Unveiling the Dynamics of Information Interplay in Supervised Learning

Kun Song [* 1] **Zhiquan Tan** [* 2] **Bochao Zou** [1] **Huimin Ma** [† 1] **Weiran Huang** [† 3 4]

Abstract

In this paper, we use matrix information theory as an analytical tool to analyze the dynamics of the information interplay between data representations and classification head vectors in the supervised learning process. Specifically, inspired by the theory of Neural Collapse, we introduce matrix mutual information ratio (MIR) and matrix entropy difference ratio (HDR) to assess the interactions of data representation and class classification heads in supervised learning, and we determine the theoretical optimal values for MIR and HDR when Neural Collapse happens. Our experiments show that MIR and HDR can effectively explain many phenomena occurring in neural networks, for example, the standard supervised training dynamics, linear mode connectivity, and the performance of label smoothing and pruning. Additionally, we use MIR and HDR to gain insights into the dynamics of grokking, which is an intriguing phenomenon observed in supervised training, where the model demonstrates generalization capabilities long after it has learned to fit the training data. Furthermore, we introduce MIR and HDR as loss terms in supervised and semi-supervised learning to optimize the information interactions among samples and classification heads. The empirical results provide evidence of the method's effectiveness, demonstrating that the utilization of MIR and HDR not only aids in comprehending the dynamics throughout the training process but can also enhances the training procedure itself.

*Equal contribution [1] University of Science and Technology Beijing [2] Department of Mathematical Sciences, Tsinghua University [3] MIFA Lab, Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University [4] Shanghai AI Laboratory. †Correspondence to: Weiran Huang <weiran.huang@outlook.com>, Huimin Ma <mhmpub@ustb.edu.cn>.

1. Introduction

Supervised learning is a significant part of machine learning, tracing its development back to the early days of artificial intelligence. Leveraging ample annotated data from large-scale datasets like ImageNet (Krizhevsky et al., 2012) and COCO (Lin et al., 2014), supervised learning has achieved outstanding performance in tasks such as image recognition (He et al., 2016; Girshick, 2015; Ronneberger et al., 2015), speech recognition (Hinton et al., 2012; Chan et al., 2016), and natural language processing (Vaswani et al., 2017), thereby advancing the development of artificial intelligence. Concurrently, with its enhanced performance in real-world applications, some interesting phenomena in supervised learning, such as Neural Collapse (Papyan et al., 2020), linear mode connectivity (Frankle et al., 2020), and grokking (Power et al., 2022) have emerged. More and more work is beginning to explore the reasons behind these phenomena.

Neural Collapse (Papyan et al., 2020) is an interesting phenomenon observed during the training process of supervised learning. In different stages of network training, features of samples within the same class become more similar in the feature space, meaning that intra-class differences decrease. At the same time, feature vectors of different classes become more distinct in feature space, leading to more significant inter-class differences. In supervised learning classification tasks, after prolonged training, a special alignment occurs, this alignment is formed between the weights of the network's final fully connected layer and the feature vectors of the classes. This indicates that for each class, the centroid of its feature vector almost coincides with the weight vector of its corresponding classifier (classification head vector).

Existing work on the theory of Neural Collapse primarily solely focuses on feature or classification head similarity, with few studies exploring the information between features and class classification heads. We introduce matrix information theory as an analytical tool, using similarity matrices constructed from sample features and class classification heads to analyze the information dynamics along the training process. According to Neural Collapse, for a well-trained model at its terminal phase, sample features and the corresponding class classification heads align well. Thus, at the stage of Neural Collapse, the similarity matrix of sample

features aligns with the similarity matrix constructed by the corresponding class classification heads. Therefore, we first theoretically calculate their matrix mutual information ratio and the matrix entropy difference ratio at the point of Neural Collapse. We find that at the point of Neural Collapse, the theoretical MIR is nearing its maximum, and the HDR reaches its theoretical minimum. Therefore, we expect an increase in MIR and a decrease in HDR during the training process. Experiments demonstrate that MIR and HDR effectively describe such phenomena during training. Motivated by the success of understanding dynamics in standard training, we also explore their uses in other settings like linear mode connectivity, label smoothing, pruning, and grokking. Compared to accuracy, MIR and HDR not only can describe the above phenomena but also possess unique analytical significance. We also explore integrating information constraints on model performance, by adding MIR and HDR as a loss term in supervised and semi-supervised learning. Experiments show that adding additional information constraints during training effectively enhances model performance, especially in semi-supervised learning with limited labeled samples, where information constraints help the model better learn information from unlabeled samples.

Our contributions are as follows:

1. Motivated by Neural Collapse and matrix information theory, we introduce two new metrics: Matrix Mutual Information Ratio (MIR) and Matrix Entropy Difference Ratio (HDR), for which we also deduce their theoretical values when Neural Collapse happens.

2. Through rigorous experiments, we find that MIR and HDR are capable of explaining various phenomena, such as the standard training of supervised learning, linear mode connectivity, pruning, label smoothing, and grokking.

3. We integrate matrix mutual information and information entropy differences as a loss term in both supervised and semi-supervised learning. Experiments demonstrate that these information metrics can effectively improve model performance.

2. Related Work

Neural network training phenomenon. Recent research has revealed several interesting phenomena that are significant for understanding the behavior and learning dynamics of neural networks. Firstly, Papyan et al. (2020) observe that in the final stages of deep neural network training, the feature vectors of the last layer tend to converge to their class centroids, and these class centroids align with the weights of the corresponding class in the final fully connected layer. This phenomenon is known as **Neural Collapse**. Neural Collapse occurs in both MSE loss and cross-entropy loss (Han et al., 2021; Zhou et al., 2022). Secondly, Frankle

et al. (2020) find that models trained from the same starting point, even when changing the input data sequence and data augmentation, eventually converge to the same local area. This phenomenon is termed **Linear Mode Connectivity**, which is influenced by architecture, training strategy, and dataset (Altıntaş et al., 2023). Lastly, Power et al. (2022) discover that after prolonged training, models can transition from simply memorizing data to inductively processing it. This phenomenon is known as the **Grokking**. Nanda et al. (2022) finds the connections of grokking on the modulo addition task with trigonometric functions.

Information theory. Traditional information theory provides a universally applicable set of fundamental concepts and metrics to understand the relationship between probability distributions and information (Wang et al., 2021). However, when dealing with high-dimensional data and complex data structures, traditional information theory tools struggle to analyze higher-order relationships within the data. As an extension and advancement of traditional information theory, matrix information theory broadens the scope of information theory to encompass the analysis of inter-matrix relationships. This enables a better understanding of the latent structures in data and more effective handling of complex relationships in high-dimensional data (Bach, 2022). There have been works that utilize matrix mutual information to analyze neural networks. Tan et al. (2023b) use matrix mutual information to study the Siamese architecture self-supervised learning methods. Zhang et al. (2023a) point out the relationship between effective rank, matrix entropy, and equiangular tight frame.

