

FASTER Recurrent Networks for Efficient Video Classification

Linchao Zhu,^{†§} Du Tran,[†] Laura Sevilla-Lara,^{†*} Yi Yang,[§] Matt Feiszli,[†] Heng Wang[†]

[†]Facebook AI [§]ReLER, University of Technology Sydney ^{*}University of Edinburgh
 {linchao.zhu, yi.yang}@uts.edu.au, lsevilla@ed.ac.uk, {trandu, mdf, hengwang}@fb.com

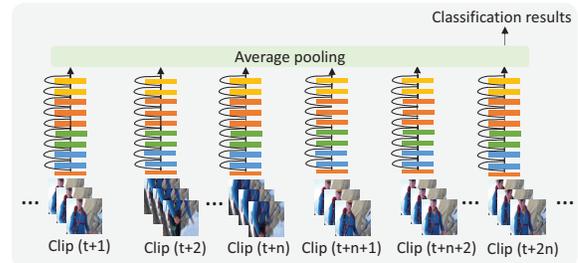
Abstract

Typical video classification methods often divide a video into short clips, do inference on each clip independently, then aggregate the clip-level predictions to generate the video-level results. However, processing visually similar clips independently ignores the temporal structure of the video sequence, and increases the computational cost at inference time. In this paper, we propose a novel framework named FASTER, *i.e.*, Feature Aggregation for Spatio-TEmporal Redundancy. FASTER aims to leverage the redundancy between neighboring clips and reduce the computational cost by learning to aggregate the predictions from models of different complexities. The FASTER framework can integrate high quality representations from expensive models to capture subtle motion information and lightweight representations from cheap models to cover scene changes in the video. A new recurrent network (*i.e.*, FAST-GRU) is designed to aggregate the mixture of different representations. Compared with existing approaches, FASTER can reduce the FLOPs by over $10\times$ while maintaining the state-of-the-art accuracy across popular datasets, such as Kinetics, UCF-101 and HMDB-51.

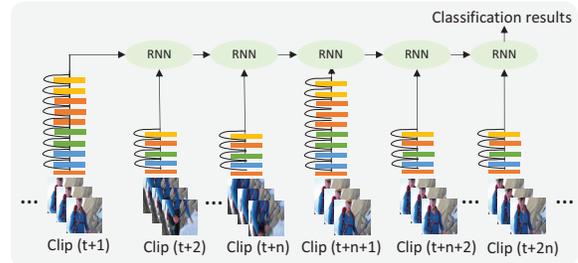
Introduction

Video classification has made tremendous progress since the popularity of deep learning. Though the accuracy of Convolutional Neural Networks (CNNs) (LeCun et al. 1998) on standard video datasets continues to improve, their computational cost is also soaring. For example, on the popular Kinetics dataset (Kay et al. 2017), the pioneering C3D (Tran et al. 2015) reported a top-1 accuracy of 63.4% with a single-clip FLOPs of 38.5G. The recent I3D (Carreira and Zisserman 2017) model improved the accuracy to an impressive 72.1% while its FLOPs also increased to 108G. Non-local networks (Wang et al. 2018) achieved the state-of-the-art accuracy of 77.7% on Kinetics with FLOPs as high as 359G. Moreover, this accuracy is achieved by sampling 30 clips from each video, which increases the computational cost by a factor of 30 at test time, as shown in Fig. 1 (a). Though modern architectures have achieved incredible accuracy for action recognition, their high computational costs prohibit them from being widely used for real-world applications

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Typical video classification framework for inference.



(b) The FASTER framework can process up to 75% of the clips using a much cheaper model without losing accuracy.

Figure 1: (a) Typical video classification framework processes each clip repeatedly with an expensive model; (b) The FASTER framework exploits the combination of expensive and cheap networks for better accuracy/FLOPs trade-offs.

such as video surveillance, or being deployed on hardware with limited computation power, like mobile devices.

Fig. 1 (a) illustrates the typical framework for video classification. Multiple clips are sampled from a video, and fed to a computationally expensive network to generate predictions for each clip. Clip-level predictions are then aggregated, often by taking the average, to form the final video-level results. To reduce the computational cost, recent efforts have focused on designing more efficient architectures (Howard et al. 2017) at the clip-level. However, less attention has been paid to the efficiency of the overall framework, including how to aggregate the clip-level predictions over time, since they have a strong temporal structure.

In this work we address the problem of reducing the computational cost of video classification by focusing on the temporal aggregation stage. In particular, we leverage the observation that video data has a strong temporal structure and is highly redundant in time. We argue that it is computationally inefficient to process many video clips close in time with an expensive video model. As adjacent clips are visually similar, computational cost can be saved by processing most of clips with a lightweight network, shown in Fig. 1 (b). A Recurrent Neural Network (RNN) is learned to aggregate the representations from both the expensive and lightweight models. We summarize the contributions of our paper in the following.

First, we propose a novel framework for efficient video classification that we call *FASTER* for Feature Aggregation for Spatio-TEmporal Redundancy (Fig. 1 (b)). *FASTER* explicitly leverages the redundancy of video sequences to reduce FLOPs. Instead of processing every clip with an expensive video model, *FASTER* uses a combination of an expensive model that captures the details of the action, and a lightweight model which captures scene changes over time, avoids redundant computation, and provides a global coverage of the entire video at a low cost. We show that up to 75% of the clips can be processed with a much cheaper model without losing classification accuracy, maintaining the state-of-the-art accuracy with $10\times$ less FLOPs.

