Review on: Visual Data Synthesis via GAN for Zero-Shot Video Classification:

In this paper, the authors have tried to come up with ways of discovering new video classification techniques using GAN since there is a huge growth of video categories. Previous work have been able to determine and classify many video categories but there are two main challenges faced. 1) They rely on large scaled labelled data which is costly to generate and some instances are quite rare. 2) Video categories are hard to grow from those particular already learned categories.

This is where Zero-shot comes in with their solutions. They gave solution to both the problems. For the first problem the authors have proposed “multi-level semantic interference approach”. It contains two opposite synthesis procedures driven by adversarial learning in which one synthesizes video feature given semantic knowledge, and the other inversely infers the semantic knowledge at both feature-level and label-level. To tackle the second problem they came up with “matching aware mutual information correlation”, which provides the synthesis procedure with informative guidance signals to overcome video degradation issue. Instead of direct feature projection, it will manually transfer knowledge. Therefore, statistical dependence within heterogeneous representations was captured and this helped bypassing the information degradation issue in typical projection-based ZSL methods.

Dataset: HMDB51 (2), UCF101 (3) Olympic Sports (4), and Columbia Consumer Video (CCV) (5) , which contained 6.7k, 13k, 783 and 9.3k videos with 51, 101, 16 and 20 categories respectively.

Zero-Shot Settings: There are two ZSL settings, namely strict setting and generalized setting (6). The former assumes the absence of seen classes at test state while the latter takes both seen and unseen data as test data during testing. They adopted strict setting for both NN search and SVM experiments.

The model was implemented with PyTorch and they used VGG-19 to extract sptial-temporal feature of videos and frame and optical flow from last pooling layer.

Benchmark: they compared their work with other ZSL methods such as Convex Combination of Semantic Embeddings (CONSE) (7), Structured Joint Embedding (SJE) (8) and Manifold regularized ridge regression (6).

State-of-the-art zero-shot video was achieved using these 4 datasets so it can be concluded that the methods were pretty accurate.

Reference:

1. Visual Data Synthesis via GAN for Zero-Shot Video Classification by Chenrui Zhang and Yuxin Peng, Institute of Computer Science and Technology, Peking University.
2. H. Kuehne, H. Jhuang, R. Stiefelhagen, and T. Serre. Hmdb51: A large video database for human motion recognition. In HPCSE. 2013.
3. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.
4. Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In ECCV, 2010.
5. Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In ICMR, 2011.
6. Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In CVPR, 2017.
7. Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zeroshot learning by convex combination of semantic embeddings. In ICLR, 2014.
8. Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In CVPR, 2015.