Agreement aware and dissimilarity oriented GLOM

Supplementary Material

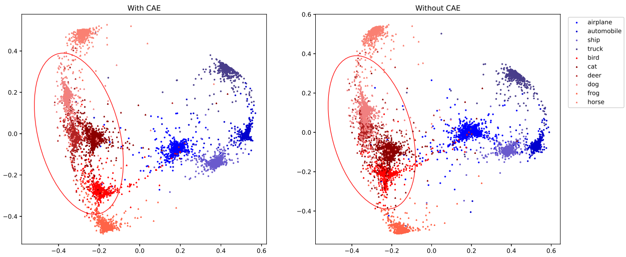


Figure 8. 2D representations extracted by our model, with and without CAE, on CIFAR-10 using PCA. The inclusion of CAE results in better separation of representations across different classes compared to the model without CAE, indicating that CAE helps model learn category-level holistic features. Besides, the representations can be grouped into 2 superclasses—*Vehicles* and *Animals*, demonstrating that the model has effectively captured meaningful relationships between the concepts of vehicles and animals.

Table 4. Sensitivity analysis for the threshold in CAE

Threshold	-0.9	-0.6	-0.3	0	0.3	0.6
TV_n	0.032	0.032	0.033	0.051	0.065	0.089

Table 5. Sensitivity analysis for the patch size in H_d

Patch Size	1	2	4	8
No. Params (Million)	72.9	6.8	2.0	1.3
Accuracy (%)	90.53	89.68	89.10	86.60

6. Datasets

Experiments are conducted on the following datasets:

CIFAR10 and CIFAR100 are widely used image classification datasets consisting of 60,000 color images of size 32x32 pixels. CIFAR-10 has 10 classes with 6,000 images per class, while CIFAR-100 has 100 classes with 600 images per class.

MNIST and FashionMNIST are widely used datasets of 70,000 grayscale images, each 28x28 pixels. MNIST contains handwritten digits (0-9) for digit classification, while FashionMNIST features images of fashion items like shirts and shoes, providing a more challenging alternative to MNIST.

SmallNORB is a dataset designed for 3D object recognition tasks, containing 48,600 grayscale images of 50 toys belonging to 5 categories: airplanes, cars, trucks, humans, and animals. The objects are captured from different lighting conditions, elevations, and azimuths, providing a variety of viewing angles for robust model training and evaluation.