Second, we design a new RNN architecture, namely *FAST-GRU*, dedicated to the problem of learning to aggregate the representations from different clip models. As shown in our experiments, the prevailing average pooling fails to encode predictions from different distributions. On the contrary, *FAST-GRU* is capable of learning integration of representations from multiple models for longer periods of time than popular RNNs, such as GRU (Cho et al. 2015) and LSTM (Hochreiter and Schmidhuber 1997). Compared with typical GRU, *FAST-GRU* keeps the spatio-temporal resolution of the feature maps unchanged, whereas GRU collapses the resolution with global average pooling. *FAST-GRU* also has a channel reduction mechanism for feature compression to reduce the number of parameters and introduce more non-linearity. As a result, the *FASTER* framework achieves significantly better accuracy/FLOPs trade-offs for video classification.

Related Work

Before the deep learning era, hand-crafted features (Wang et al. 2011; Wang and Schmid 2013; Laptev 2005; Klaser, Marszałek, and Schmid 2008; Dalal, Triggs, and Schmid 2006) were widely used for action recognition. In this section, we cover related works that apply deep learning for efficient video classification.

Clip-level Classification. Many efforts in video classification focus on designing models to classify short clips. 3D convolutions (Baccouche et al. 2011; Ji et al. 2013; Karpathy et al. 2014; Tran et al. 2015) are widely adopted for learning spatio-temporal information jointly. While these methods do improve the accuracy, they also dramatically increase com-

putational cost due to expensive 3D convolutions. Recent progress includes factorizing 3D convolution into 2D spatial convolution and 1D temporal convolution (Sun et al. 2015; Qiu, Yao, and Mei 2017; Tran et al. 2018; Xie et al. 2018; Zhou et al. 2018), and designing new building blocks to learn global information on top of 3D convolution (Wang et al. 2018; Chen et al. 2018b). Another line of research designs two-stream networks (Simonyan and Zisserman 2014; Yue-Hei Ng et al. 2015; Wang et al. 2016; Feichtenhofer, Pinz, and Zisserman 2016; Carreira and Zisserman 2017; Xie et al. 2018) that use both RGB and optical flow as input. Though two-stream networks tend to improve the accuracy, accurate optical flow algorithms are often computationally expensive.

Video-level Aggregation. To aggregate the clip-level predictions, average pooling is the most popular approach and works remarkably well in practice. Besides average pooling, (Karpathy et al. 2014) proposed convolutional temporal fusion networks for aggregation. NetVLAD (Arandjelovic et al. 2016) also has been used in (Girdhar et al. 2017) for feature pooling on videos. (Yue-Hei Ng et al. 2015) learned a global video representation by max-pooling over the last convolutional layer. (Wang et al. 2016) proposed the TSN network that divides the video into segments, which are averaged for classification. RNNs are also used for sequential learning in video classification (Srivastava, Mansimov, and Salakhudinov 2015; Yue-Hei Ng et al. 2015; Li et al. 2018). In contrast with existing work, *FAST-GRU* aggregates the representations from different models, and aims to leverage spatio-temporal redundancy for efficient video classification.

Towards Efficient Video Classification. (Zolfaghari, Singh, and Brox 2018) proposed the ECO model for online video understanding, where a 3D network is used to aggregate frame-level features. Recent work (Alwasel, Caba Heilbron, and Ghanem 2018; Wu et al. 2019; Korbar, Tran, and Torresani 2019; Bhardwaj, Srinivasan, and Khapra 2019) tried to reduce the computational cost by either reducing the number of frames that need to be processed, or sampling discriminative clips to avoid redundant computation. The *FASTER* framework is orthogonal to these approaches. Instead of applying sophisticated sampling strategies, we focus on designing a general framework with a new architecture that can learn to aggregate representations from models of different complexities.

Learning to Aggregate

As shown in Fig. 1 (b), *FASTER* is a general framework to aggregate expensive and lightweight representations from different clips. In this section, we investigate how to learn the temporal structure of clips, aggregate diverse representations to reduce computational cost. We first formulate the problem of learning to aggregate, and introduce the *FAST-GRU* architecture for aggregation. We then describe other aggregation operators used for comparison and the back-

bones for extracting the expensive and lightweight representations.

Problem Formulation

Given a sequence of N clips from a video, we denote their feature representations as \mathbf{x}_t , where $t \in [0 \dots N - 1]$. The problem of learning to aggregate can be formulated as:

$$\mathbf{o}_t = f(\mathbf{o}_{t-1}, \mathbf{x}_t), t \in [1 \dots N - 1], \quad (1)$$

where \mathbf{o}_{t-1} encodes the historical information before the current clip \mathbf{x}_t , and \mathbf{o}_0 equals to \mathbf{x}_0 . Note that Eq. 1 is exactly a recurrent neural network. All previous features $\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_t$ are recursively encoded into \mathbf{o}_t . The FASTER framework is suitable for online applications, which requires to provide classification results at any time.

Note also FASTER does not make any assumption about the form of \mathbf{x}_t . In this paper, we use the feature map from the last convolutional layer of the clip models, which is a tensor of shape $l \times h \times w \times c$, as shown in Fig. 2. l is the temporal length of \mathbf{x}_t , whereas $h \times w$ is the spatial resolution and c is the number of channels. At any time t , \mathbf{x}_t can come from either the expensive or lightweight model.

Once all the features are aggregated, we apply global average pooling to \mathbf{o}_{N-1} , then feed it to a fully-connected layer, which is trained with SoftMax loss to generate classification scores.

