

# SMART Frame Selection for Action Recognition

Shreyank N Gowda<sup>1</sup>

Marcus Rohrbach<sup>2</sup>

Laura Sevilla-Lara<sup>1</sup>

<sup>1</sup>University of Edinburgh

<sup>2</sup>Facebook AI Research

## Abstract

Action recognition is computationally expensive. In this paper, we address the problem of frame selection to improve the accuracy of action recognition. In particular, we show that selecting good frames helps in action recognition performance even in the trimmed videos domain. Recent work has successfully leveraged frame selection for long, untrimmed videos, where much of the content is not relevant, and easy to discard. In this work, however, we focus on the more standard short, trimmed action recognition problem. We argue that good frame selection can not only reduce the computational cost of action recognition but also increase the accuracy by getting rid of frames that are hard to classify. In contrast to previous work, we propose a method that instead of selecting frames by considering one at a time, considers them jointly. This results in a more efficient selection, where “good” frames are more effectively distributed over the video, like snapshots that tell a story. We call the proposed frame selection SMART and we test it in combination with different backbone architectures and on multiple benchmarks (Kinetics, Something-something, UCF101). We show that the SMART frame selection consistently improves the accuracy compared to other frame selection strategies while reducing the computational cost by a factor of 4 to 10 times. Additionally, we show that when the primary goal is recognition performance, our selection strategy can improve over recent state-of-the-art models and frame selection strategies on various benchmarks (UCF101, HMDB51, FCVID, and ActivityNet).

## Introduction

Video processing is computationally expensive. At the same time, the amount of video content being generated is increasing fast and constitutes a large part of the computation of many big social media platforms. Traditionally, most efforts in action recognition have focused on improving accuracy by creating larger architectures. These architectures take as input either a frame or a set of frames (also called a *clip*) and produce a prediction. These predictions are then aggregated over time. The frames or clips are either sampled densely (Simonyan and Zisserman 2014; Yue-Hei Ng et al. 2015) or randomly (Wang et al. 2016).

---

Copyright © 2021, Association for the Advancement of Artificial Intelligence ([www.aaai.org](http://www.aaai.org)). All rights reserved.

Videos, however, provide an opportunity for reducing computational cost in multiple ways. First, videos contain highly temporally redundant data, making it easier to skip parts without losing much information (Fan et al. 2018). Second, some parts of a video can be more discriminative than others, due to their content, or other phenomena like blur, occlusions, etc. Supporting this intuition, Huang et al. (2018) show experimentally that using an oracle to make an optimal selection of frames (or clips), produces more accurate classification results than using the entire video. Additionally, Sevilla-Lara et al. (2019) show that many action classes in standard datasets do not require motion or temporal information to be identified. For a human observer, a few still frames are often discriminative enough. This suggests that large parts of a video can be discarded.

Several recent works(Korbar, Tran, and Torresani 2019; Wu et al. 2019c; Zhu et al. 2019) have successfully leveraged these principles to reduce computational cost at test time. These methods have used a common strategy: they use an inexpensive way to decide which regions of the video are important and discriminative, and only process those with an expensive method. This general problem has been referred to as frame or clip selection. While very successful, most frame and clip selection methods have focused on a particular domain of action recognition, namely long, and frequently sparse videos with a typical length of a minute or more, e.g. ActivityNet(Caba Heilbron et al. 2015), Sports1M(Karpathy et al. 2014), FCVID(Jiang et al. 2017a), Youtube 8M(Abu-El-Haija et al. 2016). This is indeed the domain where discarding portions of a video is easier and has potentially the largest effect. In contrast, the problem of frame selection in short videos of a few seconds remains much less explored, probably due to its difficulty.

In this paper we propose a method to do frame selection in the core, standard activity classification setting of trimmed video clips. Part of the challenge in this setting is that “good” frames are often temporally close together within a video. Since most existing frame selection methods consider the value of choosing a frame one at a time, the selected frames tend to only represent part of the action. In other words, the diversity of frames, and their ability to tell a story are disregarded. We also show that using language features along with visual features helps improve the performance.

To handle these challenges, we propose a model that, in

addition to considering the discriminative value of a single frame, also considers its relation to others in a video. We do this by using an attention and a relational network (Meng et al. 2019; Sung et al. 2018), that examines the value of frames jointly. We learn our Sampling through Multi-frame Attention and Relations in Time, which we dub the *SMART* selection network.

We test our SMART frame selection network on several trimmed action recognition datasets, including Something-something, UCF101 and subsets of Kinetics. We observe that in all of them the proposed method outperforms the baselines, including using the full video, while reducing the computational cost by a factor of 4 to 10, depending on the dataset. We also test the proposed method on the untrimmed setting in ActivityNet and FCVID, where we get higher accuracies than all previous work on frame selection. Further, we extend our frame selection approach to select frames that are then passed at test time to deep action recognition models and show that we obtain state-of-the-art results on UCF101 and HMDB51 which are trimmed video datasets, showing that frame selection can be an important step to improve accuracy in trimmed action recognition.

