Course5
Transformative intelligence

Video Ontology:
Applying Deep Learning for Video Understanding

David Yan, Ph.D.

AI Scientist, Course5 Intelligence

Introduction

The past decade has seen rapid advancements in the area of computer vision, with deep neural networks, particularly convolutional neural networks (CNNs), achieving remarkable results in multiple areas of image processing. The breakthrough success of deep learning for images has sparked more recent interest in using deep learning for video processing. This is currently an active area of research with many challenges and applications.

The key difference between video and image analysis is, of course, the addition of time. Whereas an image model might be able to detect objects in an image, classify them, and provide information about their relative positions, a video also contains information about individual and relative motion of objects, and dynamic changes in objects and scenes over time. For example, actions that may be ambiguous based on a single frame (sitting down vs. standing up) are obvious from a sequence of frames. The key goal in video analysis is to efficiently and accurately capture these dynamics.

Video ontology is a broad term that encompasses several related aspects of video analysis and understanding:

- **Scene and action recognition:** classification of scenes and actions from a video at a frame level. This includes classification into setting (indoor vs. outdoor, office vs. playground), overall action (concert, basketball game), and granular/human action (hula-hooping, horseback-riding) categories.

- **Captioning:** Automatic text generation based on a sequence of frames, in order to identify relationships and semantic content.

- **Action localization:** Spatial and temporal segmentation of actions

- **Visual relation detection (VRD):** A binary or multi-object extension of action localization, visual relation detection seeks to identify relationships between pairs or groups of actors and objects in a video. VRD can be used to generate scene graphs, which are graphical models.

In this white paper, we will discuss the applications, techniques, challenges, and future work related to video ontology.

Applications

At a fundamental level, video ontology allows a computer to understand and describe the contents of a video, by creating representations of the objects, attributes and relationships present. Building a model for video understanding using ontologies such as scene classification, action localization and relationship detection is crucial to visual reasoning and language tasks like captioning, visual question answering [1], scene structuring and parsing [2], and person and object search [3]. These lead to many exciting potential applications, both for real-time inference, such as:

- Analysis of security and surveillance videos, particularly event detection and abnormal event detection
- Video visual saliency detection [4]
- Human-machine interaction in robotics and consumer electronics
- Human behavior analysis
- Motion prediction and tracking, for example for autonomous vehicles

As well as offline use-cases such as:

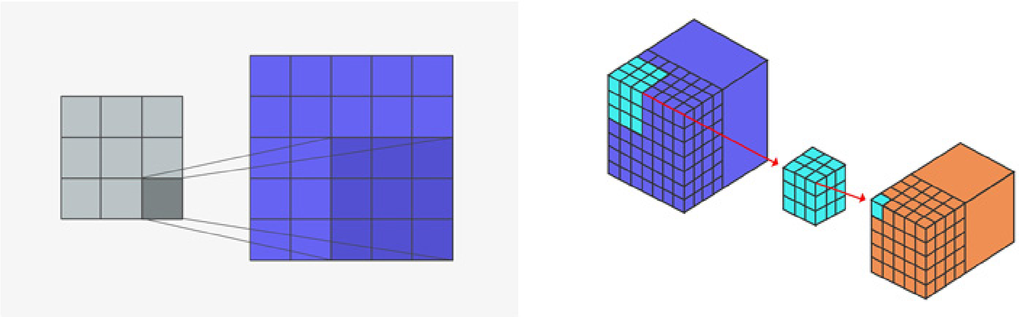
- Batch analysis on large video libraries for indexing, recommendation, and retrieval
- Video content optimization; for example, using video ontological AI variables to provide predictive insights on advertisements

Techniques and Recent Advancements

Depending on the task, video ontology requires a diverse set of models and techniques. Here are a few of the important aspects:

Feature extraction

For extracting effective feature descriptors from videos, one traditional method has been to use local, pre-processed feature descriptors, such as HOG (Histogram of Oriented Gradients), SURF (Speeded-up Robust Features), etc., and their extensions to multiple frames such as HOF (Histogram of Optical Flow), and MBH (Motion Boundary Histograms) [5, 6]. However, nearly all modern video reasoning models use some form of convolutional neural network for feature extraction. In particular, networks using 3D convolutional layers have attracted recent research interest. [7] Where 2D convolutional layers commonly used in image processing operates in an image's two spatial dimensions, 3D convolutional layers also operate over a third dimension. When frame-level features are concatenated along a time axis, 3D convolutions allow the network to capture spatio-temporal dynamics. Recent architectures use a mixture of 2D and 3D convolutional branches, and multiple "streams", which use RGB and dense optical flow features as inputs, sampled from frames at different times in a video clip. Recent work also attempts to leverage the much more comprehensive research into image classification by using pre-training and transfer learning from image