FAST-GRU

To implement the aggregation function in Eq. 1, we propose the FAST-GRU architecture shown in Fig. 2 (b). While FAST-GRU is based on GRU, we highlight their differences in the following.

3D Convolutional RNN. RNNs are usually used to model a sequence of inputs represented as 1D feature vectors. These feature vectors can be extracted by applying global average pooling on convolutional feature maps. While these representations are very compact, they do not contain any spatio-temporal information due to pooling. The FASTER framework requires the RNN to handle representations generated by different models. Thus keeping the spatio-temporal resolution of the feature maps may lead to better performance for aggregation. We empirically show that incorporating spatio-temporal information is beneficial for long sequence modeling, which has been largely ignored in recent video pooling methods based on RNNs.

Instead of applying a fully-connected layer to the pooled 1D features, we use a 3D $1 \times 1 \times 1$ convolution and keep the original resolution of the feature maps in state transitions, as shown in Fig. 2 (b). The matrix multiplication in gating functions is also replaced with 3D $1 \times 1 \times 1$ convolutions, which enable feature gating in each spatio-temporal location.

Bottleneck Gating. In RNNs, the importance of the current input \mathbf{x}_t is decided by a subnetwork conditioned on the historical information encoded in \mathbf{o}_{t-1} . The subnetwork learns to combine the state \mathbf{o}_{t-1} and input \mathbf{x}_t with different

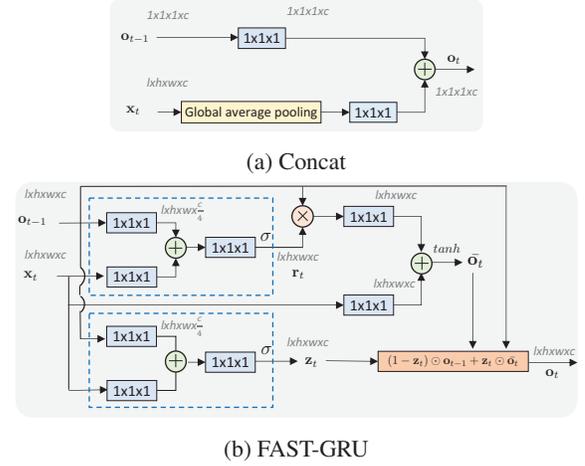


Figure 2: (a) The ‘‘Concat’’ baseline simply concatenates \mathbf{x}_t with \mathbf{o}_{t-1} ; (b) Unlike typical GRU collapses the resolution of \mathbf{x}_t with global average pooling, FAST-GRU keeps the resolution of \mathbf{x}_t to learn spatio-temporal information. We also introduce bottleneck structures (in dash rectangles) in the gate functions to reduce the number of parameters and increase non-linearity.

gating mechanisms. Different gates are designed for different functionality for sequential modeling in RNNs. Here we first briefly describe the original GRU architecture, then introduce the bottleneck gating mechanism we proposed.

In GRU, the subnetwork first concatenates \mathbf{x}_t and \mathbf{o}_{t-1} , then project them to a vector of the same dimension, *i.e.*, the number of channels c . There are two gates in GRU, *i.e.*, the read gate \mathbf{r} and update gate \mathbf{z} , which are defined as

$$\begin{aligned} \mathbf{r}_t &= \sigma(\mathbf{G}_{rx}\mathbf{x}_t + \mathbf{G}_{ro}\mathbf{o}_{t-1}), \\ \mathbf{z}_t &= \sigma(\mathbf{G}_{zx}\mathbf{x}_t + \mathbf{G}_{zo}\mathbf{o}_{t-1}), \end{aligned} \quad (2)$$

where \mathbf{G}_* are $c \times c$ matrices, and σ is the sigmoid function. After the calculation of \mathbf{r}_t and \mathbf{z}_t , The output of GRU is

$$\begin{aligned} \bar{\mathbf{o}}_t &= \tanh(\mathbf{V}_{\bar{o}}\mathbf{x}_t + \mathbf{V}_{\bar{o}}(\mathbf{r}_t \odot \mathbf{o}_{t-1})), \\ \mathbf{o}_t &= (1 - \mathbf{z}_t) \odot \mathbf{o}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{o}}_t, \end{aligned} \quad (3)$$

where \mathbf{V}_* are $c \times c$ matrices, and \odot represents element-wise multiplication. Both \mathbf{G}_* and \mathbf{V}_* are learnable parameters.

Inspired by ResNet (He et al. 2016), we introduce additional bottleneck structures to increase the expressiveness of the subnetwork in FAST-GRU, shown in the dash rectangles of Fig. 2 (b). In the bottleneck layer, \mathbf{x}_t and \mathbf{o}_{t-1} are projected to a lower dimensional feature space after concatenation. This compresses the number of channels to reduce parameters and introduces more non-linearity. The compressed features are then recovered to the original dimension c with another projection. In practice, channel compression is implemented by $1 \times 1 \times 1$ convolution to reduce the dimension by a factor of r . After that, there is an ReLU followed by $1 \times 1 \times 1$ convolution to recover the dimensionality. The read and update gates of FAST-GRU are defined

as:

$$\begin{aligned}
 \mathbf{r}'_t &= \text{ReLU}(\mathbf{U}_{rx} * \mathbf{x}_t + \mathbf{U}_{ro} * \mathbf{o}_{t-1}), \\
 \mathbf{z}'_t &= \text{ReLU}(\mathbf{U}_{zx} * \mathbf{x}_t + \mathbf{U}_{zo} * \mathbf{o}_{t-1}), \\
 \mathbf{r}_t &= \sigma(\mathbf{W}_{r'} * \mathbf{r}'), \\
 \mathbf{z}_t &= \sigma(\mathbf{W}_{z'} * \mathbf{z}'),
 \end{aligned}
 \tag{4}$$

where \mathbf{U}_* are $1 \times 1 \times 1$ convolutions to compress the number of channels by a factor of r , and \mathbf{W}_* are $1 \times 1 \times 1$ convolutions to recover the dimensionality. Our gates are generated by taking the jointly compressed features \mathbf{r}'_t and \mathbf{z}'_t , which enables more powerful gating for feature aggregation. We empirically set r to be 4, which gives good results in our experiments.

Other Aggregation Operator Variants

Besides FAST-GRU, we also consider other baselines for the learning-to-aggregate problem for a comprehensive comparison. We detail each baseline in the following.

Average pooling is the most popular approach for aggregation. Despite its simplicity, it performs remarkably well in practice. In the context of our FASTER framework, we just average the classification scores from each clip regardless of whether the scores are generated by the expensive or lightweight model. The *concat* baseline is illustrated in Fig. 2a, where \mathbf{x}_t and \mathbf{o}_{t-1} are concatenated together, and then projected to a joint feature space. It is defined as:

$$f(\mathbf{o}_{t-1}, \mathbf{x}_t) = \text{ReLU}(\text{BN}(\mathbf{W}\mathbf{o}_{t-1} + \mathbf{U}\mathbf{x}_t)), \tag{5}$$

where \mathbf{W} and \mathbf{U} are learnable parameters. Batch normalization (Ioffe and Szegedy 2015) and ReLU are also applied in the *concat* baseline to further improve the performance.

Popular RNNs such as LSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2015) are the most related baselines. These are the go-to methods for tasks that involve sequential modeling (Sutskever, Vinyals, and Le 2014), such as language or speech. For the *LSTM* baseline, we use the variant that consists of three gates and an additional cell between time steps (Graves 2013). LSTM has a forget gate that can reset the history for the current input \mathbf{x}_t . For video classification, this could be useful in the case of different camera shots or unrelated actions. The *GRU* baseline we used has two gates, *i.e.*, the read and update gate as we described in the previous section. The state transition function uses the update gate to perform a weighted average of the historical state \mathbf{o}_{t-1} and the current input \mathbf{x}_t .

Clip-level Backbones

FASTER is a general framework and does not make any assumption about the underlying clip-level backbones. We can potentially choose any popular networks, such as I3D (Carreira and Zisserman 2017), R(2+1)D (Tran et al. 2018) or non-local network (Wang et al. 2018). We implement FASTER based on R(2+1)D because it is one of the state-of-the-art methods, and its Github repository¹ provides a family of networks that can be used as the expensive and lightweight models. In particular, we choose R(2+1)D with

¹<https://github.com/facebookresearch/VMZ>

Table 1: Clip-level backbones for extracting the expensive and lightweight representations. The FLOPs of R(2+1)D-50 is about $10\times$ of R2D-26.

Layers	R2D-26	R(2+1)D-50	Output sizes $L \times H \times W$
conv ₁	$8 \times 7 \times 7, 64$ stride 8, 2, 2	$1 \times 7 \times 7, 45$, stride 1,2,2 $3 \times 1 \times 1, 64$, stride 1,1,1	R2D : $\frac{L}{8} \times 112 \times 112$ R(2+1)D : $L \times 112 \times 112$
pool ₁	$1 \times 3 \times 3$ max stride 1, 2, 2		R2D : $\frac{L}{8} \times 56 \times 56$ R(2+1)D : $L \times 56 \times 56$
res ₂	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 1 \times 3 \times 3, 144 \\ 3 \times 1 \times 1, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	R2D : $\frac{L}{8} \times 56 \times 56$ R(2+1)D : $L \times 56 \times 56$
res ₃	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 1 \times 3 \times 3, 288 \\ 3 \times 1 \times 1, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	R2D : $\frac{L}{8} \times 28 \times 28$ R(2+1)D : $\frac{L}{2} \times 28 \times 28$
res ₄	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 1 \times 3 \times 3, 576 \\ 3 \times 1 \times 1, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	R2D : $\frac{L}{8} \times 14 \times 14$ R(2+1)D : $\frac{L}{4} \times 14 \times 14$
res ₅	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ 1 \times 3 \times 3, 512 \\ 1 \times 1 \times 1, 2048 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ 1 \times 3 \times 3, 1152 \\ 3 \times 1 \times 1, 512 \\ 1 \times 1 \times 1, 2048 \end{bmatrix} \times 3$	R2D : $\frac{L}{8} \times 7 \times 7$ R(2+1)D : $\frac{L}{8} \times 7 \times 7$
global average pooling, fc			

50 layers as the expensive model, and R2D with 26 layers as the lightweight model, as detailed in Table 1.

Comparing with the original R(2+1)D, we make two changes to further improve its performance. First, we replace convolutional blocks with bottleneck blocks, which have been widely used in the family of ResNet architectures and shown to both reduce computational cost and improve accuracy. Second, we insert a max-pooling layer after conv₁, which enables R(2+1)D to support a spatial resolution of 224×224 without significantly increasing its FLOPs.

