ITI-CERTH participation in TRECVID 2020

Konstantinos Gkountakos, Damianos Galanopoulos, Marios Mpakratsas, Despoina Touska, Anastasia Moumtzidou, Konstantinos Ioannidis, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris

Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{gountakos, dgalanop, mbakratsas, destousok, moumtzid, kioannid, heliasgj, stefanos, bmezaris, ikom} @iti.gr

Ad-hoc Video Search

Approach:

- Attention-based [1] dual-encoding neural network [2]
- The network encodes video-caption pairs into a common feature subspace
- Attention mechanisms are utilized for efficient textual and visual representation
- Two similar modules, trained in parallel:
  - The visual content
  - The natural language textual content
- Multi-level encoding for both modules
  - Mean pooling
  - Attention-based bi-GRU
  - bi-GRU-CNN
- Improved marginal ranking loss

Setup:

- Training datasets:
  - **TGIF** contains approximately 100k short animated GIFs and one description per each
  - **MSR-VTT** consisting of 10k short video clips, and 20 descriptions for every clip
- Evaluation dataset:
  - **V3C1** consisting of 7,475 videos and 1,082,659 video shots
- Video keyframe representation
  - Pre-trained ResNet-152 trained on the ImageNet-11k dataset
- Textual embeddings
  - Word2Vec trained on the English tags of 30K Flickr images
  - BERT trained on Wikipedia

Submission:

- AVS 2020 main task, evaluation on 20 textual queries
- AVS progress subtask, evaluation on 10 textual queries
- Mean Extended Inferred Average Precision (**MXinfAP**)

Run:

- One run for the main task and one run for the progress subtask
- Each run combines multiple training configurations in a late fusion scheme

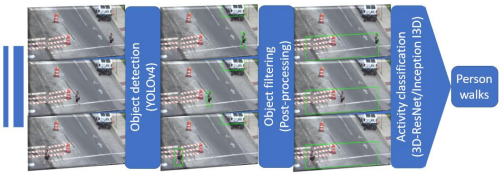
Results:

Task	Main	Progress
MXinfAP	0.202	0.159

Activities in Extended Video

Approach:

- Two-step approach:
  - Detection of objects in order to extract the **Extended Activity Bounding Boxes** (EABBox) using YOLOv4
  - Classification of each EABBox using 3D-CNNs, **3D-ResNet** and **Inception I3D,** in order to recognize the activities



Setup:

- **ActEV** contains 64, 54 and 246 videos for train, validate and evaluate sub-sets, respectively
- **Train** and **validate** activity classifier using 4311 and 3521 activities extracted from the corresponding sub-sets in order to assign an activity label to each EABBox
- Two post-processing steps were implemented to investigate four different system setups
  - **Post-processing 1** (Ps1): Examine the effect of merging overlapped EABBoxes
  - **Post-processing 2** (Ps2): Examine the immobility of objects at short intervals of time and discard EABBox with duration less than 20 frames

Submission:

- Four systems were evaluated:
  - **CERTH-ITI-I3D_base:** Ps1, I3D
  - **CERTH-ITI-YRW16:** Ps1, 3D-ResNet-50, weighted cross-entropy loss
  - **CERTH-ITI-YR16:** Ps1, CERTH-ITI-YRW16, considering activities more than 20 frames
  - **CERTH-ITI-P:** Ps2, CERTH-ITI-YR16

Results:

System Name	PARTIAL AUDC
CERTH-ITI-I3D_base	0.93125
CERTH-ITI-YRW16	0.88530
CERTH-ITI-YR16	0.88511
CERTH-ITI-P	**0.86576**

Disaster Scene Description and Indexing

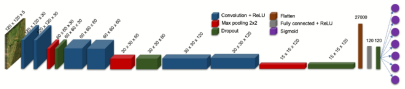
Identification of natural disasters in videos

Dataset features

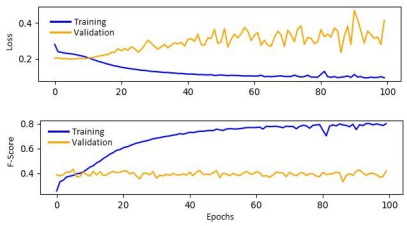
- Dataset images tagged with 32 concepts related to disaster scenes
- Greatly imbalanced dataset of high-res UAV images of different resolution and orientation
  - Resize images on same resolution preserving aspect ratio

Multi-label classification Framework

- Train a VGG-like Deep Convolutional Neural Network



- Input: RGB UAV image
- Output: Set of 32 concepts related to disaster scenes



Run:

- Consider only dataset of humanly annotated images for model training
- Extracted keyframes of 1,825 videos with a video segmentation service
- Resize images similarly to training dataset
- The mean of the probability values of the keyframes extracted for each video
- Return top-1000 results for each class

ITI-CERTH 1: Video shot annotation using VGG like DCNN predictions to extract among 32 TRECVID concepts on keyframes of videos

Results:

Submitted Run	ITI-CERTH 1
Multi-label 32 classes (F-score)	0.076

[1]. Damianos Galanopoulos and Vasileios Mezaris. Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In Proc. of the ACM Int. Conf. on Multimedia Retrieval, (ICMR '20), 2020.
[2]. Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and XunWang. Dual Encoding for Zero-Example Video Retrieval. In Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

This work was partially supported by the European Commission under contracts H2020-832876 aqua3S, H2020-786731 CONNEXIONs, H2020-833115 PREVISION and H2020-780656 ReTV.