Towards Efficient Coarse-to-Fine Networks for Action and Gesture Recognition

Niamul Quader[1], Juwei Lu[1], Peng Dai[1], and Wei Li[1]

Huawei Noah's Ark Lab, Canada
{niamul.quader1,juwei.lu,peng.dai,wei.li.crc}@huawei.com

1 Ablation study on Jester gesture recognition dataset (Contd.)

We implemented a number of multi-stream C2F implementations using training concepts from prior multi-stream ConvNets for action/gesture recognition (e.g. lateral connection similar to [2] applied in C2F pathways). However, none of these prior training methods worked on a straightforward multi-stream C2F ConvNet - accuracy of action/gesture recognitions was consistently lower than that of baseline ConvNet (Figure 1). This is likely due to the lack of complementary information in C2F ConvNet multi-stream pathways compared to prior multi-stream networks where different streams are specialized on different aspects of the input signal (e.g. spatial and motion stream in [8]). In contrast, our proposed $C2F_C$ is an effective C2F implementation that considerably improves ConvNet accuracy over baseline ConvNet (Figure 1).

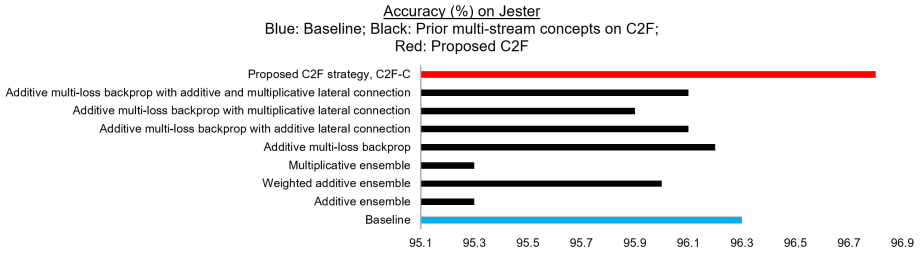

Fig. 1. Accuracy of C2F with different training approaches on Jester gesture recognition dataset.

2 Action/gesture recognition at coarse pathways

We hypothesized that many action/gesture classes can be inferred at coarse pathways of our C2F networks. In this supplementary document, we try to identify which of the action/gesture classes are more prone to be identified in coarser pathways and therefore can benefit more from the coarse-exit scheme. More specifically, we probe into two separate ratios for each of the gesture/action classes: First, the ratio of exits in the coarse, mid and $C2F_{A,En}$ pathways when approximately 80% - 85% of all videos in a dataset are classified in either the

coarse or mid pathways; second, the ratio of misclassified videos in each of the coarse, mid and C2F-En pathways compared to all the misclassified videos. We provide separate figures containing the above-mentioned ratios (plotted as percentages) in the supplementary materials for each of the following datasets: Jester [9], Something-SomethingV1[3], Something-SomethingV2[3], Kinetics[1].

We find that the ratio of early exits for most action/gesture classes at the coarsest pathway is higher than the corresponding ratio of misclassified videos in the coarsest pathway for those classes, meaning that the coarsest pathway is handling a larger percentage of the videos belonging to a class and is not failing as much as the $C2F_{A,En}$ pathway. A likely reason for the above-mentioned behavior is that the coarser pathways only infer on the relatively simpler videos and invoke the finer pathways for the more difficult videos.

Select examples where the coarsest pathway does well:
1. On the 'Pulling hand in' class in the Jester dataset, the coarsest pathway recognizes about 68% of the videos, but out of all the misclassifications the coarsest pathway is responsible for only 17% misclassifications.
2. On the 'Pretending to put something next to something in' class in the Something-Something-V1 dataset, the coarsest pathway recognizes about 48% of the videos, but out of all the misclassifications the coarsest pathway is responsible for only 8% misclassifications.
3. On the 'Lifting a surface with something on it but not enough for it to slide down' class in the Something-Something-V2 dataset, the coarsest pathway recognizes about 48% of the videos, but out of all the misclassifications the coarsest pathway is responsible for only 11% misclassifications.
4. On the 'Playing squash or racquetball' class in the Kinetics dataset, the coarsest pathway recognizes about 98% of the videos, but has no misclassification at all.

Out of the above examples, 'Playing squash or racquetball' is particularly interesting since almost all of the videos belonging to this class were identified at the coarsest pathway and there were 0 misclassifications at that pathway. Similar to this class, if a task at hand requires recognizing only a subset of actions/gestures, we think it is worthwhile to cross-check with our findings and see if $S-1$ of those classes (where S is the total number of classes) can be reliably recognized at coarser scales. In such a scenario where $S-1$ classes can be reliably recognized at coarser scale inputs, an action/gesture classifier could be built on coarse inputs requiring fewer computation resources and also requiring less expensive cameras.

3 Test results on Jester

$C2F_{C+,En}$ on a 30-crops softmax-averaged evaluation (similar to [4]) on the Jester dataset yields a state-of-the-art top-1 test accuracy of 97.09% compared to test results reported in refereed articles (Table 1).