For the lightweight R2D model, bottleneck layers are used in the same way as R(2+1)D. To reduce the FLOPs of R2D, a temporal stride of 8 is used in conv₁, which effectively reduces the temporal length of the clip by a factor of 8. Unlike R(2+1)D, R2D only has 26 layers to further reduce the computational cost.

To integrate R(2+1)D-50 and R2D-26 in FASTER, we simply use the two backbones as feature extractors. In Table 1, the output from res₅ generates a feature map of size $\frac{L}{8} \times 7 \times 7 \times 2048$, which is used as input \mathbf{x}_t to the RNNs.

Experimental Setups

In this section, we describe the experimental setups, *i.e.*, the datasets, training and test protocols for both the clip-level backbones and FASTER framework.

Datasets. We choose the Kinetics (Kay et al. 2017) dataset as the major testbed for FASTER. Kinetics is among the most popular datasets for video classification. To simplify, all reported results on Kinetics are trained from scratch, without pre-training on other datasets (*e.g.*, Sports1M or ImageNet). Kinetics has 400 action classes and about 240K training videos. We report top-1 accuracy on the validation set as labels on the testing set is not public available. We also report results on UCF-101 (Soomro, Zamir, and Shah 2012) and HMDB-51 (Kuehne et al. 2011). These datasets are much smaller, thus we use Kinetics for pre-training and report mean accuracy on three testing splits.

Table 2: Accuracy of clip-level backbones on Kinetics. The significant performance gap between the two backbones offers a great opportunity to explore better accuracy/FLOPs trade-offs when combining them.

Model	Input	Video@1	GFLOPs×Clips
R2D-26	8-frames	63.3	3.2×32
	16-frames	65.5	6.0×16
	32-frames	66.7	12.7×8
R(2+1)D-50	8-frames	70.1	30.0×32
	16-frames	72.3	60.0×16
	32-frames	74.5	119.9×8

Table 3: Comparison of different architectures for aggregation on Kinetics. The clip length L is 8. For all the methods, only the first clip is processed by the expensive model, and the remaining clips are processed by the lightweight model.

Model	Number of Clips				
	2	4	8	16	32
Avg. pool	60.7	63.9	64.2	64.0	63.8
Concat	61.6	65.9	66.3	62.3	58.2
LSTM	61.3	66.2	66.3	66.3	64.7
GRU	61.1	65.6	66.5	66.1	65.8
FAST-GRU	61.1	65.6	67.2	67.4	67.2

Due to limited GPU memory, we train the FASTER framework in a two-stage process. First, two clip-level backbones are trained by themselves separately. In the second stage, two backbones are fixed and we train different RNNs using the feature maps extracted by the backbones.

Setups for clip-level backbones. We mostly follow the procedure in (Tran et al. 2018) to train the clip-level backbones except the following two changes. First, we scale the input video whose shorter side is randomly sampled in [256, 320] pixels, following (Wang et al. 2018). Second, we adopt the cosine learning rate schedule (Loshchilov and Hutter 2016). During training, we randomly sample L consecutive frames from a given video. For testing, we uniformly sample N clips to cover the whole video, and average pool the classification scores of each clip, as shown in Table 2. We fix the total number of frames processed to be 256, *i.e.*, $N \times L = 256$. As the average length of Kinetics videos is about 10 seconds in 30 FPS, 256 frames give us enough coverage over the whole video. We only use RGB frames as input as computing optical flow can be very expensive.

Setups for learning to aggregate. To train different RNNs for learning to aggregate, we randomly sample N clips, and each clip has L consecutive frames. As mentioned above, the clip-level backbones are fixed in the second stage due to limited GPU memory. The initial state \mathbf{o}_0 is set to be the features from the first clip \mathbf{x}_0 . The training procedure is similar to the first stage except we train much fewer epochs. For testing, the N clips are uniformly sampled to cover the whole video.

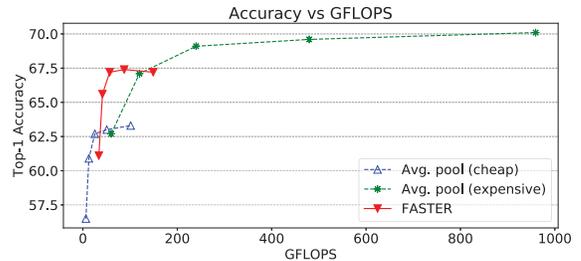


Figure 3: FASTER vs. typical video classification frameworks with average pooling. We show accuracy vs. GFLOPs of different methods as a function of the number of clips. We compare the results of FAST-GRU from Table 3 with average pooling of all expensive (Avg. pool (expensive)) and lightweight models (Avg. pool (cheap)).

Experimental Results

In this section, we demonstrate the advantage of FASTER and discuss accuracy/FLOPs trade-offs in different settings on Kinetics. We also compare FASTER with the state of the art on Kinetics, UCF-101, and HMDB-51.

Accuracy of the backbone architectures. Table 2 summarizes the top-1 accuracy of the two clip-level backbones on Kinetics. We consider the clip length L to be 8, 16 or 32 frames, and sample 32, 16 or 8 clips, respectively. This is to ensure that we always process the same number of frames under different settings. Note that the GFLOPs of R(2+1)D-50 is about 10× of R2D-26, and the accuracy of R(2+1)D-50 is about 7% better than R2D-26. The significant performance gap between the two backbones gives us the opportunity to explore different trade-offs in the following experiments. For the same backbone, longer clip length L often leads to better accuracy. But we can not train R(2+1)D-50 with $L = 64$ frames due to GPU memory limit.