Semi-supervised learning. Semi-supervised learning focuses on how to train a better model using a small number of labeled data and a large number of unlabeled data (Sohn et al., 2020; Zhang et al., 2021; Chen et al., 2023; Tan et al., 2023c; Wang et al., 2023; Tan et al., 2023a; Zhang et al., 2023b). FixMatch (Sohn et al., 2020) ingeniously integrate consistency regularization with pseudo-labeling techniques. MixMatch (Berthelot et al., 2019b) amalgamates leading SSL methodologies, achieving a substantial reduction in error rates and bolstering privacy protection. FlexMatch (Zhang et al., 2021) introduces Curriculum pseudo-labeling to improve semi-supervised learning by dynamically adapting to the model's learning status, showing notable efficacy in challenging scenarios with limited labeled data. Softmatch (Chen et al., 2023) efficiently balances the quantity and quality of pseudo-labels in semi-supervised learning, demonstrating significant performance improvements in diverse applications including image and text classification. FreeMatch (Wang et al., 2023) innovates in semi-supervised learning by self-adaptively adjusting confidence thresholds and incorporating class fairness regularization, significantly outperforming existing methods in scenarios with scarce la-

beled data. How to more accurately utilize the information of unlabeled data remains an important problem in the field of semi-supervised learning.

3. Preliminaries

3.1. Supervised classification problem

Given a labeled dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $y_i \in \{1, 2, \cdots, C\}$ is the class label. In this paper, we mainly consider training an image classification problem by concatenation of a deep neural network h and a linear classifier. The linear classifier consists of a weight matrix $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^{C \times 1}$. Denote $\mathbf{W}^T = [w_1 \cdots w_C]$. The training loss is the cross-entropy loss.

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

where p is the true probability distribution, and q is the predicted probability distribution.

3.2. Matrix entropy and mutual information

The following definitions of matrix entropy and matrix mutual information are taken from paper (Skean et al., 2023).

Definition 3.1 (Matrix entropy). Suppose a positive-definite matrix $\mathbf{K} \in \mathbb{R}^{d \times d}$ which $\mathbf{K}(i, i) = 1$ ($1 \leq i \leq d$). The matrix entropy is defined as follows:

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

In the following we assume that $\mathbf{K}_j \in \mathbb{R}^{d \times d}$ which $\mathbf{K}_j(i, i) = 1$ ($1 \leq i \leq d, j = 1, 2$).

Definition 3.2 (Matrix mutual information). The matrix mutual information is defined as follows:

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

where $\odot$ is the Hardmard product.

Based on the two definitions above, we can introduce the following concepts, which measure the normalized information interactions between matrices.

Definition 3.3 (Matrix mutual information ratio (MIR)). The matrix mutual information ratio is defined as follows:

$$\mathrm{MIR}(\mathbf{K}_1, \mathbf{K}_2) = \frac{\mathrm{MI}(\mathbf{K}_1, \mathbf{K}_2)}{\min\{\mathrm{H}(\mathbf{K}_1), \mathrm{H}(\mathbf{K}_2)\}}.$$

Definition 3.4 (Matrix entropy difference ratio (HDR)). The matrix entropy difference ratio is defined as follows:

$$\mathrm{HDR}(\mathbf{K}_1, \mathbf{K}_2) = \frac{|\mathrm{H}(\mathbf{K}_1) - \mathrm{H}(\mathbf{K}_2)|}{\max\{\mathrm{H}(\mathbf{K}_1), \mathrm{H}(\mathbf{K}_2)\}}.$$

4. Theoretic Insights in Supervised Learning

4.1. Neural collapse

Neural Collapse (NC) is an interesting phenomenon (Papyan et al., 2020) appeared at the terminal phase of the classification problem. We will briefly summarize the 3 most important NC conditions for our paper as follows.

Denote $\mu_G = \frac{\sum_{i=1}^n h(\mathbf{x}_i)}{n}$ and $\mu_c = \frac{\sum_{y_i=c} h(\mathbf{x}_i)}{\#\{y_i=c\}}$ be the global mean and class-wise mean respectively. Then we can define $\tilde{\mu}_c = \mu_c - \mu_G$.

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

(NC 2) $\cos(\tilde{\mu}_i, \tilde{\mu}_j) = \frac{C}{C-1}\delta_j^i - \frac{1}{C-1}$, where cos is the cosine similarity and δ_j^i is Kronecker symbol.

(NC 3) $\frac{\mathbf{W}^T}{\|\mathbf{W}\|_F} = \frac{\mathbf{M}}{\|\mathbf{M}\|_F}$, where $\mathbf{M} = [\tilde{\mu}_1 \cdots \tilde{\mu}_C]$.

In this paper, the matrics used in the matrix information quantities usually is the similarity (gram) matrix. For ease of exposition, we introduce a standard way of constructing a similarity (gram) matrix as follows.

Definition 4.1 (Construction of similarity (gram) matrix). Given a set of representations $\mathbf{Z} = [\mathbf{z}_1 \cdots \mathbf{z}_N] \in \mathbb{R}^{d \times N}$. Denote the l_2 normalized feature $\hat{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}$, $\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_1 \cdots \hat{\mathbf{z}}_N]$. Then gram matrix is defined as $\mathbf{G}(\mathbf{Z}) = \hat{\mathbf{Z}}^T \hat{\mathbf{Z}}$.

Note that Neural Collapse conditions impose structural information on the weight matrix and class means, we provide the matrix mutual information ratio and matrix entropy difference ratio of this imposed structure in Theorem 4.2.

Theorem 4.2. *Suppose Neural collapse happens. Then* $\mathrm{HDR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = 0$ *and* $\mathrm{MIR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)}$.

The proof can be seen in Appendix A.1. As the linear weight matrix $\mathbf{W}$ can be seen as (prototype) embedding for each class. It is natural to consider the mutual information and entropy difference between sample embedding and label embedding. We discuss this in the following Corollary 4.3.

Corollary 4.3. *Suppose the dataset is class-balanced, $\mu_G = 0$ and Neural collapse happens. Denote $\mathbf{Z}_1 = [h(\mathbf{x}_1) \cdots h(\mathbf{x}_n)] \in \mathbb{R}^{d \times n}$ and $\mathbf{Z}_2 = [w_{y_1} \cdots w_{y_n}] \in \mathbb{R}^{d \times n}$. Then $\mathrm{HDR}(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ and $\mathrm{MIR}(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)}$.*

Remark: Note $\frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)} \approx \frac{1}{C-1} + \frac{(C-2)\log(C-1)}{(C-1)\log(C-1)} = 1$ and MIR, HDR $\in [0, 1]$. These facts make quantities obtained by Theorem 4.2 and 4.3 very interesting, as HDR reaches the minimum possible value and MIR approximately reaches the highest possible value.

4.2. Some theoretical insights for HDR

Mutual information is a very intuitive quantity in information theory. On the other hand, it seems weird to consider the difference of entropy, but we will show that this quantity is closely linked with comparing the approximation ability of different representations on the same target.

