TOWARDS ROBUST FEATURE LEARNING WITH T-VFM SIMILARITY FOR CONTINUAL LEARNING

Bilan Gao
Department of Artificial Intelligent
Chung-Ang University
Seoul, Korea
bilan2020@cau.ac.kr

YoungBin Kim
Department of Imaging Science and Arts
The Graduate School of Advanced Imaging Science,
Multimedia and Film,
Chung-Ang University
Seoul, Korea
ybkim85@cau.ac.kr

ABSTRACT

Continual learning has been developed using standard supervised contrastive loss from the perspective of feature learning. Due to the data imbalance during the training, there are still challenges in learning better representations. In this work, we suggest using a different similarity metric instead of cosine similarity in supervised contrastive loss in order to learn more robust representations. We validate the our method on one of the image classification datasets Seq-CIFAR-10 and the results outperform recent continual learning baselines.

1 INTRODUCTION

Continual learning is the research direction that mainly tackles the problem of catastrophic forgetting while a model learns new data consistently. To mitigate the catastrophic forgetting phenomenon, one strategy is to enhance the learned representation while training. An effective approach for achieving this is to apply supervised contrastive learning (Cha et al., 2021; Mai et al., 2021a;b; Han & Guo, 2021; Davari et al., 2022). Neglecting the consideration of the imbalance phenomenon between current training samples and replayed samples can cause loss of intra-class information while learning representations.This is because cosine similarity within the supervised contrastive loss, can extract features with large within-class variance, resulting in degraded model performance. To enhance the stability and reliability of feature extraction, we drew inspiration from (Kobayashi, 2021), we propose to use a similarity function based on the von Mises-Fisher distribution, further extended with a student t-distribution(t-vMF) to replace cosine similarity. We adopt the method on asymmetric supervised contrastive loss(Cha et al., 2021), and conduct experiments under offline continual learning settings, and the results show effective improvement on prediction accuracy.

2 T-VMF SIMILARITY ADOPTED IN CONTINUAL REPRESENTATION LEARNING

An overview of asymmetric supervised contrastive loss L_{ASC}(Cha et al., 2021). A base encoder $f(\cdot)$ receives as input two augmented versions $\tilde{x}_i$ and $\tilde{x}_p$ from data sample x with it's label y. To using an inner product to measure distances, this representations are further mapped to a feature space with projection model $g(\cdot)$. The final extracted features can be written as $z_i = g(f(\tilde{x}_i))$ and $z_p = g(f(\tilde{x}_p))$. Differ from (Mai et al., 2021a), L_{ASC} selects anchor samples only from current training data rather than consider all the samples in the training stage. Let S denotes all the indices of the current learning samples in a single batch and $P(i)$ denotes the indices of all positive samples that are distinguished from i. L_{ASC} can be defined as follows:

$$L_{ASC} = \sum_{i \in S} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(sim(z_i \cdot z_p)/\tau)}{\sum_{a \in A(i)} exp(sim(z_i \cdot z_a)/\tau)}), \tag{1}$$

where $sim(z_i \cdot z_p)$, a cosine similarity function can be written as $= \|z_i\| \|z_p\| cos\theta$ with an angle θ between z_i and z_p. And τ is a temperature hyperparameter. $P(i)$ denotes the indices of all positive

Table 1: Classification accuracy for Seq-CIFAR-10. All results are averaged over ten independent trials. Best performance are marked in bold.

Buffer	Baselines	Class Incremental	Task Incremental
200	HAL (Chaudhry et al., 2020)	32.36 ± 2.70	82.51 ± 3.20
	DER (Buzzega et al., 2020)	61.93 ± 1.79	91.40 ± 0.92
	$DER++$ (Buzzega et al., 2020)	64.88 ± 1.17	91.92 ± 0.60
	Co^2L (Cha et al., 2021)	65.57 ± 1.37	93.43 ± 0.78
	Ours	$\mathbf{67.98 \pm 1.29}$	$\mathbf{95.10 \pm 0.42}$
500	HAL(Chaudhry et al., 2020)	41.79 ± 4.46	84.54 ± 2.36
	DER (Buzzega et al., 2020)	70.51 ± 1.67	93.40 ± 0.39
	$DER++$ (Buzzega et al., 2020)	72.70 ± 1.17	93.88 ± 0.60
	Co^2L (Cha et al., 2021)	$\mathbf{74.26 \pm 0.77}$	$\mathbf{95.90 \pm 0.26}$
	Ours	71.68 ± 0.41	95.51 ± 0.19

samples that are distinguish from i; it is equivalent to $\{p \in A(i) : \tilde{y}_p = \tilde{y}_i\}$. $|P(i)|$ is the number of elements in set $P(i)$.