FAST-GRU vs. other RNN variants. We compare the accuracy of different architectures for the learning-to-aggregate problem across a wide range of number of clips ($N = \{2, 4, 8, 16, 32\}$). The first clip is always processed with the expensive model and the following $N - 1$ clips are processed with the cheap model. We use clip length $L = 8$ frames. This allows us to observe the behavior of different architectures in shorter and longer sequences. The results are shown in Table 3. For “Avg. pool”, we simply average the classification scores of all the clips. The performance of “Avg. pool” saturates as more cheap clips are added. Less accurate predictions dominate the final score, since the accuracy of 16 clips (*i.e.*, 64.0%) is quite similar to the accuracy of 32 clips (*i.e.*, 63.8%). This suggests that the expensive representations from the first clip are not fully utilized.

“Concat” outperforms “Avg. pool” by 2.1% when $N = 8$, showing the benefits of learning temporal structure. However, as N increases, the performance of “Concat” decreases, and by $N = 32$, the accuracy is 8.1% lower than $N = 8$. This shows that learning to aggregate over a longer

Table 4: Optimal ratio between the expensive and cheap clips. For a fixed number of total frames, we vary the number of clips N and clip length L , and measure accuracy and GFLOPs on Kinetics. Empty cells correspond to combinations that are not feasible. “E” denotes the expensive model, while “C” denotes the cheap model. GFLOPs combine the costs of the cheap, expensive and FAST-GRU models together.

#E :#C	8-frames×32-clips		16-frames×16-clips		32-frames×8-clips	
	V@1	GFLOPs	V@1	GFLOPs	V@1	GFLOPs
All E	70.6	982.4	72.9	980.2	74.9	979.2
1:1	70.5	553.6	72.9	552.0	74.6	550.4
1:3	70.1	339.2	72.1	337.6	73.3	336.0
1:7	69.6	230.4	71.7	230.4	71.3	228.8
1:15	69.0	176.0	69.9	176.0		
1:31	67.2	150.4				

time period is difficult. Interestingly, “Concat”, and all recurrent networks have similar performances when the number of clips is small ($N = 2, 4$), showing that the advantage of RNNs is not evident for short sequences. As the number of clips increases ($N \geq 16$), recurrent networks show their strength in modeling long sequences and perform better than the simple “Concat” baseline.

Of all the methods, FAST-GRU achieves the best performance for long sequences. For example, FAST-GRU outperforms LSTM by 2.5% and GRU by 1.4% when $N = 32$. The results show that our proposed FAST-GRU is beneficial for long sequence learning, even when sequence samples come from different distributions.

FASTER vs. average pooling. How well does the FASTER framework do with respect to the prevailing average pooling? Recall that our main focus is not only to increase accuracy, but also to reduce computational cost. For this, we compare the results of FAST-GRU from Table 3 with the traditional framework (Fig. 1 (a)), where all clips are cheap (or all clips are expensive) and aggregated with average pooling. These can be considered as the lower and upper bound of the FASTER framework. We plot the GFLOPs vs. accuracy comparisons in Fig. 3. We use clip length $L = 8$ frames, and consider the number of clips $N = \{2, 4, 8, 16, 32\}$ as in Table 3. When $N = 4$, it takes 119 GFLOPs for “Avg. pool (expensive)” to achieve 67.1%. FASTER can achieve comparable results with only half the GFLOPs. Comparing with “Avg. pool (cheap)” when $N = 16$, FASTER is over 4% better with similar amount of GFLOPs.

Note that there is still an accuracy gap between FASTER and “Avg. pool (expensive)”. Next, we will introduce more expensive clips and study the optimal ratio between the expensive and cheap clips, achieving superior results.

Better accuracy/FLOPs trade-offs. There are three parameters in our framework that affect accuracy and computational cost: the clip length (L), the number of clips (N), and the proportion of expensive and lightweight clips used as input pattern (#E :#C). We now provide an empirical study

Table 5: Comparisons of runtime between different methods. The results of FASTER is reported in the 32-frames×8-clips setting.

Model	Time (secs)	GFLOPs	Accuracy
All C (Avg pool)	15.1	101.3	66.7
All E (Avg pool)	75.7	959.3	74.5
FASTER32	35.3	550.4	74.6

Table 6: Compare different architectures as the expensive clip-level backbone. Our FASTER framework can generalize to different architectures, and R3D even performs slightly better than R(2+1)D.

#E :#C	32-frames×8-clips			
	R(2+1)D-50		R3D-50	
	V@1	GFLOPs	V@1	GFLOPs
All E (Avg. pool)	74.5	959.3	74.4	942.7
1:1 (FAST-GRU)	74.6	550.4	75.3	542.1
1:3 (FAST-GRU)	73.3	336.0	73.9	331.9

of these parameters to achieve the best trade-off. Since we are interested in the behavior of FASTER, we “fix” the input data by keeping the number of frames constant, which spans the entire video. Thus, $L \times N = 256$ for all settings. Given N clips, we experiment with input patterns of $1 : x$. It is the ratio between the number of expensive clips and cheap clips, $x \in \{1, 3, 7, 15, 31\}$. For example, when $x = 3$ and $L = 16$, there will be 4 expensive clips and 12 cheap clips. Additionally, we evaluate the case where all the inputs are expensive clips ($x = 0$).

