Enhancing Fine-Tuning in Low Data Regime by Increasing Representation Entropy During Pre-Training Phase

Jaeill Kim[1*], Jungwook Shin[1*], Wonjong Rhee[1,2]
[1] Department of Intelligence and Information, Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence (IPAI), Seoul National University
Seoul, Republic of Korea
jaeill0704@snu.ac.kr, jungwook.shin@snu.ac.kr, wrhee@snu.ac.kr

***Abstract*—The strategy of training a new model through the fine-tuning of pre-trained model has gained considerable prominence due to its potential to reduce training costs or enhance performance. However, accomplishing both of these objectives concurrently presents a noteworthy challenge. To address the challenge, this study adopts two entropy metrics, which can be seamlessly integrated into the cross entropy loss. Our approach aims at increasing representation entropy during pre-training phase, leading to an increased information encoded in the representation of a pre-trained model. This increased information can then be effectively harnessed during subsequent fine-tuning phase. In experiments, we substantiate the reliability of the two adopted entropy metrics in accurately quantifying representation entropy. Moreover, we demonstrate the effectiveness of the two metrics in enhancing the performance of fine-tuning, particularly in low data regime.**

***Index Terms*—Fine-Tuning, Pre-Trained Model, Low Data Regime, Shannon Entropy, Von Neumann Entropy**

I. Introduction

Since the introduction of large pre-trained models in deep learning, such as CLIP [8] and LLaMA [9], the strategy of training a new model by fine-tuning a pre-trained model has gained significant traction in recent studies, attributed to its ability to reduce training costs or enhance performance. However, the concurrent achievement of the dual goals concerning cost reduction and performance enhancement is challenging. This is because, in order to enhance performance, it is necessary to fine-tune with a more extensive dataset, which unavoidably leads to an increase in the training costs.

To address the challenge, this study adopts two entropy metrics proposed in [2], denoted by $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$, and employs them as representation regularization losses to increase the information encoded in representation of a pre-trained model. Specifically, we conjecture that by increasing information encoded in the representation during pre-training phase, the subsequent fine-tuning phase can effectively harness the increased information, thereby leading to an enhanced performance under conditions of very limited data availability, referred to as the term *low data regime*.

[*]Equal contribution.

In experiments, we first substantiate the reliability of the two entropy metrics in accurately quantifying the true entropy values of representations, with $S(\mathcal{C}_{\text{auto}})$ being more effective metric. To evaluate fine-tuning in low data regime, we initiate our evaluation by pre-training models, with and without the incorporation of the entropy regularizations. Subsequently, fine-tuning is conducted on the pre-trained models. The results demonstrate that our methods significantly improves the performance of fine-tuning, particularly in low data regime, with $S(\mathcal{C}_{\text{auto}})$ being more effective regularizer.

II. Preliminaries on Two Entropy Metrics

Quantifying entropy in high-dimensional vector spaces has been recognized as a challenging problem [3]. To address the challenge, we adopt two entropy metrics proposed in [2]. The first metric quantifies the Shannon entropy, whereas the second metric quantifies the von Neumann entropy.

A. $H_{NCE}(\mathbf{Z})$: A Metric for Quantifying Shannon Entropy

In this section, we summarize the details of the Shannon self-information proposed in [2], which quantifies the Shannon entropy of deep representations, denoted by $H_{NCE}(\mathbf{Z})$. This metric utilizes the well-known lower bound mutual information (MI) estimator, InfoNCE, which is generally known as a low variance and bias estimator for large sample sizes [5], [7].

