A video indexer for searching macroscopic features with deep learning

Gary Liu (jl25)

Introduction

This project explores the viability of a video indexing program that searches macroscopic features in videos with deep learning. More precisely, we would like to find out frames in an input video with overall effect that best match a given adjective description.

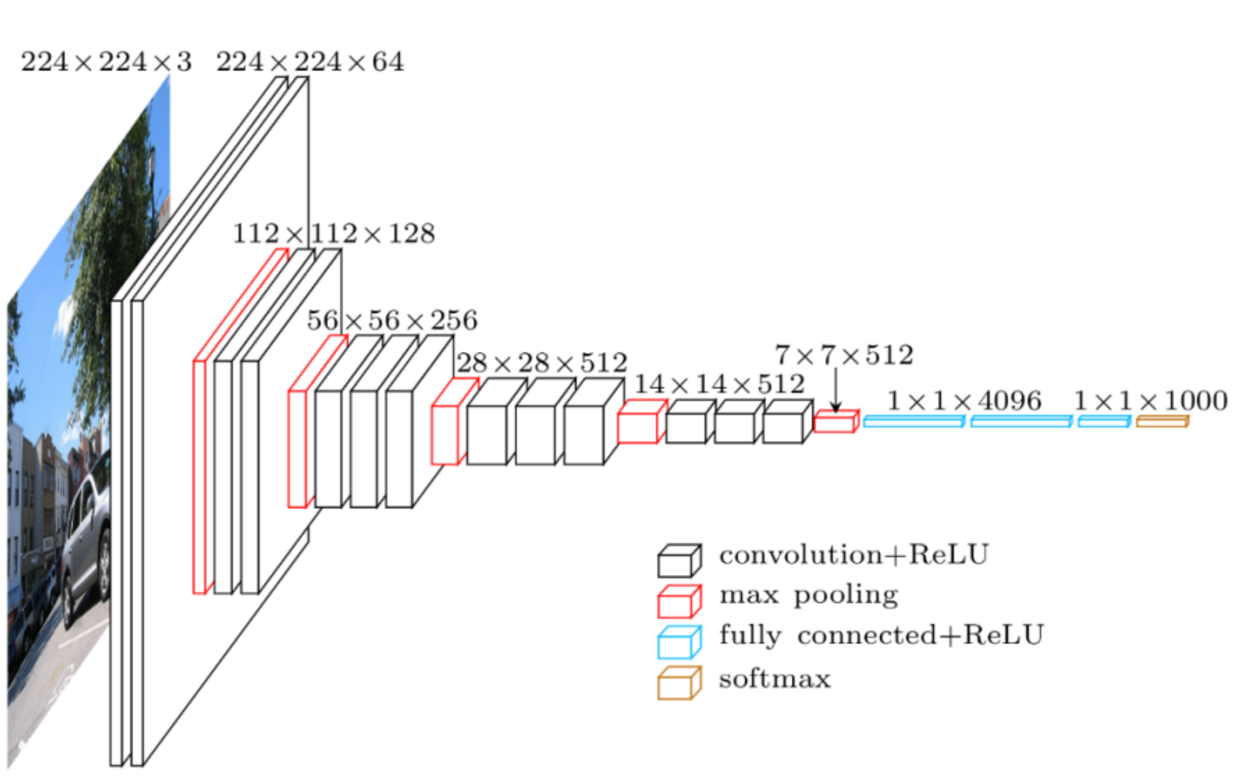
Method

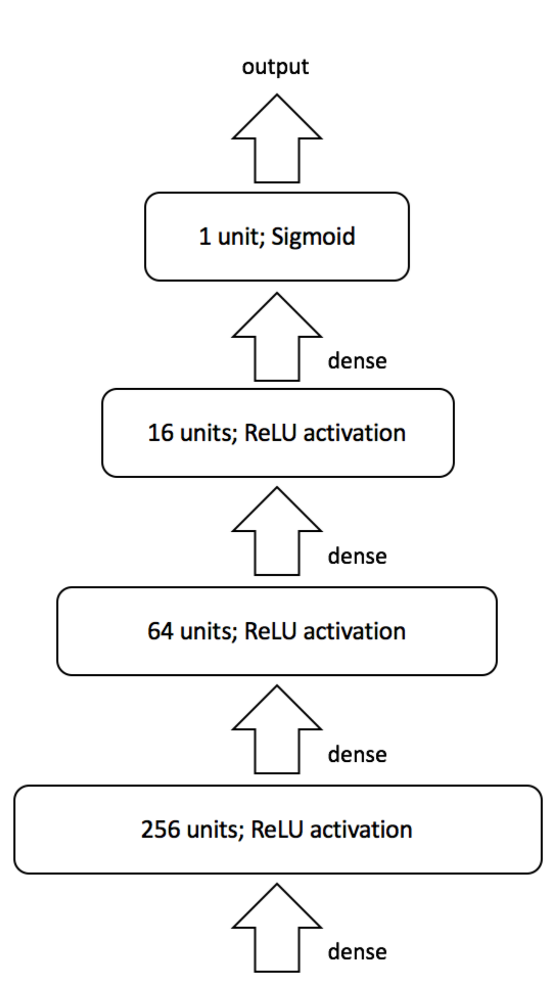
Performance metric: precision and recall

Steps
- Design and implement a model that binary classifies images
- Find positive and negative example images of the keyword as training data and train the binary classifier
- Parse the video into frames at a fixed interval and let the classifier predict whether each frame matches the keyword

Model
- A multi-layer fully-connected network on top of VGG-16 architecture
- Reuse bottleneck features of VGG-16 trained on ImageNet

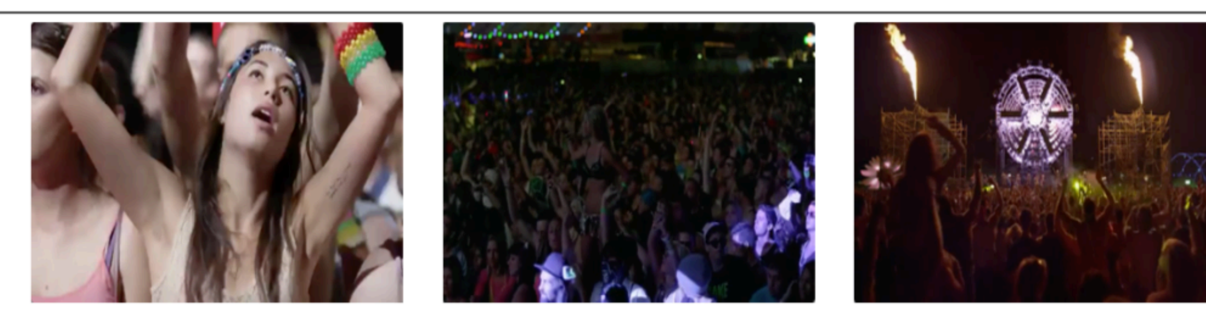
Data
- Training data retrieved from Google image searched with the keyword, the antonym of the keyword, and random images
- Testing video used documentary Timescapes, which features scenes that are easy to describe macroscopically





Result

Searches with selected keywords successfully recognize corresponding frames. The table below shows search keywords and exemplary frames recognized as matching the keywords in the video.

Keyword	Recognized frames
arid	
crowded	
mountainous	
starry	

Discussion

- Despite the fact that this model recognizes most of the matching frames (high recall rate), the result also includes many irrelevant frames (low precision rate).
  - In videos that we search through, positive frames should occur much less frequent than negative frames, but our selection of training data has nearly equivalent amount of positive and negative examples.
  - Negative examples should be characterized by "cannot be described by the keyword", not "can be described by the antonym of the keyword". The latter will introduce a large confusion region in the middle of the two classes. This analysis motivates us to collect more random images to use as negative examples in the training stage.
- Some negative examples retrieved from the image search engine are actually more likely to be positive examples.
  - Due to the fact that search engines may encounter problem generating counter-examples.
  - With such noise in the training data, the model can still reach a near-zero training error, so we speculate that some overfitting is present.

References

K. Simonyan, A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]
B. Chen, J. A. Ting, B. Marlin, and N. de Freitas. Deep learning of invariant spatio-temporal features from video. In NIPS Deep Learning and Unsupervised Feature Learning Workshop, 2010.
Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. 2011. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11). IEEE Computer Society, Washington, DC, USA, 3361-3368. DOI=http://dx.doi.org/10.1109/CVPR.2011.5995496