7. Training settings

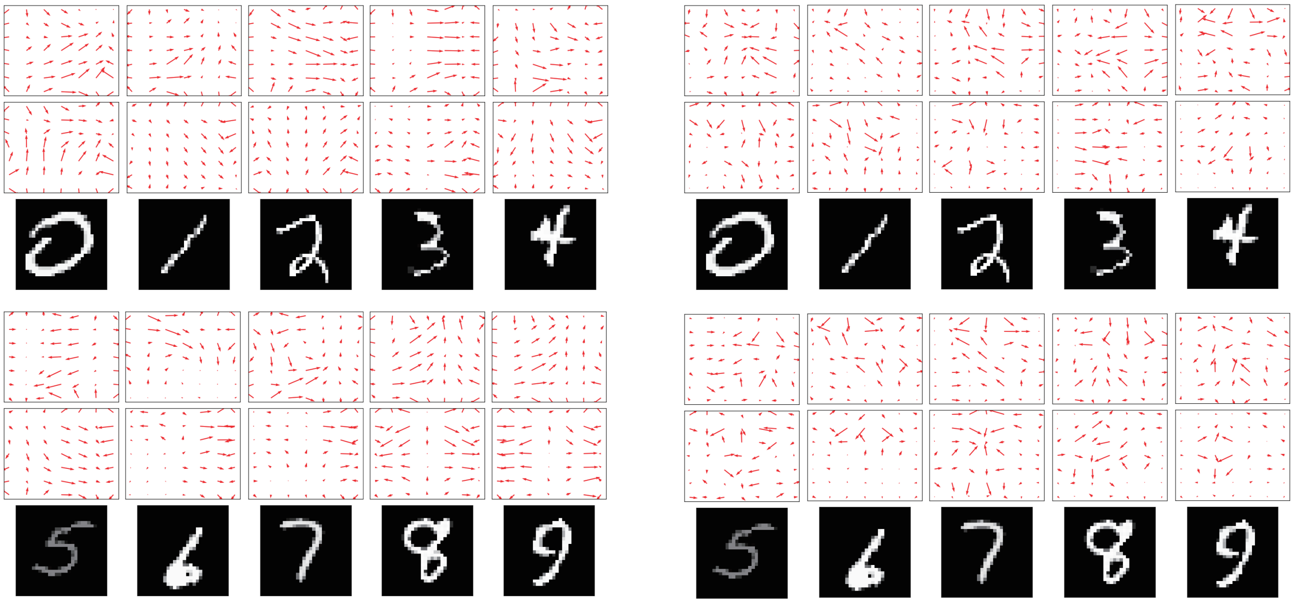
Regarding data augmentation, we follow the random augmentation strategy outlined in [5], along with applying data normalization. Additionally, images from the SmallNORB dataset are randomly cropped to a size of 32 × 32, while image resolutions in other datasets remain unchanged. Regarding the model setup, the patch size r is set to 2, resulting in 8 × 8 columns for CIFAR-10, CIFAR-100, and SmallNORB, and 7 × 7 columns for MNIST and FashionMNIST. For both pre-training and training, we use the following settings: 200 epochs, batch size of 1024, Adam optimizer, cyclic learning rate ranging from [0.004, 0.01], embedding size per level $d = 128$, time steps $T = 4$, contrastive feature dimension of 512, and downsampling size $p = 4$. The number of levels K is set to 2 in all cases except for the visual ablation experiment for CAE, as shown in Fig. 6, where K is increased to 5 to better illustrate the emerging island of different levels.

8. Additional Results

Additional results are given about threshold sensitivity test, holistic feature visualization, representation visualization using arrows.

Sensitivity test for threshold and patch size. We provide a sensitivity analysis for the threshold in CAE and the patch size in H_d on CIFAR-10. As shown in Tab.4 and Tab.5, TV_n decreases as the threshold decreases and may approach a lower bound. Increasing the patch size reduces the number of parameters, with a slight drop in classification accuracy. In our model, we set the threshold for bottle-level embeddings slightly higher than that for top-level embeddings. The patch size is chosen for a trade-off between accuracy and efficiency.

Holistic feature representation visualization. To gain deeper insights into the impact of CAE on holistic feature representation, we compare the 2D representations extracted by our model, with and without CAE, using Principal Component Analysis (PCA) [12] to project the output features from a multidimensional space into a 2D space for visualization. As illustrated in Fig. 8, the CIFAR-10 data is categorized into 10 distinct classes, which can be attributed to contrastive learning pretraining. Moreover, the inclusion of CAE results in better separation of representations across different classes compared to the model without CAE. This improvement stems from CAE's ability to enforce agreement between high-level features across different locations of the same object, enabling the model to learn category-level holistic features rather than relying on low-level fea-



With CAE

Without CAE

Figure 9. The representation visualization of the 10 digits (0–9) in MNIST using arrows for model with and without CAE. The direction and length of each arrow represent the gradient direction and magnitude of an embedding at a specific location, computed via the Sobel operator.

tures from specific regions. Additionally, these classes can be further grouped into two superclasses—*Vehicles* and *Animals*. The clear separation of these superclasses suggests that the model effectively captures meaningful relationships between these concepts.

The representation visualization using arrows. Here, we visualize all digits from 0 to 9. The implementation details are as follows: First, the Sobel operator is applied to compute the gradients in the x and y directions. Then, the gradient direction is calculated using the arctangent function, and the magnitude is determined using the Euclidean norm. Next, points are uniformly sampled from the feature map, and the arrow lengths are normalized to enhance visualization. Finally, arrows representing the gradient directions are drawn and overlaid on the background image to clearly illustrate the gradient flow. As shown in Fig. 9, the arrows in the model incorporating CAE exhibit greater consistency.