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

Abstract. State-of-the-art approaches to video-based action and gesture recognition often employ two key concepts: First, they employ multistream processing; second, they use an ensemble of convolutional networks. We improve and extend both aspects. First, we systematically yield enhanced receptive fields for complementary feature extraction via coarse-to-fine decomposition of input imagery along the spatial and temporal dimensions, and adaptively focus on training important feature pathways using a reparameterized fully connected layer. Second, we develop a 'use when needed' scheme with a 'coarse-exit' strategy that allows selective use of expensive high-resolution processing in a data-dependent fashion to retain accuracy while reducing computation cost. Our C2F learning approach builds ensemble networks that outperform most competing methods in terms of both reduced computation cost and improved accuracy on the Something-Something V1, V2, and Jester datasets, while also remaining competitive on the Kinetics-400 dataset. Uniquely, our C2F ensemble networks can operate at varying computation budget constraints.

Keywords: Action and gesture recognition, spatiotemporal coarse to fine decomposition, weight reparameterization, budgeted computation

1 Introduction

The human visual system appears to process information across multiple spatial and temporal scales in a coarse-to-fine (C2F) fashion [22]. A key advantage of such a strategy is that perception can be achieved without reliance on the most expensive high resolution processing, as analysis can exit when the needed level of precision has been achieved. Additionally, if all actions do not require equal resolutions, there is no reason to use equal input resolutions to identify all actions.

Although deep learning approaches have achieved enormous success in reliably recognizing action/gesture using ConvNets, e.g. [28, 9, 44, 34, 36, 42, 17, 4, 3, 38] and at very low latency, e.g. [17, 36, 13, 44], all of these approaches still require large computation costs (FLOPs) and energy at runtime. For instance, one of the most efficient and high performing deep Convolutional neural network (ConvNets) for action/gesture recognition [17] required 33 GFLOPs of computations for a single recognition. With this method, if we estimate a 6 GFLOPs

per Joule [1] energy consumption, then gesture/action recognition at every 0.1 second interval for around 15 minutes will completely deplete a powerful mobile battery with a capacity of 50,000 Joules. Additionally, such ConvNets cannot adapt to varying FLOPs constraints (e.g. mobile switching from normal mode to battery saver mode) - a problem that prior works attempt to address by retraining a family of ConvNets with the same baseline (e.g. TSM-ResNet50 family [17] - each of the ConvNets in the family operates on different discrete FLOPs). Overall, there is a need for action/gesture recognizing ConvNets that can operate efficiently and accurately at continuously varying computational cost constraints.

Based on this motivation, we present the first C2F cascaded ensemble network for action/gesture (Fig. 1). At inference, the C2F network starts with attempting to recognize action/gesture using a coarse resolution version of the input video clip. This coarse stream acts as a fast recognition pathway requiring a small number of computations. Only when the coarse pathway is not confident at recognizing an action/gesture, does the next finer resolution pathway gets invoked in a C2F manner (Fig. 1).

While computation costs can be decreased with a C2F scheme built on a baseline ConvNet (e.g. baseline TSM [17] used in each of the C2F pathways), identifying a C2F feature fusion scheme and a multi-pathway loss formulation that enable the C2F ensemble to have better accuracy than the baseline ConvNet is challenging. This is likely due to the following two reasons: First, in contrast to prior multi-scale ConvNets on action/gesture recognition (e.g. SlowFast [4]), the coarse pathways of the C2F are downsampled input videos containing only a subset of information compared to the input at finest pathway - so, it is more difficult to extract complementary information from the coarser inputs; second, the choices on C2F feature fusion strategy and multi-pathway loss formulation can cause some pathways dominating over the rest in the C2F ensemble (e.g. bidirectional feature fusion leading to spatial stream dominating motion stream in [5]). To address the above challenges, we propose the following:
1. A sub-network that fuses features from the ends of each of the C2F pathways using a reparameterized fully connected (FC) layer that adaptively excites gradient flow along the more important feature pathways during training.
2. An end-to-end differentiable multi-loss formulation to train the C2F network.
3. Another multi-loss formulation to train the C2F network when the baseline finest pathway ConvNet is already trained.

Using the above technical contributions, this paper shows the potency of C2F networks in action/gesture recognition by providing the first set of C2F benchmarks that considerably improve baseline ConvNet performances (i.e., improved accuracy and reduced computation cost). Additionally, we implement a coarse-exit scheme that is controllable by a hyperparameter, and implement a controller that adjusts the hyperparameter such that the C2F network operates at a computation budget assigned externally. The ability to operate continuously on a cost-accuracy curve (e.g. Fig 2) has considerable technical benefits - we no longer need to retrain different models to fit varying computational budget constraints.

