Information bottleneck predicts neural collapse leads to generalization in supervised contrastive learning

Anonymous Authors[1]

Abstract

Neural collapse describes the geometry of activation in the last layer of a deep neural network that has been trained beyond zero error. Open questions are whether neural collapse helps generalization, and if so how training beyond zero error improves test performance. We propose that neural collapse may support good generalization performance by compressing the representation of the target categories. This compression may improve test performance by reducing the overlap between outputs. To investigate whether such a compact representation exists and quantify how this could aid generalization, we modeled neural collapse as an information bottleneck (IB) problem. We hypothesize that neural collapse corresponds, generally, to a special case of IB, deterministic IB (DIB). Recent research has shown that two deep neural networks independently trained with the same contrastive loss objective are linearly identifiable, meaning that the resulting representations are equivalent up to a matrix transformation. This linear identifiability allows us to approximate the DIB objective by modeling the dependency between two learned representations as a Gaussian distribution, and quantify compression and neural collapse. We show that supervised contrastive learning with a ResNet50 backbone improves test accuracy while compressing classification information into a K-dimensional Gaussian distribution (e.g., K=10 for CIFAR10). We observe a similar compression when we use ResNet50 pretrained with the full ImageNet32 to represent its subsets with fewer classes, and to perform zero-shot transfer learning on CIFAR10. Discovering a mathematical explanation for neural collapse can help build better training algorithms.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Deep neural networks trained for classification exhibit an intriguing geometry, "neural collapse" (NC). As a deep neural network is trained past the point where the training error falls to zero, the class clusters in the learned representation can collapse to their means. In this case, classification reduces to simply finding the closest class clusters in the representation space. As a result, the class means and linear classifiers form what is called a K-simplex equiangular tight frame (ETF) (Strohmer & Heath, 2003) (K is the number of classes). This phenomenon was first observed in deep neural networks trained with both cross entropy (Papyan et al., 2020) and mean-square-loss (Han et al., 2022). Recent work has shown that contrastive-loss-trained models, especially supervised contrastive learning (Khosla et al., 2020), demonstrate neural collapse similar to cross-entropy-trained models (Fang et al., 2021).

However, whether neural collapse leads to good generalization in all cases is unknown. Although Papyan et al. (2020) showed that training beyond zero error improves test accuracy, recent work (Hui et al., 2022) has demonstrated that neural collapse may not always occur on the test set. This work also shows that more collapse may impair transfer learning performance. Here, we argue that neural collapse corresponds to finding a compact representation of the classification labels Y for test samples, and therefore neural collapse helps generalization. To show this, we use the information bottleneck (IB) method to examine and quantify the representations that exhibit neural collapse. IB is a theoretical framework that explicitly trades off input compression with the retention of relevant information. In this case, that relevant information is the label used for classification. In this work, we connect neural collapse with a specific form of IB, the deterministic information bottleneck (DIB) problem, that lends itself to rigorous analytic treatment.

Because the information bottleneck (IB) method and its variations are generally analytically intractable, previous work used variational approximations of IB (VIB Alemi et al. (2016)) to explore how and what deep neural networks encode. Contrastive-loss-trained deep neural networks may be a special case where more precise IB results can be in-

structive. Recently, advances in nonlinear independent component analysis (ICA) (Hyvarinen et al., 2018; Hyvarinen & Morioka, 2016) suggested that contrastive loss-trained networks may have an additional desirable property: they approximate a global optimum in the representation space. This means that networks trained with this kind of loss may be especially interpretable as they all obtain the same optimal solution. Specifically, Roeder et al. (2020) crystallized this notion and demonstrated that models of the same architecture trained with contrastive loss are linearly identifiable, i.e., their learned representations are equivalent only up to a trivial matrix transformation. This linear identifiability allows us to approximate the optimal solution corresponding to our DIB problem that may characterize neural collapse in supervised contrastive learning. Both DIB and IB have connections to clustering. DIB corresponds to hard clustering whereas IB corresponds to soft clustering. In this paper, we hypothesize that DIB corresponds to neural collapse because of its connection to hard clustering (see discussion in Sections 2.2.2 and 2.3 for more details).

Our contributions: We focus on neural collapse in supervised contrastive learning (Khosla et al., 2020) to use tools from information theory to quantify compression in the output representation. Models trained with contrastive loss do not explicitly optimize classification (Chen et al., 2020a;b; Khosla et al., 2020). Instead, they minimize a supervised version (Khosla et al., 2020) of the information noise contrastive estimation (InfoNCE) loss (van den Oord et al., 2018) in order to maximize the information between the learned representation and the input. Using information theoretic tools to study neural collapse in contrastive learning can leverage this training feature to quantify compression and connect compression directly to generalization. By characterizing neural collapse in supervised contrastive learning as an information bottleneck problem, we discover that:

- Theoretical contribution I: We frame neural collapse as an information bottleneck problem. We show that the optimal solution of a deterministic information bottleneck (DIB) also corresponds to variability collapse (NC1 introduced in (Papyan et al., 2020)).
- Theoretical contribution II: DIB is generally analytically and computationally intractable. However, we show that we can leverage the linear identifiability of contrasive learning to approximate the DIB optimal solutions in supervised contrastive learning. We show that this approximation is DIB-optimal after the emergence of neural collapse.
- Empirical contribution I: We show that as test accuracy improves, the learned representation compresses more classification-relevant information into a K-dimensional representation that is an optimal to fit a K-simplex ETF.
- Empirical contribution II: After the emergence of neural collapse, the K-dimensional compressed representations not only retain most of the generalization performance, but we also observe K-simplex ETF within the compressed representations (in both classification and zero-shot transfer learning).

2. Theoretical Derivation

2.1. Notation

We use X to represent input images and Y to represent their labels. Then, Y_{train} and Y_{test} refer to training labels and test labels, respectively. Throughout the paper, $P(\cdot)$, $Q(\cdot)$, $H(\cdot)$, and $I(\cdot)$ refer to probability distributions, learned probability distributions, entropy, and mutual information, respectively. We also use $Z_{i=1,2}$ to represent learned representations from models (1 and 2) with the same architecture but independently trained. To discuss the linear identifiability of supervised contrastive loss, we use $f(\cdot)$ to denote the data representation and $g(\cdot)$ to denote the context representation, respectively. When constructing compressed representations based on DIB (Strouse & Schwab, 2016) (i.e., we try to compress Z_2 using Z_1), we refer to the resulting DIB optimal representations as T. All datasets D we used here are balanced datasets (equal number of samples for each label). We use K to denote the number of classes in a dataset.

2.2. Previous work in supervised contrastive learning

2.2.1. Supervised contrastive-learning-trained deep neural networks are linearly identifiable

Given a dataset D with input x and target y, a general deep neural network learns an empirical distribution $p_D(y|x)$. Previous work from nonlinear ICA provided sufficient conditions for the learned representation of $p_D(y|x)$ to be linearly identifiable (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2018). Namely, given two estimates θ and θ' for a data model P_D, if P_D is identifiable, then $p_{D,\theta} = p_{D,\theta'} \rightarrow \theta = \theta'$. Recent work (Roeder et al., 2020) extended this notion of linear identifiablity to a broad model family that uses contrastive loss as the objective. Formally, linearly identifiable models can learn representations that are equivalent up to a linear transformation.

Definition 2.1. If $\theta \overset{L}{\sim} \theta'$, then there exists an invertible matrix M such that $q_\theta(x) = Mq'_\theta(x)$

In contrastive-loss-trained neural networks, the loss functions include two representations: a data representation f_θ and a context representation g_θ. Linear identifiability for contrastive learning indicates that there $\exists M$ and M', $f_\theta(x) = Mf'_\theta(x)$ and $g_\theta(x) = M'g'_\theta(x)$. Both M and

M' are invertible matrices of rank K when a deep neural network is trained for classification with a K-class dataset.

For the supervised contrastive loss, z_i is the representation of the data i, z_p is the representation of other positive samples (sharing the same class label $P(i)$), and z_a is the representation of negative samples in the same minibatch. The loss function can be written as:

$$p_\theta(y|x,S) = \sum_{i\in S} -\log\left[\frac{1}{P(i)}\sum_{i\in P(i)}\frac{\exp(z_i\cdot z_p)}{\sum_{i\in A(i)}(z_i\cdot z_a)}\right], \tag{1}$$

where the data representation is $f(\cdot) = z_i$, context representation is $g(\cdot) = z_p$ and $\sum_{y'\in S}\exp f_\theta(x)^T g_\theta(y')$ is $\sum_{i\in A(i)}(z_i\cdot z_a)$. Therefore, if we have another model (i.e., $f'(\cdot) = z_i'$, $g'(\cdot) = z_p'$) trained in parallel, there exist matrices M and M' such that $z_i = M\times z_i'$ and $z_p = M'\times z_p'$.

2.2.2. Neural collapse in supervised contrastive learning

Neural collapse refers to a phenomenon observed after training error goes to zero in a deep neural network whose objective is input classification (Papyan et al., 2020; Han et al., 2022). It is hotly debated whether continuing to train neural networks after the training error goes to zero improves generalization performance (Hui et al., 2022). This phase has been dubbed the "terminal phase of training" (TPT) (Ma et al., 2017; Belkin et al., 2018b;a; 2019). In general, during TPT, feature activation in the final layers collapse to a distinct cluster (or single point in extreme cases) for all samples belonging to the same class. This is the so-called "variability collapse" (NC1 introduced in (Papyan et al., 2020)). If the training dataset is balanced, for example there are n samples in each class for all K classes, then the K different class means form an equiangular tight frame (ETF). The ETF formed by the class means has a self-dual property; the optimal classifiers for each class collapse with their class means. As a result, the classification task becomes simply locating the closest class in the feature space for each test sample (NC4 in (Papyan et al., 2020)). This makes classification robust against random and adversarial Noise (Papyan et al., 2020). We hypothesize that this is the essential element leading to good generalization. In contrastive learning, supervised contrastive loss is minimized instead of directly optimizing the model's classification performance. Because we only have access to the feature activation z, two (NC1 and NC4) neural collapse phenomena discussed above (out of four outlined in (Papyan et al., 2020) are applicable (see Appendix A.2 for more details).

