Video BagNet: short temporal receptive fields increase robustness in long-term action recognition

Anonymous VIPrior submission

Paper ID 10

Abstract

Previous work on long-term video action recognition relies on deep 3D-convolutional models that have a large temporal receptive field (RF). We argue that these models are not always the best choice for temporal modeling in videos. A large temporal receptive field allows the model to encode the exact sub-action order of a video, which causes a performance decrease when testing videos have a different sub-action order. In this work, we investigate whether we can improve the model robustness to the sub-action order by shrinking the temporal receptive field of action recognition models. For this, we design Video BagNet, a variant of the 3D ResNet-50 model with the temporal receptive field size limited to 1, 9, 17 or 33 frames. We analyze Video BagNet on synthetic and real-world video datasets and experimentally compare models with varying temporal receptive fields. We find that short receptive fields are robust to sub-action order changes, while larger temporal receptive fields are sensitive to the sub-action order.

1. Introduction

Long-term action videos naturally have different sub-action combinations and orders. For instance, the action of 'making coffee' may contain either order of 'add sugar, add milk', or 'add milk, add sugar', or people can drink their coffee black. With such diversity in sub-action orders it is nearly impossible to sample representative data containing all possible permutations for training a long-term action recognition classifier. Thus, the training set in current long-term classification datasets like MultiTHUMOS [30] and Charades [24] may contain different sub-action orders than the test set. The specific sub-action order and duration is exploited by current video action recognition models due to their large temporal receptive field size. Thus, If the models encode the specific sub-action order at training time, it might cause misclassification of a video action when the sub-action order differs at test time.

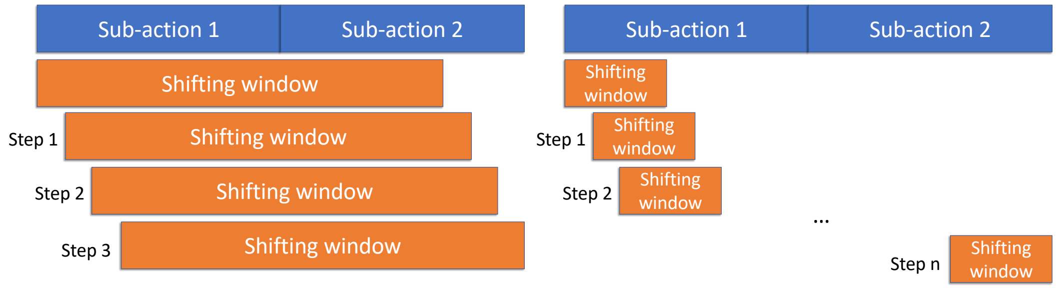
In this paper, we focus on encoding sub-action *order*. We refer to the *temporal receptive field* (RF) as the number of input frames within a shifting kernel that a network can make use of in its last convolutional layer. Usually, the last convolutional layer is followed by global temporal pooling, which collapses the temporal dimension into one unit, and a final fully connected layer. These operations do not affect the temporal RF size and the sensitivity to order, as they cannot model temporal dependencies. For this reason, we do not consider the final pooling and classification layers in our calculation of the temporal RF size. Networks with temporal RF size larger than the sub-action duration (as shown in Fig. 1 (a)) might overfit on the exact sub-action order seen at training time. In cases where the available training samples are not sufficiently representative of all possible sub-action orders, misclassifications occur at test time.

We introduce Video BagNet, a model with a small temporal RF size that is less sensitive to the exact sub-action order. Our model is inspired by BagNet [2], which reduces the spatial receptive field size for easier network interpretation. We use Video BagNet to investigate the role of the temporal RF in encoding the sub-action order. Our proposed Video BagNet is modified from 3D ResNet-50 [8]. We reduce the temporal RF size by shrinking the kernels in the temporal dimension and using less down-sampling. As shown in Fig. 1 (b), our Video BagNet with small temporal RF sizes is less sensitive to the exact sub-action order by seeing occurrences of single sub-actions rather than the combinations of ordered sub-actions. This results in better sub-action detection performance than 3D ResNet-50 on our synthetic *Directional Moving MNIST* dataset and MultiTHUMOS. We also provide a measurement of model sensitivity to the sub-action order.