## Related Work

The field of action recognition is wide, and includes a large variety of subproblems, and families of methods. Here we focus on the two areas within action recognition that are most relevant to our work: frame selection as well as attention and relational models.

**Frame Selection.** Selecting important frames for action recognition is a relatively new area. Many approaches have successfully trained a reinforcement learning (RL) agent approach that examines one frame at a time, to predict how many frames can be skipped.

AdaFrame(Wu et al. 2019c) leverages RL, in combination with an LSTM that is augmented with memory that helps providing context information for selecting frames to use. Given a frame, it generates a prediction of the action class and it decides which frame to observe next and computes the expected reward of seeing more frames. FastForward(Fan et al. 2018) is an end-to-end reinforcement learning approach. It consists of two sub networks: an adaptive stop network and fast forward network. The adaptive stop network can either let the frame sampling continue or stop. The fast forward network has a set of several actions (going backwards or going forward with varying seconds). The RL agent learns to skim through the video.

FrameGlimpse(Yeung et al. 2016) follows the intuition that detecting an action is dependent on observation and refinement. Based on this, FrameGlimpse relies on a recurrent neural network (RNN) based agent that observes and decides where to look next. Given the current frame, the agent also decides whether to emit a prediction based on a confidence score. If the agent is not confident enough then it decides to look ahead.

Multi-agent Reinforcement Learning (MARL)(Wu et al. 2019a) formulates the frame sampling procedure as multiple parallel Markov decision processes which aim at picking frames by gradually adjusting an initial sampling. They have

a context-aware observation network which jointly models context information among nearby agents and historical states of a specific agent. They also have a policy network which generates a probability distribution over a predefined action space.

SCSampler(Korbar, Tran, and Torresani 2019) is a lightweight clip-sampler that can efficiently obtain the most salient temporal clips within a long video. They sample features directly from compressed videos and also from the audio obtained from the video. Attention aware sampling (AAS) (Dong, Zhang, and Tan 2019) uses an agent which discards irrelevant frames using attention. They consider the frame selection procedure as a Markov decision process and train an agent without extra labels through deep reinforcement learning.

While all these approaches showed great results, they have mostly focused on the scenario of untrimmed videos. SCSampler does however report results on Kinetics(Carreira and Zisserman 2017), however, it requires audio as an extra modality. Untrimmed videos contain significant parts of unnecessary data and discarding them is easier than discarding frames from trimmed videos. In contrast to previous work, we propose a method that instead of selecting frames by considering one at a time, considers them jointly.

**Attention and Relational Models.** The concept of attention was introduced by Bahdanau, Cho, and Bengio (2014) for the objective of machine translation. This concept of attention is based on the concept that the neural network will learn how relevant different samples are regarding the desired output state in a sequence, or image regions. These values of importance are specified as weights of attention and are generally calculated at the same time as other model parameters trained for a specific goal.

Attention has been used in first person action recognition by having a joint learning of gaze and actions (Li, Liu, and Rehg 2018), by using object-centric attention (Sudhakaran and Lanz 2018) or via event modulated attention (Shen et al. 2018). The use of attention to weigh spatial regions representative of a particular task was done by generating spatial attention masks implicitly by training the network with video labels (Sharma, Kiros, and Salakhutdinov 2015; Zhang et al. 2018; Girdhar and Ramanan 2017). Temporal attention was used for action recognition by detecting change in gaze (Piergiovanni, Fan, and Ryoo 2017; Shen et al. 2018).

LRCN (Donahue et al. 2016) introduced a simple LSTMs for frame-aggregation across time for action recognition. Non-local Networks (Wang et al. 2018) introduce a residual self attention block in convolutional networks to aggregate information across all temporal and/or spatial locations.

Inspired by the relation-net(Sung et al. 2018), relation attention was proposed to deal with the task of facial emotion recognition (Meng et al. 2019). They believed that having a global level representation of features in addition to the local level representation helps obtain better results. We improve upon this approach by adding relation-temporal attention to add a global representation to our temporal attention.

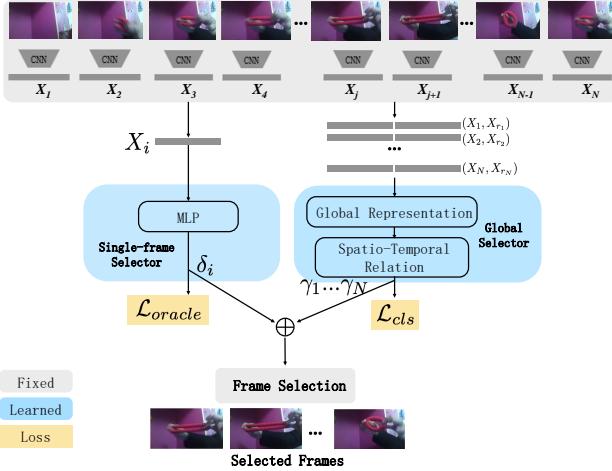


Figure 1: Overview of the SMART frame selection.