2.3. Contribution I: Deterministic information bottleneck characterizes neural collapse

From an information theoretic perspective, training beyond when a deep neural network achieves zero training error means that the mutual information between the training labels and the learned representation remains constant at its maximum once the deep neural network enters TPT, $I(Y_{train};Z)\to H(Y_{train})$. If $I(Y_{train};Y_{test})\to H(Y_{test})$, the training labels contain sufficient predictive power for the test labels, we will have $I(Y_{test},Z)\to H(Y_{test})$. Meanwhile, variability collapse means the stochastic mapping $H(Z|Y_{train})\to 0$. This corresponds to $I(Y_{train};Z) = H(Z) - H(Z|Y_{train})\to H(Z)$.

Combining the above observations, we use $I(Z;Y_{train}) = I_0$ to denote TPT. The constant I_0 indicates that the learned representation can obtain no more information about the training labels from the input. In this case, the optimization within neural collapse corresponds to both minimizing the entropy of the learned representation and maximizing the information about the test labels. If $I(Y_{train};Y_{test})\to H(Y_{test})$ (when test accuracy approaches 98% or higher, as shown for CIFAR10 in Section 3.1), minimizing the entropy of the learned representation dominates the training after the model reaches neural collapse. This is a special case of the information bottleneck method (Tishby et al., 2000), known as the deterministic information bottleneck (Strouse & Schwab, 2016). Formally, we have

$$\begin{aligned}\min_{w,b,\xi}\quad & H(Z)\\ \text{s.t.}\quad & I(Z;Y_{train}) = I_0\end{aligned} \tag{2}$$

After the emergence of neural collapse, one may not see any change in the training error. Therefore, we assume that the information between the learned representation and the input is a constant. We can rewrite the above DIB problem, with T as the optimal compression, as

$$q^*_{DIB}(t|x) = \arg\min_{q(t|x)} H(T) - \beta I(T;Y) \tag{3}$$

The solution of objective 3 has the following form ((Strouse & Schwab, 2016; 2017), see Appendix A.4 for details):

$$q^*_{DIB}(t|x) = \delta(t - t^*(x)) \tag{4}$$

where

$$t^*(x) = \arg\max_t \log q(t) - \beta D_{KL}[p(y|x)|q(y|t)] \tag{5}$$

We use T to denote the compressed representation that characterizes $I(X;Y)$. This optimal solution $\delta(\cdot)$ in Equation 4 has a statistical structure compatible with the variability collapse as $H(Z)\to 0$ (see Fig. 1). Our hypothesis is that neural collapse compresses learned representations while improving generalization. This hypothesis is supported by our empirical finding in Section 3.3. In addition, our result in Appendix A.10.3 suggests that the compressed representations still approximate the K-simplex ETF corresponding to the respective datasets. In the following section, we will discuss its connection with other applications of the information bottleneck method in contrastive learning.

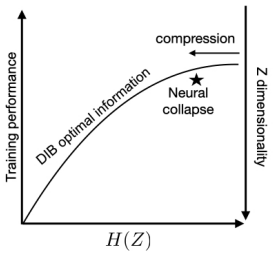


Figure 1. DIB ties compression to better generalization. The DIB information curve represents optimal classification across $H(Z)$. We hypothesize that neural collapse exists in the elbow of the DIB curve, where compression has little effect on training performance but improves test performance by simplifying the representation to improve stability to random or adversarial noise.

2.4. The relationship between supervised contrastive loss and DIB objectives within neural collapse

Within the context of supervised contrastive learning (Khosla et al., 2020), an encoder model is trained using supervised contrastive loss only and a linear classifier is usually trained independently after the completion of model training (Chen et al., 2020b;a; Khosla et al., 2020). Fang et al. (2021) proposed the following tractable objective as a surrogate that recapitulates the neural collapse in supervised contrastive learning:

$$\begin{aligned} &\min_{H} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n} L_c(z_{k,i}, y_k) \\ &\text{s.t.} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} ||z_{k,i}||^2 \leq c_0, \end{aligned} \tag{6}$$

As proved in both Fang et al. (2021); Zhu et al. (2021), the above objective 6 reaches an optimum (or equivalently exhibits neural collapse) when $z^*_{k,i} = \sqrt{c_0} m^*_k$, where m^*_k forms a K-simplex equiangular tight frame (ETF) (Papyan et al., 2020; Han et al., 2022). If the optimization algorithm uses a sufficiently large batch size such that $\log N > I(X;Z)$, then minimizing the above supervised contrastive loss is approximately equivalent to maximizing the mutual information between the input and the learned representation (van den Oord et al., 2018; Hjelm et al., 2019). This is also the InfoMax principle (Tian et al., 2019; Bachman et al., 2019; Wu et al., 2020; Hjelm et al., 2019). If we further assume that $p(Z)$ has a special form (see Assumption 2.2), then we can approximate the solution of the corresponding DIB objective. However, minimizing a lower bound of mutual information may not guarantee strong generalization (Tschannen et al., 2020). Instead of relying on any variational lower bounds, we directly use the constraint $I(Y_{train}; Z) = I_0$ to characterize the terminal phase of training and focus on minimizing $H(Z)$.

2.5. Contribution II: Approximation of the DIB solution within supervised contrastive learning

Because DIB has an intrinsic connection to geometric clustering (Strouse & Schwab, 2017) and the variability collapse of class clusters is a key feature of neural collapse, we can approximate the optimal solution of DIB by leveraging linear identifiability between parallel trained models and assuming that the correlation within class clusters becomes linear as training progresses. This proposition is supported by our empirical findings in Appendix A.10.2.

2.5.1. Gaussian mixture distribution with closed-form entropy

Assumption 2.2. After the emergence of neural collapse, a Gaussian distribution can approximate the feature activation of all samples belonging to the same class.
We observe that in neural collapse, the overall distribution of $p(z)$ has a unique Gaussian mixture structure, consisting of K distinct mixture components whose variances are close to zero (thus they have minimal or no overlap). The entropy has a closed form $H(z) = \sum_{i=1}^{K} \left[-\frac{1}{K} \log \frac{1}{K} - H(z_i)\right]$. Variability collapse corresponds to $H(z) \to 0$.

2.5.2. Approximation of DIB objective between linear identifiable representations using meta-Gaussian Information bottleneck

Linear identifiability between contrastive learning models implies the existence of a matrix A such that $Z_1 = A \times Z_2$. We propose using a noise model $Z_1 = A \times Z_2 + \xi$ to describe the actual mapping, as linear identifiability is never deterministic. We assume ξ is independent noise with a Gaussian distribution $N(0, \Sigma)$.

In Table 3.2, we show that linearly identifiable contrastive learning models achieve comparable generalization performance by correctly predicting the same subset of test samples. This compels us to investigate the structure of the the mutual information $I(Z_1; Z_2)$. The probability distribution that characterizes $I(Z_1; Z_2)$ is the copula $c(Z_1, Z_2)$ (Ma & Sun, 2008). A copula is the joint distribution of U_1 and U_2, where $U_{1,2}$ are rank transformed $Z_{1,2}$, $U_1 = P(Z_1 \leq Z^*)$. U is the short notion for cumulative distributions $F(z_1), \cdots, F(z_n)$ (i.e., $F(z_i) = p(z_i \leq z_0)$). Therefore, $p(Z_1, Z_2) = c(U_1, U_2) \prod_{i=1,2} Z_i$. Formulating $I(Z_1; Z_2)$ with a copula only requires replacing the covariance $Z_{1,2}$ with the copula correlation $corr(U_1, U_2)$. We show the derivation based on Ma & Sun (2008); Rey et al. (2014) in the Appendix. Here, we use

$$\begin{aligned} I(Z_1; Z_2) &= D_{KL}(p(z_1, z_2) || \prod_{i=1,2} p(z_1) p(z_2)) \\ &= \int c_{u_1,u_2} \log c_{(u_1,u_2)} du_1 du_2 = H(c_{u_1,u_2}) \end{aligned} \tag{7}$$

Linear identifiability between Z_1 and Z_2 suggests that U_1,

U_2 are also linearly dependent (both are characterizing the same general correlation structure), $U_1 = A' \times U_2 + \xi$. As a result, we can use a Meta-Gaussian information bottleneck (Rey & Roth, 2012) to approximate the linearly identifiable portion of $c(Z_1, Z_2)$. Such a Meta-Gaussian information bottleneck (MGIB (Rey et al., 2014) has computationally tractable optimal solution as the Gaussian information bottleneck (Chechik et al., 2005). DIB is a general objective that, in theory, characterizes neural collapse, but its solution is intractable. MGIB and GIB are special cases for which solutions can be calculated. Linear identifiability in deep neural networks trained with supervised contrastive learning allows us to approximate the intractable DIB solution using MGIB/GIB approximation. Formally speaking, if we define the information bottleneck problem between two learned representations Z_1 and Z_2 as

$$\mathcal{L}_{p(t|Z_1),\beta} = I(Z_1; T) - \beta I(Z_2; T) \tag{8}$$

Then the following proposition holds when $Z_1 = A \times Z_2 + \xi$.

