**Evaluation**

Poselets from various action images were obtained from PANDA framework. Features were extracted from full images for all the actions and their corresponding poselets obtained previously using RCNN framework. The number of poselets obtained from each image varied and also each poselet has an id. PANDA framework provides 150 different categories of poselets (human face, human body, human legs, car etc.) for attributes classification. Therefore in current research, each poselet obtained will be confined to one of the 150 poselets defined previously. Because of computational limitations, only top poselets corresponding to each image were extracted. Number of features obtained for each image is 4096. Features from top 5 poselets corresponding to same image were concatenated with the features of the full or original image and was then used for creating classification model.

Two different paradigms were adopted to create the models. In the first paradigm the total number of features were 618,496 (150\*4096 for 150 possible poselets + 1\*4096 for original image). As previously mentioned, only top 5 poselets are used for action classification therefore only 5\*4096 out of 150\*4096 features have values and rest all of the 145\*4096 features are left empty as they are not present in the current image. In second paradigm, the total of features for each image becomes 24,576 (1\*4096 + 5\*4096) only that is only features with the poselets were used to represent the images rest all were discarded.

One-against-one SVM model with polynomial kernel of order four was then trained to classify different actions. Because it takes a lot of time to extract features from each image using RCNN, the project is limited to only four actions as compared to 10 available in the VOC 2012 data. 10-fold cross validation was performed to analyze the performance of the adopted approach. The evaluation matrix includes: accuracy, average precision and the quality of poselets obtained.

**Evaluation** (by nehchal)

We test our results on the standard Pascal VOC 2012. It contains 10 action categories (how many images of each categories). However, due to constrained resources of computational power and time, we test our approach on 4 action categories - jumping reading, riding horse and running. 10-fold cross validation was performed to analyze the performance of our approach. The evaluation matrix has been provided in the table.

**Observation and Results**

From Figure 1, it is clear that the Paradigm 1 did not do well with mean average precision of 44.65 %, which is approximately 40.6 % lesser than mean of other alternative methods and an overall accuracy of 49.6 %. In most of the actions, the current study performed worse except for reading action, where it gave 53.8 % better than the mean of other methods. Overall this approach of concatenating all the features in the order of the poselet id makes the data very sparse and at a time only 4% of all the features are available and rest are absent. The total number of features in this method is 618,496 while the number of images for each action in around 100. Therefore, the data is underfit with respect to the features. This issue can be resolved by applying SVD or performing dimensionality reduction. The dominating features would be reflected and it can be further used for classification.

Figure 1: Paradigm 1 Results (Average Precision) vs. Base Paper and Other Researches

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Action** | **Jumping** | **Reading** | **Riding Horse** | **Running** | **mAP** |
| **Stanford** | 75.7 | 44.4 | 94.2 | 87.6 | 75.5 |
| **Oxford** | 77.0 | 39.5 | 95.9 | 87.7 | 75.0 |
| **Oquab et al.** | 74.8 | 45.3 | 95.0 | 86.5 | 75.4 |
| **Base Paper R-CNN + fc6 SVM** | 76.2 | 42.2 | 94.3 | 87.0 | 74.9 |
| **PANDA + R-CNN + SVM** | 20.7 | 67.0 | 35.0 | 55.9 | 44.65 |

Figure 2: Paradigm 2 Results (Average Precision) vs. Base Paper and Other Researches

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Action** | **Jumping** | **Reading** | **Riding Horse** | **Running** | **mAP** |
| **Stanford** | 75.7 | 44.4 | 94.2 | 87.6 | 75.5 |
| **Oxford** | 77.0 | 39.5 | 95.9 | 87.7 | 75.0 |
| **Oquab et al.** | 74.8 | 45.3 | 95.0 | 86.5 | 75.4 |
| **Base Paper R-CNN + fc6 SVM** | 76.2 | 42.2 | 94.3 | 87.0 | 74.9 |
| **PANDA + R-CNN + SVM** | 81.2 | 68.6 | 63.3 | 75.62 | 72.3 |

Paradigm 2 performed fairly well with an average precision of 72.3 % which is close to the mean(75.4 %) of other competitive methods. The model also had an overall accuracy of 74.4% which is nearly 50% higher than that of the Paradigm 1. Secondly, the current methodology outperforms all other methods in detecting jumping action and reading action, while it under-performed for actions running and riding-horse. The improvement in jumping action is 6.7% over the average precision by other methods. For class reading the improvement obtained is around 62% than other methods, but there is similar reduction for action riding-horse. Overall the model performed really well and it is because the number of features is much lower than Paradigm 2 and the data does not underfit.

Experimental Setup

Pose Estimation evaluation (top 5). How are they tested?

You can add about results on pose estimation and visualization.