Given two random variables $\mathbf{V}_1$ and $\mathbf{V}_2$, the definition of InfoNCE is formulated as follows:

$$I(\mathbf{V}_1;\mathbf{V}_2) \geq \log N + \mathbb{E}\left[\log \frac{e^{h(\boldsymbol{v}_{1,i},\boldsymbol{v}_{2,i})}}{\sum_{j=1}^{N} e^{h(\boldsymbol{v}_{1,i},\boldsymbol{v}_{2,j})}}\right] \triangleq I_{NCE}(\mathbf{V}_1;\mathbf{V}_2), \quad (1)$$

where N denotes the sample size, $(\boldsymbol{v}_{1,i}, \boldsymbol{v}_{2,i})$ are sampled from the joint distribution $p(\mathbf{V}_1, \mathbf{V}_2)$, and $\boldsymbol{v}_{2,j}$ are sampled from the marginal distribution $p(\mathbf{V}_2)$. $h(\cdot,\cdot)$ is a critic parameterized by neural networks [7]. Notably, $I_{NCE}(\mathbf{V}_1;\mathbf{V}_2) = \log N - \mathcal{L}_{NCE}$, where $\mathcal{L}_{NCE} = -\mathbb{E}\left[\log \frac{e^{h(\boldsymbol{v}_{1,i},\boldsymbol{v}_{2,i})}}{\sum_{j=1}^{N} e^{h(\boldsymbol{v}_{1,i},\boldsymbol{v}_{2,j})}}\right]$ represents the widely used contrastive loss [5] in machine learning.

To quantify the Shannon entropy of a representation matrix $\mathbf{Z}$ $(= [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_N]^T)$, $H_{NCE}(\mathbf{Z})$ utilizes the following

property of self-information, where the mutual information of a random variable with itself is the entropy of the random variable [1].

$$H(\mathbf{Z}) = I(\mathbf{Z}; \mathbf{Z}). \tag{2}$$

Based on Eq. 1 and Eq. 2, a variational entropy estimator, which is denoted by $H_{NCE}(\mathbf{Z})$, can be defined as below:

$$H(\mathbf{Z}) = I(\mathbf{Z}; \mathbf{Z}) \geq I_{NCE}(\mathbf{Z}; \mathbf{Z}) \triangleq H_{NCE}(\mathbf{Z}). \tag{3}$$

B. $S(\mathcal{C}_{auto})$: A Metric for Quantifying Von Neumann Entropy

In this section, we summarize another entropy metric proposed in [2], which quantifies the von Neumann entropy of deep representations, denoted by $S(\mathcal{C}_{\text{auto}})$.

To quantify the von Neumann entropy of a representation matrix $\mathbf{Z}$ $(= [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_N]^T)$, the autocorrelation matrix of deep representation $\mathbf{Z}$, which is denoted by $\mathcal{C}_{\text{auto}}$, needs to be defined as below:

$$\mathcal{C}_{\text{auto}} \triangleq \sum_{i=1}^{N} \frac{1}{N} \boldsymbol{h}_i \boldsymbol{h}_i^T, \tag{4}$$

where $\boldsymbol{h}_i \triangleq \boldsymbol{z}_i / \|\boldsymbol{z}_i\|_2$. Note that $tr(\mathcal{C}_{\text{auto}}) = 1$ and $\mathcal{C}_{\text{auto}} \geq 0$.

Then, $S(\mathcal{C}_{\text{auto}})$ can be computed using the eigenvalues λ_j of the autocorrelation matrix $\mathcal{C}_{\text{auto}}$ as follows:

$$S(\mathcal{C}_{\text{auto}}) \triangleq -\sum_{j} \lambda_j \log \lambda_j. \tag{5}$$

In quantum information theory [4], von Neumann entropy has been proven as a lower bound for Shannon entropy, which is formulated below:

$$H(\mathbf{Z}) \geq S(\mathcal{C}_{\text{auto}}). \tag{6}$$

III. Experiments

While the efficacy of $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$ as representation regularizers has been substantiated for various tasks, including domain generalization, meta-learning, self-supervised learning, and GAN, by previous work [2], their reliability in quantifying representation entropy has not been probed. Additionally, their effectiveness in enhancing fine-tuning performance has not yet been examined.

A. Reliability in Quantifying Representation Entropy

To investigate the reliability of the two entropy metrics in quantifying the representation entropy, we conduct experiments where the true entropy values are known.

