Class-Incremental Grouping Network for Continual Audio-Visual Learning (Supplementary Material)

Shentong Mo[†]
Carnegie Mellon University

Weiguo Pian[†]
University of Texas at Dallas

Yapeng Tian[*]
University of Texas at Dallas

In this supplementary material, we provide the significant differences between our CIGN and the recent grouping work, GroupViT [3], and more experiments on the depth of transformer layers and grouping strategies. In addition, we validate the effectiveness of learnable audio-visual class tokens in learning disentangled class-incremental audio-visual representations for continual audio-visual learning and report quantitative comparison results of various buffer sizes.

1. Significant Difference from GroupViT and CIGN

Compared to GroupViT [3] on image segmentation, our CIGN has three significant recognizable characteristics to address continual cross-modal learning from incremental categories of audio-visual pairs, which are highlighted as follows:

1) **Incremental-Constraint on Audio-Visual Category Tokens.** The major difference is that we have learned disentangled audio-visual class tokens for each audio category, *e.g.*, 100 audio-visual category tokens for 100 categories in the VGGSound-100 benchmark. During training, each audio-visual class token does not learn semantic overlapping information among each other, where we apply the cross-entropy loss $\sum_{i=1}^{C} \text{CE}(\mathbf{h}_i^t, \mathbf{e}_i^t)$ on each category probability $\mathbf{e}_i^t$ with the disentangled constraint $\mathbf{h}_i^t$ at current task t. Meanwhile, we apply a Kullback-Leibler (KL) divergence loss $\text{KL}(\mathbf{c}_i^t || \mathbf{c}_i^{t-1})$ to eliminate forgetting old class tokens $\{\mathbf{c}_i^t\}_{i=1}^{C}$ at task $t-1$. However, the number of group tokens used in GroupViT is a hyper-parameter, and they must tune it carefully across each grouping stage.

2) **Audio-Visual Continual Grouping.** We propose the audio-visual continual grouping module for extracting individual semantics with class-aware information from incremental audio-visual pairs. However, GroupViT utilized the grouping mechanism on only patches of images without explicit category-aware tokens involved. Therefore, GroupViT can not be directly transferred to incremental audio-visual samples for solving the new continual audio-visual learning problem. Moreover, they used multiple grouping stages during training, and the number of grouping stages is a hyper-parameter. In our grouping module, only one audio-visual incremental grouping stage with disentangled and incremental audio-visual class tokens is enough to capture disentangled audio-visual representations in the multi-modal incremental semantic space.

3) **Incremental Audio-Visual Class as Weak Supervision.** We leverage the incremental audio-visual category at the current task as the weak supervision to address continual audio-visual learning problem from class-incremental audio-visual samples, while GroupViT used a trivial contrastive loss to match the global visual representations with pre-trained text embeddings. In this case, GroupViT required a large batch size for self-supervised training on large-scale language-visual pairs. In contrast, we do not need unsupervised learning on the large-scale simulated audio-visual data with extensive training costs.

2. Depth of Transformer Layers and Continual Grouping Strategies

The depth of transformer layers and continual grouping strategies used in the proposed AVCG affect the extracted and grouped representations for continual audio-visual learning from incremental cross-modal pairs (*i.e.*, image and audio). To explore such effects more comprehensively, we varied the depth of transformer layers from $\{1, 3, 6, 12\}$ and ablated the continual grouping strategy using Softmax and Hard-Softmax. During training, to make Hard-Softmax differentiable, we applied the Gumbel-Softmax [1, 2] as the alternative. We report the comparison results of continual audio-visual performance on the VGGSound-100 benchmark in Table 1. When the depth of transformer layers is 3 and using Softmax in AVCG, we achieve the best class-incremental learning performance regarding all metrics. With increased depth from 1 to 3, the proposed CIGN consistently increases performance as better disentangled audio-visual representations are extracted from encoder features of the class-incremental audio-visual samples. Nevertheless, increasing the depth from 3 to 12

[*]Corresponding author, [†]Equal contribution.

Depth	AVCG	Audio		Visual		Audio-Visual	
		Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)
1	Softmax	42.25	13.09	47.26	10.63	51.25	9.31
3	Softmax	**45.83**	**10.21**	**49.52**	**8.83**	**55.26**	**6.52**
6	Softmax	44.62	11.52	48.63	9.55	53.21	7.58
12	Softmax	43.91	12.36	48.01	10.12	52.53	8.72
3	Hard-softmax	40.59	15.16	43.72	13.52	48.95	11.36

Table 1. Exploration studies on the depth of self-attention transformer layers and continual grouping strategies in Audio-Visual Continual Grouping (AVCG) module.

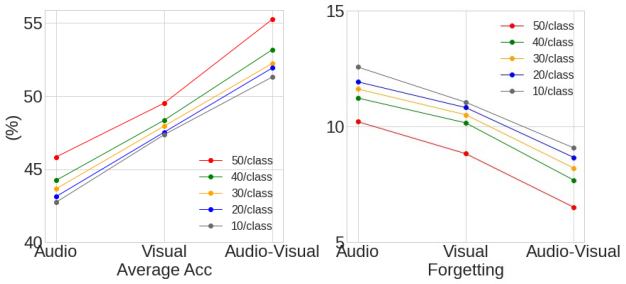

Figure 1. Impact of buffer size on the performance of Average Acc (Left) and Forgetting (Right) for continual audio-visual learning.

will not continually improve the class-incremental result since three transformer layers might be enough to extract the learned class-aware representations for audio-visual continual grouping with only one grouping stage. Furthermore, replacing Softmax with Hard-Softmax significantly deteriorates the performance of all metrics, which indicates the importance of the proposed AVCG in extracting disentangled audio-visual representations with class-incremental category-aware semantics from the audio-visual pairs.

3. Quantitative Validation on Audio-Visual Category Tokens

Learnable audio-visual incremental category tokens are essential to aggregate audio-visual representations with category-aware semantics from incremental audio-visual samples. We calculate the Precision, Recall, and F1 scores of audio-visual classification using these representations across training iterations to validate the rationality of learned audio-visual category token embeddings. All these metrics are observed to rise to 1, which indicates that each audio-visual category token learned disentangled information with incremental category-aware semantics. These quantitative results further demonstrate the effectiveness of audio-visual category tokens distillation in the continual audio-visual grouping for extracting disentangled audio-visual representations from class-incremental audio-visual samples for continual audio-visual learning.

4. Quantitative Comparison on Buffer Size

To quantitatively demonstrate the effectiveness of buffer size in continual audio-visual learning, we varied the buffer size per class from $\{10, 20, 30, 40, 50\}$, and report the comparison results in Figure 1. As can be seen, the proposed CIGN achieves the best performance of average accuracy and forgetting when we use 50 audio-visual samples for each incremental category. These results demonstrate the importance of caching samples from previous classes for continual audio-visual learning.

References

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 1

[2] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017. 1

[3] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 1