2 Related work

Multi-stream processing in action/gesture recognition: Multi-stream learning is frequently used in state-of-the-art ConvNets for action/gesture recognition (e.g. [14, 27, 5, 3, 12]). For example, the first two-stream network architecture [27] and some of its variants [5, 6, 3] have achieved competitive results by combining ConvNet activations that are generated from a spatial stream and a temporal stream. The spatial stream usually has a single crop of the RGB video with a small number of frames, whereas the temporal stream has the corresponding optical flow [27, 5, 3] or a large number of RGB frames with low resolution [4]. In contrast to multiple streams having complementary inputs, a 3D C2F decomposition of the input video results in coarse pathway inputs that are spatiotemporal downsampled signals of the finest pathway input; therefore, a C2F decomposition lacks complementary information at different streams compared to other multi-stream structures [14, 27, 4, 3]. Despite this lack of complementary information at multi-stream inputs in a C2F network, we show that C2F schemes can be very potent in action/gesture recognition by providing a number of benefits: improved accuracy, reduced computation cost, and budgeted inference.
C2F recognition to increase accuracy and reduce FLOPs: C2F-guided processing has a long history in spatial visual processing [25, 24, 15, 16] - collecting features from coarser pathways computed at a cheaper cost and fusing them with finer pathways can increase accuracy without adding much computational burden. These potential improvements in performance come from extracting features at multiple scales and establishing a hierarchy of structural features extracted at those multiple scales [15, 16]. In action/gesture recognition, many research investigated use of multiple temporal and/or spatial resolutions for increasing accuracy [17, 36, 42, 4, 14]; however, none have looked into the 3D spatiotemporal decomposition of the input video for C2F feature extraction. This could be likely due to the lack of complementary information in C2F pathways and due to the difficulty in formulating multi-loss joint optimization for multi-stream pathways consisting of greater than two streams. Also, use of C2F prediction cascades [35, 29, 39, 11] and dynamic inference approaches [18, 40] to reduce computational complexity is common in object classification/detection tasks; however, their use in action/gesture recognition is rare. Overall, the utility of combining considerably cheaper 3D spatio-temporal coarse pathways and expensive finer scale pathways in a cascade setting and the implementation of budgeted inference are missing in action/gesture recognition tasks.

3 Technical approach

We begin by documenting the layout for a generic C2F network (Fig. 1). Each of the streams in the C2F ensemble network can be any ConvNet (e.g. TSM [17], R3D [33], etc.) that works on a clip of video with either dense sampling [33, 41] or strided sampling [17, 42, 36] and recognizes actions/gestures. The C2F ensemble can be composed of an arbitrarily large number of C2F pathways (N). In this

paper, we use $N = 3$ and we call the coarsest pathway C, the finer pathway M and the finest pathway F (Fig. 1). We decompose input video spatiotemporally that compensates for decaying effective receptive field of convolution kernels [20] (Sec. 3.1), fuse features for complementary processing between the C2F pathways (Sec. 3.2), and formulate two joint optimization schemes for the alternative scenerio where the baseline ConvNet (i.e. the ConvNet at the finest pathway) is either pre-trained or not pre-trained for the action/gesture recognition task at hand (Sec. 3.3 and 3.4). For C2F inference, we implement a C2F coarse-exit scheme with a controllable hyperparameter and a controller that can operate the C2F network in a given budget computation cost (Sec. 3.5).

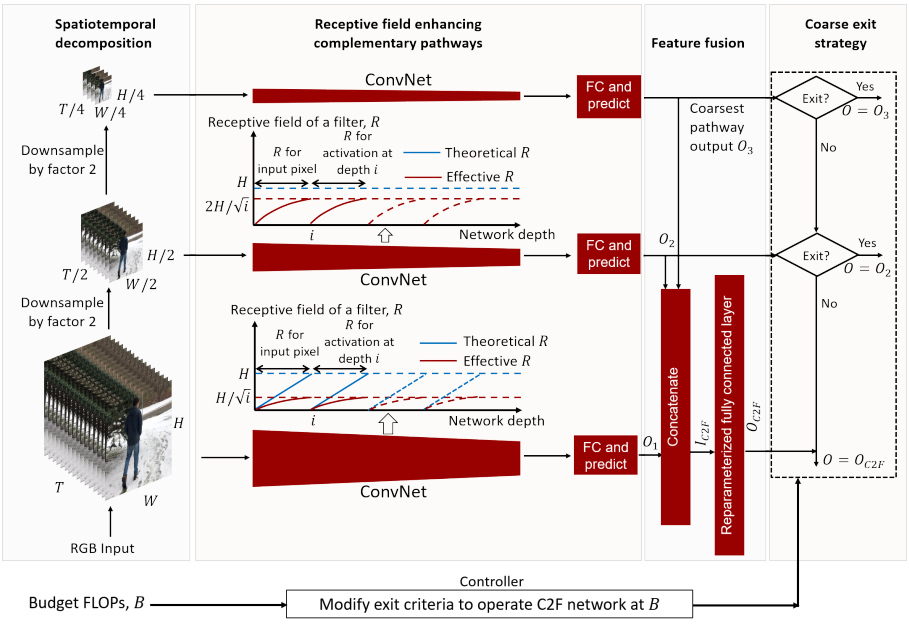

Fig. 1. The layout of our C2F network. Sampled frames of an input video are downsampled in both the spatial and temporal domains. A filter kernel at a depth i has a progressively larger receptive field for coarser pathways relative to the original input video size. This is demonstrated here with insert graphs that assume the ConvNets used in F, M and C all have same-sized filter sizes (size>1) and that the size of the activation maps does not change at different network depth. C2F features are fused using a reparemterized FC layer. During inference, a coarse-exit scheme encourages recognition output (O) at coarser scales. Additionally, for a given computation budget B, a controller modifies the coarse-exit decision criteria such that C2F operates close to the given budget.

3.1 Enhanced receptive fields via spatiotemporal decomposition

A recent study showed that using the full image rather than center cropping improves accuracy [17] suggesting that pixels in a frame that are located far away from the center could provide valuable information as well and that enhancing receptive field of convolution kernels can be important for action/gesture recognition. Therefore, we propose to systematically enhance receptive fields by spatiotemporally downsizing the input video clip in a C2F manner while compensating for the decreased resolution in the coarser pathways by complementary processing of higher resolution inputs at the finer pathways (Fig. 1). This enhanced receptive field also partially compensates for the $1/\sqrt{(i)}$ decaying effective receptive field in ConvNets [20] where i is the number of layers between the input and the convolution kernel. An illustration of compensating for the decaying receptive field is shown in Fig. 1.