Proposition 2.3 (Optimality of Meta-Gaussian Information bottleneck)**.** *Consider learned representations Z_1 and Z_2 with a Gaussian covariance structure and arbitrary margins (Rey et al., 2014; Rey & Roth, 2012)*

$$F_{Z_1,Z_2}(z_1, z_2) \sim C_G(F_{Z_1}, F_{Z_2})), \tag{9}$$

(see Equation 40 in the Appendix for details) where $F(Z) = F_{Z_{1,i}}$ or $F_{Z_{2,i}}$ are the marginal distributions of Z_1, Z_2 and C_G is a Gaussian copula parameterized by a correlation matrix G. The optimum of the minimization problem 18 is obtained for $T \in \mathcal{T}$, where $\mathcal{T}$ is the set of all random variables T such that (X, Y, T) has a Gaussian copula and T has Gaussian margins.

We provide a sketch of the proof for proposition 2.3 in the Appendix.

We use the following theorem from Chechik et al. (2005) to describe the structure of T because the optimal solution T from the Meta Gaussian information bottleneck (MGIB) is also the optimal solution for the corresponding Gaussian information bottleneck. U is the short notion for cumulative distributions $F(z_1), \cdots, F(z_n)$. $\Phi(\cdot)$ is the univariate Gaussian quantile function applied to each component, $\Phi^{-1}(u) = (\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n))$.

Theorem 2.4 (Optimal solution for the Gaussian Information Bottleneck)**.** *The optimal projection $T = A' \times U_1 + \xi$ for a given tradeoff parameter β is given by $\xi = I_x$ and*

$$A = \begin{cases} [0^T; \cdots; 0^T] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 v_1^T, 0^T; \cdots; 0^T] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 v_1^T, \alpha_2 v_2^T, 0^T; \cdots; 0^T] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \textit{otherwise}, \end{cases} \tag{10}$$

where $v_1^T, \cdots, v_n^T$ are left eigenvectors of $\Sigma_{x|y}\Sigma_x^-1$ sorted by their corresponding ascending eigenvalues $\lambda_1, \cdots, \lambda_n$, $\beta_i^c = \frac{1}{1-\lambda_i}$ are critical β values, α_i are coefficients defined by $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i = v_i^T \Sigma_x v_i$, 0^T is an n-dimensional row vector of zeros, and semicolons separate rows in the matrix A.

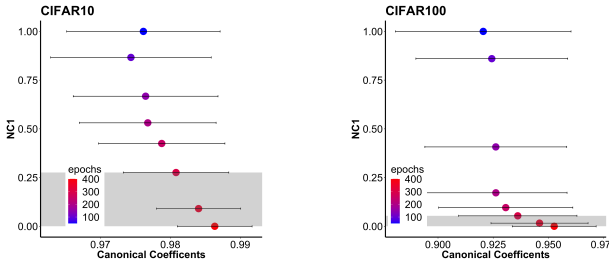
The i_{th} critical point β_i of the Gaussian information bottleneck is closely related with the i^{th} canonical correlation coefficient (CCA), $\beta_i = \frac{1}{1-\lambda_i}$. The difference between the compressed representations using the CCA or the Gaussian information bottleneck shows up in the different scaling of the respective eigenvectors. The CCA scaling is $\sqrt{1-\lambda_1}, \cdots, \sqrt{1-\lambda_n}$ whereas the scaling for IB is $\alpha_i \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$. α_i emphasizes the difference between consecutive λ instead of a single λ. Each α_i indicates, from an information-theoretic perspective, how much relevant information a given eigenvector contributes to the compressed representation T. In Fig. 11 (Appendix A.10.5), we show that the optimal representation for CIFAR10 has each of the 10 eigenvectors contributing nearly equally. This geometry is similar to that desired for a K-simple ETF.

3. Experimental Verification

We first show that as neural collapse happens, the linear identifiability between learned representations for datasets with few classes gets better (Fig. 3.1). This lets us use the meta-Gaussian information bottleneck (which we introduced in Section 2.5.2) between the representations learned by two independently trained ResNet50 backbones to approximate the DIB optimal compression solution. For datasets with a small number of classes (CIFAR10 and CIFAR100), we find that the DIB optimal representations compress the majority of the classification information into a K-dimensional representation (K is the number of classes; for CIFAR10, K=10). In addition, we observe that as generalization (measured by test accuracy) improves, more classification-relevant information is compressed into this K-dimensional representation (Fig. 3 and Tables 5 and 6 in Appendix A.10.1). This ties improved generalization to greater compression (Fig. 4). While we do not observe neural collapse in deep neural networks trained on the entire ImageNet32 dataset, we find that these deep neural networks require significantly higher dimensions (~ 70) to compress its subsets with fewer classes (e.g., Imagenette and Imagewoof) or CIFAR10 in a transfer learning scenario.

3.1. Linear identifiability improves *after* the emergence of neural collapse

We define neural collapse at the embedding layer (of a ResNet50 backbone) when we see the training accuracy go beyond 98%. We do not use 100% because previous work showed that popular datasets contain unexpected la-



(a) (b)

Figure 2. The emergence of neural collapse, shown via variability collapse (NC1), improves linear identifiability. We measure the linear identifiability between two learned representations using the average of their CCA coefficients (Raghu et al., 2017). The X-axis shows the variance of K-leading CCA coefficients for K-class datasets (e.g., K=10 for CIFAR-10 and 100 for CIFAR100). The Y-axis is the metric for variability collapse. Each dot is the mean of the CCA and the error bar shows the respective standard deviation. Colors are epochs. We also shade the area with less than 2% training error (the terminal phase of training (Ekambaram et al., 2017; Müller & Markert, 2019)). a) CIFAR-10; b) CIFAR-100;

beling inconsistencies (Ekambaram et al., 2017; Jadari & Nyberg, 2019; Müller & Markert, 2019) and we obtain near the state-of-the-art test accuracy (Table 2) with ResNet50 when training accuracy is above 98%. We observe that linear identifiability between learned representations improves significantly following the emergence of neural collapse (shaded regions in Fig. 2). We did not observe a significant improvement in linear identifiability prior to neural collapse. After neural collapse, we find that the linear identifiability of the learned representation improves in terms of both higher means and a smaller standard deviation between relevant canonical coefficients. This may correlate with the emergence of the K-simplex ETF (K=10 for CIFAR10, 100 for CIFAR100, respectively). We evaluate this hypothesis in the next section.

In addition, we observe that the linear classification performance of models trained with ImageNet32 stabilizes around 81%, while their test performance is comparable to the state-of-the-art using a vision transformer (Yang et al., 2022). This behavior does not fit the definition of the terminal phase of training (training error goes to zero) originated from cross entropy trained models. However, it is consistent with our information theoretical interpretation of neural collapse, i.e., the model does not retain additional information about the label from its input. We interpret this result as an outcome of training an encoder with the supervised contrastive loss (van den Oord et al., 2018; Khosla et al., 2020), as opposed to training linear classification directly (linear classifiers are usually trained separately after the training of the encoder finishes). In the following sections, we demonstrate that deep neural networks trained with the

DATA SET	CIFAR-10	CIFAR-100	IMAGENET32
TRAIN	98.8	98.5	81.1
TEST	95.8	76.0	56.8

Table 1. **Performance of ResNet50 trained with supervised contrastive loss**: Training errors after 400 epochs.

DATA SET	CIFAR-10	CIFAR-100	IMAGENET32
MODEL 1	95.4	75.7	56.8
MODEL 2	95.8	76.0	56.5
BOTH	93	69.2	56.4

Table 2. **Test performance shared by two deep neural networks trained in parallel**: Linearly identifiable representations from ResNet50 trained with CIFAR10, CIFAR100 and ImageNet32 learn similar decision boundaries for classifications.

full ImageNet32 contain similar learned representations to those trained with datasets of fewer classes.

3.2. $I(Z_1; Z_2)$ contains the majority of the Classification-relevant information

An important advantage of linear identifiability is that all models may converge to equivalent global optimal solutions (Hyvarinen et al., 2018; Roeder et al., 2020). If this is the case, then the learned representation is stable in the sense that its performance on downstream tasks is consistent despite random initialization. This theoretical prediction is validated in Table 2 by demonstrating that two models of the same architecture trained in parallel with a supervised contrastive learning objective have comparable generalization performance and make comparable decisions on the test dataset. This observation also suggests that the mutual information between the learned representations Z_1 and Z_2 contains the majority of the classification-relevant information. Next, we investigate the structure of the correlation between Z_1 and Z_2.

3.3. Supervised contrastive learning compresses classification into a K-dimensional representation as generalization improves

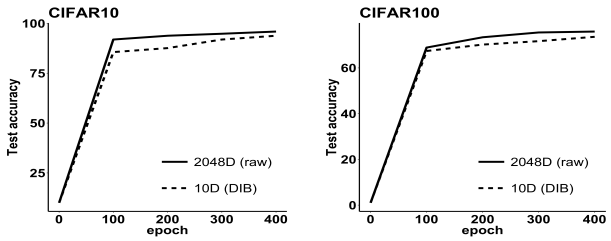
According to prior research (Zhu et al., 2021; Graf et al., 2021), learning a K-simplex ETF requires at least K dimensions. Given that, following the emergence of neural collapse, clusters of different classes become non-overlapping in the high-dimensional encoder space, we question whether the improvement in linear identifiability brought by neural collapse concentrates on the first K leading dimensions. The insight from Table 2 also prompts us to question whether the correlation structure within $I(Z_1; Z_2)$ becomes more linearly similar. Fig.6 in Appendix A.10.2 shows that the CCA coefficients between the raw learned representation and its

DATA SET	CIFAR10	CIFAR100
RAW	95.9	75.9
RANKED	95.8	75.6
RANKED K-DIB	93.8	73.5

Table 3. Test accuracy using DIB optimal representations. K=10 for CIFAR10 and K=100 for CIFAR100

rank-transformed version, after 400 epochs of training.

