View-Category Interactive Sharing Transformer for Incomplete Multi-View Multi-Label Learning

Supplementary Material

6. Algorithm

In this section, we provide the additional explanation of our method. In Algorithm 1, the detailed training process is presented.

Algorithm 1 Training process of VIST

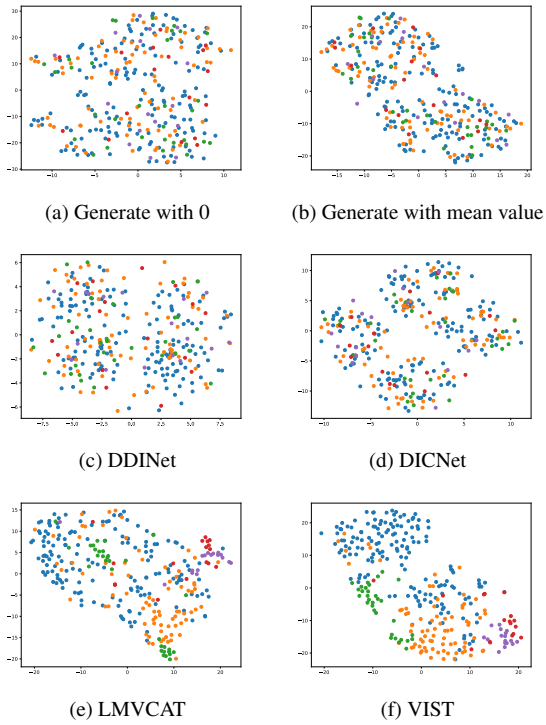
Input: Incomplete multi-view data $\{\mathbf{X}_v\}_{v=1}^{V}$, label matrix $\mathbf{Y}$, missing-view indicator matrix $\mathbf{W}$, missing-label indicator matrix $\mathbf{U}$.

Output: Final prediction $\hat{\mathbf{y}}$.

Initialize the parameters of VIST and set hyperparameters α, β, γ, k and training epochs E_{tr}.
$t = 0$.
while $t < E_{tr}$ **do**
  for $v = 1 : V$ **do**
    Embed the multi-view data $\mathbf{X}_v$ to get embedding vector $\mathbf{e}_v$.
  end
  Stack the embedding vectors to obtain origin multi-view embedding $\mathbf{E}$.
  Generate the missing views by sampling from distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ with Eq. (9) and Eq. (10).
  Calculate the loss functions of contrastive learning $\mathcal{L}_{pos}$ by Eq. (12) and $\mathcal{L}_{neg}$ by Eq. (13).
  Get multi-view embedding $\tilde{\mathbf{E}}$ by Eq. (1), advanced multi-view embedding $\bar{\mathbf{E}}$ by Eq. (7) and calculate $\mathcal{L}_{aux}$ using Eq. (17).
  Calculate the loss functions for view-category consistency guided embedding enhancement $\mathcal{L}_c$ using Eq. (18).
  Get the final prediction $\hat{\mathbf{y}}$ by Eq. (19).
  Calculate multi-label classification loss $\mathcal{L}_m$ using Eq. (20) and masked asymmetric loss $\mathcal{L}_a$ using Eq. (22).
  Calculate the overall loss function $\mathcal{L}_o$ using Eq. (23).
  Update the parameters.
  $t = t + 1$.
end

7. Additional Visualizations

We carry out a series of experiments to visualize the generated complete multi-view data in the Corel5k datasets with 50% missing instances. Due to the category correlation within multi-label datasets, we opt to select five categories with semantically distant labels for the data visualization using t-SNE. These categories are water, beach, coyote, cars and hats. The results, as depicted in Fig. 6, indicate a pronounced enhancement in the distinction of margins between different classes. This improvement is attributable to the VIST's capability to effectively explore and utilize the correlations existing among various views and categories.



(a) Generate with 0 (b) Generate with mean value

(c) DDINet (d) DICNet

(e) LMVCAT (f) VIST

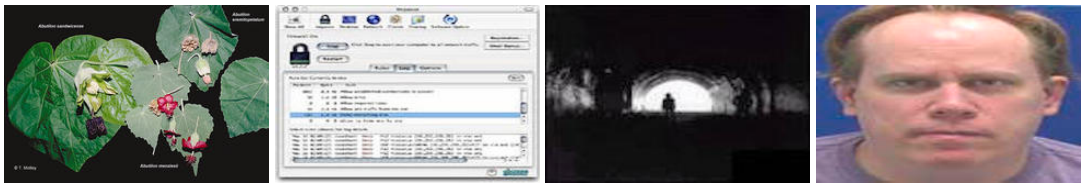
Figure 6. The results of different value of hyperparameter α and k on different datasets. β and γ are respectively set to default values of 0.1 and 0.1.