For ease of theoretical analysis, in this section, we consider the MSE regression loss.

The following Lemma 4.4 shows that the regression of two sets of representations $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to the same target $\mathbf{Y}$ are closely related. And the two approximation errors are closely related to the regression error of $\mathbf{Z}_1$ to $\mathbf{Z}_2$.

Lemma 4.4. *Suppose* $\mathbf{W}_1^*, \mathbf{b}_1^* = \arg\min_{\mathbf{W},\mathbf{b}} \|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_1 + \mathbf{b}\mathbf{1}_N)\|_F$. *Then* $\min_{\mathbf{W},\mathbf{b}} \|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_2 + \mathbf{b}\mathbf{1}_N)\|_F \leq \min_{\mathbf{W},\mathbf{b}} \|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_1 + \mathbf{b}\mathbf{1}_N)\|_F + \|\mathbf{W}_1^*\|_F \min_{\mathbf{H},\eta} \|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F$.

The proof can be found in Appendix B.1. From Lemma 4.4, we know that the regression error of $\mathbf{Z}_1$ to $\mathbf{Z}_2$ is crucial for understanding the differences of representations. We further bound the regression error with rank and singular values in the following Lemma 4.5.

Lemma 4.5. *Suppose* $\mathbf{Z}_1 = [\mathbf{z}_1^{(1)} \cdots \mathbf{z}_N^{(1)}] \in \mathbb{R}^{d' \times N}$ *and* $\mathbf{Z}_2 = [\mathbf{z}_1^{(2)} \cdots \mathbf{z}_N^{(2)}] \in \mathbb{R}^{d \times N}$ *and* $rank(\mathbf{Z}_1) > rank(\mathbf{Z}_2)$. *Denote the singular value of* $\frac{\mathbf{Z}_1}{\sqrt{N}}$ *as* $\sigma_1 \geq \cdots \geq \sigma_N$. *Then* $\min_{\mathbf{H},\eta} \frac{1}{N} \|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F^2 \geq \sum_{j=rank(\mathbf{Z}_2)+2}^{rank(\mathbf{Z}_1)} (\sigma_j)^2$.

The proof can be found in Appendix B.2. The bound given by Lemma 4.5 is not that straightforward to understand. Assuming the features are normalized, we successfully derived the connection of regression error and ratio of ranks in Theorem 4.6.

Theorem 4.6. *Suppose* $\|\mathbf{z}_j^{(1)}\|_2 = 1$, *where* $(1 \leq j \leq N)$. *Then lower bound of approximation error can be upper-bounded as follows:* $\sum_{j=rank(\mathbf{Z}_2)+2}^{rank(\mathbf{Z}_1)} (\sigma_j)^2 \leq \frac{rank(\mathbf{Z}_1) - rank(\mathbf{Z}_2) - 1}{rank(\mathbf{Z}_1)} \leq 1 - \frac{rank(\mathbf{Z}_2)}{rank(\mathbf{Z}_1)}$.

The proof can be found in Appendix B.3. From (Wei et al., 2024; Zhang et al., 2023a), $\exp(\mathrm{H}(\mathbf{G}(\mathbf{Z})))$ is a good approximate of $rank(\mathbf{Z})$. Then we can see that $\frac{rank(\mathbf{Z}_2)}{rank(\mathbf{Z}_1)} \approx \exp(\mathrm{H}(\mathbf{G}(\mathbf{Z}_2)) - \mathrm{H}(\mathbf{G}(\mathbf{Z}_1)))$, making the difference of entropy a good surrogate bound for approximation error.

5. Information Interplay in Supervised Learning

Inspired by matrix information theory and Neural Collapse theory, we focus more on the consistency between sample representations and class classification heads. We determine the relationships among samples by constructing a similarity matrix of the representations of dataset samples.

According to NC1 and NC3, the similarity matrix between samples approximates the similarity matrix of the corresponding class centers, which is also the similarity matrix of the corresponding weights in the fully connected layer. Therefore, under Neural Collapse, the similarity relationship among samples is equivalent to the similarity relationship of the corresponding category weights in the fully connected layer. Our analysis, grounded in matrix information theory, primarily concentrates on the relationship between the representations of samples and the weights in the fully connected layer. Due to constraints in computational resources, we approximate the dataset's matrix entropy using batch matrix entropy.

Our models are trained on CIFAR-10 and CIFAR-100. The default experimental configuration comprise training the models with an SGD optimizer (momentum of 0.9, weight decay of $5e^{-4}$), an initial learning rate of 0.03 with cosine annealing, a batch size of 64, and a total of 2^{20} training iterations. The backbone architecture is WideResNet-28-2 for CIFAR-10 and WideResNet-28-8 for CIFAR-100.

5.1. Information interplay during standard supervised learning process

According to Neural Collapse, during the terminal stages of training, sample features align with the weights of the fully connected layer. Theorem 4.2 indicates that during the training process, MIR increases to its theoretical upper limit, while HDR decreases to 0. We plot the model's accuracy on the test set during the training process, as well as the MIR and the HDR between data representations and the corresponding classification heads. As shown in Figure 1, on CIFAR-10 and CIFAR-100, the accuracy and MIR exhibit almost identical trends of variations. In most cases, both accuracy and MIR increase or decrease simultaneously, and MIR consistently shows an upward trend, having its trajectory toward its theoretical maximum value. As shown in Figure 2, during the training process, in most instances, accuracy and HDR show opposite trends, with HDR continually decreasing and even nearing its theoretical minimum
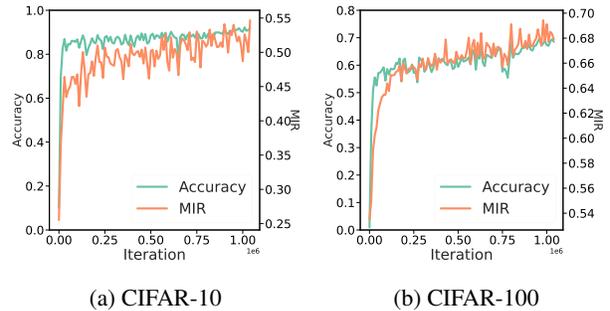
Figure 1. Accuracy and MIR on the test set during training.

value of 0 on CIFAR-100. In summary, MIR and HDR effectively describe the process of training towards Neural Collapse.

(a) CIFAR-10

(b) CIFAR-100
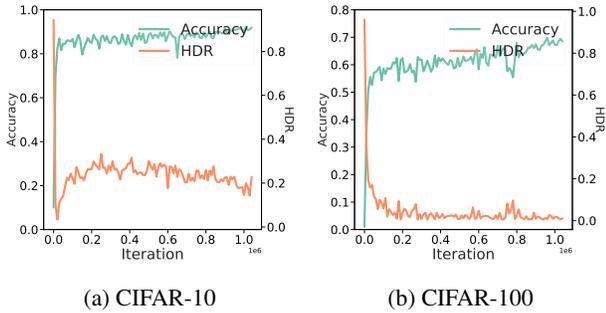
Figure 2. Accuracy and HDR on the test set during training.

