Supplementary Material for *Alignment-Uniformity aware Representation Learning for Zero-shot Video Classification*

Shi Pu[1*] Kaili Zhao[2*] Mao Zheng[1]
[1]Tencent AI Platform Department [2]Beijing University of Posts and Telecom.
{shipu, moonzheng}@tencent.com kailizhao@bupt.edu.cn

Abstract

This material includes full derivations that cannot be fitted to the main paper due to the limited space. In specific, we first clarify the formulation of $\mathcal{L}^{self}$ in the main paper, and then derive the upper bounds of $\mathcal{L}^{self}$ and $\mathcal{L}^{sup}$, at last, illustrate how the visual centers w_{y_k} approach visual features v_{y_k}.

1. The formulation of $\mathcal{L}^{self}$ in the main paper

Reviewing the literature in self-supervised learning, we observe that most works [2, 5, 7] formulate the original self-supervised contrastive loss as follows:

$$\mathcal{L}_{\text{ori}} = -\log\left[\frac{\exp\left[\lambda \text{sim}(f_i, f_j)\right]}{\sum_{k=1}^{2N} \mathbb{1}_{k\neq i} \exp\left[\lambda \text{sim}(f_i, f_k)\right]}\right]. \tag{a}$$

Given N and the augmented samples (*i.e.*, overall 2N samples), there are 1 positive pair in the numerator, the other 1 positive and $2(N-1)$ negatives in the denominator. Even there is one positive pair included in the denominator of $\mathcal{L}_{\text{ori}}$, the methods [3, 11] consider that only the $2(N-1)$ negatives contribute to the uniformity property, and propose that the positive pair in the numerator relates with the alignment property. Thus, when discussing the two properties, we formulate $\mathcal{L}^{self}$ as Eq. 1 of the main paper, which removes the positive pair in the denominator. Furthermore, [12] justifies that optimizing the $\mathcal{L}^{self}$ even with a small batch size is comparable with $\mathcal{L}_{\text{ori}}$ that requires a large batch size for allocating enough negatives. Thus, we set the latest $\mathcal{L}^{self}$ as our objectives in Section 3.2 of the main paper.

When diving into the supervised contrastive loss $\mathcal{L}^{sup}$, we observe existing works, MUFI [8] and ER [1], neglect the similarities and differences between $\mathcal{L}^{sup}$ and $\mathcal{L}^{self}$. To clarify the superiority of $\mathcal{L}^{sup}$, we derive the upper bounds of the two losses, and summarize the advantages of $\mathcal{L}^{sup}$ in Section 3.2 of the main paper. To sum up, we justified that $\mathcal{L}^{sup}$ is more feasible for zero-shot video classification.

2. Upper bounds of $\mathcal{L}^{self}$ and $\mathcal{L}^{sup}$

In Section 3.2 of the main paper, we present the upper bounds of $\mathcal{L}^{self}$ and $\mathcal{L}^{sup}$. In this section, we perform the full derivations which are based on the upper bounds of LSE and SP_λ in Eq. 2 in the main paper:

$$\begin{aligned} \text{LSE}(x) &= \log(\sum_{x\in\mathcal{X}} \exp(x)), \\ &\leq \log(n \exp(\max_{x\in\mathcal{X}}(x))), \\ &= \max_{x\in\mathcal{X}}(x) + \log(n), \end{aligned} \tag{b}$$

*These authors contributed equally.

$$\begin{aligned}\text{SP}_\lambda(x) &= \frac{1}{\lambda}\log(1+\exp(\lambda x)),\\ &= \frac{1}{\lambda}\text{LSE}(\lambda y_{y\in\{x,0\}}),\\ &\leq \max[x,0] + \frac{\log(2)}{\lambda},\end{aligned} \tag{c}$$

where n is the number of x in $\mathcal{X}$. We derive the upper bound of $\mathcal{L}^{self}$ as follow:

$$\begin{aligned}\mathcal{L}^{self} &= -\log[\frac{\exp[\lambda\text{sim}(v_{y_i}, s_{y_i})]}{\sum_{y_j\in\mathcal{Y}\setminus y_i}\exp[\lambda\text{sim}(v_{y_i}, s_{y_j})]}],\\ &= \lambda(-\text{sim}(v_{y_i}, s_{y_i}) + \frac{1}{\lambda}\text{LSE}(\lambda\text{sim}(v_{y_i}, s_{y_j})_{y_j\in\mathcal{Y}\setminus y_i})),\\ &\leq \lambda(\text{sim}_{\max} - \text{sim}(v_{y_i}, s_{y_i}) + \frac{\log(K-1)}{\lambda}),\end{aligned} \tag{d}$$

where K is the number of y in $\mathcal{Y}$, $\text{sim}_{\max} = \max_{y_j\in\mathcal{Y}\setminus y_i}\text{sim}(v_{y_i}, s_{y_j})$. The upper bound of $\mathcal{L}^{sup}$ is derived as:

$$\begin{aligned}\mathcal{L}^{sup} &= -\log[\frac{\exp[\lambda\text{sim}(v_{y_i}, s_{y_i})]}{\sum_{y_j\in\mathcal{Y}}\exp[\lambda\text{sim}(v_{y_i}, s_{y_j})]}],\\ &= \lambda\text{SP}_\lambda[-\text{sim}(v_{y_i}, s_{y_i}) + \frac{1}{\lambda}\text{LSE}(\lambda\text{sim}(v_{y_i}, s_{y_j})_{y_j\in\mathcal{Y}\setminus y_i})],\\ &\leq \lambda\text{SP}_\lambda[\text{sim}_{\max} - \text{sim}(v_{y_i}, s_{y_i}) + \frac{\log(K-1)}{\lambda}],\\ &\leq \lambda\max[\text{sim}_{\max} - \text{sim}(v_{y_i}, s_{y_i}) + \frac{\log(K-1)}{\lambda}, 0] + \log(2).\end{aligned} \tag{e}$$

