Research of Event Reconstruct Based on Multi-View Contrastive Learning (Student Abstract)

Yuefeng Ma, Zhongchao He, Shumei Wang

School of Computer Scienct, Qufu Normal University
Rizhao, Shandong Province,276827, China
rzmyf1976@163.com

Abstract

The proliferation of social media exacerbates information fragmentation, posing challenges to understanding public events. We address the problem of event reconstruction with a novel Multi-view Contrast Event Reconstruction (MCER) model. MCER maximizes feature dissimilarity between different views of the same event using contrastive learning, while minimizing mutual information between distinct events. This aggregates fragmented views to reconstruct comprehensive event representations. MCER employs momentum and weight-sharing encoders in a three-tower architecture with supervised contrastive loss for multi-view representation learning. Due to the scarcity of multi-view public datasets, we construct a new Mul-view-data benchmark. Experiments demonstrate MCER's superior performance on public data and our Mul-view-data, significantly outperforming selfsupervised methods by incorporating supervised contrastive techniques. MCER advances multi-view representation learning to counter information fragmentation and enable robust event understanding.

Introduction

The rise of self-media enables individuals to narrate public events, but information fragmentation results from real-time, high-volume brief messages. Sensationalism and distortion exacerbate this. Aggregating cross-source information is key to reconstruction. With recent advances in deep multi-view representation learning, including convolutional and generative adversarial networks, we can categorize fragmented images by events and counter information fragmentation for comprehensive public event understanding,as shown in Figure 1.

Convolutional neural networks enable multi-view representation learning through one-view-one-network and multi-view-one-network frameworks. The former extracts features from each view separately before fusing them; Yang et al.(Yang et al. 2018) proposed MCEA combining multi-view CNN, autoencoder, and classifier for 3D shape learning. The latter inputs multiple views into the same network; Wang et al.(Ahmad et al. 2021) developed a context-aware 3D CNN combining multi-level information for COVID screening. Autoencoders are used for multi-view representation learning by mapping views to a common embedding; Park et al.(Park et al. 2020) maximized mutual information between patches in contrastive learning. Generative adversarial networks like the two-pathway model by Huang et al.(Huang et al. 2017) synthesize photorealistic frontal views by capturing global and local patterns. In summary, deep networks enable multi-view representation learning through view-specific feature extraction, common embedding mapping, and joint generative modeling.

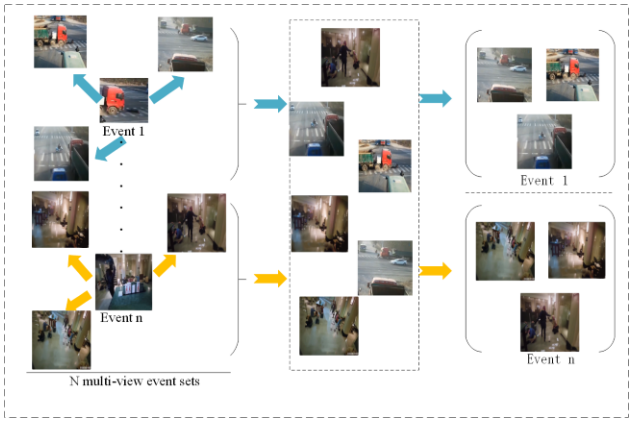

Figure 1: For the given N public events, each event has a collection of M perspectives. The main task of MCER is to integrate and process the fragmented information scattered on social media, in order to reconstruct the entire scene. Here we show partial event set images on the Mul-View-data dataset.

In summary, multi-view image classification has made progress by modeling the overall scene, but this fails for event reconstruction which requires correlating local fragments across views. To enable event reconstruction from incomplete multi-view data, we propose a multi-view contrastive learning framework to maximize mutual information between local features of the same event while minimizing it for different events. Our key contributions are:

- Applying contrastive learning on multi-view images to maximize mutual information between different representations of the same event across views.