5.2. Information interplay in linear mode connectivity

Linear mode connectivity (Frankle et al., 2020) suggests that under specific datasets and experimental setups, models initialized with the same random parameters will be optimized near the same local optimal basin, even if the order of training data and data augmentation differs. We investigate the behaviors of MIR and HDR under the setting of linear mode connectivity. We initialize models with the same random parameters and train them using different data sequences and random augmentations. Subsequently, we linearly interpolate these two checkpoints and obtain a new model $h = (1 - \omega) \cdot h_1 + \omega \cdot h_2$, where h_1 and h_2 are the two ckeckpoints and ω is the interpolation weight. Then we test these models on a test set for accuracy, MIR, and HDR.

We conduct experiments on CIFAR-10 and CIFAR-100. As shown in Figure 3a and 3b. On CIFAR-100, the performance of models obtained along the interpolation line is close, aligning with the linear mode connectivity. At this point, MIR and HDR remain almost unchanged. However, on CIFAR-10, the models do not exhibit linear mode connectivity. When the value of interpolation weight is between

(a) CIFAR-10
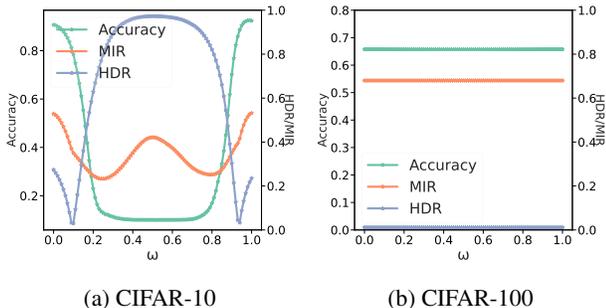
(b) CIFAR-100

Figure 3. Accuracy, MIR, and HDR of models interpolated with different weights on the test set.

(a) lr: $3e^{-2}$

(b) lr: $3e^{-3}$

(c) lr: $3e^{-4}$
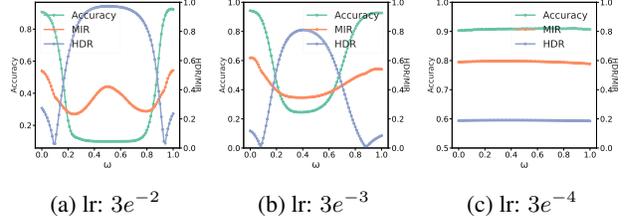
Figure 4. Accuracy, MIR, and HDR of models interpolated with different weights on CIFAR-10 test set.

0.4 and 0.6, the performance of the interpolated models even drop to that of random guessing. Surprisingly, at this time, MIR shows an additional upward trend. Moreover, when the value of interpolation weight is close to 0 and 1, despite a slight decrease in performance, HDR also decreases. Although we find it difficult to explain this anomaly, it does demonstrate that HDR and MIR have distinctive attributes compared to the accuracy metric, presenting an intriguing avenue for further exploration.

Altıntaş et al. (2023) point that linear mode connectivity is related to the experimental configuration. Therefore, we posit that the performance decline of the interpolated model on CIFAR-10 is associated with an excessively high learning rate. In the training phase, models navigate the loss landscapes in search of minimal values, and two models with linear mode connectivity are optimized near the same local optimum. When the learning rate is too high, different training sample ordering and data augmentations lead to directing model optimization towards distinct regions within the loss landscapes. We experiment with different learning rates on CIFAR-10 and test their linear mode connectivity. It is observed that as the learning rate decreased, fluctuations in accuracy, MIR, and HDR also reduced. When the learning rate is lowered to $3e^{-4}$, the model demonstrate linear mode connectivity on CIFAR-10. This suggests that HDR and MIR are also effective in describing linear mode connectivity when it exists.

5.3. Information interplay in label smoothing

Label smoothing (Szegedy et al., 2016) is a widely used technique in deep learning. It improves the generalization of the model by setting smoothness of labels. $y' = (1 - \epsilon) \cdot y + \frac{\epsilon}{C}$, where ϵ is the smoothness and y is the one-hot label. We train models with various smoothness levels to explore their impact on accuracy, HDR, and MIR. As shown in Figure 5, the variation in accuracy, MIR, and HDR are minimal, indicating that HDR and MIR can effectively describe the performance of label smoothing technique.

5.4. Information interplay in model pruning

We would like to use MIR and HDR to understand why the pruning technique is effective in maintaining relatively high accuracy. We apply standard unstructured pruning to models: for a well-trained model, we determine the number of parameters to prune, denoted as k. The model's parameters are sorted in descending order based on their absolute values, and the smallest k parameters are removed. Subsequently, the remaining parameters are fine-tuned again on the dataset (Han et al., 2015).

We prune the model under various sparsity levels and extract features using the model before and after pruning. We calculate the MIR and HDR of the features extracted by the model before and after pruning. As shown in Figure 6, even when the model sparsity is 90%, the features extracted by the pruned model maintain a high MIR with those extracted before pruning. At various sparsity ratio, the variations in MIR and HDR for models after pruning are little compared to these before pruning, the performance differences of the models before and after pruning are also not significant. This indicates that fine-tuning the pruned subnetwork can restore adequate information extraction capabilities.

5.5. Information interplay in grokking

In supervised learning, training models on certain datasets can result in an anomalous situation. Initially, models quickly learn the patterns of the training set, but at this point, their performance on the test set is very poor. As training continues, the models learn representations that can generalize to the test set, a phenomenon referred to as Grokking(Nanda et al., 2022). We aim to explore the information interplay in Grokking. Following Nanda et al. (2022); Tan & Huang (2023), we train a transformer to learn modular addition $c \equiv (a + b) \pmod{p}$, with p being 113. The model input is "$a\ b\ =$", where a and b are encoded into p-dimensional one-hot vectors, and "$=$" is used to signify the output value c. Our model employ a single-layer ReLU transformer with a token encoding dimension of 128

(a) CIFAR-10   (b) CIFAR-100
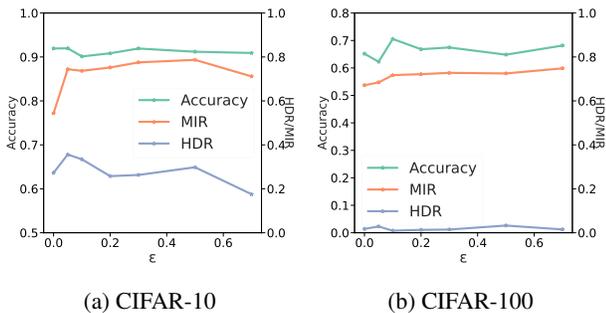
Figure 6. Accuracy, MIR, and HDR of the features extracted by the model before and after pruning.

to learn positional encodings, four attention heads each of dimension 32, and an MLP with a hidden layer of dimension 512. We train the model using full-batch gradient descent, a learning rate of 0.001, and an AdamW optimizer with a weight decay parameter of 1. We use 30% of all possible inputs (113×113 pairs) for training data and test performance on the remaining 70%.