At the 400th epoch, we observed high performance on both training and testing datasets for CIFAR10, CIFAR100 (Table 2 and 3.2). Therefore, we hypothesize that this is the stage where the models approximate optimal solutions for the deterministic information bottleneck (DIB) objectives. In general, the DIB objective is intractable. However, the linear similarity we observed in Appendix A.10.2 enables us to construct a Meta-Gaussian information bottleneck (MGIB) between two learned representations. We hypothesize that the close similarity between the rank transformed data and the raw data suggests that the optimal solution to Meta-Gaussian information bottleneck may capture the statistical structure within the K-dimensional subspace for the emerging K-simplex ETF ($K = 10$ for CIFAR10, $K = 100$ for CIFAR100). In Table 3, we show that the classification performance using the MGIB-optimal representation, which is only K-dimensional, retains nearly all the classification performance achievable from the raw, 2048 dimensional learned representation.

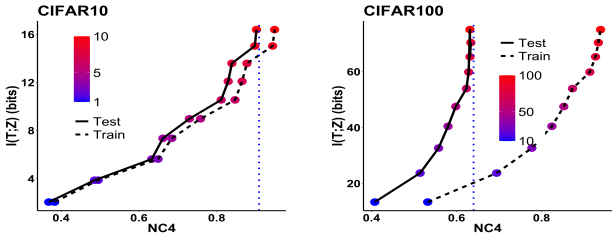


(a) (b)

Figure 3. Test accuracy improves in both the raw 2048D embedding Z and the K-IB compressed representations.

Fig.3 demonstrates that ResNet50 trained with supervised contrastive loss compresses most of the information relevant to classification into the DIB optimal K-dimensional representations while improving generalization (shown as test accuracy). In addition, Tables 5 and 6 show that the test accuracy of the raw 2048D representation varies little during neural collapse (identified as the shadowed regions in Fig.3.1 around 300-400 epochs), whereas the test accuracy of the K-dimensional DIB optimal representations improves by approximately 2.2% for both CIFAR10 and CIFAR100. This suggests that neural collapse corresponds to the compression of classification-relevant information.

3.4. The DIB optimal representations encourage classification to become finding the nearest class cluster



(a) (b)

Figure 4. Within the K dimensions, the compressed representations obtained from the Meta-Gaussian Information Bottleneck (MGIB) exhibit nearest center classification (the NC4 introduced in (Papyan et al., 2020), shown in both the train and test sets); each of the MGIB dimensions contributes to both adding more information to $I(T;Z)$ and improving NC4 (as shown by the improvement of NC4 on the test dataset). The vertical blue lines are the upper bounds of NC4 measurements on the test sets, using the full 2048D output representation from the ResNet50 backbone. The x-axis shows the percentage of samples that can be correctly classified by finding the closest class cluster. The y-axis shows the mutual information between 2048D embedding from ResNet50 and the compressed representation T, obtained from solving the meta-Gaussian information bottleneck between Z_1 and Z_2.

We investigate whether geometry of the compressed MGIB-optimal representation (solutions for $I(Z_1;T) - \beta I(T;Z_2)$) also encourages classification to become finding the nearest class cluster. This is referred to as "NC4" in the original papers on neural collapse (Papyan et al., 2020). As more dimensions are added to the optimal representation, we observe that classification becomes finding the closest class cluster (shown in Fig. 4) for more samples in both the train and test sets. Combining with our finding of the K-simplex ETF in these compressed MGIB-optimal representations (see Appendix A.10.3), we show that these optimal K-dimensional representations derived from a Meta-Gaussian information bottleneck also exhibit neural collapse. Notably, the percentage of simplification to the nearest class center (NC4 in Papyan et al. (2020); Fang et al. (2021); Zhu et al. (2021)) reaches its maximum at $K - 1$ for CIFAR10 and ~ 70 for CIFAR100 (adding more dimensions do not improve NC4). Using CIFAR10, we find that each IB dimension corresponds to similar phase transition coefficients, and contributes nearly equally to the improvement of NC4 and the addition of more relevant information denoted by $I(T;Z)$. These results show that the emergence of neural collapse within supervised contrastive learning results in classification relevant information being compressed into a 10D Gaussian distribution. For CIFAR100 (illustrated

in Fig.4), we find that adding dimensions beyond 70 does not contribute to the improvement of NC4. Because these MGIB optimal representations are the results of significant compression from the 2048D embedding layer, these representations correspond to an enhancement of generalization, as predicted from the generalization theorem of the information bottleneck method (Shamir et al., 2010) (see Appendix A.5). We further show that neural collapse results in similar scaling of individual feature vectors. The resulting MGIB optimal representations show compatible geometries with the desired equiangular tight frame (Fig. 11). These geometries are not present in the singular value decomposition of a learned representation Z (Fig.12).

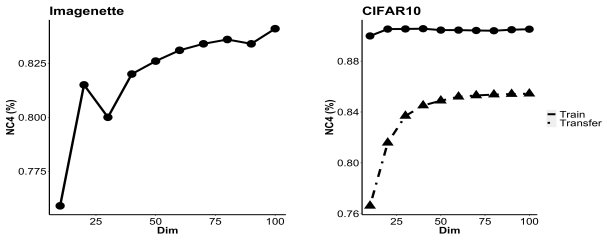
3.5. ImageNet32 trained models learn noisy K'-dimensional linear representations for the K-simplex ETF

Fig.6 demonstrates that ImageNet32-trained ResNet50 exhibit linear similarity only in the first few dimensions (less than K=1000). We investigate whether a ResNet50 backbone pretrained with the full ImageNet32 can obtain representations for K'-simplex ETF for a subset of the ImageNet32 with K' classes (K' is significantly less than 1000). Here we use Imagenette and Imagewoof (Howard, 2022). There are ten classes in these subsets. Imagenette contains a variety of classes (e.g., tench, English springer, cassette player), whereas Imagewoof contains only dog breeds (e.g., Australian terrier, Border terrier, Samoyed); therefore, Imagewoof is more difficult.

Table 4 shows that ImageNet32 trained models reach performance close to neural collapse, using the full 2048D embedding. However, Fig.5 shows that neural collapse for these two subsets requires more dimensions than the number of classes. For example. Imagenette uses 60-70 dimensions to reach the saturation level of NC4. Although 10-dimensional MGIB compression retains the majority of the NC4, it is necessary to use 70 dimensions to achieve 93.5% of the classification performance. This suggests that the learned representations from ImageNet32 may contain a noisy representation of the K'-ETF. To reconstruct the K'-ETF, one may need to learn a higher dimensional Gaussian distribution between a pair of models. We observe that this behavior is extendable to zero-shot transfer learning. The transfer learning for CIFAR10 retains 93.2% classification performance using 70 dimensions in its MGIB optimal representations, whereas only 10 dimensions are needed when a ResNet50 is directly trained for CIFAR10. For the more challenging Imagewoof, we show that the DIB optimal representation retains most of the compressible information (A.10.4), but the compression is lossy compared to Imagenette. This may echo previous results that all 2048 embedding dimensions are important for the overall performance of ImageNet (Jing et al., 2021).

DATA SET	IMAGENETTE	IMAGEWOOF	CIFAR10
RAW	96.9	90.66	95.8
RANKED (2048D)	96.8	87.99	93.9
IB (70D)	93.5	74.7	93.2

Table 4. Classification performance on subsets of ImageNet (Imagenette and Imagewoof) and zero-shot transfer learning (CIFAR10). Imagewoof is a challenging dataset. The 70D-IB compresses the majority of compressible classification relevant information (see Table A.10.4 in Appendix)



(a) (b)

Figure 5. a) learned representations for subset of ImageNet32 contains a noisier K'-ETF; b) The zero-shot transfer learning for CIFAR10 needs nearly 70 dimensions to exhibit neural collapse and retain classification performance.

4. Conclusion

This paper demonstrates that neural collapse may be quantified by correspondence with compression in the deterministic information bottleneck (DIB). Although DIB is generally intractable, the linear identifiability in contrastive learning enables us to approximate its optimal DIB representation. We observe that the compressibility of learned representations improves test accuracy. This demonstrates that the emergence of neural collapse corresponds to learning a more compact representation that improves generalization performance. We also discover that neural collapse occurs in the optimal DIB representations. Because these representations have only K dimensions and it takes at least K dimensions to fit a K-simplex ETF, our results connect good generalization with near-optimal compression. While we do not observe neural collapse with deep neural networks pretrained on the full ImageNet32 dataset, we observe compression, particularly during zero-shot transfer learning with CIFAR10. This suggests that, broadly, neural collapse in supervised contrastive learning may lead to improved generalization via compression.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *Proceedings of the*

International Conference on Learning Representations (ICLR) 2017, December 2016.

Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, jan 1972. doi: 10.1109/tit.1972.1054753.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. June 2019.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, jan 1989. doi: 10.1016/0893-6080(89)90014-2.

Belkin, M., Hsu, D., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. June 2018a.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? June 2018b.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116:15849–15854, August 2019. ISSN 1091-6490. doi: 10.1073/pnas.1903070116.

Ben-Shaul, I. and Dekel, S. Nearest class-center simplification through intermediate layers. January 2022.

Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, jul 1972. doi: 10.1109/tit.1972.1054855.

Braun, L., Dominé, C. C. J., Fitzgerald, J. E., and Saxe, A. M. Exact learning dynamics of deep linear networks with prior knowledge. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=lJx2vng-KiC.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for gaussian variables. *J. Mach. Learn. Res.*, 6:165–188, dec 2005. ISSN 1532-4435.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. February 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 22243–22255. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf.

Ekambaram, R., Goldgof, D. B., and Hall, L. O. Finding label noise examples in large scale datasets. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, oct 2017. doi: 10.1109/smc.2017.8122985.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences of the United States of America*, 118, October 2021. ISSN 1091-6490. doi: 10.1073/pnas.2103091118.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. Dissecting supervised contrastive learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Haeffele, B. D. and Vidal, R. Global optimality in tensor factorization, deep learning, and beyond. June 2015.

