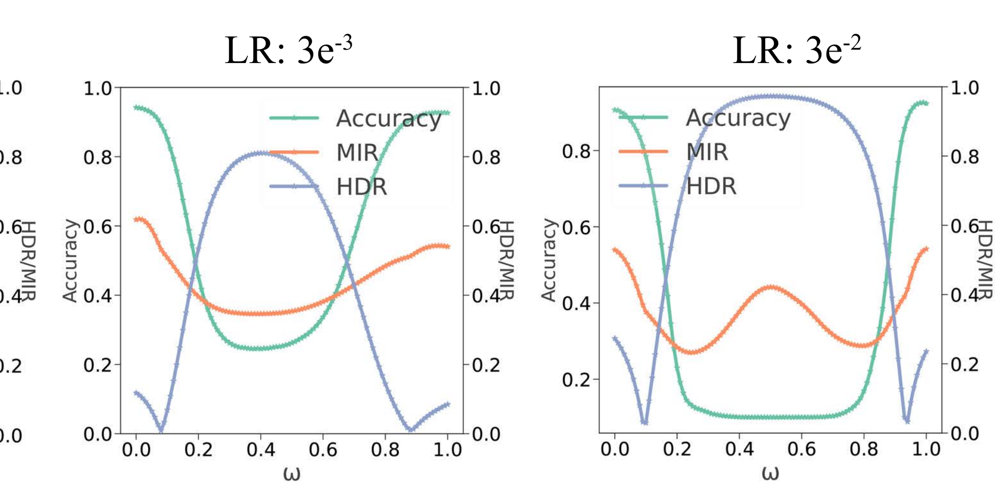
Accuracy and HDR on the test set during training

Information Interplay in Linear Mode Connectivity

$$h = (1-\omega)\cdot h_1 + \omega \cdot h_2$$

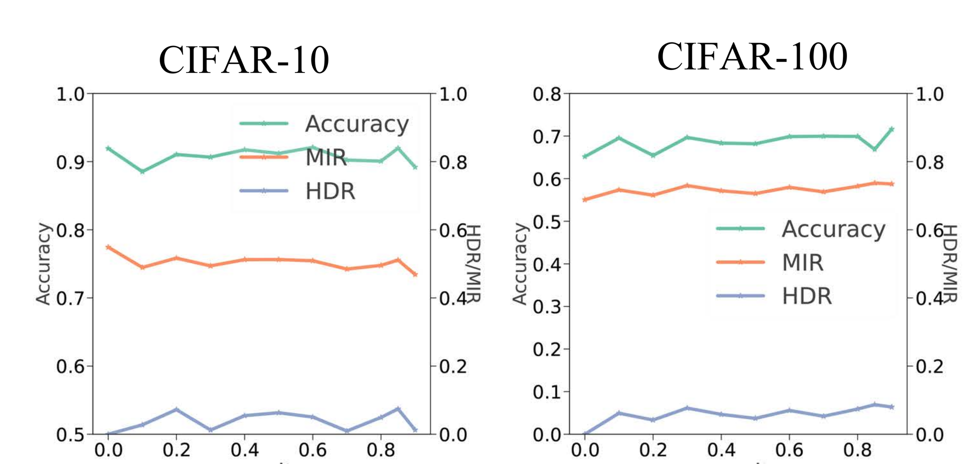


Metrics with different weight on the test set (LR: 3e-4)



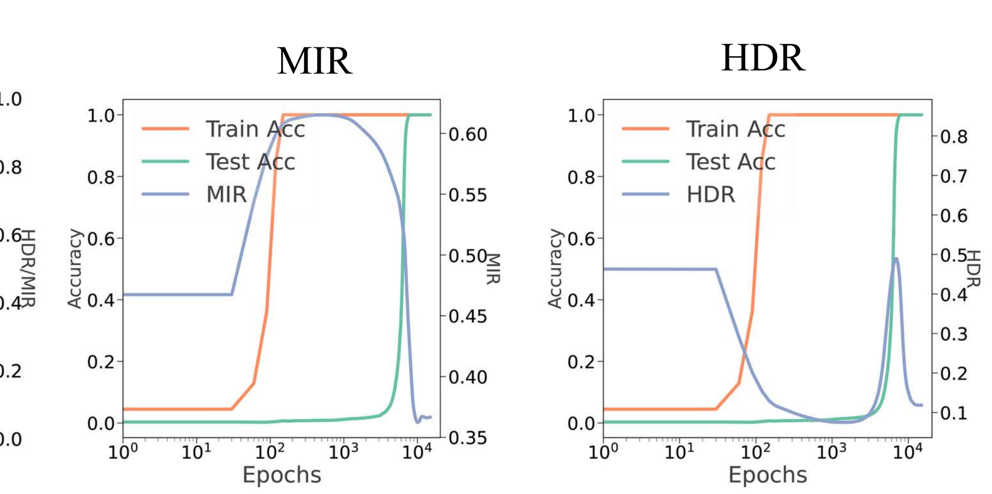
Models trained with different learning rate on the test set