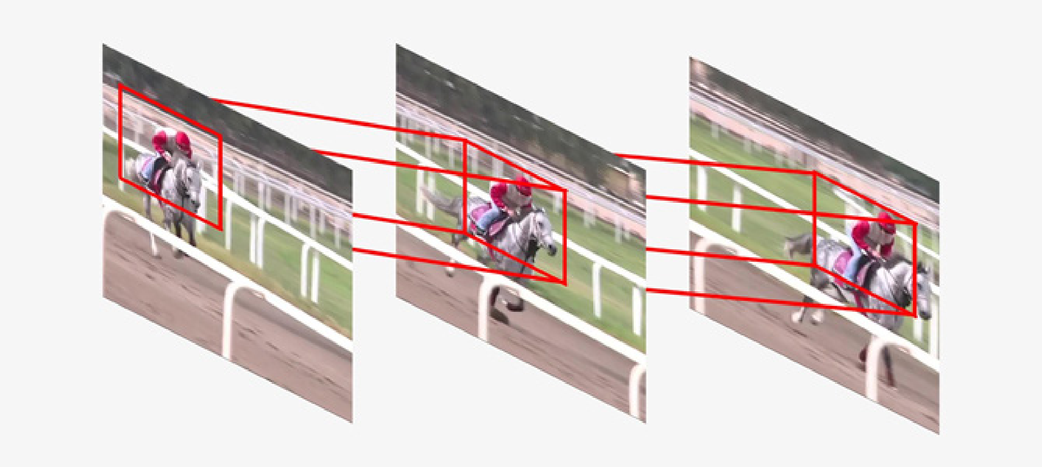
Diagrammatic representations of
two dimensional (left) and three dimensional (right) convolutions

Captioning

On top of feature extraction, captioning typically uses some form encoder-decoder network, where a CNN acts as the feature encoder, and the encoder is typically some form of RNN (Recurrent Neural Network), especially LSTM (Long Short-Term Memory) networks, which have seen many successes in the field of image captioning. [9]

Action localization

Action localization and visual relation detection have the additional challenge of requiring object localization in both time and space, similar to the related field of visual object tracking. Such a model must learn to extract both local appearance and dynamics. Object bounding boxes for images are generalized to tubelets in videos, which can be thought of as a sequence of bounding boxes in consecutive frames. Tubelets can be generated using image-level models, which are then linked based on position and class, though videos present additional challenges such as motion blur, partial occlusion, changes in camera angle, and so forth. Much of current research into video action localization focuses on regressing object tubelets using video-level rather than image-level features. [10]



*A representation of linked bounding boxes between frames, creating a **tubelet***

Challenges and Future Work

Being an area of research which is still developing, there are a number of challenges facing video ontology models. Here, we highlight a few of the major challenges and some directions for future work.

Challenges

- There is still a need for larger and more comprehensive training datasets for video reasoning tasks. Image datasets underwent an enormous proliferation in the past fifteen years as CNNs revolutionized computer vision. Videos, on the other hand, still lack the large datasets, like ImageNet and COCO, that image processing enjoys, and are still undergoing growth as researchers build new large-scale benchmarks. Also, the video datasets that do exist tend to be carefully chosen and trimmed so that only videos with consistent quality which have a single, clear action are included. Less work has been done on so-called “wild” videos, where video content and quality are unpredictable.

- Video classification and relation detection are still much less accurate than their equivalent image tasks, due in part to a lack of large datasets, and also due to video analysis being inherently a more complex task. For context, state-of-the-art video action classification models such as P3D and TRN report top-1 accuracies between 45 and 55 percent on video datasets with hundreds of classes [11, 12], whereas standard image classification models such as ResNet achieve over 80 percent on much larger image datasets with tens of thousands of classes. Additionally, occlusion, motion blur, changes in viewpoint, scale, and illumination, background clutter, and dense foreground object tracking are all current challenges in video analysis.

- Adding a dimension to convolution can dramatically increase the number of trainable parameters, with commensurate increases in training time and complexity. Recent models on videos are also generally restricted to a small number of frames (usually between 5 and 15 frames), due to the long training times involved in using more video frames. Ongoing research effort is going into the development of models that reduce computational cost while maintaining accuracy.

