SwimMaster

CS230-Winter 2019

Erick Cardenas
ecasas@stanford.edu

Troy Ling
troyling@stanford.edu

Emilio Unda
eunda@stanford.edu

https://github.com/unda91/SwimMaster
https://youtu.be/O8-qavw5jcU

Introduction

Objective:
Build an action detection system that can automatically recognize swimming styles: ('Butterfly', 'Backstroke', 'Breaststroke', 'Front crawl' and 'Start').

Using:
Long Term Recurrent Convolutional Network.

- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN)
  - LSTM

Our best LRCN model consisted of a pre-trained CNN and a 4-layer LSTM model as the recurrent module, with **44%** accuracy.

Dataset

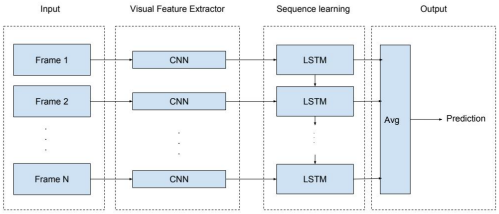
- Sports Videos in the Wild, 2015.
  - Videos shot by amateurs
    - Inconsistent and low quality
  - Contains 4200 videos
  - 244 swimming videos (119 are labeled)
  - Split into > 2600 examples

Features

Video segments:

- 1 second long (30 frames)
- 299 x 299 pixels
- RGB

Models



We implemented a Long-Term Recurrent Convolutional Network (LRCN) model. We used a pre-trained Inception V3 as the visual feature extractor. We experimented multiple LSTM models for the recurrent module. For this model, we used the categorical cross entropy as the loss function:

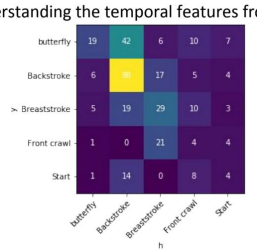
$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_j^{(i)} \log(P_j^{(i)})$$

Results

Name	Visual Feature Extractor	Sequence Learning	Output	Dev Accuracy
Deep Swim Master	Inception V3	LSTM x4 (512, 256, 128, 64)	Dense Softmax	44%
Wide Swim Master	Inception V3	LSTM x1 (1024)	Dense Softmax	34%
Convolutional Baseline	Conv (3x3x5), (5x5x10), (5x5x40)	Flatten	Dropout (0.4), Dense Softmax	28%

Discussion

Our best model achieved approximately **44%** accuracy. Throughout the project, we discovered the dataset we used was not large enough to build a robust action recognition classifier. The dataset was also imbalanced and skewed to a certain label (i.e. Backstroke), which led our models to overfit the training data. In addition, our experiments indicate that using a deeper LSTM model results in a better performance, as the model is more capable of understanding the temporal features from video data.



Future

Regarding the future work, we would try to increase the model's accuracy by using:

- Building a more robust data preprocessing pipeline and more videos.
- Modifying the model for multiple object action detection.

References

1) Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, DeanCraven **Sports Videos in the Wild (SVW): A Video Dataset for Sports Analysis Proc. International** Conference on Automatic Face and Gesture Recognition (FG 2015)
2) Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna **Rethinking the Inception Architecture for Computer** VisionCoRR, 2015
3) Donahue, Jeffrey, et al. **Long-term recurrent convolutional networks for visual recognition and description.** Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.