We propose to use the t-vMF Similarity(Kobayashi, 2021) to replace the cosine similarity in equation1. We first build a von Mises–Fisher distribution(mar; Banerjee et al., 2005) based similarity $p(\tilde{z}_i; \tilde{z}_p, \kappa) = C_\kappa exp(\kappa cos\theta)$ between features $\tilde{z}_i$ and $\tilde{z}_p$. It produces a probability density function on p dimension with with parameter κ. The parameter κ controls concentration of the distribution, and C_κ is a normalization constant. Apply a t-distributed profile function $f_t(d; \kappa) = \frac{1}{1+\frac{1}{2}\kappa d^2}$ on $p(\tilde{z}_i; \tilde{z}_p, \kappa)$, the equation of t-vFM similarity function between two features $\tilde{z}_i$ and $\tilde{z}_p$ is formalized as follows(More detail in Appendix A.1):

$$\phi_t(cos\theta; \kappa) == \frac{1 + cos\theta}{1 + \kappa(1 - cos\theta)} - 1 \tag{2}$$

Adopt equation2 in the original L_{ASC}, the loss function of our proposed method is shown as follows:

$$L_{Ours} = \sum_{i \in S} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(\phi_t(cos\theta; \kappa)/\tau)}{\sum_{a \in A(i)} exp(\phi_t(cos\theta; \kappa)/\tau)}). \tag{3}$$

The larger value κ has, that smaller compact region between two representation is.

3 Experiment and Discussion

Following prior work (Cha et al., 2021), we demonstrate the our method by training ResNet-18 (He et al., 2016) as a base encoder on Seq-CIFAR-10 dataset (Krizhevsky et al., 2009). We validate our method on class incremental learning and task incremental learning scenarios with replay buffers of size 200 and 500 (Buzzega et al., 2020).The value of κ in Equation 3 is set to 16.(More detail in Appendix A.2). As shown in Table 1, our proposed method outperforms all the baselines with buffer 200, which indicates that the t-vMF similarity shows promise. In the case of buffer size 500, even though the proposed method does not outperform all baselines, it approaches the best baselines.

4 Conclusion and Future Work

In this work, we delve deeply into continual representation learning, using t-vMF similarity to enhance feature reliability and outperform most baselines in terms of prediction accuracy. Our future aim is to further investigate alternative representation learning approaches and various metrics to improve continual representation learning.

ACKNOWLEDGEMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2022R1C1C1008534), and Institute for Information & communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program, Chung-Ang University).

URM STATEMENT

Author Bilan Gao meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Directional statistics. URL http://doi.wiley.com/10.1002/9780470316979.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(46): 1345–1382, 2005. URL http://jmlr.org/papers/v6/banerjee05a.html.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9516–9525, October 2021.

Arslan Chaudhry, Albert Gordo, David Lopez-Paz, Puneet K. Dokania, and Philip Torr. Using hindsight to anchor past knowledge in continual learning, 2020. URL https://openreview.net/forum?id=Hke12T4KPS.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16712–16721, June 2022.

Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 154–162, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/gopal14.html.

Xuejun Han and Yuhong Guo. Contrastive continual learning with feature propagation. *arXiv preprint arXiv:2112.01713*, 2021.

Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, Stéphane Gentric, and Liming Chen. von mises-fisher mixture model-based deep learning: Application to face verification. *ArXiv*, abs/1706.04264, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. doi: 10.1109/CVPR.2016.90. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90.

Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7334–7344, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Takumi Kobayashi. T-vmf similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6616–6625, June 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3589–3599, June 2021a.

Zheda Mai, Ruiwen Li, Hyunwoo J. Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3584–3594, 2021b.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Pengfei Wang, Daqing Wu, Chong Chen, Kunpeng Liu, Yanjie Fu, Jianqiang Huang, Yuanchun Zhou, Jianfeng Zhan, and Xiansheng Hua. Deep adaptive graph clustering via von mises-fisher distributions. *ACM Transactions on the Web*, 2023.

Xuefei Zhe, Shifeng Chen, and Hong Yan. Directional statistics-based deep metric learning for image classification and retrieval. *Pattern Recognition*, 93:113–123, 2019. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2019.04.005. URL https://www.sciencedirect.com/science/article/pii/S0031320319301451.