3. Visual centers W

In Section 3.3 of the main paper, we use Eq. 5 in the main paper to learn the visual centers $W = [w_{y_1}, \ldots, w_{y_K}]$ which is the parameter matrix of a linear classifier without biases. The parameter vector w_{y_k} from the linear classifier can be interpreted as the class representation of the class y_k [4, 6]. In this section, we justify how the visual centers approach visual features on a unit hypersphere during back-propagation. For convenience, we reprint Eq. 5 as follow:

$$\begin{aligned}\mathcal{L}_\text{C} &= -\log\frac{\exp[\lambda\cos(v_{y_i}, w_{y_i})]}{\sum_{y_j\in\mathcal{Y}}\exp[\lambda\cos(v_{y_i}, w_{y_j})]},\\ &= -\log\frac{\exp[\lambda\frac{v_{y_i}}{\|v_{y_i}\|}\times\frac{w_{y_i}}{\|w_{y_i}\|}]}{\sum_{y_j\in\mathcal{Y}}\exp[\lambda\frac{v_{y_i}}{\|v_{y_i}\|}\times\frac{w_{y_j}}{\|w_{y_j}\|}]}.\end{aligned} \tag{f}$$

Then we derive the gradient of $\mathcal{L}_C$ with respect to $\frac{w_{y_k}}{\|w_{y_k}\|}$ as follow:

$$\frac{\partial\mathcal{L}_\text{C}}{\partial\frac{w_{y_k}}{\|w_{y_k}\|}} = \begin{cases}\lambda(P_{ik}-1)\frac{v_{y_i}}{\|v_{y_i}\|}, & i = k\\ \lambda P_{ik}\frac{v_{y_i}}{\|v_{y_i}\|}, & i \neq k\end{cases}, \tag{g}$$

where, $P_{ik} = \frac{\exp[\lambda\cos(v_{y_i}, w_{y_k})]}{\sum_{y_j\in\mathcal{Y}}\exp[\lambda\cos(v_{y_i}, w_{y_j})]} \in [0, 1]$. During the back-propagation, $\mathcal{L}_\text{C}$ encourages that changing $\frac{w_{y_k}}{\|w_{y_k}\|}$ to $\frac{\bar{w}_{y_k}}{\|\bar{w}_{y_k}\|} = \frac{w_{y_k}}{\|w_{y_k}\|} - l\cdot\frac{\partial\mathcal{L}_\text{C}}{\partial\frac{w_{y_k}}{\|w_{y_k}\|}}$ where l is the learning rate. We compute $\cos(v_{y_i}, \bar{w}_{y_k})$ as follow:

$$\cos(v_{y_i}, \bar{w}_{y_k}) = \begin{cases}\cos(v_{y_i}, w_{y_k}) + l\cdot\lambda(1-P_{ik}), & i = k\\ \cos(v_{y_i}, w_{y_k}) - l\cdot\lambda P_{ik}, & i \neq k\end{cases}, \tag{h}$$

Eq. h shows that $\frac{w_{y_k}}{\|w_{y_k}\|}$ approaches the visual feature $\frac{v_{y_k}}{\|v_{y_k}\|}$ (i.e., $\cos(v_{y_k}, \bar{w}_{y_k}) \geq \cos(v_{y_k}, w_{y_k})$) and stays away from $\frac{v_{y_i}}{\|v_{y_i}\|}, i \neq k$ (i.e., $\cos(v_{y_i}, \bar{w}_{y_k}) \leq \cos(v_{y_i}, w_{y_k})$) during the back-propagation. After a number of training iterations, we can treat $\frac{w_{y_k}}{\|w_{y_k}\|}$ as the visual center of all $\frac{v_{y_k}}{\|v_{y_k}\|}$ even if it may not be exactly the mean of all $\frac{v_{y_k}}{\|v_{y_k}\|}$ due to the effect of hard positive/negative samples to P_{ik} [9, 10].

References

[1] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, 2021. 1

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1

[3] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *arXiv preprint arXiv:2011.02803*, 2020. 1

[4] Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. Explicit interaction model towards text classification. In *AAAI*, 2019. 2

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[6] Ofir Press and Lior Wolf. Using the output embedding to improve language models. *arXiv:1608.05859*, 2016. 2

[7] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *CVPR*, 2021. 1

[8] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xiao-Ping Zhang, Dong Wu, and Tao Mei. Boosting video representation learning with multi-faceted integration. In *CVPR*, 2021. 1

[9] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, pages 926–930, 2018. 3

[10] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 3

[11] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 1

[12] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. *arXiv:2110.06848*, 2021. 1