## SMART Frame Selection

The proposed approach is designed to use a small portion of the overall computational cost in selecting the best frames. These frames will then be classified using a more computationally expensive model. Therefore, we use a very lightweight representation of the frames as input to the SMART frame selection model.

The model consists of two streams. The first considers the information of the frames one at a time, and outputs a score  $\delta_i$  for each frame, which represents how useful the frame is for classification. The second stream considers the entire video at a time. It takes as input pairs of frames, and uses an attention and relational network to also obtain a score  $\gamma_i$  of how useful these pairs of frames are. Both scores are then multiplied, to obtain a final score of how good each frame is. Given a budget of  $n$  frames, we now select the top  $n$  frames with the highest discriminative score, and use an expensive, high quality classifier for the final prediction. An overview of the method can be seen in Fig. 1. We now describe each of the components in detail.

### Feature Representation

We choose the lightweight MobileNet(Sandler et al. 2018) to extract the visual features of each frame, to minimize the computational cost of this stage. We also make the observation that, in addition to the visual features, we can use language features associated with the content of the frame. The intuition behind this is to enrich the representation with terms that are related to the content of the image. One could imagine that if for an action class like “kayaking”, having associated words like *water*, *boat*, or *paddle* can help discrimination in cases where the kayak is not as apparent visually. We run Mobilenet pre-trained on Imagenet on the frames, and take the top 10 Imagenet classes with highest probability. The names of these classes are then embedded with a pre-trained GloVe (Pennington, Socher, and Manning 2014) over Wikipedia 2014 and average over the 10 classes. The language embedding is then concatenated with the vi-

sual features resulting in a feature vector  $X_i$  for each frame  $i$ .

### Single-frame Selector

This stream is designed to be extremely fast. We build on the observation from Huang et al. (2018) that an oracle that looks at the predictions from an expensive network, and selects the frames with the highest confidence for the ground truth class, actually outperforms using the entire video for prediction. Thus, we use a simple multi-layer perceptron (MLP) that takes as input a feature vector  $X_i$ , and computes the confidence of the classification for the ground truth. This MLP has 2 layers, and is trained using the oracle mentioned before wherein each frame outputs the probability of that frame with respect to the ground truth class. At training time we can obtain the ground truth probability of each frame using an expensive model trained on the dataset we are looking at. The model is trained on that. At test time  $\delta_i$  is predicted by the trained model as the importance score of a particular frame.

### Global Selector

The multi-frame discriminator is designed to use information across frames for selection. This is done by first obtaining a global representation of the video using an attention model over the entire video. Given this global representation, the temporal relationships across frames are learned using a relation model and a long short-term memory (LSTM) network. While lightweight and easy to learn, this network provides information about how useful frames are when considered globally. The global selector uses a relational model to learn temporal relationships across frames over the entire video. This produces an inexpensive global representation of the video.

**Pairs of frames.** Consider an input sequence  $X = (X_1, \dots, X_N)$ ,  $X_i$  represents the concatenated visual and categorical features in frame  $i$  and  $N$  represents the total number of frames. For each frame, we concatenate a second, randomly selected frame,  $X_r^i, r \in \{1, \dots, N\}$ . The random frame is always chosen from the subsequent set of frames to capture the temporal changes that occur in actions. Some actions will be most recognizable when these pair of frames are only a few frames apart, while others will be more recognizable when they are further apart. This random choice allows the model to be flexible and capture the temporal changes in different classes. The input to the attention model is the concatenation of both vectors  $Z_i = [X_i : X_r^i]$ . The output of the network are a set of temporal relation-attention weights  $\gamma_1, \gamma_2, \dots, \gamma_N$ . This helps our model to obtain temporal information.

**Attention Module.** The coarse self-attention weights  $\alpha_i$  are first calculated using a fully connected layer and a sigmoid function (Meng et al. 2019). The mathematical representation is in Eq. 1, where  $U$  are network parameters. We now aggregate the input features using these self-attention weights. We do this in order to obtain a global representation  $Z'$  of the frame features, as in Eq. 1.

Self-attention weights are learned using individual frames with the help of non-linear mapping. To obtain a more reli-

able form of attention, we need both local and global features to be used.  $Z'$  is aggregated from all local features and hence contains the global information of the video. Hence, by using  $Z'$  we can further refine the attention weights by modeling the relationship between local frame features and  $Z'$ .

$$\alpha_i = \sigma(Z_i U) \quad \text{and} \quad Z' = \frac{\sum_{i=1}^N \alpha_i Z_i}{\sum_{i=1}^N \alpha_i} \quad (1)$$