To generate representation with known entropy values, we train ResNet-18 models until training loss converges toward zero. In this case, by neural collapse phenomenon [6], within-class variability of last-layer training representation collapses to zero and the individual representations collapse to their class-mean vectors. Consequently, the entropy of the representation can be computed using the following formula:

$$H(\mathbf{Z}) = \sum_{c=1}^{K} -p_c \log p_c, \tag{7}$$

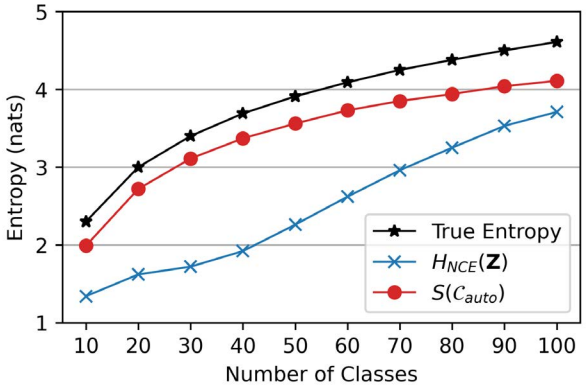

Fig. 1: Effectiveness of the two entropy metrics in quantifying representation entropy. ResNet-18 models are trained with a subset of various sizes of classes in CIFAR-100 and until training loss converges to zero. Both $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$ values are computed for the penultimate representation.

where p_c represents the proportion of class c within the data distribution encompassing K classes. It is important to note that when the representation matrix $\mathbf{Z}$ is generated from a dataset with uniformly distributed classes, the entropy $H(\mathbf{Z})$ becomes equal to $\log(K)$.

In Fig. 1, the computed values of $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$ for the representation matrix $\mathbf{Z}$ are shown alongside the true entropy values computed using Eq. 7. The results present that both $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$ consistently underestimate their respective true entropy values, aligning with the theoretical bounds outlined in Eq. 3 and Eq. 6, respectively. Furthermore, the findings suggest that $S(\mathcal{C}_{\text{auto}})$ potentially serves as a more accurate estimator compared to $H_{NCE}(\mathbf{Z})$, as it demonstrates a closer alignment with the true entropy values (with respective average biases of 1.32 nats for $H_{NCE}(\mathbf{Z})$ and 0.37 nats for $S(\mathcal{C}_{\text{auto}})$).

B. Impact of Increasing Representation Entropy During Pre-Training Phase on the Performance of Fine-Tuning In Low Data Regime

To investigate the impact of increasing representation entropy during pre-training phase on the performance of fine-tuning in low data regime, a comprehensive analysis is conducted. This analysis involves the division of both the CIFAR-100 and ImageNet-100 datasets into two distinct subsets, where the first 25 classes are utilized for pre-training, and the last 75 classes are utilized for fine-tuning.

During pre-training phase, ResNet-18 models are trained from scratch using a subset of the first 25 classes. Within this context, three distinct loss functions are taken into account. As a baseline, the $\mathcal{L}_{\text{CrossEntropy}}$ loss is employed. In order to increase the representation entropy, auxiliary losses in the form of $H_{NCE}(\mathbf{Z})$ or $S(\mathcal{C}_{\text{auto}})$ are incorporated, with their formulations given by:

TABLE I: Impact of increasing representation entropy on the performance of the pre-trained models. ResNet-18 models are pre-trained by three different losses with a subset of the first 25 classes in CIFAR-100 and ImageNet-100.