Fig. 2. Ratio of early exits and ratio of misclassified videos for the Jester dataset. Here, red bars correspond to coarsest pathway, yellow bars correspond to mid pathways and green bars correspond to the $C2F_{A,En}$ pathway.

Table 1. Top-1 test and validation accuracy of C2F-En+ and competitive refereed methods that reported test results for the Jester dataset [9]. C2F-En+ beats the closest competitor [7] with a misclassification rate reduction by around 14%.

Method	Test-1 Acc.(%)	Val-1 Acc.(%)	△Test-1
C2F-En+	**97.09**	**97.28**	**-0.0**
Deformable ResNeXt3D [7]	96.60	-	-0.49
CPNet [6]	96.56	96.70	-0.53
MFF [5]	96.28	96.33	-0.81

Fig. 3. Ratio of early exits and ratio of misclassified videos for the Something-Something V1 dataset. Here, red bars correspond to coarsest pathway, yellow bars correspond to mid pathways and green bars correspond to the $C2F_{A,En}$ pathway.

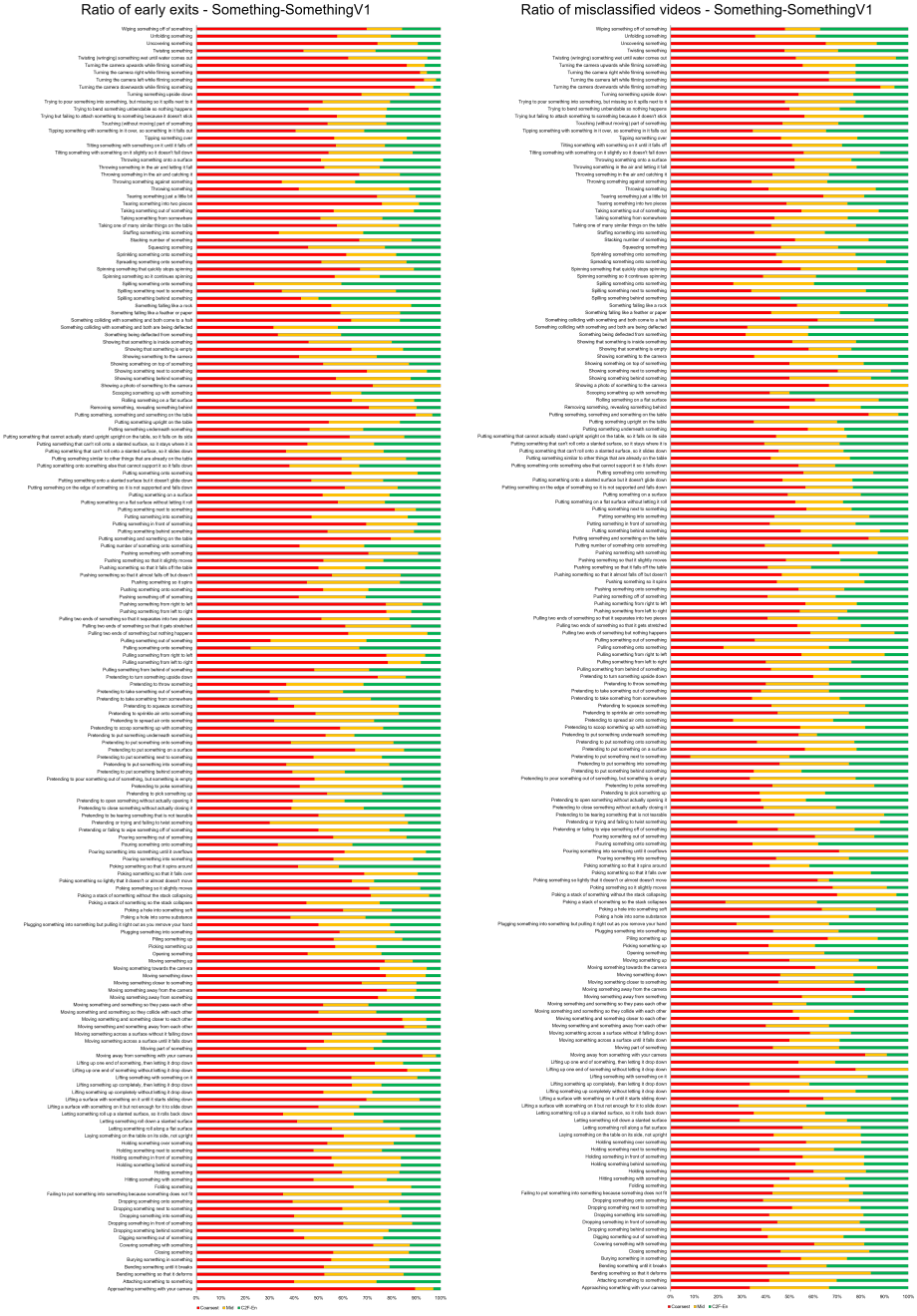

Fig. 4. Ratio of early exits and ratio of misclassified videos for the Something-Something V2 dataset. Here, red bars correspond to coarsest pathway, yellow bars correspond to mid pathways and green bars correspond to the $C2F_{A,En}$ pathway.