3.2 Feature fusion using reparameterized FC layer

We concatenate the pre-softmax outputs from each of the C2F pathways (Fig. 1), and fuse these accumulated features (I_{C2F}) through a reparameterized FC layer that adaptively excites gradient flow along the more important features of I_{C2F} during training. Intuitively, for a particular output node of the C2F ensemble (e.g. output node representing 'moving hand from right to left' gesture), there are likely some I_{C2F} features that are more important than others (e.g., corresponding nodes for 'moving hand', 'right to left', 'left to right' in each of the C2F pathways more important than the other nodes). Adaptively exciting gradient flow along these important nodes leads to larger gradient backpropagation along all the learnable ConvNet parameters that contributed to these nodes. We achieve this by modifying each of the weights (w) of the FC layer such that:

$$w_r = 0.5 \times [((2-\beta) \times w)^{\circ 1} + (\beta \times w)^{\circ 3}], \tag{1}$$

$$\nabla_{w_{new}} = 0.5 \times [(2-\beta) \times \nabla_w) + 2\beta^3(w^{\circ 2} \times \nabla_w)], \tag{2}$$

where $\circ$ denotes Hadamard power, ∇_w is the backpropagated gradient on w if the above reparameterization was not applied and β is a hyperparameter. $\beta \in \mathbb{R} : \beta \in [0, 2]$, and $\nabla_{w_{new}}$ is the backpropagated gradient of w. For any $\beta > 0$, w_r will have relatively larger magnitude values for larger magnitude w values, and the backpropagated gradient ($\nabla_{w_{new}}$) for higher valued weights will also be higher. Higher values of β will further encourage this asymmetrical gain in magnitude, and at $\beta = 0$ this asymmetrical gain in magnitude disappears. Notably, since this weight reparameterization is only done during training, no computation cost is added during inference compared to a normal FC layer.

3.3 Multi-loss paradigm of C2F

The C2F network need to be trained such that each pathway becomes reliable for action/gesture recognition by itself and also provides complementary features

for use by the finer pathways. We do this by formulating a joint optimization formulation with the multi-loss function,

$$L = \sum_{n=1}^{N} \alpha L_n + (1-\alpha) L_{C2F}, \tag{3}$$

where L_n and L_{C2F} are the softmax cross-entropy losses comparing ground truth $\hat{O}$ with O_n and O_{C2F}, respectively, O_n is the output at pathway n, O_{C2F} is the output after the reparameterized FC layer, and $\alpha \in \mathbb{R} : \alpha \in [0, 1]$. A high value of α will have the C2F ensemble network focus only on optimizing each of the pathways, whereas a low value of α will have the C2F network focus more on extracting complementary information for improving O_{C2F}. Since we are motivated to improve performances of coarser pathways so we can exit early and save computation costs, we use a high value of $\alpha = 0.9$. Also, note that eq. (3) is differentiable, so our joint optimization method is end-to-end differentiable and can be trained together.

3.4 Multi-loss paradigm of C2F with pre-trained F

With availability of high performing networks that are already trained on large datasets such as the Kinetics dataset [3], we propose extending the student-teacher learning paradigm [10] to a classroom learning paradigm - here, F (i.e. the finest pathway ConvNet) teaches the coarser pathways, and additionally a classroom (the reparameterized FC layer layer of C2F) learns both from the students and the teacher to perform better than F. Similar to a student-teacher learning, the F sub-network is no longer trained and is only used for teaching the students (i.e., the coarser pathways). To optimize all C pathways and the reparameterized FC layer, we train the C2F network by minimizing the following multi-loss function:

$$L_d = \sum_{n=1}^{N} \alpha L_{n,KLD} + \sum_{n=1}^{N} (1-\alpha/2) L_n + (1-\alpha/2) L_{C2F}, \tag{4}$$

where $L_{n,KLD}$ is the Kullback-Leibler divergence between the distributions of p_n/τ and p_F/τ, p_n is the softmax output of the nth scale, p_F is the softmax output of F, τ is a temperature parameter that softens the distributions between p_F and p_n [10], and α is a hyperparameter [10]. The primary difference in L_d from the original knowledge distillation scheme [10] is the $(1-\alpha/2)L_{C2F}$ term that encourages each of the coarse scales to provide some complementary information to F that may help in improving overall performance of O_{C2F}.

3.5 Coarse-exit and budgeted inference

For fast inference, C2F includes a coarse-exit decision stage that encourages using low average computation costs while retaining high accuracy. Inference starts with forward propagation along the coarsest stream (Fig. 1). To ensure

that recognition at this coarsest stream is performed accurately, action/gesture recognition is done only when the softmax output $s_N > T$, where s_N is the maximum value in p_N and T is a hyper-parameter controlled externally. However, since softmax outputs tend to be too confident in their predictions and since they are not well-calibrated for uncertainty measures [8], we adjust s_N using a global training accuracy context as follows:

$$s_N^C = 1 - (1 - s_N) \times e_N/e_1, \tag{5}$$

where e_N is the training misclassification rate for the coarsest scale and e_1 is the training misclassification rate at the end of the C2F ensemble. This coarse-exit strategy is repeated and finer streams in the ensemble are only invoked when the coarse-exit in the coarser stream fails (Fig. 1).

Inference stage C2F can also have a budget FLOPs ($B \in \mathbb{R} : B \in [f_C, f_{C2F}]$) as input, where f_C is the FLOPs of C pathway and f_{C2F} is the computation cost of the C2F ensemble for a single recognition. We implement a controller that continuously modifies T as:

$$T = T_{av} + (B - f_{av}) * (f_{av} - f_{C2F})/(T_{av} - 1.0), \tag{6}$$

where f_{av} is the running average FLOPs and T_{av} is the average of previous r recognitions (default r=100). Only when C2F is operating at desired budget (i.e. $B - f_{av} = 0$), T does not update from T_{av}.