2. Related Work

2.1. Temporal extent of recent models for action recognition

Recent action recognition architectures can model long temporal extents [11, 14, 19, 28, 29, 31]. This is achieved



(a) Model with large temporal RF (b) Model with small temporal RF

Figure 1. Large (a) versus small (b) temporal RF compared to the sub-action duration. The temporal RF size in the last convolutional layer is represented by the size of the convolutional shifting windows. Models with large temporal RF see sub-actions in ordered co-occurrences, while models with small temporal RF are more likely to see single sub-action occurrences. Because of this, models with small temporal RFs encode sub-action occurrences but not strict sub-action orders.

through two main approaches. The first one is by extending the temporal receptive field of convolutional models, either by stacking strided convolutional layers, thus making the model deeper [3, 27], or by harnessing auxiliary temporal modules [11, 12, 28]. The second approach is by means of transformer architectures, whose design entails a temporal receptive field which spans over the whole input duration [1, 18, 21]. Large temporal extents make it possible to learn dependencies in videos over time. This allows for modeling the order of the sub-actions that are seen at training time, which is considered useful to capture the inner structure of complex, long-term activities [13].

However, models with large temporal RF have a drawback: they are prone to overfitting on the order when the available training data is limited [7]. This is the case for most of the current long-term action recognition datasets, which only consist of a few hundred or thousand videos [16, 25, 30]. In this work, we investigate whether modeling large temporal extents is always beneficial to solve long-term action recognition. In particular, we investigate whether models with large temporal RF overfit on the order of the sub-actions seen at training time, causing misclassifications at test time.

2.2. Order invariant networks

In [12], it is empirically shown that the classification performance of order-aware methods drops significantly when new sub-action orders are presented at test time. On the other hand, order invariant methods, like ActionVLAD [6], are robust to sub-actions permutations. Hussein *et al.* [13] propose a permutation invariant convolutional module, PIC, to model temporal dynamics in long-range activities. The PIC module performs self-attention across pre-extracted visual features and can be stacked on top of convolutional backbones. PIC is robust to sub-action permutation compared to ordered-aware convolutional baselines [11], while maintaining a large temporal RF.

Our approach deviates from ActionVLAD and PIC. While ActionVLAD is completely order unaware, we maintain order information within short receptive fields. This allows modeling fine-grained motions, which is proven beneficial for action recognition [10, 23]. Differently than PIC, we investigate sensitivity to sub-action order by looking at the temporal RF size of spatio-temporal convolutional networks, commonly used as backbones in long-term action recognition models [11, 12, 28]. Our method only requires simple modification to the spatio-temporal convolutional networks.

2.3. Reducing the receptive field size: BagNet

Our idea of reducing the temporal receptive field size is inspired by Brendel *et al.* [2], who investigated how bag-of-local-features can be used for image classification. Bag-of-local-features can be obtained by restricting the spatial receptive field of the image classifier to a small number of pixels. In Brendel *et al.*'s model, the *BagNet*, this is achieved by replacing a set of 3×3 convolutions with 1×1 convolutions and removing the first downsampling layer. The property of this architecture is that the image feature representation is given by a collection of local features, corresponding to small image patches, that do not take into account the global spatial structure. Surprisingly, ignoring global structures does not hurt substantially the classification accuracy of BagNet. Using bag-of-local-features has been taken on for other visual classification tasks. Some examples are exploring local features for face anti-spoofing [22], and predicting the histogram of visual words of a discretized image as part of a self-supervision task [5]. To the best of our knowledge, our method is the first work that relies on bag-of-temporal-features models to learn video representations.