**Relation Module.** We can add a sample concatenation and another fully connected layer (Sung et al. 2018) to estimate a relation-attention weight  $\beta$ .  $\Theta_1$  is a parameter of the fully connected layer and  $\sigma$  represents the sigmoid function. Using this we have obtained frame attention weights. However, we also want temporal attention weights. We use an LSTM to capture sequential per frame changes. The input to the LSTM at each time step is the dynamic weighted sum using the relational self-attention weights ' $\omega_t$ '. This is represented in Eq. 2.

$$\beta_i = \sigma([Z_i : Z']^T \Theta_1) \quad \text{and} \quad \omega_t = \sum_{i=1}^t \beta_i Z_i \quad (2)$$

The temporal attention weights are then calculated as shown in Eq. 3 and Eq. 4. It is dependent on the previous time step output of the LSTM and the input at that time step.  $b$  is a bias vector.

$$h_t, m_t = \text{LSTM}(\omega_t, h_{t-1}, m_{t-1}) \quad (3)$$

$$\lambda_t = \text{softmax}(Vh_t + b) \quad (4)$$

To compute the relational-temporal weights, we follow the procedure used to obtain relational-frame attention weights, as in Eq. 5. Here  $\Theta_2$  is simply a network parameter.

$$Z'' = \frac{\sum_{t=1}^N \lambda_t \omega_t}{\sum_{t=1}^N \lambda_t} \quad \text{and} \quad \gamma_t = \sigma([\omega_t : Z'']^T \Theta_2) \quad (5)$$

Using these  $\gamma_t$  we can obtain an attended content vector  $c_t$  at time 't' using Eq. 6. Here  $h_i$  refers to the hidden state of the LSTM at  $i$ . For classification,  $c_t$  is fed into an MLP to generate the predicted label  $y$ . Overall, this module aims to minimize the loss  $\mathcal{L}_{cls}$  that is described in Eq. 7, given ground truth labels  $\hat{y}_t$ . Steps to calculate all the attention weights and intermediaries can be seen in Figure 2.

$$c_t = \sum_{i=1}^t \gamma_i h_i \quad (6)$$

$$\mathcal{L}_{cls} = - \sum_{i=1}^C \hat{y}_i \log(y_i) + \varepsilon \sum_i \sum_j \Theta_{i,j}^2 \quad (7)$$

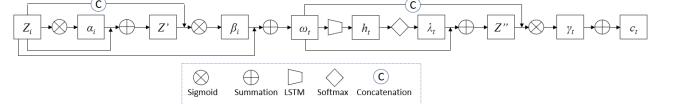


Figure 2: Steps involved in calculating the attention weights and intermediaries involved.

## Experimental Analysis

In this section we describe all experiments that we conduct to test the behavior of the proposed SMART frame selection network. In the qualitative results analysis subsection we describe the ablation experiments that led to the specific design of the network, justify each of the components and measure their impact. We analyze the behavior of the frame selector components individually (single frame and global selection). We show the generality of the proposed method on several datasets. Finally, we compare to other state-of-the-art frame sampling methods in the untrimmed setting, showing that the proposed method still produces higher accuracy.

### Experimental Setup

**Datasets.** We use 6 of the most popular benchmark datasets throughout our experimental analysis. We use the Something-something-v2 dataset (Goyal et al. 2017) for our extensive ablation study. The purpose of the ablation study is to drive the design choices through experimental evidence. We choose this for the ablation study because we think that it contains the types of actions where relations of frames over time matter more. This allows us to truly evaluate the effect of the global model that we propose. In particular, the action classes in this dataset are designed to focus on the action, (eg.: “put something”) instead of on an object (eg.: “playing guitar”). As a result, actions tend to have more temporal structure, and relations across frames may matter more.

The Something-something dataset has a total of 168,913 training videos and 24,777 validation videos with a total of 174 classes. After the ablation study, we show the generality of our approach by testing it in other datasets as well. The Kinetics(Carreira and Zisserman 2017) dataset is one of the most widely used large-scale datasets in action recognition.

We use two subsets (Sevilla-Lara et al. 2019) of Kinetics that have been identified as containing mostly temporal information and mostly static information. In our experiments we refer to these as Kinetics-Temporal and Kinetics-Static. These subsets were created using a human perceptual test, where users are asked to identify the class of a video where the frames are not in order, therefore removing temporal information. Static classes are those that users could identify without temporal information. Temporal classes are those that users were not able to identify when the frames were not in order. Each of the two splits contains 32 classes. The temporal subset consists of 26509 videos and the static subset consists of 23675. For our generality tests, we also use the well-known UCF101(Soomro, Zamir, and Shah 2012) dataset which contains 101 classes and about 13K videos.

We also extend our approach as a pre-processing step for more complex models and compare performances on

HMDB51 which contains 51 classes and 6849 video clips along with UCF101. Previous frame selection for action recognition have focused on untrimmed videos. In order to compare with them, we use ActivityNet(Caba Heilbron et al. 2015) and FCVID(Jiang et al. 2017b). ActivityNet consists of 19994 videos, and contains 200 classes. As the testing labels are not available publicly, the reported performances are on the validation set. FCVID is made up of 91, 223 videos taken from YouTube having an average duration of 167 seconds, and these are annotated into 239 classes.