Model Pruning



Accuracy, MIR and HDR between the features extracted by model before and after pruning with different sparsity

Grokking



Accuracy, MIR and HDR during grokking

Improving Supervised Learning

Maximizing Mutual Information $\mathcal{L} = \mathcal{L}_s - \lambda_{mi}\cdot \mathbf{MI}(\mathbf{G}(f), \mathbf{G}(V))$

Minimizing Entropy Difference $\mathcal{L} = \mathcal{L}_s + \lambda_{id}\cdot |\mathbf{H}(\mathbf{G}(f)) - \mathbf{H}(\mathbf{G}(V))|$

Table 2. Results for fully supervised learning

Datasets	CIFAR-10	CIFAR-100
Fully supervised	95.35	80.77
Ours (MIR)	95.52	80.81
Ours (HDR)	**95.57**	**80.96**

Improving Semi-Supervised Learning

Maximizing Mutual Information $\mathcal{L} = \mathcal{L}_{ssl} - \lambda_{mi}\cdot \mathbf{MI}(\mathbf{G}(f'), \mathbf{G}(V'))$

Minimizing Entropy Difference $\mathcal{L} = \mathcal{L}_{ssl} + \lambda_{id}\cdot |\mathbf{H}(\mathbf{G}(f')) - \mathbf{H}(\mathbf{G}(V'))|$

Table 1. Error rates (100% - accuracy) on CIFAR-10/100, and STL-10 datasets for state-of-the-art methods in semi-supervised learning. Bold indicates the best performance, and underline indicates the second best.

Dataset	CIFAR-10			CIFAR-100		STL-10	
# Label	10	40	250	400	2500	40	1000
Π Model (Rasmus et al., 2015)	79.18±1.11	74.34±1.76	46.24±1.29	86.96±0.80	58.80±0.66	74.31±0.85	32.78±0.40
Pseudo Label (Lee et al., 2013)	80.21± 0.55	74.61±0.26	46.49±2.20	87.45±0.85	57.74±0.28	74.68±0.99	32.64±0.71
VAT (Miyato et al., 2018)	79.81± 1.17	74.66±2.12	41.03±1.79	85.20±1.40	48.84±0.79	74.74±0.38	37.95±1.12
MeanTeacher (Tarvainen & Valpola, 2017)	76.37± 0.44	70.09±1.60	37.46±3.30	81.11±1.44	45.17±1.06	71.72±1.45	33.90±1.37
MixMatch (Berthelot et al., 2019b)	65.76± 7.06	36.19±6.48	13.63±0.59	67.59±0.66	39.76±0.48	54.93±0.96	21.70±0.68
ReMixMatch (Berthelot et al., 2019a)	20.77± 7.48	9.88±1.03	6.30±0.05	42.75±1.05	26.03±0.35	32.12±6.24	6.74±0.17
UDA (Xie et al., 2020)	34.53± 10.69	10.62±3.75	5.16±0.06	46.39±1.59	27.73±0.21	37.42±8.44	6.64±0.17
FixMatch (Sohn et al., 2020)	24.79± 7.65	7.47±0.28	5.07±0.05	46.42±0.82	28.03±0.16	35.97±4.14	6.25±0.33
Dash (Xu et al., 2021)	27.28± 14.09	8.93±3.11	5.16±0.23	44.82±0.96	27.15±0.22	34.52±4.30	6.39±0.56
MPL (Pham et al., 2021)	23.55± 6.01	6.93±0.17	5.76±0.24	46.26±1.84	27.71±0.19	35.76±4.83	6.66±0.00
FlexMatch (Zhang et al., 2021)	13.85± 12.04	4.97±0.06	4.98±0.09	39.94±1.62	26.49±0.20	29.15±4.16	5.77±0.18
FreeMatch (Wang et al., 2023)	8.07± 4.24	4.90±0.04	4.88±0.18	37.98±0.42	26.47±0.20	15.56±0.55	5.63±0.15
OTMatch (Tan et al., 2023c)	4.89± 0.76	4.72±0.08	4.60±0.15	37.29±0.76	26.04±0.21	**12.10±0.72**	5.60±0.14
SoftMatch (Chen et al., 2023)	4.91± 0.12	4.82±0.09	**4.04±0.02**	37.10±0.07	26.66±0.25	21.42±3.48	5.73±0.24
FreeMatch + Maximizing Mutual Information (Ours)	<u>4.87± 0.66</u>	<u>4.66± 0.13</u>	<u>4.56± 0.15</u>	**36.41± 1.91**	**25.77± 0.35**	16.61± 1.19	**5.24 ± 0.17**
FreeMatch + Minimizing Entropy Difference (Ours)	**4.69± 0.16**	**4.63± 0.25**	4.60± 0.15	<u>37.31± 1.96</u>	<u>25.79± 0.41</u>	<u>14.93 ± 3.28</u>	<u>5.30 ± 0.18</u>

WeChat **Paper**