3.6 Protocols

Our C2F architecture is generic and can be applied to different baseline ConvNets. In this work, we use C2F with TSM [17] as its baseline ConvNet. We chose TSM since it has been dominating the leaderboard in the Something-Something V1 and V2 datasets [31, 32], and is considerably more efficient compared to most competitive ConvNets for action/gesture recognition [17]. We investigate improvements achieved by extending TSM baseline to a C2F architecture on four standard action/gesture recognition datasets, and compare them against the improvements achieved by applying the popular non-local blocks (NL) [37] with TSM baseline. Additionally, we compare C2F-extended TSM with other competitive ConvNets on Jester and Something-Something V1, V2 datasets.

Datasets: For gesture recognition, we evaluate C2F ensemble networks on the Jester dataset [30], which is currently the largest publicly available dataset for gesture recognition. It has ∼119K training videos, ∼15K validation videos, and a total of 27 different gesture and no-gesture classes. For action recognition, we train and evaluate C2F on the Something-Something V1 and V2 datasets [7], and the Kinetics-400 dataset [3]. The Something-Something V1 dataset has ∼86K training videos and ∼12K validation videos of humans interacting with everyday objects in some pre-defined action-sets (e.g. pushing some object from left to right, etc.). The V2 dataset expands on the V1 dataset with a total of ∼169K training videos and ∼25K validation videos. Activity recognition in these videos is challenging since it requires identifying objects as well as the sets of

movements resulting in a particular interaction. We also evaluate C2F on the Kinetics-400 [3], which is another large-scale dataset with 400 classes of actions that are less sensitive to temporal movements of objects [17].

Training and testing: F in our C2F ensemble networks are instantiated with TSM-ResNet50 [17]. We denote a trained C2F that has not used action-pretrained F (i.e., not pretrained on the action recognition task at hand) at initialization as $C2F_A$, and we denote a trained C2F that used action-pretrained F and trained using the classroom learning paradigm as $C2F_C$. For the latter, F is initialized with ImageNet pretraining using only 224×224 resolution. All other pathways also use the same ImageNet pretraining, so multiple ImageNet pretraining at different scales is not necessary. Coarse pathways cause small FLOPs and memory burden on the overall $C2F_C$ architecture (see Table 1). Additionally, instantiating a coarse pathway with a heavier network causes less computational overhead compared to instantiating F with a heavier network. Due to this relatively smaller computation overhead, we also investigate heavier networks (TSM-ResNet101) at the coarser pathways while still using the pretrained TSM-ResNet50 as F in our $C2F_C$ architecture. For convenience, we will call this modified architecture $C2F_{C+}$. For all ConvNet results we report in this section and subsequent sections, we use subscript En to denote use of the entire ensemble network during inference and we use subscript Ex to denote use of coarse-exit scheme during inference.

We sample 16 strided frames for the Something-Something and Jester datasets, and use 8 densely sampled frames for the Kinetics dataset similar to [17]. For the coarser pathways we use 8 frames. We do not decompose the spatiotemporal video to less than 8 frames, since we found empirically that using 4 frames or less can deteriorate C pathway performances. We used data augmentation similar to [36], used a batch size of 64, and optimized with stochastic gradient descent [23] having momentum 0.9 and a weight decay of 5e-4. For the classroom learning, F is not trained any further from its pretrained form. Training was done for 50 epochs - here, learning rate was initialized at 0.02 and reduced thrice with a factor of 10^{-1} after 20^{th}, 40^{th} and 45^{th} epochs. To prevent overfitting, we also added extra dropout layers with dropout rate of 0.5 before each of the final FC layers in O_n. To compare with TSM baseline (Table 1), we use the single center-crop evaluation followed in [17] - here, the center crop has 224/256 of the shorter side of a frame. To compare with other methods (Table 2), we follow [37, 13] where the center-crop has the entire shorter side which is then resize to 224×224 for Something-Something V1/V2 and Kinetics datasets and resized to 128×128 for the Jester dataset. We note that single center-crop action/gesture recognition accuracy are likely to approximate accuracy in practical settings that will likely not use multiple crops to ensure efficiency. We also report multi-crop evaluations using settings used in [17] for each of the datasets. On all experiments, we use $\beta = 1.0$ for the reparameterized FC layer, and $\alpha = 0.1$ and $\tau = 6.0$ for the classroom learning paradigm based on empirical observations on the Jester dataset.

Table 1. Effect of extending TSM baseline with NL [37] or our C2F architectures. Memory usage was computed based on a inference batch size of 1. The C2F ensembles perform better than TSM with NL, except on the Kinetics dataset where using C2F with TSM-NL has the highest accuracy.