As shown in Figure 7, we plot the accuracy of both the training and test sets during the grokking process, as well as the variation in MIR and HDR between the representation and the fully connected layer. It can be observed that, in the early stages of training, the model quickly fits the training data and achieves 100% accuracy on the training set. However, at this point, the performance on the test set is nearly equivalent to that of random guessing. As training continues, the model gradually shows generalization capability on the test set, ultimately achieving 100% accuracy, which is a hallmark of grokking. Figure 7 also reveals a clear two-phase variation in both MIR and HDR between data representation and the weight of fully connected layer. Initially, similar to fully supervised learning, MIR increases, while HDR decreases. However, as training proceeds, MIR begins to decrease, and HDR starts to increase, indicating the model is seeking new optimal points. After the model achieves the grokking, MIR reaches its lowest, and HDR rapidly declines from its highest point. The experiments demonstrate that HDR and MIR exhibit distinct phenomena in two stages, suggesting that information metrics can

(a) CIFAR-10   (b) CIFAR-100

Figure 5. Accuracy, MIR, and HDR under different smoothness levels.
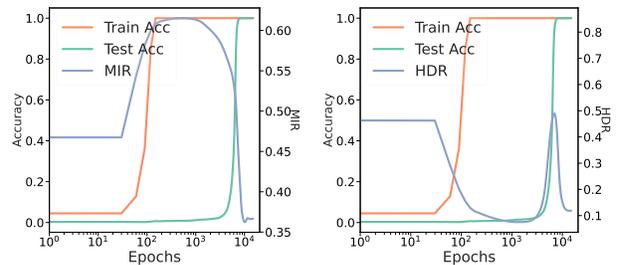
Figure 7. Accuracy, MIR, and HDR during Grokking.

6

describe the grokking phenomenon, providing a basis for further research.

6. Improving Supervised and Semi-Supervised Learning With Information Interplay

6.1. Pipeline of supervised and semi-supervised learning

In this section, we introduce how to apply matrix information entropy in supervised and semi-supervised learning. In supervised learning, we train the nerual network h and classifier $\mathbf{W} \in \mathbb{R}^{C \times d}$ on the dataset $\mathcal{D}_L = \{(x_i, \ y_i)\}_{i=0}^{N_L}$ consisting of N_L samples. h is used to extract data features $f \in \mathbb{R}^D$, and $\mathbf{W}$ classifies the extracted features. The model is optimized using the following cross-entropy loss.

$$\mathcal{L}_s = \frac{1}{B} \sum_{i=1}^{B} \mathcal{H}(y_i, p(\omega(x_i))),$$

where B represents the batch size, $\mathcal{H}$ denotes the cross-entropy loss, $p(\cdot)$ refers to the model's output probability of a sample, and ω means random data augmentation.

Compared to supervised learning, semi-supervised learning includes an additional unlabeled dataset $\mathcal{D}_U = \{u_i\}_{i=0}^{N_U}$ which contain N_U unlabeled data and utilizes it to assist in optimizing the model. In the processing of unlabeled data, we adopt the approach outlined in Freematch (Wang et al., 2023). This involves generating pseudo-labels through weak data augmentation and selecting data based on a probability threshold. The model is then employed to extract features from strongly augmented data for the computation of cross-entropy loss in conjunction with the pseudo-labels. The formulaic representation of the training objective for unlabeled data is as follows:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{i=1}^{\mu B} \mathbb{I}\left(max(q_i) > \tau\right) \cdot \mathcal{H}\left(\hat{q}_i, Q_i\right),$$

where q_i and Q_i correspond to $p(y|\omega(u_i))$ and $p(y|\Omega(u_i))$, respectively. The term $\hat{q}_i$ refers to one-hot pseudo-labels generated from q_i. The symbol $\mathbb{I}(\cdot > \tau)$ denotes the indicator function applied to values surpassing the threshold τ. Furthermore, ω and Ω are used to distinguish between weak and strong data augmentation.

In addition, Freematch incorporates a fairness objective to predict each class with uniform frequency.

$$\mathcal{L}_f = -H\left(\text{SumNorm}\left(\frac{p_1}{hist_1}\right), \text{SumNorm}\left(\frac{p_2}{hist_2}\right)\right).$$

SumNorm $= (\cdot)/\sum(\cdot)$. p_1 and p_2 refer to the average predictions of the model under weak and strong augmenta-

tion, respectively. Likewise, $hist_1$ and $hist_2$ indicate the histogram distributions resulting from weak and strong augmentation, respectively.

The overall objective is

$$\mathcal{L}_{ssl} = \mathcal{L}_s + \lambda_u \mathcal{L}_u + \lambda_f \mathcal{L}_f,$$

where λ_u and λ_f represent the weight for $\mathcal{L}_u$ and $\mathcal{L}_f$.

6.2. Insights from information interplay

In a batch of labeled data $\{(x_i, y_i)\}_{i=1}^{B} \in \mathcal{D}_L$, h extracts feature representations, denoted as $f \in \mathbb{R}^{B \times D}$. In Neural Collapse theory, the representation of each sample's class center aligns with the classifier weight of the respective category, i.e., $V_i = W_{y_i}$. In the case of unlabeled data $\{u_i\}_{i=1}^{\mu B} \in \mathcal{D}_U$, we select sample features f' from μB samples with pseudo-label probabilities greater than τ. i.e., $f' = \{f_i \in f | \mathbb{I}\left(max(q_j) > \tau\right)\}$, and obtain the corresponding class centers $V' = W_{y_i'}$, where y_i' is the pseudo label of f'.

Maximizing mutual information. As shown in Figure 1, during the model training process, the mutual information between a batch's data features f and the corresponding class weights V increases. Therefore, we add an additional loss term to increase the mutual information between them. For supervised learning, the final optimization objective is

$$\mathcal{L} = \mathcal{L}_s - \lambda_{mi} \cdot \text{MI}\left(\mathbf{G}(f), \mathbf{G}(V)\right).$$

For semi-supervised learning, the final optimization objective is

$$\mathcal{L} = \mathcal{L}_{ssl} - \lambda_{mi} \cdot \text{MI}\left(\mathbf{G}(f'), \mathbf{G}(V')\right),$$

where λ_{mi} is the weight for the mutual information.

Minimizing entropy difference. As depicted in Figure 2, throughout the training phase, the disparity in information entropy between a batch's data features f and the associated category weights V diminishes in tandem with an increase in accuracy. Consequently, it is feasible to introduce an auxiliary loss component within the training regime to further mitigate this entropy discrepancy. In the context of supervised learning, the ultimate optimization target is delineated as

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

Regarding semi-supervised learning, this target shifts to

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

wherein λ_{id} signifies the weight for entropy difference.

