Augmented Human Action Classification with Joint Estimation via double CNN

Colin Zheng, Yixian Li, Tianheng Shi

Stanford University

Stanford CS231N

Background

- Human action classification is a challenging task with wide applications.
- Potential classification accuracy increase with deeper understanding of human pose
- Include human joint keypoint as input along with image features to increase classifier performance

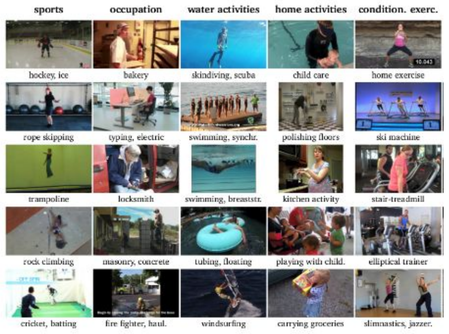
Problem Statement

Increase classifier accuracy with two step models:

1. **Joint Localization:** Identify the pose key-points from the input images through an upstream model
2. **Human Action Recognition:** Incorporate keypoint info in classifier
   a. Linear joint embedding
   b. Convnet joint embedding **[1]**

Dataset

- MPII Dataset

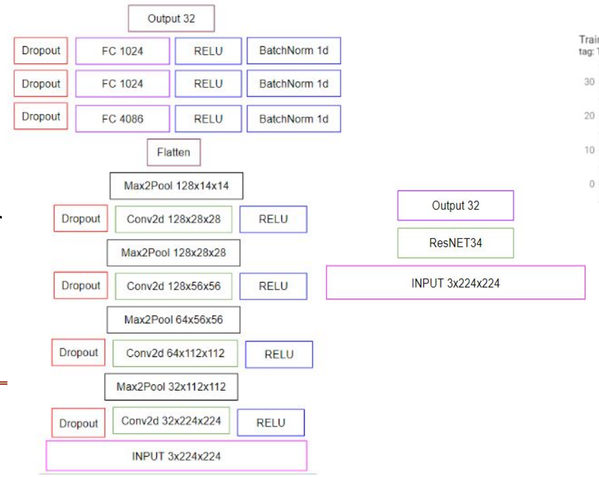


- 16 Joint pairs labels
- 20 Activity Category labels
  - Filtered into 15 Classes due to data imbalance

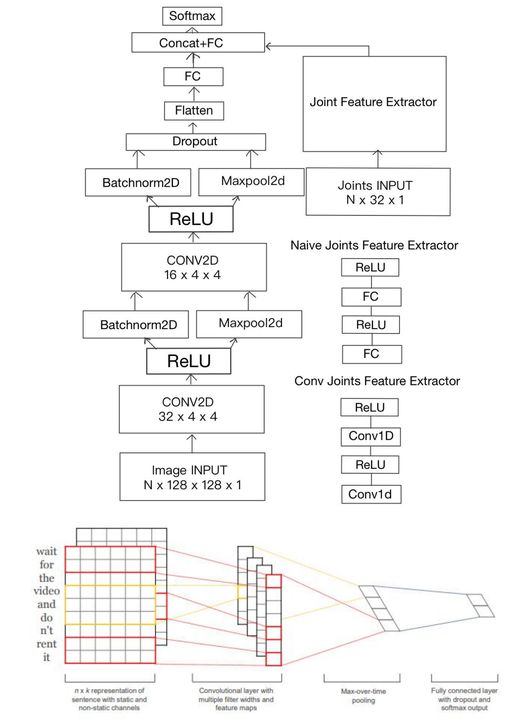
Data Preprocessing

- Image
  - Resized to 3x224x224
  - Zero-center per channel, per image
- Joints information
  - x, y, and visibility flag (bool)
  - x, y are resized in range 224, and then subtracted by 112
- Data Distribution
  80% training, 10% validation, 10% test

Joint Localization Models

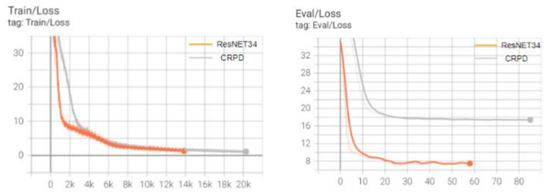


Classification Models



Results

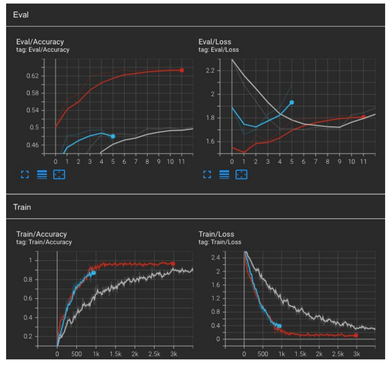
- **Joint Localization**



Part/Model	CDRP	ResNET34	Hourglass
Head	23.84	30.46	95.46
Shoulder	36.98	43.83	90.06
Elbow	30.30	37.91	81.44
Wrist	21.18	27.52	75.40
Hip	31.71	39.44	80.82
Knee	24.10	34.02	75.54
Ankle	18.21	26.99	70.67
Mean	26.00	34.37	81.46

Matric Use: PCKh@0.5

- **Classification**



- Blue: Baseline
- Gray: Naive Joints Feature Extractor
- Red: Convolutional Feature Extractor

Model	Train Loss	Train Acc	Val Loss	Val Acc
Baseline	0.542	0.8321	1.869	0.4834
LFE	0.1522	0.9531	1.715	0.5219
CFE	0.1419	1.0	1.458	0.6353

Discussion

- **Joints Localization**
  - Baseline Model
    - Low training loss, high val loss, low val accuracy.
  - ResNET34
    - Higher val accuracy and lower val loss.
    - Predicted joints may be off the ground truth, but the skeleton indicates the human action.
  - 8 Stacked Hourglass
    - benchmark, high accuracy, better joint prediction.
- **Classification**
  - Baseline Model
    - Overfitting
    - Data Imbalance
    - Human Body too small
  - Naive Joints Feature addition
    - Minor enhancements
    - Pure joints information without correlations
  - 1D ConvNets
    - Joints connection information preserved

Future Work

- Better Dataset
- Improve accuracy of localization using different models and loss.
- Transformer Model for Self-attention on joints information for classifier

[1] Yoon Kim. Convolutional neural networks for sentence clas-sification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, 2014

Stanford University