CSC 249/449 Machine Vision: Final Project

Term: Spring 2020
Instructor: Prof. Chenliang Xu
TA: Jing Shi, Zhiheng Li, Haitian Zheng, Tianyou Xiao, Yihang Xu, Sizhe Li, Weitao Tan, Xiaoning Guo
Due Date: 11:59 p.m., 04/23/2020

In this final project, you are going to build deep learning models for a task on A2D dataset [6], 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 is the task you need to work on. For more details, please refer to *README.md* in *csc_249_final_proj_a2d_cls*.

Multi-Label Actor-Action Classification (100 pts)

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. Note that you should **NOT** use ground-truth bounding boxes or semantic segmentation map in training or testing stage.

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

Submission:

Your submission should contain the followings:

- code - The folder of the implementation of your method.
- 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.). You also need to prepare a figure about the overview your method (You can refer to Figure 1 in [5]).
  - Novelty of your method. Note that this cannot be trivial (e.g., more training epochs, larger learning rate). **Methods without good novelty will not receive good grades.**
  - Performance on validation set.
- presentation.pdf (or presentation.ppt(x)) - This file is a presentation version of the write up. Method description, novelty, and performance should be reported in this file.
- Prediction result on testing set. Please refer to *README.md* of the 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] J. Chen, Z. Li, J. Luo, and C. Xu. Learning a weakly-supervised video actor-action segmentation model with a wise selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.