Han, X., Papyan, V., and Donoho, D. L. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Howard, J. Imagewang. https://github.com/fastai/imagenette/, 2022.

Hui, L., Belkin, M., and Nakkiran, P. Limitations of neural collapse for understanding generalization in deep learning. February 2022.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. May 2016.

Hyvarinen, A., Sasaki, H., and Turner, R. E. Nonlinear ica using auxiliary variables and generalized contrastive learning. May 2018.

Jadari, S. and Nyberg, J. F. Finding mislabeled data in datasets. 2019.

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. Understanding dimensional collapse in contrastive self-supervised learning. *ICLR 2022*, October 2021.

Kawaguchi, K. Deep learning without poor local minima. May 2016.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. April 2020.

Laurent, T. and von Brecht, J. Deep linear neural networks with arbitrary loss: All local minima are global. December 2017.

Liang, S., Sun, R., Lee, J. D., and Srikant, R. Adding one neuron can eliminate all bad local minima. May 2018.

Ma, J. and Sun, Z. Mutual information is copula entropy. August 2008.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. December 2017.

Müller, N. M. and Markert, K. Identifying mislabeled instances in classification datasets. *2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019*, December 2019. doi: 10.1109/IJCNN.2019.8851920.

Papyan, V., Han, X. Y., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences of the United States of America*, 117:24652–24663, October 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. June 2017.

Rey, M. and Roth, V. Meta-gaussian information bottleneck. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/3cef96dcc9b8035d23f69e30bb19218a-Paper.pdf.

Rey, M., Roth, V., and Fuchs, T. Sparse meta-gaussian information bottleneck. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 910–918, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/rey14.html.

Roeder, G., Metz, L., and Kingma, D. P. On linear identifiability of learned representations. July 2020.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. December 2017.

Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, jun 2010. doi: 10.1016/j.tcs.2010.04.006.

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Sklar, M. Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

Strohmer, T. and Heath, R. Grassmannian frames with applications to coding and communication. January 2003.

Strouse, D. and Schwab, D. J. The deterministic information bottleneck. April 2016.

Strouse, D. and Schwab, D. J. The information bottleneck and geometric clustering. *Neural Computation 31 (2019) 596-612*, December 2017.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. June 2019.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. April 2000.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. July 2018.

Wieczorek, A., Wieser, M., Murezzan, D., and Roth, V. Learning sparse latent representations with the deep copula information bottleneck. *Conference track - ICLR 2018*, April 2018.

Wu, M., Zhuang, C., Yamins, D., and Goodman, N. D. On the importance of views in unsupervised representation learning. 2020.

Yang, X., Shih, S.-M., Fu, Y., Zhao, X., and Ji, S. Your vit is secretly a hybrid discriminative-generative diffusion model. August 2022.

Yun, C., Sra, S., and Jadbabaie, A. Global optimality conditions for deep neural networks. July 2017.

Yun, C., Sra, S., and Jadbabaie, A. Small nonlinearities in activation functions create bad local minima in neural networks. February 2018.

Zhou, M., Ge, R., and Jin, C. A local convergence theory for mildly over-parameterized two-layer neural network. February 2021.

Zhu, Z., Soudry, D., Eldar, Y. C., and Wakin, M. B. The global optimization geometry of shallow linear neural networks. May 2018.

Zhu, Z., Ding, T., Zhou, J., Li, X., You, C., Sulam, J., and Qu, Q. A geometric analysis of neural collapse with unconstrained features. May 2021.

A. Appendix

A.1. Supervised Contrastive Learning belongs to a model family that is linearly identifiable

In (Roeder et al., 2020), they showed that the model family with contextualized representation for the input data is generally identifiable in the function space (i.e., any pairs of learned representations initialized with different random seeds are only different by up to a matrix transformation). This model family has its parametric probability defined by combining a target context variable y, the corresponding observed data x and a set S (containing both positive samples with the true label y and negative samples with labels y':

$$p_\theta(y|x,S) = \frac{\exp\left(f_\theta(x)^T g_\theta(y)\right)}{\sum_{y'\in S}\exp f_\theta(x)^T g_\theta(y')} \tag{11}$$

$f_\theta(x)$ is generalized to a data representation function, $g_\theta(y)$ is generalized to a context representation function, and $\Sigma_{y'\in S}\exp f_\theta(x)^T g_\theta(y)$ is the normalization factor. It is a constant if we have a fixed batch size.

They showed that the original contrastive loss function, i.e., contrastive predictive coding (CPC), proposed in van den Oord et al. (2018) belongs to the above model family. Specifically, if we let $X = x_1 \cdots, x_N$ be a set of N random samples obtaining one positive sample from $p(x_{t+k}|c_t)$ and $N-1$ samples from the input marginal distribution $p(x_{t+k})$. Then, contrastive predictive coding optimizes the following loss function:

$$-\mathbb{E}_X\left[\log\frac{l_k(x_{t+k},c_t)}{\sum_{x_j\in X} l_k(x_j,c_t)}\right] = -\mathbb{E}_X\left[\log\frac{\exp\left(z_{t+k}^T W_k c_t\right)}{\sum_{x_j\in X}\exp\left(z_j^T W_k c_t\right)}\right] \tag{12}$$

Where $f(\cdot) = z_{t+k}$, $g(\cdot) = W_k c_t$, $\sum_{y'\in S}\exp f_\theta(x)^T g_\theta(y')$ is $\sum_{x_j\in X}\exp\left(z_j^T W_k c_t\right)$.

Supervised contrastive loss (Khosla et al., 2020) is a variation of CPC. The only difference is instead of using an additional encoder to generate context encoding c_t, supervised contrastive loss uses the output features of the same encoder using other positive samples to generate context encoding, i.e., $g_\theta(x) = z_p$. Therefore, we can rewrite the supervised contrastive loss (Khosla et al., 2020) in the following form

$$p_\theta(y|x,S) = \sum_{i\in S} -\log\left[\frac{1}{P(i)}\sum_{i\in P(i)}\frac{\exp\left(z_i\cdot z_p\right)}{\sum_{i\in A(i)}(z_i\cdot z_a)}\right] \tag{13}$$

Where $f(\cdot) = z_i$, $g(\cdot) = z_p$, $\sum_{y'\in S}\exp f_\theta(x)^T g_\theta(y')$ is $\sum_{i\in A(i)}(z_i\cdot z_a)$.

This shows that the supervised contrastive loss belongs to the model family defined in Equation 11.

A.2. Neural collapse and K-simplex ETF

The seminal work of (Papyan et al., 2020) is the first to provide a mathematically elegant insight on the behaviors of deep neural networks for classification. It showed that the last layer features and classifiers in deep neural networks exhibit the following four properties:

NC1 Variability collapse: The within-class variation of the last-layer features becomes 0. This corresponds to the phenomenon that features collpase to their class means;

NC2 All class means collapse to vertices of a simplex equiangular tight frame (ETF) up to scaling

NC3 Up to scaling, the last-layer classifiers each collapse to the dual of the corresponding class means.

NC4 When the network performs inference on a test example, its decision collapses to simply choosing the class with the closest Euclidean distance between its class mean and the activations of the test example.

Definition A.1 (K-Simplex ETF)**.** A standard Simplex ETF is a collection of points in $\mathbb{R}^K$ specified by the columns of

$$M = \sqrt{\frac{K}{K-1}}\left(I_k - \frac{1}{K}1_k 1_K^T\right) \tag{14}$$

where $I_K \in \mathbb{R}^{K\times K}$ is the identity matrix, and $1_K \in \mathbb{R}^K$ is the all ones vector. Alternatively, we may rewrite the Equation 14 as:

$$M^T M = MM^T = \frac{K}{K-1}(I_K - \frac{1}{K}1_K 1_K^T) \tag{15}$$

. Following the notion introduced in (Papyan et al., 2020; Fang et al., 2021; Zhu et al., 2021), we consider general Simplex ETF as a collection of points in $\mathbb{R}^d$ specified by the columns of $\sqrt{\frac{K}{K-1}}P\left(I_k - \frac{1}{K}1_k 1_K^T\right)$, where (i) when $d \geq K$, $P \in \mathbb{R}^{d\times K}$ is an orthonormal matrix, i.e., $P^T P = I_K$, and (ii) when $d = K-1$, P is chosen such that $\left[P^T \frac{1}{\sqrt{K}}1_K\right]$ is an orthonormal matrix.

A.3. Neural collapse with supervised contrastive loss

The remarkable development of deep learning over the past decade has generated architectures with growing sizes and intricate details. No theory yet exists that can rigourously reason why these models work. Recently, Fang et al. (2021) proposed a mathematically tractable surrogate model, i.e., Layer-Peeled model, that can effectively explain and predict common patterns of deep neural networks. This model is derived by isolating the topmost layer (hence the name Layer-Peeled model), followed by imposing constraints with respect to weight decay or normalization applied during training. This is a top-down approach. It is in contrast to the conventional bottom-up approach that studies the feature representation within a deep neural network starting from the input (Yun et al., 2018; 2017; Haeffele & Vidal, 2015; Baldi & Hornik, 1989; Kawaguchi, 2016; Safran & Shamir, 2017; Laurent & von Brecht, 2017; Zhu et al., 2018; Zhou et al., 2021; Liang et al., 2018; Braun et al., 2022). The underlying reasoning is that modern deep networks are often highly overparameterized with the capacity to learn any representations, so that the last-layer features can be regarded as the output of a universal function approximator such that they approximate, or interpolate, any point in the feature space.