3. Method

We study how the size of the temporal RF effects model sensitivity to sub-action order. To this end, we compare

	3D ResNet-50 (RN)	*Video BagNet-1/9/17/33 (BN)*	
# parameters for 3 classes	46.2 M	45.9/46.7/45.6/46.5 M	Output sizes $T \times S^2$
conv1	$7 \times 7^2, 64$, stride $(1, 2, 2)$	$1/3/3/3 \times 7^2, 64 \times k$, stride $(1, 2, 2)$	*RN*: 64×32^2 *BN*: 64×32^2
downsampling	Max pool (3, 3, 3), stride 2	Max pool (1, 3, 3), stride (1, 2, 2)	*RN*: 32×16^2 *BN*: 62×16^2
conv2_x	$\begin{bmatrix} 1 \times 1^2, 64 \\ 3 \times 3^2, 64 \\ 1 \times 1^2, 64 \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 256 \\ 3 \times 3^2, 64 \\ 1 \times 1^2, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1^2, 64 \times k \\ 1/3/3/3 \times 3^2, 64 \times k \\ 1 \times 1^2, 64 \times k \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 256 \times k \\ 1/1/1/1 \times 3^2, 64 \times k \\ 1 \times 1^2, 64 \times k \end{bmatrix} \times 2$	*RN*: 32×16^2 *BN*: 60×16^2
conv3_x	$\begin{bmatrix} 1 \times 1^2, 256 \\ 3 \times 3^2, 128 \\ 1 \times 1^2, 128 \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 512 \\ 3 \times 3^2, 128 \\ 1 \times 1^2, 128 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1^2, 256 \times k \\ 1/3/3/3 \times 3^2, 128 \times k \\ 1 \times 1^2, 128 \times k \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 512 \times k \\ 1/1/1/1 \times 3^2, 128 \times k \\ 1 \times 1^2, 128 \times k \end{bmatrix} \times 3,$	*RN*: 16×8^2 *BN*: 29×8^2
conv4_x	$\begin{bmatrix} 1 \times 1^2, 512 \\ 3 \times 3^2, 256 \\ 1 \times 1^2, 256 \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 1024 \\ 3 \times 3^2, 256 \\ 1 \times 1^2, 256 \end{bmatrix} \times 5$	$\begin{bmatrix} 1 \times 1^2, 512 \times k \\ 1/1/3/3 \times 3^2, 256 \times k \\ 1 \times 1^2, 256 \times k \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 1024 \times k \\ 1/1/1/1 \times 3^2, 256 \times k \\ 1 \times 1^2, 256 \times k \end{bmatrix} \times 5$	*RN*: 8×4^2 *BN*: 14×4^2
conv5_x	$\begin{bmatrix} 1 \times 1^2, 1024 \\ 3 \times 3^2, 512 \\ 1 \times 1^2, 512 \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 2048 \\ 3 \times 3^2, 512 \\ 1 \times 1^2, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1^2, 1024 \times k \\ 1/1/1/3 \times 3^2, 512 \times k \\ 1 \times 1^2, 512 \times k \end{bmatrix}, \begin{bmatrix} 1 \times 1^2, 2048 \times k \\ 1/1/1/1 \times 3^2, 512 \times k \\ 1 \times 1^2, 512 \times k \end{bmatrix} \times 2$	*RN*: 4×2^2 *BN*: 6×2^2
	Average pool, n_classes-d fc, softmax		

Table 1. Network architectures: 3D ResNet-50 (*RN*) vs Video BagNet-1, 9, 17 and 33 (*BN*). In the first row, we report the number of parameters. The next rows correspond to the network layers, which contain convolutions and downsampling. For the convolutional layers, we report the kernel size $T \times S^2$, in the temporal (T) and spatial (S^2) dimensions, and the number of channels. The rightmost column of the table reports the output sizes at each layer, given an input clip of size 64×64^2. The convolutional blocks follow the structure of ResNet Bottleneck blocks [9]. We widen the channels of Video BagNet with factor k, equal to 1.40, 1.40, 1.35 and 1.25, to keep the number of parameters comparable among the different models. In both architectures, each layer is followed by Batch Norm [15] and a ReLU [17].

long-term action recognition performance of 3D convolutional networks with variable temporal RF size.

3.1. Video BagNet

Inspired by the 2D BagNet for image classification [2], we design Video BagNet, a 3D convolutional network that reasons over short temporal extents. The key idea behind Video BagNet is to harness bag-of-feature representations for video classification. Specifically, the word vocabulary is composed of short video segments. Although this representation does not allow to model long-term temporal dependencies, it prevents learning strict temporal orders that can lead to the misclassification of a video if unseen permutations between sub-actions occur at test time.