A Appendix

A.1 More detail explanation on T-vFM similarity

T-vFM similarity has the foundation on a similarity function that based von Mises–Fisher distribution(mar; Banerjee et al., 2005), a von Mises–Fisher distributed similarity which produces a probability density function on p dimension with parameter κ between two features $\tilde{z}_i$ and $\tilde{z}_p$ in d dimension can be formulated as:

$$p(\tilde{z}_i; \tilde{z}_p, \kappa) = C_\kappa exp(\kappa \tilde{z}_p^\top \tilde{z}_i) \tag{4}$$

$$= C_\kappa exp(\kappa cos\theta). \tag{5}$$

The parameter κ controls concentration of the distribution, and C_κ is a normalization constant. To find out the part that mainly support the distribution $p(\tilde{z}_i; \tilde{z}_p, \kappa)$, define a profile function f_e with d and κ, then the profile function $f_e(d; \kappa)$ is defined as

$$f_e(d; \kappa) = exp(-\frac{1}{2}\kappa d^2). \tag{6}$$

Thus, Equation 5 can be reformulated by using a profile function(Equation 6) as follows

$$p(\tilde{z}_i; \tilde{z}_p, \kappa) = C_\kappa exp(\kappa - \frac{1}{2}\kappa \|\tilde{z}_i - \tilde{z}_p\|^2) \tag{7}$$

$$= C_\kappa^{'} f_e(\|\tilde{z}_i - \tilde{z}_p\|; \kappa). \tag{8}$$

Observing from Equation 8, the vMF similarity is primarily distinguished by $f_e(\|\tilde{z}_i - \tilde{z}_p\|; \kappa)$, which means the vMF similarity can formally be re-defined as an alternative for $cos\theta$ as

$$\phi_e(cos\theta; \kappa) = 2\frac{f_e(|\tilde{z}_i - \tilde{z}_p\|; \kappa) - f_e(2; \kappa)}{f_e(0; \kappa) - f_e(2; \kappa)} - 1 \tag{9}$$

$$= 2\frac{exp(\kappa cos\theta) - exp(-\kappa)}{exp(\kappa) - exp(-\kappa)} - 1 \in [1, -1]. \tag{10}$$

Equation 10 demonstrates von Mises–Fisher distributed similarity between two features $\tilde{z}_i$ and $\tilde{z}_p$ with an important parameter κ. According to its function that introduced above, changing different value of κ can dynamically balance the trade-off between learning discriminative features for each task and minimizing interference between new task samples and learned task samples.
A previous work Van der Maaten & Hinton (2008) suggested using a heavy tail distribution to better capture distinguishing features. Modify the profile function(Equation 6) with a heavy-tailed student T distribution, the new profile function can be updated as follows:

$$f_t(d; \kappa) = \frac{1}{1 + \frac{1}{2}\kappa d^2}. \tag{11}$$

Due to the property of t-distribution, it makes the similarity function less sensitive to small changes in the learned representations. We reformulate vMF similarity with Equation 10, the formal definition of t-vFM similarity between features $\tilde{z}_i$ and $\tilde{z}_p$ is given as follows:

$$\phi_t(cos\theta; \kappa) = 2\frac{f_t(|\tilde{z}_i - \tilde{z}_p\|; \kappa) - f_t(2; \kappa)}{f_t(0; \kappa) - f_t(2; \kappa)} - 1, \tag{12}$$

$$= 2\frac{\frac{1}{1+\kappa(1-cos\theta)} - \frac{1}{1+2\kappa}}{1 - \frac{1}{1+2\kappa}} - 1, \tag{13}$$

$$= \frac{1 + cos\theta}{1 + \kappa(1 - cos\theta)} - 1. \tag{14}$$

As shown in Equation 14, t-vFM similarity can be simply calculated with $cos\theta$ and κ. Since $cos\theta \in [-1, +1]$, $f_e(\|\tilde{z}_i - \tilde{z}_p\|; \kappa)$ is resized to be adaptable with $cos\theta$. Applying Equation 14 to asymmetric contrastive loss, our proposed loss can be modified from Equation 1 as follows:

$$L_{Ours} = \sum_{i \in S} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(\phi_t(cos\theta; \kappa)/\tau)}{\sum_{a \in A(i)} exp(\phi_t(cos\theta; \kappa)/\tau)}). \tag{15}$$

A.2 Experimental Details

Dataset and architecture In the training stage, we split the CIFAR-10 dataset into five distinct sets of samples,with each set consisting of two different classes. For architecture, we use a non-pretrained ResNet-18 as a base encoder for extracting features followed by a 2-layer projection MLP to map these representations to a latent space of 128 dimensions (Khosla et al., 2020).