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)	$79.18_{\pm1.11}$	$74.34_{\pm1.76}$	$46.24_{\pm1.29}$	$86.96_{\pm0.80}$	$58.80_{\pm0.66}$	$74.31_{\pm0.85}$	$32.78_{\pm0.40}$
Pseudo Label (Lee et al., 2013)	$80.21_{\pm0.55}$	$74.61_{\pm0.26}$	$46.49_{\pm2.20}$	$87.45_{\pm0.85}$	$57.74_{\pm0.28}$	$74.68_{\pm0.99}$	$32.64_{\pm0.71}$
VAT (Miyato et al., 2018)	$79.81_{\pm1.17}$	$74.66_{\pm2.12}$	$41.03_{\pm1.79}$	$85.20_{\pm1.40}$	$48.84_{\pm0.79}$	$74.74_{\pm0.38}$	$37.95_{\pm1.12}$
MeanTeacher (Tarvainen & Valpola, 2017)	$76.37_{\pm0.44}$	$70.09_{\pm1.60}$	$37.46_{\pm3.30}$	$81.11_{\pm1.44}$	$45.17_{\pm1.06}$	$71.72_{\pm1.45}$	$33.90_{\pm1.37}$
MixMatch (Berthelot et al., 2019b)	$65.76_{\pm7.06}$	$36.19_{\pm6.48}$	$13.63_{\pm0.59}$	$67.59_{\pm0.66}$	$39.76_{\pm0.48}$	$54.93_{\pm0.96}$	$21.70_{\pm0.68}$
ReMixMatch (Berthelot et al., 2019a)	$20.77_{\pm7.48}$	$9.88_{\pm1.03}$	$6.30_{\pm0.05}$	$42.75_{\pm1.05}$	$26.03_{\pm0.35}$	$32.12_{\pm6.24}$	$6.74_{\pm0.17}$
UDA (Xie et al., 2020)	$34.53_{\pm10.69}$	$10.62_{\pm3.75}$	$5.16_{\pm0.06}$	$46.39_{\pm1.59}$	$27.73_{\pm0.21}$	$37.42_{\pm8.44}$	$6.64_{\pm0.17}$
FixMatch (Sohn et al., 2020)	$24.79_{\pm7.65}$	$7.47_{\pm0.28}$	$5.07_{\pm0.05}$	$46.42_{\pm0.82}$	$28.03_{\pm0.16}$	$35.97_{\pm4.14}$	$6.25_{\pm0.33}$
Dash (Xu et al., 2021)	$27.28_{\pm14.09}$	$8.93_{\pm3.11}$	$5.16_{\pm0.23}$	$44.82_{\pm0.96}$	$27.15_{\pm0.22}$	$34.52_{\pm4.30}$	$6.39_{\pm0.56}$
MPL (Pham et al., 2021)	$23.55_{\pm6.01}$	$6.93_{\pm0.17}$	$5.76_{\pm0.24}$	$46.26_{\pm1.84}$	$27.71_{\pm0.19}$	$35.76_{\pm4.83}$	$6.66_{\pm0.00}$
FlexMatch (Zhang et al., 2021)	$13.85_{\pm12.04}$	$4.97_{\pm0.06}$	$4.98_{\pm0.09}$	$39.94_{\pm1.62}$	$26.49_{\pm0.20}$	$29.15_{\pm4.16}$	$5.77_{\pm0.18}$
FreeMatch (Wang et al., 2023)	$8.07_{\pm4.24}$	$4.90_{\pm0.04}$	$4.88_{\pm0.18}$	$37.98_{\pm0.42}$	$26.47_{\pm0.20}$	$15.56_{\pm0.55}$	$5.63_{\pm0.15}$
OTMatch (Tan et al., 2023c)	$4.89_{\pm0.76}$	$4.72_{\pm0.08}$	$4.60_{\pm0.15}$	$37.29_{\pm0.76}$	$26.04_{\pm0.21}$	$\mathbf{12.10_{\pm0.72}}$	$5.60_{\pm0.14}$
SoftMatch (Chen et al., 2023)	$4.91_{\pm0.12}$	$4.82_{\pm0.09}$	$\mathbf{4.04_{\pm0.02}}$	$37.10_{\pm0.07}$	$26.66_{\pm0.25}$	$21.42_{\pm3.48}$	$5.73_{\pm0.24}$
FreeMatch + Maximizing Mutual Information (Ours)	$\underline{4.87_{\pm0.66}}$	$\underline{4.66_{\pm0.13}}$	$\underline{4.56_{\pm0.15}}$	$\mathbf{36.41_{\pm1.91}}$	$\mathbf{25.77_{\pm0.35}}$	$16.61_{\pm1.19}$	$\mathbf{5.24_{\pm0.17}}$
FreeMatch + Minimizing Entropy Difference (Ours)	$\mathbf{4.69_{\pm0.16}}$	$\mathbf{4.63_{\pm0.25}}$	$4.60_{\pm0.15}$	$\underline{37.31_{\pm1.96}}$	$\underline{25.79_{\pm0.41}}$	$\underline{14.93_{\pm3.28}}$	$\underline{5.30_{\pm0.18}}$

6.3. Performances on supervised and semi-supervised learning

In our effort to conduct a fair comparison between our proposed method and existing methodologies, we meticulously designed our experiments building upon previous scholarly work. TorchSSL (Zhang et al., 2021), a sophisticated code-base encompassing a wide array of semi-supervised learning techniques as well as supervised learning implementations, was employed as our foundational code base. This enables us to implement our algorithm effectively and assess its performance on well-established datasets like CIFAR-10, CIFAR-100, and STL-10. In the realm of supervised learning, our unique loss components are applied to annotated data, facilitating the computation of both mutual information loss and information entropy difference loss. For semi-supervised learning scenarios, these loss components are extended to unlabeled data, enhancing the calculation of these loss metrics during the unsupervised learning phase. We use an SGD optimizer, configured with a momentum of 0.9 and a weight decay parameter of $5e^{-4}$. The learning rate was initially set at 0.03, subject to cosine annealing. We report the performance metrics over several runs of seeds. The batch size are maintained at 64 across a comprehensive 1,048,000 iterations training regimen. Concerning model architecture, WideResNet-28-2, WideResNet-28-8, and WideResNet-37-2 are respectively chosen for datasets CIFAR-10, CIFAR-100, and STL-10,

We train supervised and semi-supervised learning models using mutual information and information entropy difference as constraints in the loss function. Table 1 and 2 present the performance of semi-supervised and supervised learning, respectively. It is observed that applying mutual information and information entropy constraints led to a slight improve-

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**

ment in supervised learning performance. We believe this is because sufficient labeled data provides adequate information constraints, leading to only a modest enhancement in performance. However, in semi-supervised learning, in most settings, maximizing mutual information and minimizing information entropy resulted in the best or second-best performance. Additionally, our method consistently outperformed our baseline, FreeMatch, across various settings. This suggests that in situations with insufficient labeled samples, additional information constraints can more effectively improve model performance.

