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

Anonymous authors
Paper under double-blind review

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 a type of research that helps a single model learn new tasks consistently, with less forgetting of the learned knowledge. The catastrophic forgetting problem in continual learning can be broken down into two main questions: How to learn representations that are meaningful and reliable, and how to keep learned features for future training stages? Standard supervised contrastive loss has been used or modified in recent works (Cha et al., 2021; Mai et al., 2021a;b; Han & Guo, 2021; Davari et al., 2022) so that it can learn better features from the point of view of representation learning. However, the data imbalance problem has not been considered when the standard supervised contrastive loss is used. Motivated by a recent work on feature learning (Kobayashi, 2021), we propose replacing the original cosine similarity function with a similarity function based on the von Mises-Fisher distribution, further extended with a student t-distribution(t-vMF). This allows learning more robust and useful representations.

2 T-VMF SIMILARITY ADOPTED IN CONTINUAL REPRESENTATION LEARNING

Given the representations z_i and z_p extracted from a base encoder, the similarity between them in the supervised contrastive loss can be calculated as

$$sim(z_i \cdot z_p) = \|z_i\|\|z_p\|cos\theta, \tag{1}$$

where the angle between z_i and z_p is θ. To formulate the similarity in a von Mises-Fisher (mar; Banerjee et al., 2005) manner, we let the angle θ be a core metric on a unit hyper-sphere. The representation are displayed with von Mises-Fisher(vMF) distribution as

$$\begin{aligned} p(\tilde{z}_i; \tilde{z}_p, \kappa) &= C_\kappa exp(\kappa \tilde{z}_p^\top \tilde{z}_i) \\ &= C_\kappa exp(\kappa cos\theta), \end{aligned} \tag{2}$$

where the normalization constant is denoted as C_κ. And κ is a concentration parameter that controls the concentration of the distribution. Equation 2 can be reformulated by using a profile function $f_e(d;\kappa) = exp(-\frac{1}{2}\kappa d^2)$. Then, vMF similarity between the two

$$\phi_e(cos\theta;\kappa) = 2\frac{f_e(|\tilde{z}_i - \tilde{z}_p\|;\kappa) - f_e(2;\kappa)}{f_e(0;\kappa) - f_e(2;\kappa)} - 1 \tag{3}$$

$$= 2\frac{exp(\kappa cos\theta) - exp(-\kappa)}{exp(\kappa) - exp(-\kappa)} - 1 \in [1, -1]. \tag{4}$$

Table 1: Classification accuracy for Seq-CIFAR-10. All results are averaged over ten independent trials. The performance of the our method is 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)	74.26 ± 0.77	95.90 ± 0.26
	Ours	$\mathbf{71.68 \pm 0.41}$	$\mathbf{95.10 \pm 0.42}$

Previous work Van der Maaten & Hinton (2008) suggest to use a heavy tail distribution to better capturing distinguishing features. With a new heavy-tailed student T distribution profile function $f_e(d;\kappa) = exp(-\frac{1}{2}\kappa d^2)$, the t-vMF similarity is formulated as follows:

$$\phi_t(cos\theta;\kappa) = 2\frac{f_t(|\tilde{z}_i - \tilde{z}_p\|;\kappa) - f_t(2;\kappa)}{f_t(0;\kappa) - f_t(2;\kappa)} - 1, \tag{5}$$

$$= 2\frac{\frac{1}{1+\kappa(1-cos\theta)-\frac{1}{1+2\kappa}}}{1-\frac{1}{1+2\kappa}} - 1, \tag{6}$$

$$= \frac{1+cos}{1+\kappa(1-cos\theta)} - 1. \tag{7}$$

We implement the t-vMF similarity in the asymmetric supervised contrastive loss Cha et al. (2021), which counts the previous task samples as negative samples in the next learning stage.

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 to extract features for the 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). We choose $\kappa = 16$. For evaluation, we use an additional linear classifier trained only on the final task samples and the buffer data with learned representations. As shown in Table 1, the 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 examine a different perspective of learning representations in continual learning. The initial results show promise for strengthening feature reliability by using the t-vMF similarity instead of the normal cosine similarity. Moreover, the proposed method outperforms most of the baselines in prediction accuracy. Finally, we aim to investigate other representation learning approaches to aid feature learning in continual learning.

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.

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

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.

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.

A APPENDIX

A.1 ASYMMETRIC SUPERVISED CONTRASTIVE LOSS

Self-supervised contrastive loss One of the most popular contrastive losses is introduced by the work Chen et al. (2020). Given sample data x and its label y, create two random augmentation views of sample x, denoted as $\tilde{x}_i$ and $\tilde{x}_p$. Only the augmented view of data can be counted as a positive sample. We get representation vectors h_i and h_p by putting augmented data samples into an encoder network $h_i = f(\tilde{x}_i)$. Then, there is a projection model $g(\cdot)$ that maps the representation vectors into the feature space; get $z_i = g(h_i)$ and $z_p = g(h_p)$. Thus, the loss functions between two positive samples (anchor i and its positive $p(i)$) can be shown as follows:

$$L_{ssc} = -\sum log \frac{exp(sim(z_i \cdot z_{p(i)})/\tau)}{\sum_{a \in A(i)} exp((sim(z_i \cdot z_a)/\tau)}), \tag{8}$$

where $sim(z_i \cdot z_p)$ is cosine similarity between z_i and z_p, τ is a temperature hyper-parameter, and $A(i) = \frac{I}{\{i\}}$.

Supervised contrastive loss Supervised contrastive loss was recently introduced (Khosla et al., 2020). Differing from self-supervised contrastive loss, supervised contrastive loss function consider samples with same class label as positive samples as well. For ease of notation and understanding, we display the formula with the same notation as above. The formula for the supervised contrastive loss is defined as follows:

$$L_{sc} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(sim(z_i \cdot z_p)/\tau)}{\sum_{a \in A(i)} exp((sim(z_i \cdot z_a)/\tau)}), \tag{9}$$

where $P(i)$ denotes the indices of all positive 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 $P(i)$.

Asymmetric supervised contrastive loss Based on the supervised contrastive loss, Asymmetric supervised contrastive loss uses samples from the current task only as anchors and uses the previous task sample buffer as negative samples. The asymmetric supervised contrastive loss can be defined as follows:

$$L_{sc} = \sum_{i \in S} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(sim(z_i \cdot z_p)/\tau)}{\sum_{a \in A(i)} exp((sim(z_i \cdot z_a)/\tau)}). \tag{10}$$

S in equation 10 is denoted as all the indices of the current learning samples in a single batch.