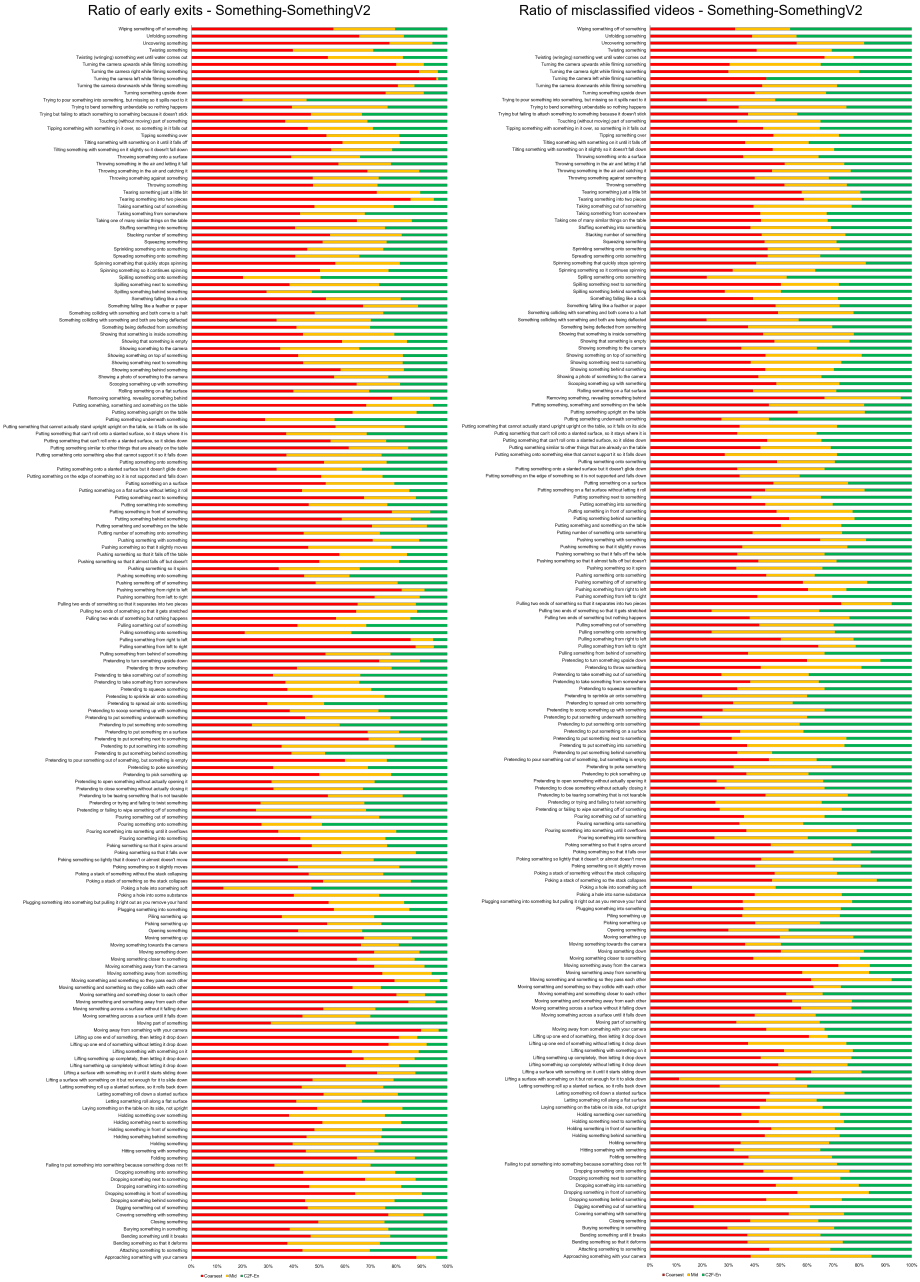

Fig. 5. Ratio of early exits and ratio of misclassified videos for the Kinetics dataset (first 200 classes). Here, red bars correspond to coarsest pathway, yellow bars correspond to mid pathways and green bars correspond to the $C2F_{A,En}$ pathway.

Fig. 6. Ratio of early exits and ratio of misclassified videos for the Kinetics dataset (last 200 classes). Here, red bars correspond to coarsest pathway, yellow bars correspond to mid pathways and green bars correspond to the $C2F_{A,En}$ pathway.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019)
3. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV. vol. 1, p. 3 (2017)
4. Jiang, B., Wang, M., Gan, W., Wu, W., Yan, J.: Stm: Spatiotemporal and motion encoding for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2000–2009 (2019)
5. Kopuklu, O., Kose, N., Rigoll, G.: Motion fused frames: Data level fusion strategy for hand gesture recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 2103–2111 (2018)
6. Liu, X., Lee, J.Y., Jin, H.: Learning video representations from correspondence proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4273–4281 (2019)
7. Shi, L., Zhang, Y., Hu, J., Cheng, J., Lu, H.: Gesture recognition using spatiotemporal deformable convolutional representation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1900–1904. IEEE (2019)
8. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
9. The 20bn-jester dataset v1: *The 20BN-jester Dataset V1*, [Accessed 8th November 2019]