Results are shown in Table 4. As expected, a higher ratio of expensive models leads to a higher accuracy. For example, when $L = 8$, there is a large gap of 1.8% when x goes from 31 to 15, with a modest increase of 25.6 GFLOPs. The gap is less obvious for higher ratios of expensive models. A remarkable observation is that FASTER achieves 70.1% when $x = 3$, which is the same accuracy as the average pooling baseline with all expensive clips from Table 2. But FASTER only consumes 35% of its GFLOPs. Even more, for $x = 1$, the proposed method outperforms the average pooling baseline in Table 2, at 57% of the cost. This shows the benefit of learning temporal aggregation of hybrid inputs.

We also observe that for a fixed cost, it is often beneficial to use longer clip length L . For example, for a fixed proportion of cheap models $x = 3$, the accuracy of 32-frames×8-clips is higher than 16-frames×16-clips, and that in turn is considerably higher than 8-frames×32-clips. In other words, it is better to have fewer clips with more frames. However, for higher ratios of cheap clips (e.g., $x = 7$), this pattern does not hold.

We choose the parameter setting $L = 16$ and $x = 7$ as the best inexpensive configuration, and call this setting FASTER16 (i.e., $L = 16, x = 7$). It indicates that it is possible to achieve better performance with shorter clip length L , which reflects the importance of learning a good aggregation function. When $L = 32$ and $x = 1$, the proposed framework achieves 74.6% accuracy with 550.4 GFLOPs. We denote this setting as FASTER32 (i.e., $L = 32, x = 1$) and it is the best model when the cost is around ~ 550 GFLOPs.

Table 7: Comparisons to the state-of-the-art methods on Kinetics. The inputs for all the methods are RGB frames. For “GFLOPs \times clips”, we report the cost of a single clip and the number of clips used. “N/A” indicates the number of clips are not reported in the paper.

Model	Top-1	ImageNet pre-train	GFLOPs \times clips
I3D (Carreira and Zisserman 2017)	72.1	✓	108 \times 4
S3D (Xie et al. 2018)	72.2	✓	66.4 \times N/A
MF-Net (Chen et al. 2018b)	72.8	✓	11.1 \times 50
A ² -Net (Chen et al. 2018a)	74.6	✓	40.8 \times 30
S3D-G (Xie et al. 2018)	74.7	✓	71.4 \times N/A
NL I3D-50 (Wang et al. 2018)	76.5	✓	282 \times 30
NL I3D-101 (Wang et al. 2018)	77.7	✓	359 \times 30
I3D (Carreira and Zisserman 2017)	68.4	-	108 \times 4
STC (Diba et al. 2018)	68.7	-	N/A \times N/A
ARTNet (Wang et al. 2017)	69.2	-	23.5 \times 250
S3D (Xie et al. 2018)	69.4	-	66.4 \times N/A
ECO (Zolfaghari, Singh, and Brox 2018)	70.0	-	N/A \times N/A
R(2+1)D-34 (Tran et al. 2018)	72.0	-	152 \times 115
FASTER16	71.7	-	14.4 \times 16
FASTER32	75.3	-	67.7 \times 8

Table 8: Comparisons on UCF-101 and HMDB-51. “I” denotes pre-training on ImageNet, while “K” denotes pre-training on Kinetics.

Model	Pre-train I	Pre-train K	UCF-101	HMDB-51
STC (Diba et al. 2018)	-	✓	95.8	72.6
ARTNet (Wang et al. 2017)	-	✓	94.3	70.9
ECO (Zolfaghari, Singh, and Brox 2018)	-	✓	93.6	68.4
R(2+1)D-34 (Tran et al. 2018)	-	✓	96.8	74.5
I3D (Carreira and Zisserman 2017)	✓	✓	95.6	74.8
S3D (Xie et al. 2018)	✓	✓	96.8	75.9
FASTER32	-	✓	96.9	75.7

FASTER also works when all input features come from the same network. When the inputs are all expensive features, FASTER outperforms the average pooling baseline from Table 2 in all settings, with a negligible increase of FLOPs from FAST-GRU.

Runtime vs. FLOPs. We measure the runtime speed of different methods on a TITAN X GPU with an Intel i7 CPU. We sum up the runtime over 100 Kinetics videos. The results are listed in Table 5, which confirms that the theoretical FLOPs are consistent with the runtime for our experiments. The reduction of FLOPs is roughly proportional to the decrease of runtime.

Generalize to different backbones. Our FASTER framework does not make any assumption about the clip-level backbones. To demonstrate that FASTER can generalize to different backbones, we replace the expensive model, *i.e.*, R(2+1)D-50, with R3D-50 (Tran et al. 2018) and keep everything else unchanged. In Table 6, R3D even works slightly better than R(2+1)D. Comparing FASTER32 with the average pooling baseline, using R3D leads to a larger improvement of 0.9%.