**Implementation Details.** As mentioned before, the lightweight features used for frame selection are computed using MobileNet(Sandler et al. 2018) and GloVe(Pennington, Socher, and Manning 2014). After the frame selection is done, we can use a more expensive and high-quality feature representation. In our experiments, we use three different backbones: ResNet-152, ResNet-101(He et al. 2016), and Inception-v3(Szegedy et al. 2017). The backbones are pre-trained either on ImageNet(Deng et al. 2009) or Kinetics. These architectures are representative of the state-of-the-art, and are chosen according to what other methods that we want to compare to have used.

We use Pytorch for implementation. All frames are resized to 224x224. We use mini-batch stochastic gradient descent, with a momentum of 0.9. We run 200 epochs on UCF101 and the Kinetics subsets, and 100 epochs on Something-something dataset and Activitynet due to the computational requirements for these larger scale datasets. We use a batch size of 128 for UCF101 and the Kinetics subsets and a batch size of 64 for the Activitynet and something-something datasets. The initial learning rate is set at 0.0001 and reduces by 10 after every 25 epochs.

**Baselines.** We compare the performance of our frame selection model with that of random and uniform frame selection. Random frame selection picks frames uniformly at random from the entire video, while uniform frame selection picks frames that are evenly spaced. Once the frames are picked, we predict an action by average pooling the predictions of every selected frame using one of the expensive backbones. In addition to these baselines, we compare to other state-of-the-art frame sampling methods, including Adaframe(Wu et al. 2019c), FastForward(Fan et al. 2018), FrameGlimpse(Yeung et al. 2016) and MARL(Wu et al. 2019a).

### Ablation Study on the SMART Frame Selection

Here, we look at the impact of the feature representation (visual and categorical), the choice of frame selector (the global multi-frame selector and the single-frame discriminator), and the use of pairs of frames. We use the Something-something-v2 (Goyal et al. 2017) dataset for this study. Table 1 shows the results.

We first test and compare the use of the simple visual features (from MobileNet), then combining them with the categorical ones (from GloVe). We use the global selector for this initial test. We observe that the addition of semantic language features helps, supporting the intuition that using words related to the content of a frame actually helps in the context of frame selection. Using that, we examine the effect

Table 1: Ablation study to determine the effect of each of the components of the SMART frame selection network. In the table, the best configuration of each section is the setting used for the section below. We use 26 frames for all experiments. Something-something-v2 dataset. 'G' represents GLOPs, 'VF' represents standalone visual features, 'SFS' and 'GS' stand for single frame selection and global selection respectively.

Method		Inc v3		Res-152	
		Acc	G	Acc	G
Baselines	Random	44.2	152	45.8	277
	Uniform	49.6	152	50.8	277
	All frames	58.8	607	60.1	1105
Features	VF	58.3	182	60.2	308
	VF + GloVe	59.2	183	60.3	309
Selector	SFS only	59.7	155	60.7	279
	GS only	59.2	183	60.3	309
	SFS + GS	60.6	184	61.0	310
<b>SMART</b>	2-frame input	<b>60.8</b>	186	<b>61.2</b>	311

of different selectors: the single-frame selector, the global selector, and the combination of both. We observe that the combination of both is the best choice, suggesting that these two selectors behave in different but complementary ways.

We also measure the impact of using pairs of frames as input to the global selector. While we use a relational component inside the selector, adding pairs would give an additional mechanism to consider frames jointly. We observe that this does indeed help. Since we use random frames, we report the average accuracy in Table 1. The standard deviation on the Something-something-v2 (Goyal et al. 2017) dataset on 10 random runs was 0.067 using Inception v3 and 0.082 using Resnet-152.

### Analysis of the Behavior of SMART Frame Selection

**Number of Selected Frames.** First, we measure the impact of selecting different number of frames. For this, we vary the number of selected frames between 10 and 50, and measure the impact on accuracy and GFLOPs and compare with random and uniform sampling. The results are in Fig. 3(a). We choose the Something-something dataset, and the Inception-v3 as backbone. We see that as the number of frames increases, the uniform and random frame selection perform strictly. The proposed method performs much better than these baselines across frames. It is also interesting that the accuracy increases and reaches a peak, and then slowly drops in performance. This behavior confirms the intuition that there is a sweet spot in the number of frames, and that using more than that, will include frames that are harder to classify, which will pollute the prediction.