Our Video BagNet is based on the 3D ResNet-50 described in Hara *et al*. [8]. We apply a set of modifications to 3D ResNet-50 to restrict the size of its temporal receptive field, while leaving the computation in the spatial dimensions unchanged. In particular, we propose four variants of Video BagNet, with temporal RF sizes of 1, 9, 17, and 33 input frames. We choose these temporal extents following the design choice of Brendel *et al*. [2] in the image domain. Video BagNet is sensitive to order within its small temporal RF, allowing for fine-grained motion modeling.

The set of modifications that we apply to 3D ResNet-50 can be summarized as follows.

First, we restrict the size of some of the convolutional kernels in the temporal dimensions. This is done to adap-

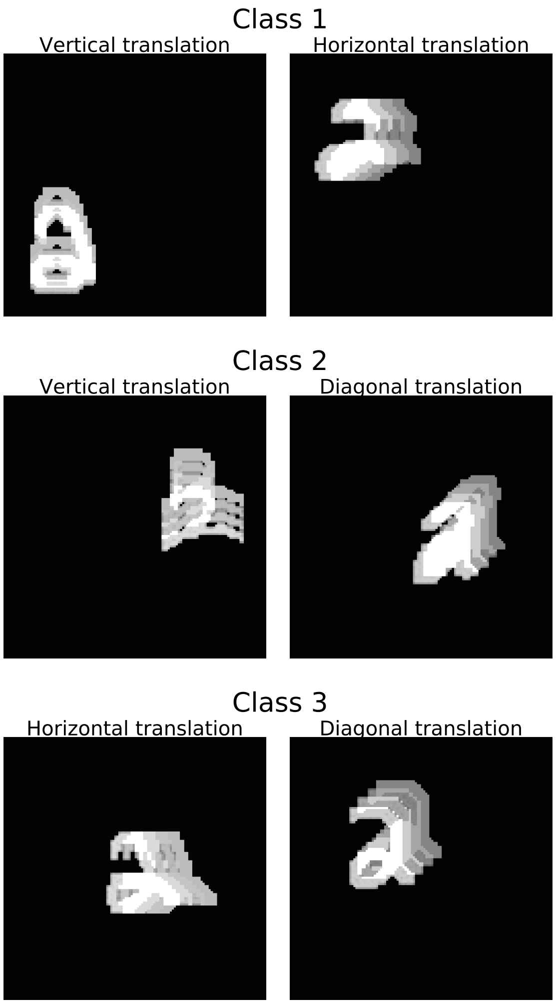


Figure 2. Example of videos of digit 2 from the *Directional Moving MNIST* dataset. The videos are composed of two sub-actions, i.e. vertical, horizontal or diagonal translation. Sub-action co-occurrences determine the video class. We explicitly superimposed multiple frames with shading to show the movement.

tively control the expansion of the RF in the temporal dimension through the convolutional layers, without changing the depth of the network. We express the size of the convolutional kernels in the temporal (T) and spatial (S^2) dimensions as $T \times S^2$. The 7×7^2 convolutional kernel in the first layer is replaced with a convolutional kernel of size 3×7^2 (1×7^2 for Video BagNet-1). In the following layers, we modify a set of 3D ResNet-50 bottleneck blocks. Bottleneck blocks consist of three consecutive convolutional layers of size $\begin{bmatrix} 1 \times 1^2, \\ 3 \times 3^2, \\ 1 \times 1^2 \end{bmatrix}$. We replace them with $\begin{bmatrix} 1 \times 1^2, \\ 1 \times 3^2, \\ 1 \times 1^2 \end{bmatrix}$.

In addition, to prevent the temporal RF size from growing in the first layer, we alter the MaxPool operator that follows layer *conv1* to perform pooling only in the spatial dimensions. To maintain a comparable amount of parameters between 3D ResNet-50 and the different Video BagNet models, we widen the number of channels. Finally, to keep the input size equal to the video length, we remove the padding. An overview of the architecture design of Video BagNet and the differences from 3D ResNet-50 is provided in Table 1.