Layer-peeled model is a surrogate model that can reproduce neural collapse phenomena in many deep neural networks. Here we introduce the layer-peeled model for contrastive loss. Given $z_{k,i}$ as the last layer features for the i-th example with label k, the layer-peeled model takes the form

$$\begin{aligned} &\min_Z \frac{1}{N}\sum_{k=1}^K\sum_{i=1}^n L_c(z_{k,i}, y_k) \\ &\text{s.t.}\frac{1}{K}\sum_{k=1}^K \frac{1}{n}\sum_{i=1}^n ||z_{k,i}||^2 \leq c_0 \end{aligned} \tag{16}$$

Where the overall loss function L_c takes the following form (including all training data)

$$\frac{1}{n}\sum_{j=1}^n -\log\left(\frac{\exp\left(z_{k,i}z_{k,j}/\tau\right)}{\sum_{k'=1}^K\sum_{l=1}^n \exp\left(z_{k,i}z_{k',j}/\tau\right)}\right) \tag{17}$$

They proved a unique theorem (Theorem 3 in (Fang et al., 2021)) showing that the supervised contrastive loss exhibits neural collapse in its last-layer features.

Because the contrastive loss function only uses the activation itself, two out of the total four neural collapse phenomena are relevant, i.e., NC1 and NC4. We use the metric introduced in both (Fang et al., 2021; Ben-Shaul & Dekel, 2022) to measure them.

NC1 **Variability collapse**: the feature activations for all examples within one class collapse to a vertex of the K-simplex ETF. If we define the covariance matrix for class c as Σ_c. Then variability collapse corresponds to $||\Sigma_c|| \to 0$.

NC4 **Classification reduces to finding the nearest center**. If there are class means for C classes, i.e., $\mu_1, \cdots, \mu_C$. Given feature activation $z_{i,c}$ (the i-th example for class c), then classification of $z_{i,c}$ reduces to $\arg\min c' ||z_{i_c} - \mu_{c'}||_2$.

A.4. Information bottleneck and Deterministic Information Bottleneck

The information bottleneck (IB) method was first proposed in Tishby et al. (2000) based on trate distortion theory (Shannon, 1948). Given input X and output Y, the information bottleneck method identifies whether a compressed representation

T retains as much information as possible about the relevant variable Y while compressing away irrelevant components of the input X. In this context, the information bottleneck shows how much information a compressed representation needs, in order to encode a specific amount of information about the features of interest, Y. We can recapitulate the above compression into the following objective function:

$$\mathcal{L}_{p(t|X),\beta} = I(X;T) - \beta I(Y;T) \tag{18}$$

Due to the data processing inequality $I(Y;T) \leq I(X;Y)$, the equality only holds when T is the exact sufficient statistic of $p(X;Y)$. The optimal solution of the above IB objective can be determined by the iterative Blahut-Arimoto (Blahut, 1972; Arimoto, 1972) algorithm:

$$q(t|x) = \frac{q(t)}{Z(x,\beta)} \exp\left(-\beta D_{KL}(p(y|x)|q(y|t))\right) \tag{19}$$

$$q(t) = \sum_x q(t|x)p(x) \tag{20}$$

$$q(y|t) = \sum_x q(x|t)p(y|x) \tag{21}$$

$Z(x,\beta)$ is the normalization constant. Note that the Blahut-Arimoto algorithm requires access to probability distributions $p(x,y)$ itself. Therefore, the IB objective is intractable in general.

If we rewrite $I(X;T) = H(T) - H(T|X)$, we observe that $H(T|X)$ measures the stochasticitiy in the mapping from X to T. However, in the neural collapse scenario, all variability goes to zero and X can only map to their corresponding class mean. This suggests that $H(T|X) = 0$ when neural collapse happens. Given this observation, we use the α-formulation to change the above Equation 18 to define the objective for the deterministic case:

$$\mathcal{L}^{DIB}_{p(t|X),\beta} = H(T) - \alpha H(T|X) - \beta I(Y;T) \tag{22}$$

The solution of the above Equation 22 is closely related to the original IB solution.

$$q_\alpha(t|x) = \frac{1}{Z(\alpha,\beta)} \exp\left[\frac{1}{\alpha}(\log q_\alpha(t) - \beta D_{KL}[p(y|x)|q_\alpha(y|t)])\right] \tag{23}$$

$$q_\alpha(t) = \sum_x p(x)q_\alpha(t|x) \tag{24}$$

$$q_\alpha(y|t) = \frac{1}{q_\alpha(t)} \sum_x p(y|x)q_\alpha(t|x)p(x) \tag{25}$$

If we take the limit $\alpha \to 0$, the argument of the exponential in $q_\alpha(t|x)$ begins to blow up. Therefore, Strouse & Schwab (2016) argue that $q(t|x)$ collapse into a delta function with a particular t^* that maximizes $\log q(t) - \beta D_{KL}[p(y|x)|p(y|t)]$. We can rewrite Equation 22 as:

$$q^*_{DIB}(t|x) = \arg\min_{q(t|x)} H(T) - \beta I(T;Y) \tag{26}$$

This change makes Equations 23 to 25 to become a hard clustering (as discussed in Strouse & Schwab (2017)):

$$q^*_{DIB}(t|x) = \lim_{\alpha\to 0} q_\alpha(t|x) = \delta(t - t^*(x)) \tag{27}$$

$$t^*(x) = \arg\max_t \log q(t) - \beta D_{KL}[p(y|x)|q(y|t)] \tag{28}$$

$$q(t) = \sum_x q(t|x)p(x) \tag{29}$$

$$q(y|t) = \frac{1}{q(t)} \sum_x q(t|x)p(x)p(y|x) \tag{30}$$

In the general neural collapse scenario, training erorr is zero, this corresponds to a β^* that encourages retaining of all information about Y. The corresponding solution is $t_1^*, \cdots, t_K^*$. Those are critcial points that add the most information about Y to T. Here, using as few critcial points in t as possible is reminiscent to using no more than K-simplex to construct the ETF for a K-class dataset. Delta functions is reminiscent to variability collapse.

A.5. Simpler T leads to better generalization for $I(Y;T)$

Now we introduce the notations needed to write down the generalization bound of the information bottleneck problem. We will use $\hat{I}(X;T)$, $\hat{I}(Y;T)$ to denote the empirically estimated $I(X;T)$ and $I(Y;T)$. The intuition behind the generalization bound is the gap between $I(X;T)$ $I(Y;T)$ and their empirical estimations $\hat{I}(X;T)$, $\hat{I}(Y;T)$. We would like such a bound to be small so we can learn T with a finite number of samples. X and Y are short notions for $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$. $p(T|X)$ is the short notion for $p(t|x_1), \cdots, p(t|x_n)$. $\hat{p}(T|Y)$ is the short notion for the estimated $\hat{p}(t|y_1), \cdots, \hat{p}(t|y_m)$. $\hat{H}(T|Y)$ is the short notion for the vector of estimated entropy $(\hat{H}(t|y_1), \cdots, \hat{H}(t|y_m))$. Then we define the variance of all elements in a vector as $V(a) = ||a - \frac{1}{m}\sum_{j=1}^n a_j||^2$ and the following auxiliary function $\phi(x)$:

$$\phi(x) = \begin{cases} 0 & x = 0 \\ x \log 1/x & 0 < x < 1/e \\ 1/e & x > 1/e \end{cases} \tag{31}$$

Now we can introduce the following theorem from (Shamir et al., 2010)

Theorem A.2 (Full generalization bound of IB)**.** *Let S be a sample of size $||S||$ drawn from the joint probability distribution $p(X,Y)$ and assume C is a small constant. For any confidence parameter $\delta \in (0,1)$, it holds with a probability of at least $1-\delta$ over sample S that for any T simultaneously,*

$$|I(X;T) - \hat{I}(X;T)| \leq \sqrt{\frac{C \log m/\delta \cdot V(H(T|X))}{||S||}} + \sum_t \phi(\sqrt{\frac{C \log m/\delta \cdot V(P(T=t|X))}{||S||}}) \tag{32}$$

and

$$|I(Y;T) - \hat{I}(Y;T)| \leq \sqrt{\frac{C \log m/\delta \cdot V(\hat{H}(T|Y))}{||S||}} + 2 \times \sum_t \phi(\sqrt{\frac{C \log m/\delta \cdot V(P(T=t|X))}{||S||}}) \tag{33}$$

Intuitively, the bounds in Theorem A.2 say that if $V(\cdot)$ are closer to zero (having a "smoother" T), we will have a tigher bound. In the extreme case if T is a uniform distribution and independent of X, i.e., $I(T;X) = \hat{I}(T;X) = 0$. The generalization bound become zero ($H(T|x) = H(T)$ and $\hat{H}(T|y) = \hat{H}(T)$. These observations also imply that generalization becomes better as T becomes less statistically dependent on X. This also motivates the following theorem, which states that T is the minimum sufficient statistic for Y based on X. The above theorem does not depend on the information bottleneck theory. It is a general bound on how well the approximated $\hat{I}(Y;T)$ compared to the actual $I(Y;T)$. Nevertheless, it helps the information bottleneck method to propose the following notion of "minimum sufficient statistic".

Theorem A.3 (Minimum sufficient statistic of T)**.** *Let X be a set of samples drawn according to a distribution determined by the random variable Y . The set of solutions T to*

$$\begin{aligned} \min_T \quad & I(X;T) \\ \textit{s.t.} \quad & I(Y;T) = \max_{T'} I(Y;T') \end{aligned} \tag{34}$$

is exactly the set of minimal sufficient statistics for Y based on the sample X.

