CSC 249/449 Machine Vision: Final Project

Term: Spring 2019
Instructor: Prof. Chenliang Xu
TA: Yapeng Tian, Jing Shi, Zhiheng Li, Yutong (Kelly) He, Chaoying Xue
Due Date: 11:59 pm, 04/20/2019

In this final project, you are going to build deep learning models for two tasks on A2D dataset [18], which contains 3782 videos from YouTube. In each video, objects are annotated with actor-action label, meaning that an actor is performing an action (e.g. dog-running). Both bounding boxes and semanctic segmentation annotations are provided. For more details of A2D dataset, please visit `http://web.eecs.umich.edu/~jjcorso/r/a2d/`.

Since A2D dataset is too large to be trained on a single GPU, you only need to use a smaller portion of A2D. Besides, template code of each task, including code of baseline model, evaluation, data loader, is also provided.

Here are tasks you need to work on. For more details of each task, please refer to *README.md* of each task's GitHub repository.

Task 1 (60 pts): Multi-Label Actor-Action Classification (Warm-Up)

Description: Build a model to predict classes of actor and action in each frame. Since some frames may have multiple actors performing different actions, this is a multi-label classification problem.

Evalutaion Metric(s): We use precision, recall, and F1-score to measure performance of trained models. The descriptions about the three metrics can be found in course slides or https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9.

Performance Expectation: We expect your model performance should be better than Precision: 23.8 Recall: 30.5 F1: 25.2.

Template Code: csc_249_final_proj_a2d_cls

You must choose **ONE** of the following tasks to complete. We will rank the teams based on your performance. The top teams will be invited to give a presentation of their methods to the class. The oral presentations will be given **extra 20 points**.

Task 2: (40 pts): Actor-Action Detection (Option 1)

Description: Build a model to

- predict boxing boxes of objects (xyhw, xy position of the bounding box with height and width)
- predict both actor and action classes for each bounding box

Evalutaion Metric(s): mean Average Precision (mAP).

Performance Expectation: mAP $\geq 22.5\%$

Code Template: csc_249_final_proj_a2d_det

Task 3: (40 pts): Actor-Action Segmentation (Option 2)

Description: Predict segmentation map of each frame. Predictions of both actor class and action class should be given at each pixel in a segmentation map.

Evalutaion Metric(s): mean accuracy and mean IoU

Performance Expectation: mean accuracy $> 32.77\%$, mean IoU $> 23.37\%$ on validation set.

Code Template: *csc_249_final_proj_a2d_seg*

Here are some tips you may consider in your model:

1. Which backbone model should be used?
2. How to leverage temporal information (e.g. multiple frames, optical flow) for action recognition?
3. For actor-action classification, is it better to decouple this problem into two independent classification problems or should we regard each actor-action pair as a unique category to do the classification? Please also note that some actor-action pairs are invalid in this dataset (e.g., adult-fly).

Besides, you may also refer to some paper listed in **References**.

Submission:

Your submission should contain the followings:

- code - The implementation of your model.
- write_up.pdf - In this file, please explain you models of each task in several aspects:
  - Method description (e.g., preprocessing method, network architecture (pretrained or not), losses, optimization method, number of iterations/epochs of convergence, hyperparameters, etc.).
  - Novelty of your method, which cannot be too simple (e.g., more training epochs, larger learning rate).
  - Performance on validation set.
- Prediction result on testing set. Please refer to *README.md* of each code template for more details.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] J. Ji, S. Buch, A. Soto, and J. Carlos Niebles. End-to-end joint semantic segmentation of actors and actions in video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 702–717, 2018.

[9] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid. Joint learning of object and action detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4163–4172, 2017.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[13] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[15] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[17] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[18] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso. Can humans fly? action understanding with multiple classes of actors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2264–2273, 2015.