- Effective data augmentation and cross-view feature incorporation to highlight common event information over irrelevant factors.
- A three-tower structure with momentum and weight-sharing encoders to maintain feature consistency.

Method

Our goal is supervised event reconstruction by modeling mutual information across multiple views. We use contrastive learning to maximize mutual information between different views of the same location in feature space, while minimizing it for different locations. Rather than two views only, we augment viewpoints via cropping.Given a dataset I with M view sets $v_1, v_2, ..., v_m$, each view vi is a random variable. Contrastive learning aligns these random variables for the same scene while dispersing different scenes. This enables reconstructing events from fragmented multi-view data.

Multi-view Contrastive Loss

For supervised contrastive learning, it is essential to allow each anchor to have multiple positive samples to effectively utilize labels and bring closer instances of the same event while separating images of different events. Equation (3) cannot handle the case where the same event contains multiple viewpoints. To address this issue, we modify equations (2) and (3) and adopt a supervised contrastive loss function, which can be expressed as:

$$L^{\text{sup}} = \sum_{i}^{2N} L_i^{\text{sup}} \tag{1}$$

$$L^{\text{sup}} = M \cdot \log \frac{\exp\left(sim(z_i \cdot z_j)\right)}{-\sum_{K=1}^{2N} 1_{i \neq k} \cdot \exp\left(sim(z_i \cdot z_j)\right)} \cdot \frac{1}{\tau} \tag{2}$$

$$M = \frac{1}{2N_{\widehat{y_i}} - 1} \sum_{j=1}^{2N} 1_{i \neq j} \cdot 1_{\widehat{y_i} = \widehat{y_i}} \tag{3}$$

where $N_{\widehat{y_i}}$ represents the total number of images in each batch that have the same label (same event) y_i as anchor i. Through Equation (5), contrastive learning can be performed between positive samples of multiple viewpoints within the same event and negative samples from other events, thereby improving the discriminative power of the model's features.

Experiments

In this section, we evaluate the proposed method on two challenging datasets, NYU Depth-V2 and our own dataset Mul-View-data, and compare it with state-of-the-art methods. We also conduct ablation studies to analyze the effects of different components of the proposed method.

When selecting positive samples, we set a threshold θ as a hyperparameter, which needs to be experimentally determined for optimal selection of positive samples. This threshold is used to compare with the Hamming distance to determine whether two samples belong to the same event. The comparison of the results obtained from the NYU Depth-V2 multi-view dataset and our own Mul-View-data dataset using SimCLR v1 is as follows:

	$\theta = 0.3$	
	NYU-data	Mul-View-data
SimCLR v1	91.34%	85.37%
MCER	94.74%	86.20%
	$\theta = 0.5$	
	NYU-data	Mul-View-data
SimCLR v1	88.34%	82.62%
MCER	86.42%	82.52%

Table 1: Comparison of accuracy between selecting positive samples with different probabilities and SimCLR on different datasets.

We optimized our model using the supervised momentum contrastive loss function and achieved the following results on the NYU-depth V2 dataset with different selections of threshold θ:

method	$\theta = 0.1$	$\theta = 0.3$
Using unsupervised contrastive loss	85.31%	82.34%
Remove MoCo	93.45%	91.74%
MCER	99.38%	96.42%

Table 2: MCER uses the same momentum encoder setup as MoCo, with supervised contrastive loss, and presents accuracy results with different probabilities of selecting positive samples.

References

Ahmad, P.; Qamar, S.; Shen, L.; and Saeed, A. 2021. Context aware 3D UNet for brain tumor segmentation. 207–218. Springer.

Huang, R.; Zhang, S.; Li, T.; and He, R. 2017. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *Proceedings of the IEEE international conference on computer vision*, 2439–2448.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345. Springer.

Yang, Z.-X.; Tang, L.; Zhang, K.; and Wong, P. K. 2018. Multi-view CNN feature aggregation with ELM auto-encoder for 3D shape recognition. *Cognitive Computation*, 10(6): 908–921.