Note that the Equation 34 is equivalent to Equation 18 by rewriting it with a Lagrangian multiplier. After the emergence of neural collapse, the distribution of T contains multiple discrete states. In this scenario, the optimal solution of objective 34 corresponds to a soft clustering of T, whose respective hard clustering is the deterministic IB solution for 26. The insight from the generalization Theorem A.2 is that the more compressed T is, the smaller the generalization gap between the estimated $\hat{I}(Y;T)$ and the true $I(Y;T)$.

A.6. Gaussian Information Bottleneck

This general IB principle is intractable. So far, close-form solution only exists at a special case: when $p(X, Y)$ is a joint Gaussian. In this case, solving Equation 18 becomes finding a matrix A such that $Y = AX + \xi$ (ξ is noise). The following theorem (Theorem 3.1 in (Chechik et al., 2005)) shows how to determine A using eigenvectors of the normalized regression matrix $\Sigma_{X|Y}\Sigma_{x^{-1}}$:

Theorem A.4 (Optimal solution for the Gaussian Information Bottleneck)**.** *The optimal projection $T = AX + \xi$ for a given tradeoff parameter β is given by $\xi = I_x$ and*

$$A = \begin{cases} [0^T; \cdots; 0^T] & 0 \leq \beta \leq \beta_1^c \\ [\alpha_1 v_1^T, 0^T; \cdots; 0^T] & \beta_1^c \leq \beta \leq \beta_2^c \\ [\alpha_1 v_1^T, \alpha_2 v_2^T, 0^T; \cdots; 0^T] & \beta_2^c \leq \beta \leq \beta_3^c \\ \vdots & \textit{otherwise} \end{cases} \tag{35}$$

where $v_1^T, \cdots, v_n^T$ are left eigenvectors of $\Sigma_{X|Y}\Sigma_X^{-1}$ sorted by their corresponding ascending eigenvalues $\lambda_1, \cdots, \lambda_n$, $\beta_i^c = \frac{1}{1-\lambda_i}$ are critical β values, α_i are coefficients defined by $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$, $r_i = v_i^T \Sigma_x v_i$, 0^T is an n-dimensional row vector of zeros, and semicolns separate rows in the matrix A.

A.7. Copula formulation for joint distributions and mutual information

In (Rey et al., 2014; Wieczorek et al., 2018), they extended the Gaussian information bottleneck to the case when $p(X, Y) = p(X)p(Y)C(X, Y)$, where $C(X, Y)$ is a Gaussian copula (Sklar, 1959).

When we look at a joint distribution $P(z_1, \cdots, z_n)$, the copula formulation (Sklar, 1959) allows us to describe the dependence structure between random variables and their margins separately. If we define the cumulative distribution (or the distribution function) as $F(z_i) = p(z_i \leq z_0)$ (the rank transformed distribution of each margin i) and use $p_i(z_i)$ to denote marginal density of an individual random variable z_i, (Sklar, 1959) shows that:

$$p(z_1, \cdots, z_n) = c(F_1(z_1), \cdots, F_n(z_n)) \prod_{i=1}^{n} p_i(z_i) \tag{36}$$

Using Equation 36, we can rewrite the multivariate mutual information $I(Z)$ as the entropy of the copula $H(c(z_1, \cdots, z_n))$ or $H(c_Z)$ ($Z = Z_1, \cdots, Z_n$):

$$I(Z) = D_{KL}(p(z_1, \cdots, z_n) || \prod_{i=1}^{n} p_i(z_i)) = \int c_Z(F(Z)) log c_Z(F(Z)) dF(Z) = -H(c_Z) \tag{37}$$

Following (Rey & Roth, 2012), Equation 37 allows us the rewrite the mutual information $I(Z_1; T)$ and $I(Z_2; T)$ in the following form: $I(Z_{1,2}; T) = -H(c_{Z_{1,2}T}) + H(c_T) + H(c_{Z1,2})$. Finding an IB optimal solution reduces to calculating an optimal copula density $c_{Z_1 T}$ with the markov chain defined as $T \to Z_1 \to Z_2$. Next we will use the known structure within $c_{Z_1 Z_2}$ based on their linear identifiablity to determine c_T.

A.8. The Meta-Gaussian information bottleneck problem between pairs of linearly identifiable models within Neural Collapse

The above section shows that an optimal compressed representation T between two learned representations Z_1 and Z_2 only depends on the correlation and thus the copula density c_{z_1, z_2}. Considering that Z_1 and Z_2 are linearly identifiable, we will use a Gaussian copula to describe their correlation structure. Given G as the correlation matrix between Z_1 and Z_2, Gaussian copula has the following density function:

$$c_G(u) = |G|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\Phi^{-1}(u)^T (G^{-1} - I)\Phi^{-1}(u)\right] \tag{38}$$

u is the short notion for cumulative distributions $F(z_1), \cdots, F(z_n)$. $\Phi(\cdot)$ is the univariate Gaussian quantile function applied to each component, i.e., $\Phi^{-1}(u) = (\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_n)$. Given that the correlation between Z_1 and Z_2 has a

Gaussian copula and their mutual information is determined by this correlation, we can simplify our analysis of $I(Z_{1,2};T)$ by tranforming Z_1 and Z_2 with their normal scores. We define the normal scores for rank transformed Z_1, Z_2 as:

$$\tilde{Z} = (\Phi^{-1} \circ F_{Z_1}(Z_1), \cdots, \Phi^{-1} \circ F_{Z_1}(Z_1)) \tag{39}$$

Note that the normal scores $\tilde{Z_1}$, $\tilde{Z_2}$ are multivariate Gaussian. However, they have the same copulas as the original Z_1 and Z_2 because copulas are invariant to monotonic transformations.

In (Rey et al., 2014; Wieczorek et al., 2018), they used these normal scores ($\tilde{Z_1}$, $\tilde{Z_2}$) to prove the following proposition:

Proposition A.5 (Optimality of Meta-Gaussian Information bottleneck)**.** *Consider learned representations Z_1 and Z_2 with a Gaussian dependence structure and arbitrary margins (Rey et al., 2014; Rey & Roth, 2012):*

$$F_{Z_1,Z_2}(z_1, z_2) \sim C_G(F_{Z_{1,1}}(z_{1,1}), \cdots, F_{Z_{1,g}}(z_{1,g}), F_{Z_{2,1}}(z_{2,1}), \cdots, F_{Z_{2,g}}(Z_{2,g})), \tag{40}$$

where $F_{Z_{1,i}}$ $F_{Z_{2,i}}$ are the marginal distributions of Z_1, Z_2 and C_G is a Gaussian copula parameterized by a correlation matrix G. Then the optimum of the minimization problem 18 is obtained for $T \in \mathcal{T}$, where $\mathcal{T}$ is the set of all random variables T such that (X, Y, T) has a Gaussian copula and T has Gaussian margins.

Lemma A.6. *$T \in \mathcal{T} \leftrightarrow (\tilde{Z_1}, \tilde{Z_2}, T)$ are jointly Gaussian.*

Proof. 1 $T \in \mathcal{T} \rightarrow (Z_1, Z_2, T)$ has a Gaussian copula. Therefore, $(\tilde{Z_1}, \tilde{Z_2}, T)$ also has a Gaussian copula. Because ($\tilde{Z_1}$, $\tilde{Z_2}$ and T all have Gaussian margins, $\tilde{Z_1}, \tilde{Z_2}, T$ are jointly Gaussian

2 If $(\tilde{Z_1}, \tilde{Z_2}, T)$ are jointly Gaussian, then $(\tilde{Z_1}, \tilde{Z_2}, T)$ has a Gaussian copula. Therefore, (Z_1, Z_2, T) also has a Gaussian copula. $(\tilde{Z_1}, \tilde{Z_2}, T)$ are jointly Gaussian also implies that T has Gaussian margins, it follows that $T \in \mathcal{T}$.

□

Proof of proposition A.5. Assume there exists $T^* \notin \mathcal{T}$, then we will have:

$$\begin{aligned} \mathcal{L}(Z_1, Z_2, T^*) = I(Z_1;T) - \beta I(Z_2;T) &< \min_{p(t|Z_1), T \in \tilde{T}} I(Z_1;T) - \beta I(T;Z_2) \\ &= \min_{p(t|\tilde{Z_1}), T \in \mathcal{T}} I(\tilde{Z_1};T) - \beta I(T;\tilde{Z_2}) \\ &= \min_{p(t|\tilde{Z_1}), (\tilde{Z_1}, \tilde{Z_2}, T) \sim \mathcal{N}} I(\tilde{Z_1};T) - \beta I(T;\tilde{Z_2}) \end{aligned} \tag{41}$$

The Equation 41 rewrites its right hand side using $(\tilde{Z_1}, \tilde{Z_2}, T)$ based on Lemma A.8. If $(\tilde{Z_1}, \tilde{Z_2}, T)$ has the same copula as $I(Z_1, Z_2, T)$ then $I(\tilde{Z_1}, T) = I(Z_1, T)$ and $I(\tilde{Z_2}, T) = I(Z_2, T)$. Similarly, we also have $I(\tilde{Z_1}, T^*) = I(Z_1, T*)$ and $I(\tilde{Z_2}, T^*) = I(Z_2, T^*)$. Then we can rewrite Equation 41 as:

$$\mathcal{L}(\tilde{Z_1}, \tilde{Z_2}, T^*) < \min_{p(t|\tilde{Z_1}), (\tilde{Z_1}, \tilde{Z_2}, T) \sim \mathcal{N}} \mathcal{L}(\tilde{Z_1}, \tilde{Z_2}, T) \tag{42}$$

Equation 42 contradicts the optimality property of Gaussian information bottleneck (Chechik et al., 2005), which dictates that the optimal solution T must be jointly Gaussian with $\tilde{Z_1}, \tilde{Z_2}$. □

Using proposition A.5, we have the following corollary:

Corollary A.7. *The optimal projection T' for $\tilde{Z_1}, \tilde{Z_2}$ is also optimal for (Z_1, Z_2) in case (Z_1, Z_2) has a Gaussian copula*

As a consequence of proposition A.5, we can obtain the optimal compressed representation between Z_1 and Z_2 using their correlation matrix P when (Z_1, Z_2) has a Gaussian copula. The correlation matrix P is parameterized by the Gaussian copula C_P and the solution following the form outlined in Theorem A.4 as the original Gaussian information bottleneck (Chechik et al., 2005).