Pre-train loss	CIFAR-100	ImageNet-100
$\mathcal{L}_{\mathsf{CrossEntropy}}$	88.36	87.36
$\mathcal{L}_{\mathsf{Shannon}}$	86.96	87.52
Diff.	-1.40	0.16
$\mathcal{L}_{\mathsf{vonNeumann}}$	**88.32**	**88.08**
Diff.	-0.04	0.72

$$\mathcal{L}_{\mathsf{Shannon}} = \mathcal{L}_{\mathsf{CrossEntropy}} - \gamma_{\mathsf{Shannon}} \cdot H_{NCE}(\mathbf{Z}), \quad (8)$$

$$\mathcal{L}_{\mathsf{vonNeumann}} = \mathcal{L}_{\mathsf{CrossEntropy}} - \gamma_{\mathsf{vonNeumann}} \cdot S(\mathcal{C}_{\text{auto}}), \quad (9)$$

where $\gamma_{\mathsf{Shannon}}$ and $\gamma_{\mathsf{vonNeumann}}$ represent regularization coefficients, and $\mathbf{Z}$ denotes a penultimate representation.

The performance evaluation of the pre-trained models, conducted prior to the fine-tuning phase, is presented in Table I. Compared to the baseline performance achieved with $\mathcal{L}_{\mathsf{CrossEntropy}}$, the results for $\mathcal{L}_{\mathsf{Shannon}}$ exhibit a marginal decline of -0.62% on average. Conversely, the results for $\mathcal{L}_{\mathsf{vonNeumann}}$ demonstrate a subtle enhancement, with an average of +0.34%.

During fine-tuning phase, we employ the pre-trained models and reinitialize the last classification layer to accommodate 75 classes. Subsequently, fine-tuning is conducted across all models, utilizing the $\mathcal{L}_{\mathsf{CrossEntropy}}$ loss function within the context of low data regime. Specifically, we undertake this process by utilizing randomly sampled subsets comprising 100%, 10%, and 1% of examples. The performance evaluation of the fine-tuned models is presented in Table II.

Primarily, the results highlight that increasing representation entropy demonstrates greater efficacy for fine-tuning, particularly in low data regime. This trend is substantiated by the observation that the difference in performance, compared to the baseline, increases as less examples are utilized for fine-tuning. Specifically, the average improvements for the utilization of 1% of examples yield 4.23% for CIFAR-100 and 5.51% for ImageNet-100, whereas the corresponding average improvements for the utilization of 100% of examples yield 2.27% for CIFAR-100 and 0.92% for ImageNet-100.

Additionally, it is noteworthy that the application of $\mathcal{L}_{\mathsf{vonNeumann}}$ during the pre-training phase consistently yields superior performance compared to the use of $\mathcal{L}_{\mathsf{Shannon}}$. This observation indeed concurs with the findings presented in [2], where the efficacy of regularizing $S(\mathcal{C}_{\text{auto}})$ outperforms that of regularizing $H_{NCE}(\mathbf{Z})$ in the domain generalization tasks.

IV. Conclusion

In this study, we adopt two entropy metrics proposed in [2], denoted by $H_{NCE}(\mathbf{Z})$ and $S(\mathcal{C}_{\text{auto}})$. Our experiments substantiate the efficacy of both metrics in not only quantifying representation entropy but also in enhancing the performance of fine-tuning, particularly in low data regime.

TABLE II: Impact of increasing representation entropy during pre-traing phase on the performance of the fine-tuned models in low data regime. The models pre-trained by three different losses are fine-tuned by a cross-entropy loss with randomly sampled 100%, 10%, and 1% of examples.

Pre-train loss	CIFAR-100			ImageNet-100		
	100%	10%	1%	100%	10%	1%
$\mathcal{L}_{\mathsf{CrossEntropy}}$	70.21	42.80	19.72	78.27	56.00	30.02
$\mathcal{L}_{\mathsf{Shannon}}$	71.56	45.91	23.87	78.57	58.51	34.79
Diff.	1.35	3.11	4.15	0.31	2.51	4.77
$\mathcal{L}_{\mathsf{vonNeumann}}$	**73.40**	**46.60**	**24.03**	**79.80**	**62.17**	**36.27**
Diff.	3.19	3.80	4.31	1.53	6.17	6.25
Avg. Diff.	2.27	3.45	4.23	0.92	4.34	5.51

Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C2007139) and in part by IITP grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

[1] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[2] Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3799–3810, June 2023.

[3] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[4] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[5] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[6] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[7] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[9] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.