Comparison to state-of-the-art. We compare to state-of-the-art methods across three most popular video datasets,

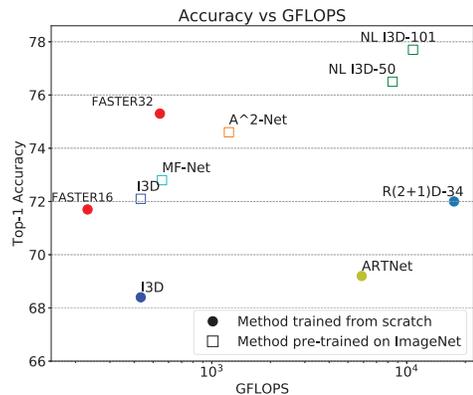


Figure 4: Log-scale GFLOPs vs. accuracy comparisons on Kinetics. In this chart, optimal methods are closer to the top-left corner. The two proposed variants of FASTER are the closest to the corner.

i.e., Kinetics, UCF-101 and HMDB-51. We only include methods that use RGB frames as inputs. Results on Kinetics are shown in Table 7. The results are also visualized in terms of accuracy vs. GFLOPs in Fig. 4. Our proposed FASTER framework achieves the best accuracy/GFLOPs trade-offs. This is specially impressive since FASTER is not pre-trained on ImageNet, and the clip-level backbones used are not the most cost effective ones. In this comparison, we use FASTER32 with R3D as its backbone. FASTER32 outperforms R(2+1)D-34 by 3.3% with only 4% of its GFLOPs. FASTER16 outperforms I3D trained from scratch by 3.3% (71.7% vs. 68.4%), using only half of I3D’s cost (432 GFLOPs). Even though we do not leverage pre-training, our framework outperforms all existing pre-trained models except NL-I3D. However, FASTER32 only uses 5% of the GFLOPs of NL-I3D. Our FASTER32 model outperforms A²-Net by 0.7%, while the cost is reduced by over 50%. The results validate the effectiveness of our framework, showing it is promising to leverage the combination of expensive and lightweight models for better accuracy/FLOPs trade-off.

Competitive results are also achieved on UCF-101 and HMDB-51 (Table 8). It shows learning the aggregation function on larger datasets can facilitate the generalization on smaller datasets. The improvements are less profound as the performance on small datasets tends to be saturated.

Conclusion

In this paper, we propose a novel framework called FASTER for efficient video classification. To exploit the spatio-temporal redundancy of the video sequence, we combine the representations from both expensive and lightweight models. We propose a recurrent unit called FAST-GRU to learn to aggregate mixed representations. Experimental results show that our FASTER framework has significantly better accuracy/FLOPs trade-offs, achieving the state-of-the-art accuracy with 10 \times less FLOPs.

References

- Alwassel, H.; Caba Heilbron, F.; and Ghanem, B. 2018. Action search: Spotting actions in videos and its application to temporal action localization. In *ECCV*.
- Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*.
- Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; and Baskurt, A. 2011. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*.
- Bhardwaj, S.; Srinivasan, M.; and Khapra, M. M. 2019. Efficient video classification using fewer frames. In *CVPR*.
- Carreira, J., and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*.
- Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; and Feng, J. 2018a. A²-nets: Double attention networks. In *NIPS*, 350–359.
- Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; and Feng, J. 2018b. Multi-fiber networks for video recognition. In *ECCV*.
- Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2015. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*.
- Dalal, N.; Triggs, B.; and Schmid, C. 2006. Human detection using oriented histograms of flow and appearance. In *ECCV*.
- Diba, A.; Fayyaz, M.; Sharma, V.; Arzani, M. M.; Yousefzadeh, R.; Gall, J.; and Van Gool, L. 2018. Spatio-temporal channel correlation networks for action classification. In *ECCV*.
- Feichtenhofer, C.; Pinz, A.; and Zisserman, A. 2016. Convolutional two-stream network fusion for video action recognition. In *CVPR*.
- Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; and Russell, B. 2017. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*.
- Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.
- Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.
- Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3d convolutional neural networks for human action recognition. *TPAMI*.
- Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *CVPR*.
- Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Klaser, A.; Marszałek, M.; and Schmid, C. 2008. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*.
- Korbar, B.; Tran, D.; and Torresani, L. 2019. Scsampler: Sampling salient clips from video for efficient action recognition. *arXiv preprint arXiv:1904.04289*.
- Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. Hmdb: a large video database for human motion recognition. In *ICCV*.
- Laptev, I. 2005. On space-time interest points. *IJCV*.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.
- Li, Z.; Gavriluyk, K.; Gavves, E.; Jain, M.; and Snoek, C. G. 2018. Videolstm convolves, attends and flows for action recognition. *CVIU* 166:41–50.
- Loshchilov, I., and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*.
- Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.
- Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Srivastava, N.; Mansimov, E.; and Salakhudinov, R. 2015. Unsupervised learning of video representations using LSTMs. In *ICML*.
- Sun, L.; Jia, K.; Yeung, D.-Y.; and Shi, B. E. 2015. Human action recognition using factorized spatio-temporal convolutional networks. In *ICCV*.
- Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.
- Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*.
- Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*.
- Wang, H., and Schmid, C. 2013. Action recognition with improved trajectories. In *ICCV*.
- Wang, H.; Kläser, A.; Schmid, C.; and Liu, C.-L. 2011. Action recognition by dense trajectories. In *CVPR*.
- Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Van Gool, L. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.
- Wang, L.; Li, W.; Li, W.; and Van Gool, L. 2017. Appearance-and-relation networks for video classification. In *CVPR*.
- Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *CVPR*.
- Wu, Z.; Xiong, C.; Ma, C.-Y.; Socher, R.; and Davis, L. S. 2019. Adafame: Adaptive frame selection for fast video recognition. In *CVPR*.
- Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*.
- Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *CVPR*.
- Zhou, Y.; Sun, X.; Zha, Z.-J.; and Zeng, W. 2018. Mict: Mixed 3d/2d convolutional tube for human action recognition. In *CVPR*.
- Zolfaghari, M.; Singh, K.; and Brox, T. 2018. Eco: Efficient convolutional network for online video understanding. In *ECCV*.