**Frame Selection Across Similar Classes.** We now plot the combined frame score from both selectors, to analyze its behavior. We plot the frame score of classes that are seman-

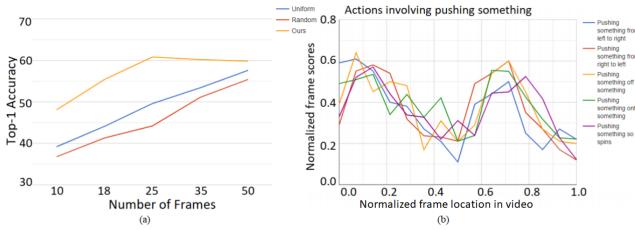


Figure 3: (a) Behavior of different sampling strategies with respect to number of frames. Orange represents SMART , blue represents uniform selection and red represents random selection (b) Comparison of the importance score of semantically similar actions. We can see a striking resemblance for all actions involving pushing.

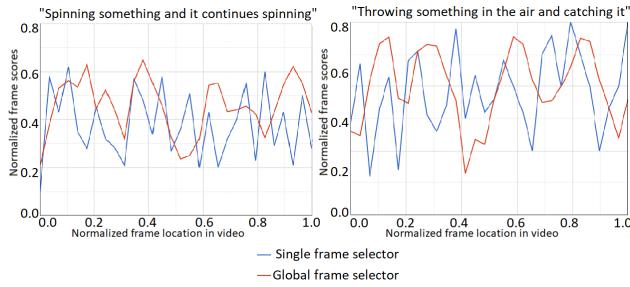


Figure 4: Graphical comparison of how the two selection modules give importance scores. We can see that each selector is giving different importance weights to different parts of the video.

tically related, to compare if frames scores are also similar. The Something-something dataset contains groups of classes that are very related. We sample 25 videos within a class, for 5 classes and average the importance scores. The plots are shown in Fig. 3. We see a strong resemblance for actions involving “pushing”, suggesting that the general structure of the action has been captured by the model.

**Selecting Frames with the Global Selector vs. the Single-frame Selector.** We measure whether the pattern of frame selection from the global selector tends to be different from the pattern from the single-frame selector. For this, we randomly sample 25 videos within a class, and score each of their frames with the two selectors. We plot the average score at each frame, in Fig. 4. Again we use the Something-something dataset and Inception-v3. While the scores from the single-frame selector change more erratically, the score from the global selector seems to be more temporally consistent. This suggests that frames scores from the global selector are actually more structured.

**Selected frames.** It is also interesting to look at the frames selected for one of the classes, in Fig. 5. Indeed, the few selected frames do tell the story of the action. The class is “pushing something from left to right”. Fig. 6 shows another example of selected frames.



Figure 5: Examples of frames not selected (top) and selected (bottom) for the class “pushing something from left to right”. Frames from (Goyal et al. 2017).



Figure 6: Examples of frames not selected (top) and selected (bottom) for the class “pulling something so that it gets stretched”. Frames from (Goyal et al. 2017).

## Quantitative results analysis

### Generality of SMART on Additional Datasets

**UCF101.** Results for UCF101 can be seen in Table 2a. As in the Something-something dataset we observe that the SMART selection outperforms the baselines of random and uniform, regardless of the number of frames. We also see that it outperforms using the full video (for all except for using 10 frames) while the “sweet spot” of number of frames is slightly larger than in the Something-something dataset. This is consistent with the fact that videos in UCF101 are about 7 seconds long, while Something-something are closer to 3 seconds. Therefore it makes sense that the proportion of “good frames” stays the same.

**Subsets of Kinetics.** We also show results on the subsampled 32 temporal classes of Kinetics and the 32 static classes (Sevilla-Lara et al. 2019). These two subsets are described in detail in the Datasets section. Results are shown in Table 2b. We see that the pattern is similar to all other experiments: SMART outperforms the other sampling baselines, and for the optimal number of frames, it outperforms the full video as well. Also the proposed method behaves slightly differently in these two subsets of Kinetics. This is consistent with the expected behavior of the proposed method, which considers the entire video globally, and is able to make selections that are more temporally aware.

### Performance on untrimmed datasets

Finally, we compare the performance of the proposed method to previous work for untrimmed video. Therefore, we test on the ActivityNet(Caba Heilbron et al. 2015) and the FCVID(Jiang et al. 2017b) datasets. We show that using fewer frames than recent approaches such as Adaframe(Wu et al. 2019c), FastForward(Fan et al. 2018),

Table 2: Baseline frame sampling techniques vs our SMART with ResNet-152 backbone; #F: #frames.

Method	Accuracy			GFLOPs			
	#F	10	26	50	10	26	50
Random	63.2	68.3	70.2	110	277	652	
Uniform	63.8	69.1	70.7	110	277	652	
<b>SMART</b>	<b>72.8</b>	<b>75.3</b>	<b>75.5</b>	164	331	706	
All frames	<b>74.6</b>	74.6	74.6	1969	1969	1969	

(a) UCF101 dataset									
Method	Temporal, Acc			Static, Acc			GFLOPs		
#F	10	26	50	10	26	50	10	26	50
Uniform	59.4	60.3	62.1	60.1	60.6	61.2	110	227	652
Random	60.1	60.8	62.7	60.3	60.9	61.7	110	227	652
<b>SMART</b>	63.1	<b>64.8</b>	<b>65.4</b>	61.4	61.9	<b>62.6</b>	185	353	728
All frames	<b>64.1</b>	64.1	64.1	<b>62.4</b>	<b>62.4</b>	62.4	2761	2761	2761