Method		Inference costs and accuracy					
Baseline	**Module added**	**Param.**	**Memory**	**FLOPs**	**Top-1 Acc.(%)**	**Top-5 Acc.(%)**	△Top-1
Something-Something V1							
TSM-ResNet50 (16F-224) [17]	None	1.0×	1.0×	1.0×	47.0	76.9	+0.0
	NL [37]	1.3×	1.24×	1.2×	41.3	72.1	-5.7
	$C2F_{A,En}$	3.0×	1.19×	1.16×	47.9	77.5	0.8
	$C2F_{C,En}$	2.0×	1.12×	1.08×	**48.4**	**78.4**	**+1.4**
	$C2F_{C,Ex}$ (T=.65)	2.0×	1.12×	**0.8**×	47.4	77.7	**+0.4**
Something-Something V2							
TSM-ResNet50 (16F-224) [17]	None	1.0×	1.0×	1.0×	61.5	87.5	+0.0
	NL [37]	1.3×	1.24×	1.2×	57.2	84.0	-4.3
	$C2F_{A,En}$	3.0×	1.19×	1.16×	62.1	**88.1**	+0.6
	$C2F_{C,En}$	2.0×	1.12×	1.08×	**62.4**	88.0	**+0.9**
	$C2F_{C,Ex}$ (T=.80)	2.0×	1.12×	**0.7**×	61.8	87.6	**+0.3**
Kinetics							
TSM-ResNet50 (Dense 8F-224) [17]	None	1.0×	1.0×	1.0×	69.8	88.3	+0.0
	NL [37]	1.3×	1.24×	1.2×	70.9	89.3	+1.1
	$C2F_{C,En}$	2.0×	1.12×	1.08×	70.5	88.9	+0.7
	$C2F_{C,En}$ + NL	2.3×	1.39×	1.28×	**71.4**	**90.0**	**+1.6**
	$C2F_{C,Ex}$ + NL (T=.0.5)	2.3×	1.39×	**0.8**×	70.1	89.0	**+0.3**
TSM-ResNet50 (Strided 16F-224) [17]	None	1.0×	1.0×	1.0×	72.4	90.8	+0.0
	$C2F_{C,En}$	2.0×	1.12×	1.28×	**73.5**	**91.4**	**+1.1**
	$C2F_{C,Ex}$ (T=.0.7)	2.0×	1.12×	**0.8**×	**72.6**	**90.9**	**+0.2**
Jester							
TSM-ResNet50 (16F-128) [17]	None	1.0×	1.0×	1.0×	96.3	99.7	+0.0
	NL [37]	1.3×	1.24×	1.2×	96.4	99.8	+0.1
	$C2F_{A,En}$	3.0×	1.19×	1.16×	**96.5**	**99.8**	**+0.2**
	$C2F_{C,En}$	2.0×	1.12×	1.08×	**96.5**	**99.8**	**+0.2**
	$C2F_{C,Ex}$ (T=.99)	2.0×	1.12×	**0.5**×	96.4	**99.8**	**+0.1**

3.7 Results

Performance improvements over baseline: We find consistent improvements with our C2F ensemble networks ($C2F_{A,En}$ and $C2F_{C,En}$ (Table 1), and $C2F_{C+,En}$ (Table 2)) over TSM baseline (Table 1). The accuracy improvements achieved over baseline ($\triangle Top1$) with $C2F_{C,En}$ are larger compared to $\triangle Top1$ achieved by adding NL [37] to baseline TSM (except for the Kinetics dataset, see Table 1), and the improvements come at less FLOPs and memory usage overhead. When 'TSM-ResNet50 with NL added' is used as F in the $C2F_C$ learning on the Kinetics dataset, the accuracy of the ensemble output (i.e., $C2F_{C,En}$ with NL added in F) improves further, suggesting that the beneficial effects of adding the NL block is complementary to extending a base network to C2F. We also find that the smallest $\triangle Top1$ is on the Jester dataset (with the highest baseline TSM accuracy) and the largest $\triangle Top1$ is on the Something-Something V1

Table 2. Comparison of $C2F_{C+,En}$ with competitive approaches that use single-crop RGB for evaluation on the Something-Something and Jester datasets. Our $C2F_{C+,En}$ and its coarse-exit scheme outperforms all competitive methods that are at FLOPs< $50G$ and at FLOPs> $50G$. For methods with *, we use our training settings on official implementations of the methods.

	Jester		Something V1			Something V2		
Model	**FLOP**	**Val1 (%)**	**FLOP**	**Val1 (%)**	**Val5 (%)**	**FLOP**	**Val1 (%)**	**Val5 (%)**
FLOPS<50G								
TRN-Multiscale-BNInception [42]	-	-	16	34.4	-	16	48.8	77.6
TSN-ResNet50 [36]	-	-	33	19.7	46.6	33	30.0	60.5
ECO [44]	-	-	32	39.6	-	-	-	-
$C2F_{C+,Ex}$	**25**	**96.8**	**15**	**42.1**	**72.0**	**11**	**54.2**	**82.0**
FLOPS>50G								
GST-ResNet50 [19]	-	-	**59**	48.6	77.9	59	62.6	87.9
TSM-ResNet50* [17]	65	96.4	65	47.1	77.0	65	61.6	87.7
STM-ResNet50 [13]	-	-	67	49.2	**79.3**	-	-	-
S3D-BNInception-G [41]	-	-	71	48.2	78.7	-	-	-
TSM-ResNet50$_{En}$* [17]	-	-	98	49.0	78.9	-	-	
ABM-AC-in_{En}-ResNet50[43]	-	-	106	46.8	-	61.2	-	
TSN-STD-ResNet50[21]	-	-	-	50.1	-	-	-	
$C2F_{C+,Ex}$	-	-	**59**	**49.3**	79.1	**56**	**62.8**	**88.0**
$C2F_{C+,En}$	-	-	85	**50.2**	**79.9**	85	**63.8**	**88.8**

dataset (with the lowest baseline TSM accuracy) - this suggests that extending a baseline network to C2F may provide more accuracy improvements for difficult datasets.

