Social Media Video Analysis
Literature Review and Tools

Brett Bass and Evan Ezell

October 12, 2018

1 Literature Review

1.1 Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos [1]

Hou admits that video data is complex and there is a lack of annotations which make it more difficult to use neural networks. This makes our problem particularly hard, especially, when predicting views versus looking at objects or motion. Action detection typically follows the following two steps: frame-level action proposal generation and association of proposals across frames. Most methods use two-stream CNN framework which handles spatial and temporal features separately. Whereas Hou uses the tube approach, an end-to-end neural net which also captures space and time. Hou test their T-CNN on trimmed and untrimmed videos and find that it significantly out performs the two-steam CNN framework. More about the project can be found at http://crcv.ucf.edu/projects/TCNN/.

1.2 Large-scale Video Classification with Convolutional Neural Networks [2]

Karpathy et al provide a dataset called the Sports-1M dataset which is over 1 million sports videos taken from Youtube. Karpathy et al look at a variety of different architectures and discuss performance on the Sports-1M dataset and the UCF-101 data set. Karpathy et al find that CNNs are better that feature-based methods for analyzing video classification. Karpathy finds that CNNs are fairly robust in regards to what fusion architecture is used. These videos have been annotated automatically with 487 sports labels. The links to the video are provided but the metadata is not included. Use the links and the Youtube API to grab the metadata for as many videos as possible where the links are still valid. More about this project can be found at https://cs.stanford.edu/people/karpathy/deepvideo/.

1.3 Deep Learning for Video Classification and Captioning [3]

Wu provides an up to date summary of what research is available in video classification. Wu discusses the differences in video classification automatically labeling videos based on their semantic concepts; and video captioning generating a complete sentence describing the video which enhances a single label. Wu discusses the limitations of CNNs and expresses the need for RNNs. However, the RNN structure did not originally enable the ability to store past information for long periods of time which lead to the introduction of long short-term memory RNNs. Wu also describes the difficulty with video classification having an insufficient number of labeled videos. Part of the success in image classification should be attributed to the large number of labeled images available to train on. Ji wrote the seminal work first using a 3D CNN model which takes video input, or layered images [4].

2 Tools

2.1 Scraping Software

2.1.1 PAFY

- https://pythonhosted.org/Pafy/pafy-attributes
- Can scrape YouTube video attributes such as title, views (target), likes, dislikes, etc.
- Can also obtain mp4 videos for analysis
- Can also obtain full playlists of items for faster download and analysis

2.1.2 YouTube 8M*

- https://research.google.com/youtube8m/index.html
- Contains all video along with classification labels
- Breaks down videos into categories which would allow us to choose large number of videos easily
- Doesnt have view counts*
- But does have links that from which we could use Pafy to scrape iteratively*

2.2 Image Extraction and Processing

2.2.1 OpenCV2

- https://opencv-python-tutroals.readthedocs.io/en/latest/index.html
- Can break videos into series of images (If that is preferred structure of neural network)

2.2.2 Pillow

- https://pillow.readthedocs.io/en/5.3.x/
- Converts Images to matrices that can be used for analysis
- Image.open -¿ array = arrary(image)
- The array will have 3 dimensions, height, width, and color which will need to be reshaped according to how the CNN will need

2.3 Machine learning and CNNs

2.3.1 Keras

- https://keras.io/
- Can build CNNs of many types and architectures
- Can use transfer learning by retraining successful CNN architectures such as Googles Inception V3
- https://keras.io/applications/
- Keras also has feature extraction capabilities which may allow us to perform analysis more efficiently than looking at entire images

2.3.2 TensorFlow

- https://www.tensorflow.org/
- Keras must be run on a backend and they recommend TF

3 Video Analysis Strategies

3.1 Types of CNN architectures

- Classifying one frame at a time with CNN

- Use a 3D CNN
- Extract features from each frame with CNN and pass sequence to RNN
- Extract features from each frame with CNN and flatten and send to MLP (No need for live-time based prediction as full videos are released at once)
- https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5

3.2 Other Strategies

- Use transfer learning by utilizing currently well developed CNNs (Googles Inception V3)
- Most CNNs are used for classification, will need to use a linear activation in last layer to have a continuous output if we want to predict something like views
- Use classification for videos to determine occurrences of certain items to see what items in videos have the highest popularity
- These classifications could then be added as features describing occurrence or number of occurrences (scaled by video length)

References

[1] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (T-CNN) for action detection in videos. *CoRR*, abs/1703.10664, 2017. URL `http://arxiv.org/abs/1703.10664`.

[2] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Rahul Li Fei-Fei. Large-scale video classification with convolutional neural networks. pages 1725–1732. IEEE, 2014. ISBN 9781479951185.

[3] Zuxuan Wu, Ting Yao, Yanwei Fu, and Yu-Gang Jiang. Deep learning for video classification and captioning. *CoRR*, abs/1609.06782, 2016. URL `http://arxiv.org/abs/1609.06782`.

[4] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.59.