EPOCH	100	200	300	400
CIFAR10 TRAIN	90.2	94.6	97.8	99.2
CIFAR10 TEST	91.9	93.8	94.8	96.0
CIFAR100 TRAIN	75.6	87.8	95.8	98.5
CIFAR100 TEST	68.9	73.4	75.5	75.9

Table 5. Train/Test accuracy for CIFAR10/CIFAR100

EPOCH	100	200	300	400
CIFAR10 TRAIN (K-DIB)	78.4	83.1	90.9	96.8
CIFAR10 TEST (K-DIB)	85.6	87.6	91.9	94.1
CIFAR100 TRAIN (K-DIB)	66.9	78.7	90.2	95.8
CIFAR100 TEST (K-DIB)	67.4	70.2	71.3	73.5

Table 6. Train/Test accuracy for CIFAR10/CIFAR100 using the K-dimension DIB optimal representation

A.9. Training details

We used the publicly available code to train all supervised contrastive learning models https://github.com/HobbitLong/SupContrast. We use the default settings for all hyper-parameters (e.g., temp =0.1). We train all models with ResNet50 as the backbone for 400 epochs. For CIFAR10, CIFAR100, we use a batch size of 512 on a single Nvidia A100 GPU. For models trained with ImageNet32, we use an 8×A100 GPU cluster with a batch size of 4096. We obtain all our linear classification results after 20 epochs of training and report the Top-1 accuracy.

A.10. Additional Results

A.10.1. TRAIN/TEST ACCURACY FOR CIFAR10/CIFAR100

A.10.2. K-SIMPLEX ETF SHOW HIGH LINEAR CORRELATION WITHIN THE K-DIM CORRELATION STRUCTURE

For CIFAR10 and CIFAR100, we observe a close linear similarity match at the first K singular vectors between the learned representations and their rank-transformed versions. Linear similarity between rank transformed representations suggests that the correlation within $I((Z_1; Z_2)$ becomes linear. Despite the fact that the linear similarity for rank-transformed CIFAR100 is slightly lower than the linear similarity for untransformed CIFAR100, both exhibit CCA coefficients with high magnitudes (> 0.9). In addition, the CIFAR100's CCA similarity reveals a clear saturation at K=100. These observations suggest that supervised contrastive learning models are capable of learning a K-simplex ETF in a K-dimensional linear feature space. Such observations on small datasets do not scale perfectly to the ImageNet32 dataset. Although the rank-transformed representations of ImageNet32 capture the majority of the linear similarity present in the raw representations (similar to what we observed in CIFAR10 and CIFAR100), the overall CCA similarity magnitude at K = 1000 drops below 0.5. Given the difference between small (CIFAR10, CIFAR100) and large (ImageNet32) datasets, we will analyze them in independent sections below.

A.10.3. K-SIMPLEX ETFS ALSO EMERGE IN DIB OPTIMAL COMPRESSION SOLUTION

In this section, we show the K-simplex ETF emergences in the DIB optimal compression solution with calculations similar to Fig.2-4 in (Papyan et al., 2020). We also find that the K-simplex ETFs using the DIB optimal compression solution is comparable to that using the raw 2-48D ResNet50 embedding.

A.10.4. MORE CHALLENGING DATASET (E.G., IMAGEWOOF) CONTAINS LESS COMPRESSIBLE INFORMATION

A.10.5. INFORMATION CURVE AND IB EMBEDDING GEOMETRIES FOR CIFAR10, CIFAR100

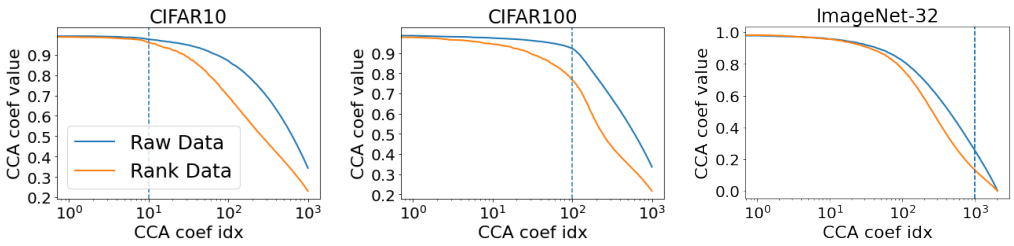


Figure 6. Canonical coefficient (CCA) similarity for pairs of supervised contrastive learning models exhibiting neural collapse using CIFAR10, CIFAR100, and ImageNet32. The blue lines are the CCA similarity of the raw data. Orange lines are the ranked transformed data (correlation only, no information from marginals, this shows if the correlation structure within $I(Z_1, Z_2)$) is linear). The vertical lines show K for each dataset. We hypothesize that neural collapse would make the first K leading dimensions exhibit high linear similarity.

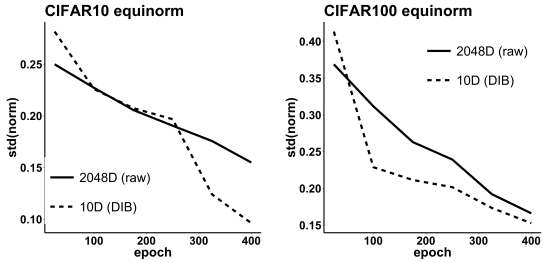


Figure 7. $Std_k(||\mu_k - \mu_{all}||_2 / Avg(||\mu_k - \mu_{all}||_2))$

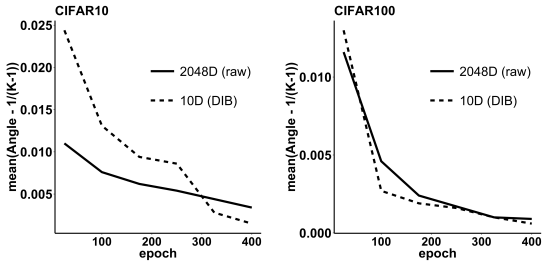


Figure 8. $\cos_\mu =< \mu_k - \mu_a ll, \mu_{k'} - \mu_{all} > /(||\mu_k - \mu_{all}||_2 ||\mu_{k'} - \mu_{all}||_2)$, μ_k is the mean of class k.μ_{all} is the global mean of all class clsuters. This figure shows the difference between the mean of $\cos_\mu$ and the maximum angle for K-simplex ETF (1/(K-1)).

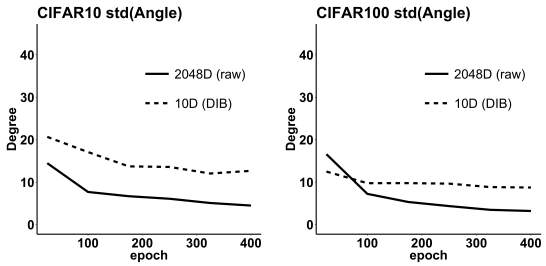


Figure 9. $\cos_\mu =< \mu_k - \mu_{all}, \mu_{k'} - \mu_{all} > /(||\mu_k - \mu_{all}||_2 ||\mu_{k'} - \mu_{all}||_2)$, μ_k is the mean of class k. μ_{all} is the global mean of all class clusters. This figure shows the standard deviation of $\cos_\mu$ or whether the class clusters have equal angle between each other.

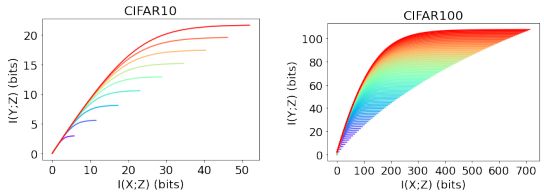


Figure 10. Information curve planes for the classification relevant representation obtained from pairs of linearly identifiable contrastive learning models trained in parallel. Each information curve (denoted by different colors) shows the optimal complexity $I(Z_1;T)$ and relevancy $I(T;Z_2)$ tradeoff of a fixed dimensionality corresponding to the information bottleneck objective $I(Z_1;T) - \beta I(T;Z_2)$ ($\beta = 500$).

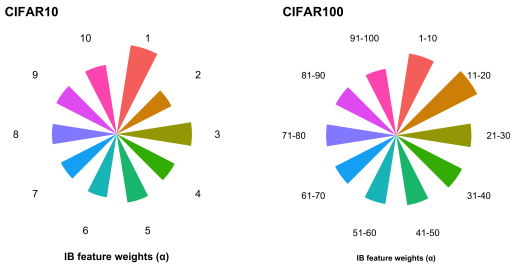


Figure 11. The relative scalings of different eigenvectors in the DIB optimal representation.

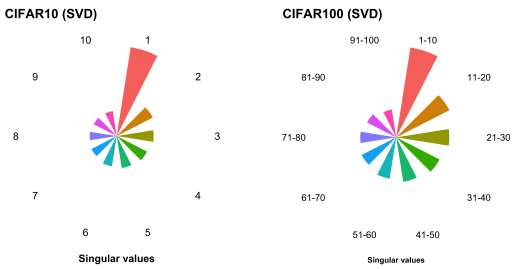


Figure 12. The relative scalings from the singular value decomposition of learned representations.

DIMENSIONS	500	1000	2048 (FULL EMBDDING)
	79.9	81.4	87.99

Table 7. **Test performance on imagewoof**: The representations for imagewoof in ImageNet32 trained models are much noisier. IB compressed representation shown in Table 4 already includes most of the compressible information for classification.