(b) Kinetics dataset subsets: Temporal and Static

FrameGlimpse(Yeung et al. 2016) we can obtain a higher accuracy. However, we access all frames which makes our approach slower than these. We also compare with LiteEval (Wu et al. 2019b) which is a lightweight action recognition model. We also compare our approach to Multi-agent Reinforcement Learning (MARL)(Wu et al. 2019a) approach and Dynamic Sampling Networks (DSN) (Zheng et al. 2020) by using a model pretrained on Kinetics for fair comparison. Table 3 shows the results.

### Extension of SMART as a pre-processing step

We look at the results of using our approach as a pre-processing step to Temporal Segment Networks (TSN) (Wang et al. 2016) and using the selected frames at inference in Table 4. To the best of our knowledge this gives us state-of-the-art results on UCF101 and HMDB51. We compare with other recent state-of-the-art approaches such as two-stream networks (Simonyan and Zisserman 2014; Gowda 2017), DynaMotion (Asghari-Esfeden, Sznaier, and Camps 2020), I3D (Carreira and Zisserman 2017) and Knowledge Integration network (KI-Net) (Zhang et al. 2020) which are among the latest state-of-the-art approaches. We also add comparison with AAS (Dong, Zhang, and Tan 2019) as a frame selection approach.

### Improving performances of other models

Here, we show that using our model to select frames and pass the selected frames at inference helps to improve the performance of models such as I3D (Carreira and Zisserman 2017), STM-ResNet (Feichtenhofer, Pinz, and Wildes 2017) and ISTPAN (Du et al. 2018). This can be seen in Table 5.

### Conclusion

We have proposed a method for frame selection in the domain of trimmed videos, that we refer to as SMART frame

Table 3: Results on ActivityNet and FCVID of the SMART frame selection. Compared to recent state-of-the-art methods, the proposed method outperforms their accuracy. #F: Number of frames, 10c corresponds to 10 clips used instead of frames

Method	Pre-trained	Backbone	ActivityNet		FCVID	
			#F	Acc	#F	Acc
FastForward	Imagenet	Inc v3	9.61	58.1	15.34	73.3
FrameGlimpse	Imagenet	VGG16	9.42	62.8	9.26	71.7
Adaframe	Imagenet	Res101	8.65	71.5	8.21	80.2
LiteEval	Imagenet	Res101	-	72.7	-	80.0
<b>SMART</b>	Imagenet	Res101	8	71.4	8	80.8
<b>SMART</b>	Imagenet	Res101	10	<b>73.1</b>	10	<b>82.1</b>
DSN	Kinetics	Res18	10c	68.0	-	-
DSN	Kinetics	Res34	10c	82.6	-	-
MARL	Kinetics	Res152	25	83.8	-	-
<b>SMART</b>	Kinetics	Res152	24	<b>84.4</b>	-	-

Table 4: Extending SMART as a pre-processing step to state-of-the-art deep learning approaches. The '+ Kinetics' indicate that the backbone is pre-trained with Kinetics.

Method	Backbone	UCF101	HMDB51
Two-stream	VGG	92.5	62.4
I3D	Inc v3	98.0	80.7
DynaMotion + I3D	Inc v3	98.4	84.2
TSN	BN-Inc	94.2	69.9
KI-Net	Res-152	97.8	78.2
AAS	TSN	94.6	71.2
<b>SMART</b>	TSN	<b>95.8</b>	<b>74.6</b>
AAS	TSN+Kinetics	96.8	77.3
<b>SMART</b>	TSN+Kinetics	<b>98.6</b>	<b>84.3</b>

Table 5: Extending SMART to other approaches

Method	UCF101	HMDB51
ISTPAN	95.5	70.7
ISTPAN + SMART	<b>96.4</b>	<b>72.1</b>
I3D	98.0	80.0
I3D + Smart	<b>98.2</b>	<b>81.1</b>
STM-Resnet	94.2	68.9
STM-Resnet + SMART	<b>94.9</b>	<b>69.7</b>

selection. The method addresses the issue of considering all frames in a video at once, instead of individually, therefore making decisions globally. The proposed method outperforms the accuracy of the baselines on 3 different action classification datasets, while it reduces the computation cost up to 4 times. Further, it outperforms recent frame selection approaches on untrimmed videos in accuracy. Also, it can be extended as a pre-processing step to obtain state-of-the-art accuracy on 2 benchmarks.

## Potential Ethical Impact

This work is about efficient and effective recognition in videos and shares benefits and concerns with other video recognition models. Being able to recognize content in videos more effectively potentially allows positive impact for users when accessing video, e.g. during search. It might also allow to more effectively remove harmful content, although we do not experiment with this kind of data in this work. However, before pursuing any such use cases it is important to analyze the models for potential algorithmic biases, either obtained during training our model or inherited from pre-trained models our approach is using.