Extra FLOPs and memory usage: The extra FLOPs of our ensemble networks are considerably reduced when our coarse-exit (Ex) scheme is invoked - e.g. using $C2F_{C,Ex}$ with $T = 0.8$ on the Something-Something V2 dataset reduces FLOPs of $C2F_{C,Ex}$ 30% below the baseline TSM while still retaining a non-trivial accuracy improvement (Table 1). Interestingly, in contrast to the trend in accuracy improvement vs. base accuracy, we see an opposite trend for computation cost reduction - easier datasets enjoy larger computation cost reductions (Table 1). Finally, the small additional memory usage of C2F networks are justifiable considering their ability to flexibly and continuously operate on the cost-accuracy curve, and their ability to operate at very low FLOPs when availability of computational resources are limited (Fig. 2). The flexibility to operate continuously on the cost-accuracy curve has enormous technical benefits. We no longer need to retrain different models to fit varying computational budget constraints - we are now able to flexibly control the C2F network to operate in varying cost constraints by simply modifying the parameter T (Fig. 2).

Single-crop RGB performances: Table 2 shows single-crop RGB performance of C2F networks compared to competitive ConvNets on Jester and Something-Something V1, V2 datasets. While TSM [17] does not exhibit the best single-crop

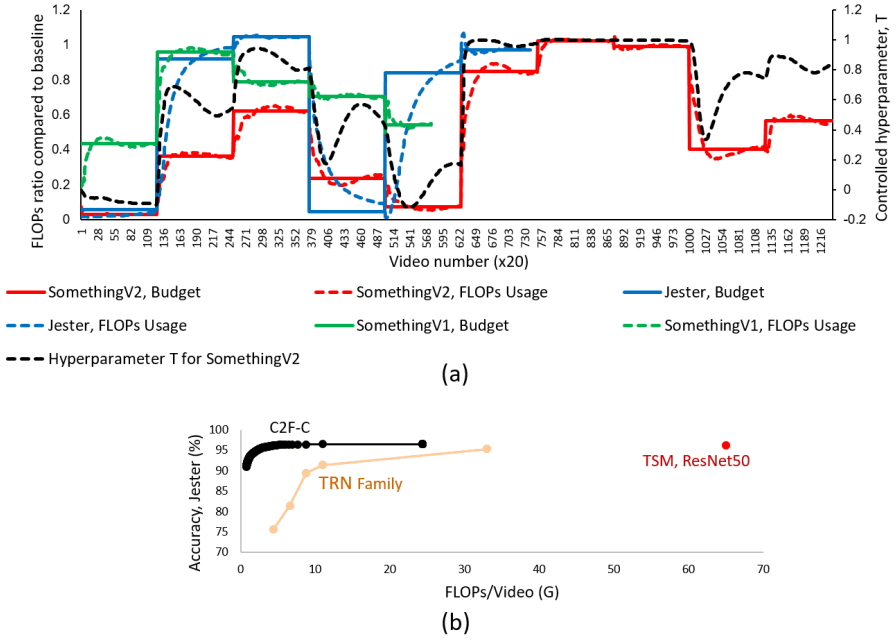

Fig. 2. (a) FLOPs of $C2F_C$ network at inference (dotted blue, green and red lines) follows closely the randomly defined budget FLOPs B that is assigned randomly at every 2500th videos for Jester and Something-Something V1 and V2 datasets (solid blue, green and red lines). The dotted black line represents the modified T values for Something-Something V2 dataset to enable the $C2F_C$ network to operate close to B. (b) $C2F_C$ network's cost-accuracy curve for Jester dataset has considerably better cost-accuracy tradeoff compared to baseline TSM and the more efficient TRN family. The various cost-accuracy operating points on the $C2F_C$ are obtained simply by adjusting the T values ($T \in \mathbb{R} : T \in [1 - e_N/e_1, 1]$) without retraining the entire model.

results among competitive methods, $C2F_{C+,En}$ built on TSM baseline considerably outperforms the rest (Table 2); thus, a different baseline to our C2F could bring even higher accuracy. The extra computation costs of $C2F_{C+,En}$ can be flexibly reduced using the coarse-exit scheme ($C2F_{C+,Ex}$). Controlling the hyperparameter T in $C2F_{C+,Ex}$ yields considerably higher performances than competing methods at various FLOPs constraints (e.g. FLOPs<50G and FLOPs>50G in Table 2). On the Jester dataset, C2F uses inputs having spatial dimension of 128×128 in contrast to 224×224 used in other methods - this makes C2F considerably more efficient for Jester dataset.

Multi-crop RGB performances: We find similar accuracy improvements with $C2F_{C,En}$ implementation on multi-crop evaluations over TSM baseline (improvement from 63.1% to 64.1% on Something-Something V2 and improvement from 75.6% to 76.0% on Kinetics). Compared to the competitive multi-resolution ap-

proach SlowFast [4], $C2F_{C,En}$ has better accuracy at lower computation cost on Kinetics - 75.6% accuracy with SlowFast4 $\times$ 16R50 at computation cost of 1083 GFLOPs/video vs. 76.0% accuracy with $C2F_{C,En,NL}$ at less than half the computation cost (460GFLOPs/video). Finally, $C2F_{C+,En}$ on a 30-crops softmax-averaged evaluation on the Jester dataset yields a top-1 test accuracy of 97.09%, beating the closest competitor [26] with a misclassification rate reduction by around 14%.

3.8 Ablation studies:

We performed a series of ablation studies, and compare results of all experiments using single-crop validation accuracy. First, we compare performances of the baseline ConvNet (TSM) with alternative multi-stream fusion schemes - (i) late fusion with ensemble of classification scores similar to [27, 4, 17, 36, 14], (ii) slow fusion with lateral connections similar to [4]. Next, we separately investigate effectiveness of - (i) the C2F decomposition compared to all fine pathways, (ii) the reparameterized FC layer compared to a vanilla FC layer, (iii) the class-learning method compared to the student-teacher [10] learning method, and (iv) the effectiveness of our generic C2F on other popular baseline ConvNets for action/gesture recognition.