7. Conclusion

In conclusion, we have made significant advancements in understanding the dynamics of supervised learning by utilizing matrix information and Neural Collapse theory. Our introduction of matrix mutual information ratio (MIR) and matrix entropy difference ratio (HDR) provide novel insights into the interplay between data representations and classification head vectors, serving as new tools to understand the dynamics of neural networks.

Through a series of rigorous theoretical and empirical analyses, we demonstrate the effectiveness of MIR and HDR in elucidating various neural network phenomena, such as

grokking, and their utility in improving training dynamics. The incorporation of these metrics as loss functions in supervised and semi-supervised learning shows promising results, indicating their potential to enhance model performance and training efficiency. This study not only contributes to the field of machine learning by offering new analytical tools but also applies matrix information to optimize supervised learning algorithms.

Acknowledgment

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Altıntaş, G. S., Bachmann, G., Noci, L., and Hofmann, T. Disentangling linear mode connectivity. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023. 2, 5

Bach, F. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2022. 2

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a. 8

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b. 2, 8

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4960–4964. IEEE, 2016. 1

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 2023. 2, 8

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pp. 3259–3269. PMLR, 2020. 1, 2, 5

Garrido, Q., Balestriero, R., Najman, L., and Lecun, Y. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International Conference on Machine Learning*, pp. 10929–10974. PMLR, 2023. 12

Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015. 1

Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015. 6

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under mse loss: Proximity to and dynamics on the central path. In *ICLR*, 2021. 2

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 1

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012. 1

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012. 1

Lee, D.-H. et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896, 2013. 8

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014. 1

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 8

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 6

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. 1, 2, 3

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021. 8

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. 1, 2

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28:3546–3554, 2015. 8

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015. 1

Skean, O., Osorio, J. K. H., Brockmeier, A. J., and Giraldo, L. G. S. Dime: Maximizing mutual information by a difference of matrix-based entropies. *arXiv preprint arXiv:2301.08164*, 2023. 3

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 8

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016. 5

Tan, Z. and Huang, W. Understanding grokking through a robustness viewpoint. *arXiv preprint arXiv:2311.06597*, 2023. 6

Tan, Z., Wang, Z., and Zhang, Y. Seal: Simultaneous label hierarchy exploration and learning. *arXiv preprint arXiv:2304.13374*, 2023a. 2

Tan, Z., Yang, J., Huang, W., Yuan, Y., and Zhang, Y. Information flow in self-supervised learning. *arXiv e-prints*, pp. arXiv–2309, 2023b. 2

Tan, Z., Zheng, K., and Huang, W. Otmatch: Improving semi-supervised learning with optimal transport. *arXiv preprint arXiv:2310.17455*, 2023c. 2, 8

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

Wang, X., Al-Bashabsheh, A., Zhao, C., and Chan, C. Adaptive label smoothing for classifier-based mutual information neural estimation. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1035–1040. IEEE, 2021. 2

Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., et al. Freematch: Self-adaptive thresholding for semi-supervised learning. In *Eleventh International Conference on Learning Representations*, 2023. 2, 7, 8

Wei, L., Tan, Z., Li, C., Wang, J., and Huang, W. Large language model evaluation via matrix entropy. *arXiv preprint arXiv:2401.17139*, 2024. 4

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020. 8

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., and Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021. 8

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2, 8

Zhang, Y., Tan, Z., Yang, J., Huang, W., and Yuan, Y. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023a. 2, 4

Zhang, Y., Yang, J., Tan, Z., and Yuan, Y. Relationmatch: Matching in-batch relationships for semi-supervised learning. *arXiv preprint arXiv:2305.10397*, 2023b. 2

Zhou, J., You, C., Li, X., Liu, K., Liu, S., Qu, Q., and Zhu, Z. Are all losses created equal: A neural collapse perspective. *NeurIPS*, 35:31697–31710, 2022. 2

Appendix

A. Detailed proofs for Neural Collapse related Theorems

Theorem A.1. *Suppose Neural collapse happens. Then* $\mathrm{HDR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = 0$ *and* $\mathrm{MIR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)}$.

Proof. By (NC 3), we know that $\mathbf{W}^T = \frac{\|\mathbf{W}\|_F}{\|\mathbf{M}\|_F}\mathbf{M}$. Noticing that $\frac{\|\mathbf{W}\|_F}{\|\mathbf{M}\|_F} > 0$, we know that $\frac{w_i}{\|w_i\|} = \frac{\tilde{\mu}_i}{\|\tilde{\mu}_i\|}$. It is then very clear that $\mathbf{G}(\mathbf{W}^T) = \mathbf{G}(\mathbf{M})$. Therefore from definition 4.1 and 3.4, it is clear that $\mathrm{HDR}(\mathbf{G}(\mathbf{W}^T), \mathbf{G}(\mathbf{M})) = 0$.

Define $\mathcal{E}(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix}$. From (NC 2), we know that $\mathbf{G}(\mathbf{W}^T) = \mathbf{G}(\mathbf{M}) = \mathcal{E}(\frac{-1}{C-1})$ and $\mathbf{G}(\mathbf{W}^T) \odot \mathbf{G}(\mathbf{M}) =$ $\mathcal{E}(\frac{1}{(C-1)^2})$. Notice that $\mathcal{E}(\alpha) = (1-\alpha)\mathbf{I}_C + \alpha\mathbf{1}_C^T\mathbf{1}_C$, we can obtain its spectrum as $1 - \alpha$ ($C - 1$ times) and $1 + (C-1)\alpha$ (1 time). Therefore, we can obtain that $\mathrm{H}(\mathbf{G}(\mathbf{W}^T)) = \mathrm{H}(\mathbf{G}(\mathbf{M})) = \log(C-1)$. And $\mathrm{H}(\mathbf{G}(\mathbf{W}^T) \odot \mathbf{G}(\mathbf{M})) = -\frac{1}{C-1}\log\frac{1}{C-1} - (C-1)\frac{C-2}{(C-1)^2}\log\frac{C-2}{(C-1)^2} = \frac{1}{C-1}\log(C-1) - \frac{C-2}{C-1}\log(C-2) + \frac{2(C-2)}{C-1}\log(C-1) = (2 - \frac{1}{C-1})\log(C-1) - \frac{C-2}{C-1}\log(C-2)$. Then then conclusion follows from definition 3.3.

$\square$

As the linear weight matrix $\mathbf{W}$ can be seen as (prototype) embedding for each class. It is natural to consider the mutual information and entropy difference between sample embedding and label embedding. We discuss this in the following theorem 4.3.

Corollary A.2. *Suppose the dataset is class-balanced,* $\mu_G = 0$ *and Neural collapse happens. Denote* $\mathbf{Z}_1 = [h(\mathbf{x}_1)\cdots h(\mathbf{x}_n)] \in \mathbb{R}^{d \times n}$ *and* $\mathbf{Z}_2 = [w_{y_1}\cdots w_{y_n}] \in \mathbb{R}^{d \times n}$. *Then* $\mathrm{HDR}(\mathbf{Z}_1, \mathbf{Z}_2) = 0$ *and* $\mathrm{MIR}(\mathbf{Z}_1, \mathbf{Z}_2) = \frac{1}{C-1} + \frac{(C-2)\log(C-2)}{(C-1)\log(C-1)}$.

