Supplementary Material

Kyle Min and Jason J. Corso

University of Michigan, Ann Arbor, MI 48109
{kylemin,jjcorso}@umich.edu

A Center Update Equation

Notations

- D: Dimension of the embedding (feature) space
- $\mathbf{c}_j^r \in \mathbb{R}^D$: center of activity class j
- $\tilde{\mathbf{c}}_j^r \in \mathbb{R}^D$: l_2-normalized center of activity class j
- $\mathbf{F}_i^r(j) \in \mathbb{R}^D$: video-level feature representation for the activity class j of the i-th training sample
- $\widetilde{\mathbf{F}}_i^r(j) \in \mathbb{R}^D$: l_2-normalized video-level feature representation
- $n_{i,j}^r = \underset{k \neq j}{\operatorname{argmin}} \mathcal{D}\big(\mathbf{F}_i^r(j), \mathbf{c}_k^r\big)$: index for the nearest negative center
- $\mathcal{D}(\cdot)$ represents the angular distance:

$$\begin{aligned}\mathcal{D}\big(\mathbf{F}_i^r(j), \mathbf{c}_j^r\big) &= \arccos\left(\frac{\mathbf{F}_i^r(j) \cdot \mathbf{c}_j^r}{\|\mathbf{F}_i^r(j)\|_2 \|\mathbf{c}_j^r\|_2}\right) \\ &= \arccos\big(\widetilde{\mathbf{F}}_i^r(j) \cdot \tilde{\mathbf{c}}_j^r\big)\end{aligned}$$

- $\mathbf{f}_i^r$: new video-level feature that attends more strongly to the background features than $\mathbf{F}_i^r$ is
- $\tilde{\mathbf{f}}_i^r$: l_2-normalized new video-level feature
- $\mathbf{y}_i \in \mathbb{R}^{N_c}$: ground-truth annotation for video-level label of the i-th sample where N_c is the number of activity classes. $\mathbf{y}_i(j) = 1$ if the activity class j is present in the sample and $\mathbf{y}_i(j) = 0$ otherwise

Previous works on center loss [3, 1, 2] suggest using an averaged gradient (typically denoted as $\Delta\mathbf{c}_j^r$) to update the centers for better stability. In this section, we show how we update the centers based on this convention. For simplicity, we first look at the RGB stream.

As mentioned in the paper, $\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}}$ and $\tilde{\mathcal{L}}^r_{\mathrm{NT}_{i,j}}$ are the loss terms inside the max operation of the i-th sample and of the j-th activity class as follows:

$$\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}} = \mathcal{D}\big(\mathbf{F}_i^r(j), \mathbf{c}_j^r\big) - \mathcal{D}\big(\mathbf{F}_i^r(j), \mathbf{c}_{n_{i,j}^r}^r\big) + m_1 \tag{1}$$

$$\tilde{\mathcal{L}}^r_{\mathrm{NT}_{i,j}} = \mathcal{D}\big(\mathbf{F}_i^r(j), \mathbf{c}_j^r\big) - \mathcal{D}\big(\mathbf{f}_i^r(j), \mathbf{c}_j^r\big) + m_2 \tag{2}$$

Let $\mathbf{g}^r_{1_{i,j}}$ and $\mathbf{g}^r_{2_{i,j}}$ be the derivatives of Eq. 1 with respect to $\tilde{\mathbf{c}}^r_j$ and $\tilde{\mathbf{c}}^r_{n^r_{i,j}}$, respectively; and let $\mathbf{h}^r_{i,j}$ be the derivative of Eq. 2 with respect to $\tilde{\mathbf{c}}^r_j$:

$$\mathbf{g}^r_{1_{i,j}} = -\frac{\widetilde{\mathbf{F}}^r_i(j)}{\sin\Big(\mathcal{D}\big(\mathbf{F}^r_i(j), \mathbf{c}^r_j\big)\Big)} \tag{3}$$

$$\mathbf{g}^r_{2_{i,j}} = \frac{\widetilde{\mathbf{F}}^r_i(j)}{\sin\Big(\mathcal{D}\big(\mathbf{F}^r_i(j), \mathbf{c}^r_{n^r_{i,j}}\big)\Big)} \tag{4}$$

$$\mathbf{h}^r_{i,j} = -\frac{\widetilde{\mathbf{F}}^r_i(j)}{\sin\Big(\mathcal{D}\big(\mathbf{F}^r_i(j), \mathbf{c}^r_j\big)\Big)} + \frac{\tilde{\mathbf{f}}^r_i(j)}{\sin\Big(\mathcal{D}\big(\mathbf{f}^r_i(j), \mathbf{c}^r_j\big)\Big)} \tag{5}$$

Then, we can represent the averaged gradient considering the three terms:

$$\Delta\tilde{\mathbf{c}}^r_j = \Delta_{\mathbf{g}^r_{1_{i,j}}} + \Delta_{\mathbf{g}^r_{2_{i,j}}} + \Delta_{\mathbf{h}^r_{i,j}} \tag{6}$$

where each term of the right hand side is given by:

$$\Delta_{\mathbf{g}^r_{1_{i,j}}} = \frac{1}{N}\left(\frac{\sum_{i:\mathbf{y}_i(j)=1} \mathbf{g}^r_{1_{i,j}}\, \delta(\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}} > 0)}{1 + \sum_{i:\mathbf{y}_i(j)=1} \delta(\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}} > 0)}\right) \tag{7}$$

$$\Delta_{\mathbf{g}^r_{2_{i,j}}} = \frac{1}{N}\left(\frac{\sum_{i:\mathbf{y}_i(n^r_{i,j})=1} \mathbf{g}^r_{2_{i,j}}\, \delta(\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}} > 0)}{1 + \sum_{i:\mathbf{y}_i(n^r_{i,j}))=1} \delta(\tilde{\mathcal{L}}^r_{\mathrm{ATCL}_{i,j}} > 0)}\right) \tag{8}$$

$$\Delta_{\mathbf{h}^r_{i,j}} = \frac{1}{N}\left(\frac{\gamma \sum_{i:\mathbf{y}_i(j)=1} \mathbf{h}^r_{i,j}\, \delta(\tilde{\mathcal{L}}^r_{\mathrm{NT}_{i,j}} > 0)}{1 + \sum_{i:\mathbf{y}_i(j)=1} \delta(\tilde{\mathcal{L}}^r_{\mathrm{NT}_{i,j}} > 0)}\right) \tag{9}$$

Here, $\delta(\mathit{condition}) = 1$ if the *condition* is true and $\delta(\mathit{condition}) = 0$ otherwise. Using the chain rule, we can find the general form of the averaged gradient as follows:

$$\Delta\mathbf{c}^r_j = \frac{I_D - \tilde{\mathbf{c}}^r_j \otimes \tilde{\mathbf{c}}^r_j}{\|\mathbf{c}^r_j\|_2} \Delta\tilde{\mathbf{c}}^r_j \tag{10}$$

where I_D is an identity matrix of dimension D and $\otimes$ denotes the outer product. Finally, the centers are updated using $\Delta\mathbf{c}^r_j$ for every iteration of the training process by a gradient descent algorithm. We can update the centers of other streams in a similar manner.

B More Qualitative Results

We provide more qualitative results of A2CL-PT. Please refer to the videos that are included in the subfolder.

References

1. He, X., Zhou, Y., Zhou, Z., Bai, S., Bai, X.: Triplet-center loss for multi-view 3d object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1945–1954 (2018)
2. Li, Z., Xu, C., Leng, B.: Angular triplet-center loss for multi-view 3d shape retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8682–8689 (2019)
3. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)