Effectiveness of multi-loss schemes: To investigate effectiveness of multi-loss schemes we compare C2F results against simple late fusion schemes with ensemble of classification scores. We investigate three simple C2F ensembling schemes: First, summing up softmax scores of each C2F pathway; second, multiplying softmax scores with training accuracy of each pathway separately and then summing up the weighted scores; third, using a product of softmax scores of each C2F pathway. On the Jester dataset, all of the above results in poorer accuracy (95.3% with additive ensemble, 96.0% with weighted additive ensemble and 95.3% with multiplicative ensemble) compared to baseline TSM with accuracy 96.3% - this could be because softmax outputs tend to be too confident in their predictions [8] such that misclassifications from coarser pathways start disabling correct recognitions at finest pathway. Additionally, end-to-end learning with multi-stream loss added and backpropagated does not improve accuracy over baseline (96.2% with such multi-loss optimization vs. 96.3% with baseline TSM).

Slow fusion with lateral connections: To encourage feature reuse of each convolutional kernel, we experiment with lateral connections between adjacent C2F streams. We only investigate lateral connections sourced at coarser streams and ending at finer streams, so we can still enable coarse-exit at inference. We laterally connect activation maps from the coarse pathway (A_C) to the activation maps in the adjacent finer pathway (A_F) in a manner similar to [5], but find deteriorating accuracy by 0.05% on the Jester dataset. This failure in improving performance is perhaps due to incorporation of noisy activation maps from the coarser pathways to the finer pathways, which outweights the benefits of lateral connections. This could perhaps be resolved with concatenation of activation maps rather than naive arithmetic lateral connections; however, since we are

more concerned with keeping overall network computation efficient, we defer that experiment for later.

C2F decomposition vs. all-fine pathways: When the coarser pathways of $C2F_C$ are fed with non-downsampled input rather than downsampled coarse input, the improvement in accuracy compared to baseline TSM ($\triangle Top1 = 1.1$) was slightly lower than with the original $C2F_C$ ($\triangle Top1 = 1.4$), suggesting that a C2F architecture not only reduces computation burden compared to a non-downsampled scheme, but also provides beneficial complementary information for the C2F ensemble.

Utility of the reparameterized FC layer: Replacing our reparameterized FC layer with a regular FC layer in the $C2F_C$ ensemble networks deteriorates accuracy - the reduction in top-1 accuracy(%) is around 0.1 for Jester, around 0.3 for both Something-Something V1 and V2 datasets, and a larger 0.6 for the Kinetics dataset. Since the Kinetics dataset has the highest number of classes among the datasets we used (400 compared to 174 of Something-Something and 27 of Jester), the number of connections in the FC layer are much larger for a $C2F_C$ instantiation on Kinetics. The reparameterized FC layer is particularly effective in such instances of large numbers of FC connections since it is able to focus more on the important FC connections. For datasets with higher number of classes (e.g. Kinetics-700 [2]), the reparameterized FC layer or a variant of the reparameterized FC layer that focuses on important FC connections will likely be crucial for a successful C2F implementation.

Comparison with student-teacher [10] learning method: In contrast to a student-teacher learning method [10], our classroom learning approach enables $C2F_{C,En}$ to perform significantly better than the teacher network F (Table 1). In terms of utility in teaching the student networks (i.e., the coarser networks of $C2F_C$), we find no significant difference in the coarsest pathway accuracy when trained with the student-teacher method (using $\alpha = 0.1$ and $\tau = 6.0$) [10] and our classroom learning method on the Jester dataset (90.4% with student-teacher vs. 90.5% with $C2F_C$) and Something-Something V1 dataset (31.0% with student-teacher vs. 30.7% with $C2F_C$). Both the classroom learning and the student-teacher learning outperforms a coarsest pathway trained separately (e.g. 90.4% with student-teacher and 90.5% with $C2F_C$ vs. 87.8% with no knowledge-distillation).

C2F effectiveness on baseline other than TSM: Our C2F architectures are generic and can be easily implemented with different baseline ConvNets. While TSM uses 2D convolution-based networks with shift operations added to infuse temporal information, many ConvNets for action/gesture recognition rely on 3D convolution-based networks [34, 3, 41, 4]. Therefore, we investigate effectiveness of C2F extension on the 3D convolution-based R3D-ResNet18 and R2plus1D-ResNet18 ConvNets [34]. $C2F_C$ with these baseline ConvNets instantiated on all three pathways outperforms the baseline ConvNets (94.2% with $C2F_C$-R3D-ResNet18 vs. 93.6% with R3D-ResNet18, 94.6% with $C2F_C$-R2plus1D-ResNet18 vs. 93.9% with R2plus1D-ResNet18).

4 Discussion

Our ablation studies showed that much of the prior popular multi-stream techniques (e.g. slow fusion with lateral connection [4], late fusion using ensemble of classification scores or end-to-end learning with multi-stream loss added and backpropagated) fail to work on C2F multi-stream ConvNets for action/gesture recognition. We conjecture that this ineffectiveness of prior multi-stream techniques primarily arise from the relative lack of complementary information at inputs of the C2F pathways compared to prior multi-stream ConvNets (e.g. complementary spatial and temporal stream in [27]). With much of the prior techniques being ineffective for C2F ConvNets on action/gesture recognition, our proposed baseline C2F ConvNet along with its effective training strategies are significant for future C2F ConvNet research on action/gesture recognition.