Furthermore, we also present visualizations comparing the performance of our method with the current state-of-the-art approaches on datasets Corel5k and Espgame, as shown in Fig. 7 and Fig. 8. It is observable that our model achieves superior performance in these instances, a result that is directly attributable to the effective interaction between the views and categories.

Labels: (people, pool, swimmers, water)	Labels: (sky, sun, clouds)	Labels: (leaf, flowers, petals)	Labels: (ruins, stone, pyramid)
DDINet: (**1**, 3, 45, 31)	DDINet: (3, 6, 10)	DDINet: (**2**, **1**, 5)	DDINet: (5, 2, 6)
DICNet: (**1**, 17, 53, 207)	DICNet: (**1**, 55, **2**)	DICNet: (3, 2, 16)	DICNet: (6, 5, 26)
LMVCAT: (**1**, 6, 15, 7)	LMVCAT: (**1**, 20, **2**)	LMVCAT: (**2**, **1**, 8)	LMVCAT: (4, **1**, 6)
VIST: (**1**, **2**, **4**, **3**)	VIST: (**1**, **3**, **2**)	VIST: (**2**, **1**, **4**)	VIST: (**2**, **1**, **5**)

Labels: (water, tree, boats, buildings)	Labels: (bear, polar, snow, head)	Labels: (cars, tracks, turn)	Labels: (tree, horses, mare, foals)
DDINet: (**2**, 21, 15, **1**)	DDINet: (18, 10, 3, 15)	DDINet: (**1**, **2**, 15)	DDINet: (8, 1, 5, 4)
DICNet: (**2**, 52, 9, **1**)	DICNet: (10, 40, 6, 29)	DICNet: (2, 8, 22)	DICNet: (**6**, 4, 5, 2)
LMVCAT: (**2**, 8, 7, **1**)	LMVCAT: (12, 8, 2, 25)	LMVCAT: (**1**, **2**, 9)	LMVCAT: (**6**, **1**, 7, 3)
VIST: (**2**, **6**, **3**, **1**)	VIST: (**2**, **3**, **1**, **11**)	VIST: (**1**, **2**, **7**)	VIST: (**6**, **1**, **4**, **2**)

Figure 7. Visualization of prediction results on Dataset Corel5k. The numbers enclosed in parentheses represent the ranking assigned by the model to the likelihood of the image belonging to each respective category, the same below.



Labels: (flower, green, leaf, plant)	Labels: (computer, screen, white, window)	Labels: (black, dark, light, man)	Labels: (eye, face, hair, man)
DDINet: (57, **1**, **2**, 19)	DDINet: (4, 5, 25, 2)	DDINet: (**1**, 16, 10, 5)	DDINet: (6, **3**, **1**, **4**)
DICNet: (36, **1**, **2**, 45)	DICNet: (**2**, **3**, 17, **1**)	DICNet: (**1**, **7**, 49, 44)	DICNet: (4, **3**, **1**, **4**)
LMVCAT: (10, **1**, **2**, 12)	LMVCAT: (**2**, **3**, 5, **1**)	LMVCAT: (**1**, **7**, **8**, 12)	LMVCAT: (**2**, **3**, **1**, **4**)
VIST: (**4**, **1**, **2**, **3**)	VIST: (**2**, **3**, **4**, **1**)	VIST: (**1**, **7**, **8**, **2**)	VIST: (**2**, **3**, **1**, **4**)

Labels: (animal, desert, head, sand)	Labels: (grass, green, tree, wood)	Labels: (family, girl, white, woman)	Labels: (brown, fence, rectangle, square)
DDINet: (7, **1**, 26, 10)	DDINet: (4, **1**, **2**, 37)	DDINet: (98, **1**, 19, **2**)	DDINet: (8, 1, 5, 4)
DICNet: (9, **1**, 45, 5)	DICNet: (19, **1**, **2**, 23)	DICNet: (75, **1**, 34, **2**)	DICNet: (**6**, 4, 5, 2)
LMVCAT: (6, 3, 17, 4)	LMVCAT: (6, **1**, **2**, 55)	LMVCAT: (43, **1**, 13, 3)	LMVCAT: (**6**, **1**, 7, 3)
VIST: (**2**, **1**, **7**, **3**)	VIST: (**3**, **1**, **2**, **17**)	VIST: (**13**, **1**, **5**, **2**)	VIST: (**1**, **9**, **4**, **5**)

Figure 8. Visualization of prediction results on Dataset Espgame.