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

Shi Pu\textsuperscript{1*} Kaili Zhao\textsuperscript{2*} Mao Zheng\textsuperscript{1}
\textsuperscript{1}Tencent AI Platform Department \textsuperscript{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 $L_{\text{self}}$ in the main paper, and then derive the upper bounds of $L_{\text{self}}$ and $L_{\text{sup}}$, at last, illustrate how the visual centers $w_{yk}$ approach visual features $v_{yk}$.

1. The formulation of $L_{\text{self}}$ 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:

$$L_{\text{ori}} = -\log\left(\frac{\exp[\lambda \text{sim}(f_i, f_j)]}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp[\lambda \text{sim}(f_i, f_k)]}\right).$$  \hspace{1cm} (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 $L_{\text{ori}}$, 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 $L_{\text{self}}$ as Eq. 1 of the main paper, which removes the positive pair in the denominator. Furthermore, [12] justifies that optimizing the $L_{\text{self}}$ even with a small batch size is comparable with $L_{\text{ori}}$ that requires a large batch size for allocating enough negatives. Thus, we set the latest $L_{\text{self}}$ as our objectives in Section 3.2 of the main paper.

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

2. Upper bounds of $L_{\text{self}}$ and $L_{\text{sup}}$

In Section 3.2 of the main paper, we present the upper bounds of $L_{\text{self}}$ and $L_{\text{sup}}$. In this section, we perform the full derivations which are based on the upper bounds of LSE and SP\textsubscript{λ} in Eq. 2 in the main paper:

$$L_{\text{SE}}(x) = \log(\sum_{x \in \mathcal{X}} \exp(x)), \hspace{1cm} (b)$$

\begin{align*}
\leq \log(n \exp(\max_{x \in \mathcal{X}}(x))), \\
= \max_{x \in \mathcal{X}}(x) + \log(n),
\end{align*}

*These authors contributed equally.
\[ SP_\lambda(x) = \frac{1}{\lambda} \log(1 + \exp(\lambda x)), \]
\[ = \frac{1}{\lambda} \text{LSE}(\lambda y_x \in \{x, 0\}), \]
\[ \leq \max[x, 0] + \frac{\log(2)}{\lambda}, \]

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

\[
\mathcal{L}^{self} = -\log[\frac{\exp[\lambda \text{sim}(v_{y_i}, s_{y_j})]}{\sum_{y_j \in Y \backslash y_i} \exp[\lambda \text{sim}(v_{y_i}, s_{y_j})]}],
= \lambda(-\text{sim}(v_{y_i}, s_{y_j}) + \frac{1}{\lambda} \text{LSE}(\lambda \text{sim}(v_{y_i}, s_{y_j}) y_j \in Y \backslash y_i)),
\leq \lambda(\text{sim}_{\text{max}} - \text{sim}(v_{y_i}, s_{y_j}) + \frac{\log(K-1)}{\lambda}),
\]

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

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

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:

\[
\mathcal{L}_C = -\log \frac{\exp[\lambda \cos(v_{y_i}, w_{y_i})]}{\sum_{y_j \in Y} \exp[\lambda \cos(v_{y_i}, w_{y_j})]},
= -\log \frac{\exp[\lambda \frac{v_{y_i}}{||v_{y_i}||} \times \frac{w_{y_i}}{||w_{y_i}||}]}{\sum_{y_j \in Y} \exp[\lambda \frac{v_{y_i}}{||v_{y_i}||} \times \frac{w_{y_j}}{||w_{y_j}||}]},
\]

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

\[
\frac{\partial \mathcal{L}_C}{\partial \frac{w_{y_k}}{||w_{y_k}||}} = \begin{cases} \lambda(P_{ik} - 1) \frac{v_{y_i}}{||v_{y_i}||}, & i = k \\ \lambda P_{ik} \frac{v_{y_i}}{||v_{y_i}||}, & i \neq k \end{cases},
\]

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

\[
\cos(v_{y_i}, \bar{w}_{y_k}) = \begin{cases} \cos(v_{y_i}, w_{y_k}) + l \cdot \lambda(1 - P_{ik}), & i = k \\ \cos(v_{y_i}, w_{y_k}) - l \cdot \lambda P_{ik}, & i \neq k \end{cases},
\]
Eq. h shows that $\frac{w_{yk}}{\|w_{yk}\|}$ approaches the visual feature $\frac{v_{yk}}{\|v_{yk}\|}$ (i.e., $\cos(v_{yk}, \bar{w}_{yk}) \geq \cos(v_{yk}, w_{yk})$) and stays away from $\frac{v_{yi}}{\|v_{yi}\|}$, $i \neq k$ (i.e., $\cos(v_{yi}, \bar{w}_{yk}) \leq \cos(v_{yi}, w_{yk})$) during the back-propagation. After a number of training iterations, we can treat $\frac{w_{yk}}{\|w_{yk}\|}$ as the visual center of all $\frac{v_{yk}}{\|v_{yk}\|}$ even if it may not be exactly the mean of all $\frac{v_{yk}}{\|v_{yk}\|}$ due to the effect of hard positive/negative samples to $P_{ik}$ [9, 10].

References