To discuss the importance of future C2F ConvNet research on action/gesture recognition, we review the technical benefits achieved with our proposed $C2F_C$ ConvNet. $C2F_C$ on TSM baseline can operate at unprecedented low FLOPs ($< 1\%$ GFLOPs) and remain competitive with methods such as [42]. In fact, some classes of action (e.g. 'Playing squash or racquetball' in the Kinetics dataset [3]) can be recognized at near-perfect accuracy at such low FLOPs (more details in the supplementary document). When varying FLOPs constraints are imposed, rather than re-designing and training different ConvNets, our proposed $C2F_C$ can operate flexibly at the desired FLOPs, and at a better cost-accuracy trade-off than its baseline ConvNet. Finally, $C2F_C$ is generic and can be built on different ConvNets, so the benefits of $C2F_C$ remain complementary to emerging standalone ConvNets for action/gesture recognition. For all of these advantages, C2F extension approaches on action/gesture recognition are likely to become valuable techniques.

5 Conclusion

'Use when needed' is an extremely effective philosophy for improving efficiency. We introduce this philosophy in the field of action recognition using C2F ConvNets. We demonstrate that fusing complementary feature in a C2F approach with a coarse-exit scheme can significantly improve ConvNet performances. While our C2F ensembles can outperform previous competing methods in terms of both increased accuracy and reduced computation costs, the C2F ensembles are particularly potent in improving accuracy for difficult datasets, in reducing computational costs for easier datasets and in operating efficiently and flexibly at a large range of computational cost constraints.

Acknowledgement

We are very grateful to Professor Richard P. Wildes (York University) who has generously supported this work by providing critical feedback and guidance during concept formulation, early experimentations and writing of earlier drafts.

References

1. Green 500 List for June 2016, howpublished = https://www.top500.org/green500/lists/2016/06/, note = Accessed 8th November 2019
2. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
3. 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)
4. 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)
5. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4768–4777 (2017)
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1933–1941 (2016)
7. 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)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1321–1330. JMLR. org (2017)
9. He, D., Zhou, Z., Gan, C., Li, F., Liu, X., Li, Y., Wang, L., Wen, S.: Stnet: Local and global spatial-temporal modeling for action recognition. arXiv preprint arXiv:1811.01549 (2018)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., Weinberger, K.Q.: Multi-scale dense convolutional networks for efficient prediction. arXiv preprint arXiv:1703.09844 **2** (2017)
12. Huang, W., Fan, L., Harandi, M., Ma, L., Liu, H., Liu, W., Gan, C.: Toward efficient action recognition: Principal backpropagation for training two-stream networks. IEEE Transactions on Image Processing **28**(4), 1773–1782 (2018)
13. 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)
14. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
15. Koenderink, J.J.: The structure of images. Biological cybernetics **50**(5), 363–370 (1984)
16. Koenderink, J.J.: Scale-time. Biological Cybernetics **58**(3), 159–162 (1988)
17. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision (2019)

18. Lin, J., Rao, Y., Lu, J., Zhou, J.: Runtime neural pruning. In: Advances in neural information processing systems. pp. 2181–2191 (2017)
19. Luo, C., Yuille, A.L.: Grouped spatial-temporal aggregation for efficient action recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5512–5521 (2019)
20. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in neural information processing systems. pp. 4898–4906 (2016)
21. Martinez, B., Modolo, D., Xiong, Y., Tighe, J.: Action recognition with spatial-temporal discriminative filter banks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5482–5491 (2019)
22. Mermillod, M., Guyader, N., Chauvin, A.: The coarse-to-fine hypothesis revisited: evidence from neuro-computational modeling. Brain and Cognition **57**(2), 151–157 (2005)
23. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)
24. Rodríguez-Sánchez, A.J., Fallah, M., Leonardis, A.: Hierarchical object representations in the visual cortex and computer vision. Frontiers in computational neuroscience **9**, 142 (2015)
25. Rosenfeld, A.: Multiresolution image processing and analysis, vol. 12. Springer Science & Business Media (2013)
26. 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)
27. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
28. Stroud, J.C., Ross, D.A., Sun, C., Deng, J., Sukthankar, R.: D3d: Distilled 3d networks for video action recognition. arXiv preprint arXiv:1812.08249 (2018)
29. Teerapittayanon, S., McDanel, B., Kung, H.T.: Branchynet: Fast inference via early exiting from deep neural networks. 2016 23rd International Conference on Pattern Recognition (ICPR) pp. 2464–2469 (2016)
30. The 20bn-jester dataset v1: *The 20BN-jester Dataset V1*, [Accessed 8th November 2019]
31. The 20BN-something-something Dataset V1: *The 20BN-something-something Dataset V1*, [Accessed 8th November 2019]
32. The 20BN-something-something Dataset V2: *The 20BN-something-something Dataset V2*, [Accessed 8th November 2019]
33. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
34. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
35. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**(2), 137–154 (2004)
36. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)

37. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7794–7803 (2018)
38. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 399–417 (2018)
39. Wang, X., Luo, Y., Crankshaw, D., Tumanov, A., Yu, F., Gonzalez, J.E.: Idk cascades: Fast deep learning by learning not to overthink. arXiv preprint arXiv:1706.00885 (2017)
40. Wang, X., Yu, F., Dou, Z.Y., Darrell, T., Gonzalez, J.E.: Skipnet: Learning dynamic routing in convolutional networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 409–424 (2018)
41. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 305–321 (2018)
42. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018)
43. Zhu, X., Xu, C., Hui, L., Lu, C., Tao, D.: Approximated bilinear modules for temporal modeling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3494–3503 (2019)
44. Zolfaghari, M., Singh, K., Brox, T.: Eco: Efficient convolutional network for online video understanding. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 695–712 (2018)