4. Experiments

4.1. Datasets

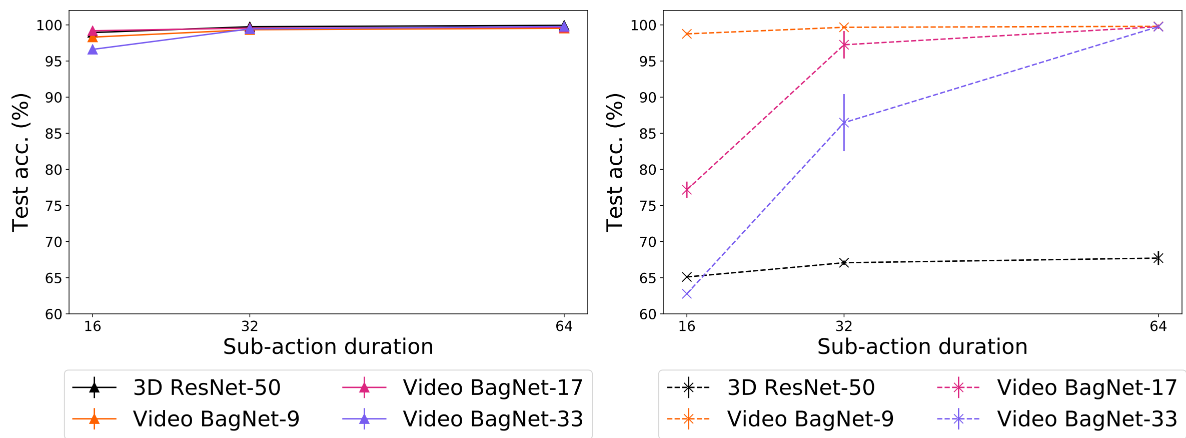
We study the effect of the temporal RF size on two long-term datasets, namely the *Directional Moving MNIST*, that we propose, and MultiTHUMOS [30]. These datasets contain multiple sub-actions and can last up to several minutes. For these datasets, the classification task consists of recognizing the sub-actions that compose the videos.

Directional Moving MNIST is a dataset composed of videos of one single moving digit, randomly sampled from the original MNIST dataset [4]. It contains 3 classes and 1000 videos per class. In this dataset, the digit translations correspond to sub-actions and the co-occurrence of two sub-actions determines the video class. More specifically, vertical and horizontal translation form class 1, vertical and diagonal translation form class 2 and horizontal and diagonal translation form class 3.

Within each class, digit appearance and starting position have been randomized. In addition, the translations occur at two possible speeds. All sub-actions have equal duration and there are no pauses between consecutive sub-actions.

One fixed sub-action order appears in the training set. At test time we use two sets: in the *test set without permutations*, the sub-action order is the same as training time; while in the *test set with permutations* the sub-action order is permuted with 50% probability. An example of the *Directional Moving MNIST* dataset is provided in Fig. 2.

MultiTHUMOS [30] is a multi-label video dataset for long-term action recognition. It is a collection of 400 complex, unconstrained, sports videos that have been densely annotated with sub-action time steps. The dataset contains a total of 65 possible sub-actions and each video contains, on average, 84.03 ± 113.56 sub-actions. The small size of the dataset prevents from training classification models using all the possible sub-action combinations and orders that usually occur in sports videos. For example, the dataset contains 20 basketball videos of which 15 videos contain the sub-actions *BasketballDribble*, *Run*, *BasketballPass*. Only 4 videos contain the order *BasketballDribble - Run - BasketballPass*.



(a) Test set with no permutations (b) Test set with permutations

Figure 3. Sensitivity to sub-action order on the *Directional Moving MNIST* dataset. Models with different temporal RF are tested on two test sets with the same order (a) and different order (b) w.r.t. training time. The models with small temporal RF compared to the sub-action duration, namely Video BagNet 9, 17 and 33, perform well on the two sets. Differently, 3D ResNet, with temporal RF larger than 100 frames, overfits the temporal order at training time and fails to classify the test set with permutations.