Proof. Denote $n_1 = \frac{n}{C}$ the number of samples in each class. Without loss of generality, assume samples are arranged as C consecutive groups, each group has samples from the same class.

Define $\mathcal{E}(\alpha) = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix}$ and $\mathbf{1}_{n_1,n_1}$ a $n_1 \times n_1$ matrix with all its element 1.

From (NC 1), we know that $\mathbf{G}(\mathbf{W}^T) = \mathbf{G}(\mathbf{M}) = \mathcal{E}(\frac{-1}{C-1}) \otimes \mathbf{1}_{n_1,n_1}$ and $\mathbf{G}(\mathbf{W}^T) \odot \mathbf{G}(\mathbf{M}) = \mathcal{E}(\frac{1}{(C-1)^2}) \otimes \mathbf{1}_{n_1,n_1}$, where $\otimes$ is the Kronecker product. By the property of the Kronecker product spectrum, the non-zero spectrum of $\mathbf{G}(\mathbf{W}^T)$ will be n_1 times the spectrum of $\mathcal{E}(\frac{-1}{C-1})$. Then this corollary follows by the same proof of the theorem 4.2. $\square$

B. Some theoretical guarantees for HDR

Mutual information is a very intuitive quantity in information theory. On the other hand, it seems weird to consider the difference of entropy, but we will show that this quantity is closely linked with comparing the approximation ability of different representations on the same target.

For ease of theoretical analysis, in this section, we consider the MSE regression loss.

The following lemma 4.4 shows that the regression of two sets of representations $\mathbf{Z}_1$ and $\mathbf{Z}_2$ to the same target $\mathbf{Y}$ are closely related. And the two approximation errors are closely related to the regression error of $\mathbf{Z}_1$ to $\mathbf{Z}_2$.

Lemma B.1. *Suppose* $\mathbf{W}_1^*, \mathbf{b}_1^* = \arg\min_{\mathbf{W},\mathbf{b}}\|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_1 + \mathbf{b}\mathbf{1}_N)\|_F$. *Then* $\min_{\mathbf{W},\mathbf{b}}\|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_2 + \mathbf{b}\mathbf{1}_N)\|_F \leq \min_{\mathbf{W},\mathbf{b}}\|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_1 + \mathbf{b}\mathbf{1}_N)\|_F + \|\mathbf{W}_1^*\|_F \min_{\mathbf{H},\eta}\|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F$.

Proof. Suppose $\mathbf{H}^*, \eta^* = \arg\min_{\mathbf{H},\eta} \|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F$. Then $\min_{\mathbf{W},\mathbf{b}} \|\mathbf{Y} - (\mathbf{W}\mathbf{Z}_2 + \mathbf{b}\mathbf{1}_N)\|_F \leq \|\mathbf{Y} - (\mathbf{W}_1^*\mathbf{H}^*\mathbf{Z}_2 + (\mathbf{b}_1^* + \mathbf{W}_1^*\eta^*)\mathbf{1}_N)\|_F \leq \|\mathbf{Y} - (\mathbf{W}_1^*\mathbf{Z}_1 + \mathbf{b}_1^*\mathbf{1}_N)\|_F + \|\mathbf{W}_1^*(\mathbf{Z}_1 - (\mathbf{H}^*\mathbf{Z}_2 + \eta^*\mathbf{1}_N))\|_F \leq \|\mathbf{Y} - (\mathbf{W}_1^*\mathbf{H}^*\mathbf{Z}_2 + (\mathbf{b}_1^* + \mathbf{W}_1^*\eta^*)\mathbf{1}_N)\|_F \leq \|\mathbf{Y} - (\mathbf{W}_1^*\mathbf{Z}_1 + \mathbf{b}_1^*\mathbf{1}_N)\|_F + \|\mathbf{W}_1^*\|_F\|\mathbf{Z}_1 - (\mathbf{H}^*\mathbf{Z}_2 + \eta^*\mathbf{1}_N)\|_F.$ $\square$

From lemma 4.4, we know that the regression error of $\mathbf{Z}_1$ to $\mathbf{Z}_2$ is crucial for understanding the differences of representations. We further bound the regression error with rank and singular values in the following lemma 4.5.

Lemma B.2. *Suppose* $\mathbf{Z}_1 = [\mathbf{z}_1^{(1)} \cdots \mathbf{z}_N^{(1)}] \in \mathbb{R}^{d' \times N}$ *and* $\mathbf{Z}_2 = [\mathbf{z}_1^{(2)} \cdots \mathbf{z}_N^{(2)}] \in \mathbb{R}^{d \times N}$ *and* $rank(\mathbf{Z}_1) > rank(\mathbf{Z}_2)$. *Denote the singular value of* $\frac{\mathbf{Z}_1}{\sqrt{N}}$ *as* $\sigma_1 \geq \cdots \geq \sigma_N$. *Then* $\min_{\mathbf{H},\eta} \frac{1}{N}\|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F^2 \geq \sum_{j=rank(\mathbf{Z}_2)+2}^{rank(\mathbf{Z}_1)} (\sigma_j)^2$.

Proof. The proof idea is similar to (Garrido et al., 2023). Suppose $\mathbf{H}^*, \eta^* = \arg\min_{\mathbf{H},\eta} \frac{1}{N}\|\mathbf{Z}_1 - (\mathbf{H}\mathbf{Z}_2 + \eta\mathbf{1}_N)\|_F^2$ and $r = rank(\mathbf{H}^*\mathbf{Z}_2 + \eta^*\mathbf{1}_N)$.

Then from Eckart–Young–Mirsky theorem $\frac{1}{N}\|\mathbf{Z}_1 - (\mathbf{H}^*\mathbf{Z}_2 + \eta^*\mathbf{1}_N)\|_F^2 \geq \sum_{j=r+1}^{N} (\sigma_j^{(1)})^2$. And note $r \leq rank(\mathbf{Z}_2) + 1$ and singular value index bigger than rank is 0. The conclusion follows. $\square$

The bound given by lemma 4.5 is not that straightforward to understand. Assuming the features are normalized, we successfully derived the connection of regression error and ratio of ranks in theorem 4.6.

Theorem B.3. *Suppose* $\|\mathbf{z}_j^{(1)}\|_2 = 1$, *where* $(1 \leq j \leq N)$. *Then lower bound of approximation error can be upper-bounded as follows:* $\sum_{j=rank(\mathbf{Z}_2)+2}^{rank(\mathbf{Z}_1)} (\sigma_j)^2 \leq \frac{rank(\mathbf{Z}_1) - rank(\mathbf{Z}_2) - 1}{rank(\mathbf{Z}_1)} \leq 1 - \frac{rank(\mathbf{Z}_2)}{rank(\mathbf{Z}_1)}$.

Proof. The proof is direct by noticing the summation of the square of singular values is 1 and we have already ranked singular values by their indexes. $\square$