Future Directions

- Development of larger training sets, with more classes and content variety; the opposite of this direction is also important, i.e. training with limited datasets

- Exploration of novel architectures – for example, graph convolutional networks have received interest for action relationship modeling [13]

- Improvement on the recurrent networks used in video captioning, by using spatio-temporal attention mechanisms [14], as well as increasing temporal resolution via. dense captioning [15]

- Examining video-wide or long-term spatio-temporal structure – Existing models are generally limited to local, clip-level representations, activities, and patterns. Some recent examples of models that exploit long-term structure are [16] and [17]

- Models that more explicitly deal with camera motion and viewpoint changes within a scene [18]

- Building models which combine video with audio, text, and/or other inputs and sensors [19]

Closing remarks

Though many challenges remain, video ontology technologies have been receiving greater momentum from academia and industry in recent years. We are starting to see video recognition and reasoning models catch up to their image-based cousins and begin to be used to solve real-world problems in many areas. As they mature, video ontology techniques such as action classification are paving the way for exciting applications in the field of video understanding.

References:

1. A. Ganesan et. al. “Video based contextual question answering." *arXiv preprint. arXiv:1804.07399* (2018).

2. Y. H. Tsai et al. "Video relationship reasoning using gated spatio-temporal energy graph." *IEEE Conf. Comput. Vision and Pattern Recognit.* 2019.

3. J. Johnson et al. "Image retrieval using scene graphs." *IEEE Conf. Comput. Vision Pattern Recognit.* 2015.

4. R. Cong et al. "Review of visual saliency detection with comprehensive information." *IEEE Trans. Circuits Syst. Video Technol.* 2018.

5. H. Wang et al. "Dense trajectories and motion boundary descriptors for action recognition." *Int. J. Comput. Vision.* 2013.

6. H. Wang and C. Schmid. "Action recognition with improved trajectories." *Proceedings of the IEEE Int. Conf. Comput. Vision.* 2013.

7. D. Tran et al. "Learning spatiotemporal features with 3D convolutional networks." *IEEE Int. Conf. Comput. Vision.* 2015.

8. J. Carreira and A. Zisserman. "Quo vadis, action recognition? A new model and the kinetics dataset." *IEEE Conf. on Comput. Vision Pattern Recognit.* 2017.

9. L. Gao et al. "Video captioning with attention-based LSTM and semantic consistency." *IEEE Trans. Multimedia.* 2017.

10. Shang, Xindi, et al. "Video visual relation detection." *25th ACM Int. Conf. Multimedia.* 2017.

11. Z. Qiu, T. Yao, and T. Mei. "Learning spatio-temporal representation with pseudo-3D residual networks." *IEEE Int. Conf. Comput. Vision.* 2017.

12. B. Zhou et al. "Temporal relational reasoning in videos." *Eur. Conf. Comput. Vision.* 2018.

13. P. Ghosh et al. "Stacked spatio-temporal graph convolutional networks for action segmentation." *arXiv preprint. arXiv:1811.10575* (2018).

14. Y. Yu et al. "Video captioning and retrieval models with semantic attention." *arXiv preprint. arXiv:1610.02947* (2016).

15. Krishna, Ranjay, et al. "Dense-captioning events in videos." *IEEE Int. Conf. Comput. Vision.* 2017.

16. G. Varol, I. Laptev and C. Schmid. "Long-term temporal convolutions for action recognition." *IEEE Trans. Pattern Anal. Mach. Intell..* 2017.

17. B. Fernando et al. "Modeling video evolution for action recognition." Proceedings of the *IEEE Conf. Comput. Vision Pattern Recognit.* 2015.

18. P. Gay, J. Stuart, and A. Del Bue. "Visual Graphs from Motion (VGfM): Scene understanding with object geometry reasoning." *Asian Conf. Comput. Vision.* 2018.

19. K. Brady et. al. "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction." *6th Int. Workshop Audio/Visual Emotion Challenge.* 2016.

About Course5 Intelligence

Course5 Intelligence enables organizations to make the most effective strategic and tactical moves relating to their customers, markets, and competition at the rapid pace that the digital business world demands. We do this by driving digital transformation through analytics, insights, and Artificial Intelligence (AI). Our clients experience higher top line and bottom line results with improved customer satisfaction and business agility. As we solve today's problems for our clients, we also enable them to reshape their businesses to meet and actualize the future.

Rapid advances in Artificial Intelligence and Machine Learning technology have enabled us to create disruptive technologies and accelerators under our Course5 Intelligence suites that combine analytics, digital, and research solutions to provide significant and long-term value to our clients.

Course5 Intelligence creates value for businesses through synthesis of a variety of data and information sources in a 360-degree approach, solution toolkits and frameworks for specific business questions, deep industry and domain expertise, Digital Suite and Research AI to accelerate solutions, application of state-of-the-art AI and next-generation technologies for cognitive automation and enhanced knowledge discovery, and a focus on actionable insight.

Visit : www.course5i.com

Course5
Transformative intelligence