## References

- Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; and Vijayanarasimhan, S. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.
- Asghari-Esfeden, S.; Sznajer, M.; and Camps, O. 2020. Dynamic Motion Representation for Human Action Recognition. In *The IEEE Winter Conference on Applications of Computer Vision*, 557–566.
- Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 961–970.
- Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Donahue, J.; Hendricks, L. A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; and Darrell, T. 2016. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- Dong, W.; Zhang, Z.; and Tan, T. 2019. Attention-Aware Sampling via Deep Reinforcement Learning for Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8247–8254.
- Du, Y.; Yuan, C.; Li, B.; Zhao, L.; Li, Y.; and Hu, W. 2018. Interaction-aware spatio-temporal pyramid attention networks for action classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 373–389.
- Fan, H.; Xu, Z.; Zhu, L.; Yan, C.; Ge, J.; and Yang, Y. 2018. Watching a small portion could be as good as watching all: Towards efficient video classification. In *IJCAI International Joint Conference on Artificial Intelligence*.
- Feichtenhofer, C.; Pinz, A.; and Wildes, R. P. 2017. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4768–4777.
- Girdhar, R.; and Ramanan, D. 2017. Attentional pooling for action recognition. In *Advances in Neural Information Processing Systems*, 34–45.
- Gowda, S. N. 2017. Human activity recognition using combinatorial Deep Belief Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1–6.
- Goyal, R.; Kahou, S. E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The “Something Something” Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Huang, D.-A.; Ramanathan, V.; Mahajan, D.; Torresani, L.; Paluri, M.; Li, F. F.; and Niebles, J. C. 2018. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7366–7375.
- Jiang, Y.-G.; Wu, Z.; Wang, J.; Xue, X.; and Chang, S.-F. 2017a. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* 40(2): 352–364.
- Jiang, Y.-G.; Wu, Z.; Wang, J.; Xue, X.; and Chang, S.-F. 2017b. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* 40(2): 352–364.
- Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; and Fei-Fei, L. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732.
- Korbar, B.; Tran, D.; and Torresani, L. 2019. SCSampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 6232–6242.
- Li, Y.; Liu, M.; and Rehg, J. M. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 619–635.
- Meng, D.; Peng, X.; Wang, K.; and Qiao, Y. 2019. frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, 3866–3870. IEEE.
- Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- Piergiovanni, A.; Fan, C.; and Ryoo, M. S. 2017. Learning latent subevents in activity videos using temporal attention filters. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.
- Sevilla-Lara, L.; Zha, S.; Yan, Z.; Goswami, V.; Feiszli, M.; and Torresani, L. 2019. Only Time Can Tell: Discovering Temporal Data for Temporal Modeling. *CoRR* abs/1907.08340. URL <http://arxiv.org/abs/1907.08340>.
- Sharma, S.; Kiros, R.; and Salakhutdinov, R. 2015. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119* .

- Shen, Y.; Ni, B.; Li, Z.; and Zhuang, N. 2018. Egocentric activity prediction via event modulated attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 197–212.
- Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, 568–576.
- 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*.
- Sudhakaran, S.; and Lanz, O. 2018. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. *arXiv preprint arXiv:1807.11794*.
- Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.
- Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- 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 *European conference on computer vision*, 20–36. Springer.
- Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.
- Wu, W.; He, D.; Tan, X.; Chen, S.; and Wen, S. 2019a. Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 6222–6231.
- Wu, Z.; Xiong, C.; Jiang, Y.-G.; and Davis, L. S. 2019b. LiteEval: A Coarse-to-Fine Framework for Resource Efficient Video Recognition. In *Advances in Neural Information Processing Systems*, 7778–7787.
- Wu, Z.; Xiong, C.; Ma, C.-Y.; Socher, R.; and Davis, L. S. 2019c. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1278–1287.
- Yeung, S.; Russakovsky, O.; Mori, G.; and Fei-Fei, L. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2678–2687.
- 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4694–4702.
- Zhang, P.; Xue, J.; Lan, C.; Zeng, W.; Gao, Z.; and Zheng, N. 2018. Adding attentiveness to the neurons in recurrent neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 135–151.
- Zhang, S.; Guo, S.; Wang, L.; Huang, W.; and Scott, M. R. 2020. Knowledge Integration Networks for Action Recognition. *arXiv preprint arXiv:2002.07471*.
- Zheng, Y.-D.; Liu, Z.; Lu, T.; and Wang, L. 2020. Dynamic Sampling Networks for Efficient Action Recognition in Videos. *IEEE Transactions on Image Processing* 29: 7970–7983.
- Zhu, L.; Sevilla-Lara, L.; Tran, D.; Feiszli, M.; Yang, Y.; and Wang, H. 2019. FASTER Recurrent Networks for Video Classification. *CoRR* abs/1906.04226. URL <http://arxiv.org/abs/1906.04226>.