Hyperparameters We train the base encoder with a 512 batch size, and 0.5 as learning rate. The temperature hyperparameter(τ) for our loss is set to 0.5.

Evaluation We follow class balance strategyCha et al. (2021) and use an additional linear classifier trained only on the final task samples and the buffer data with learned representations. All the results are reported with 10 trials.

Additional experiment on different values of κ As shown in Table 2, we report results with different κ values, and our algorithm surpasses all the baselines of recent works in the case of buffer 200 setting. In the comparison of 3 different values of κ, class incremental learning reaches the best performance when $\kappa = 32$, and task incremental learning achieves the best performance when $\kappa = 4$. Such results indicate that our method successfully learns and reduces the effect of data imbalance. Moreover, there are no big difference among all the results with different values of κ, which indicates a narrow region does not effect the performance in this case.

Table 2: Classification accuracy for Seq-CIFAR-10. All results are averaged over ten independent trials. Best performance are marked in bold. Our proposed method choose three values of κ: 4,16,32

Buffer	Baselines	Class Incremental	Task Incremental
200	HAL (Chaudhry et al., 2020)	32.36 ± 2.70	82.51 ± 3.20
	DER (Buzzega et al., 2020)	61.93 ± 1.79	91.40 ± 0.92
	$DER++$ (Buzzega et al., 2020)	64.88 ± 1.17	91.92 ± 0.60
	Co^2L (Cha et al., 2021)	65.57 ± 1.37	93.43 ± 0.78
	$Ours(\kappa = 4)$	67.94 ± 1.04	$\mathbf{95.20 \pm 0.22}$
	$Ours\ (\kappa = 16)$	67.98 ± 1.29	95.10 ± 0.42
	$Ours\ (\kappa = 32)$	$\mathbf{67.99 \pm 0.71}$	95.14 ± 0.38
500	HAL(Chaudhry et al., 2020)	41.79 ± 4.46	84.54 ± 2.36
	DER (Buzzega et al., 2020)	70.51 ± 1.67	93.40 ± 0.39
	$DER++$ (Buzzega et al., 2020)	72.70 ± 1.17	93.88 ± 0.60
	Co^2L (Cha et al., 2021)	$\mathbf{74.26 \pm 0.77}$	$\mathbf{95.90 \pm 0.26}$
	$Ours\ (\kappa = 4)$	71.51 ± 0.84	95.37 ± 0.18
	$Ours\ (\kappa = 16)$	71.68 ± 0.48	95.51 ± 0.19
	$Ours\ (\kappa = 32)$	71.32 ± 0.70	95.43 ± 0.26

A.3 Related works

Contrastive Continual Learning Looking at continual learning from the perspective of feature learning, the former can be broken down into following two main questions: How can the learning features be robust and more reliable? How can the learned features be maintained? There are many recent works that focused on apply contrastive loss to leverage more generalized presentations. Self-supervised contrastive loss(Chen et al., 2020) was adopted in these works (**??**), which increased models' performance without using labels. Supervised contrastive loss(Khosla et al., 2020) was used in many recent works(Mai et al., 2021a; Cha et al., 2021; Han & Guo, 2021; Davari et al., 2022) to make samples that belong to the same class get closer to each other. Supervised contrastive loss was used in continual learning in the study presented by (Mai et al., 2021a), then feature propagation was introduced with supervised contrastive loss to tackle multiple continual learning tasks(Han & Guo, 2021). In (Davari et al., 2022), controlling feature forgetting was suggested using a linear classifier before and after a new task started. Another recent work(Cha et al., 2021) improved supervised contrastive loss for better feature learning, and it also introduced a relation distillation to maintain the learned representation. There haven't been many works that consider a better approach on calculating similarity under the data imbalance environment. In our work, we investigate a different type of similarity function that adopts in asymmetric supervised contrastive loss(Cha et al., 2021) to diminish intra-class variance.

von Mises-Fisher Distribution In recent years, von Mises–Fisher Distribution(mar) is one of the probability distributions on sphere that has been applied to many fields in deep learning. One of the previous works(Gopal & Yang, 2014) presented a von Mises–Fisher distribution-based clustering method, which is more effective on high-dimensional data. Different types of von Mises–Fisher distribution based clustering approaches have been adopted for different tasks such as semantic segmentation(Hwang et al., 2019) and graph neural network(Wang et al., 2023). Using von Mises–Fisher distribution on metric learning or loss functions (Hasnat et al., 2017; Zhe et al., 2019; Kobayashi, 2021) showed significant improvement in performance as well.