4.2. The size of the temporal RF affects model sensitivity to sub-action order

We design a simple controlled experiment to investigate whether spatio-temporal models encode the sub-action order through their temporal RF. For this, we deploy the *Directional Moving MNIST* dataset. We vary the size of sub-actions to relate it with different temporal RF sizes. Specifically, we use sub-action duration of 16, 32 or 64 frames and temporal RF size equal to 217 frames for 3D ResNet-50 and 9, 17 and 33 frames for our Video BagNet.

The results of this experiment are summarized in Fig. 3. Irrespectively of the temporal RF size and the sub-action duration, all the models perform well when the order of sub-actions of the training and test sets match, that is in the *test set without permutations*. However, on the *test set with permutations*, the models with large temporal RF size compared to the sub-action duration, e.g. 3D ResNet-50, and, in some instances, Video BagNet-17 and Video BagNet-33, perform poorly. In particular, 3D ResNet-50 always achieves an accuracy of $\approx 66\%$, which is equivalent to classifying correctly the videos with no permutations ($\approx 50\%$ of the *test set with permutations*) and randomly the videos with sub-action permutations. Our Video BagNet-9, which has the shortest temporal RF among the analyzed models, performs above 98.5% on all the different test videos.

These results show that sensitivity to sub-action order depends on the sub-action duration and temporal RF size. We quantify the sensitivity to order by relating the sub-action size to the temporal RF size. For this, we analyze the convolutional shifting windows in the last convolutional layer of the 3D ResNet-50 and Video BagNet models, represented in Fig. 1. In particular, we measure the sensitivity by a ratio of the amount of shifting windows that contain single sub-actions (*# single sub-action windows*) over the total amount of convolutional windows (*# total windows*). When the ratio is high, the sensitivity to the sub-action order is low.

As shown in Fig. 1, models with very large temporal RF size, like 3D ResNet-50, always see sub-action co-occurrences rather than single sub-actions. Therefore, in Fig. 4, their ratio *# single sub-action windows / # total windows* is always low, which leads to low performance on the test sets with permutations. On the other hand, models with small temporal RF size, e.g. Video BagNet-9, have a large ratio of *# single sub-action windows / # total windows* and low sensitivity to the sub-action order, achieving good performance on the test set with permutations.

4.3. Small vs. large temporal RF for long-term video action recognition

In our controlled experiment, we show that models with large temporal RF encode the sub-action order at training time. We argue that this causes misclassification when the distributions of sub-actions order are different in the training and test sets. This is the case for the commonly used MultiTHUMOS dataset, which only consists of 400 videos with high variability in sub-actions composition and order.

We evaluate the effect of the temporal RF size on MultiTHUMOS. Again, we deploy 3D ResNet-50 and Video BagNet with temporal RF 1, 9, 17 and 33. We train the models from scratch, without using either pre-training or data augmentation. We train with 512 input frames, with batch size 4. We do this to limit the computational effort of our experiments. Since we train the models from scratch and without data augmentation, our results are not compara-

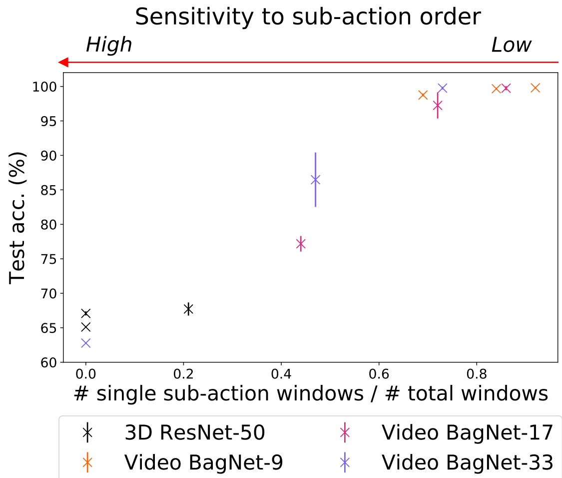


Figure 4. Accuracy on the *Directional Moving MNIST* test set with permutations in terms of models sensitivity to sub-action order. Sensitivity to sub-action order depends on the sub-action duration and temporal RF size, as shown in Fig. 1. It can be expressed by counting the amount of convolutional shifting windows that contain single sub-actions (*# single sub-action windows*) over the total convolutional windows (*# total windows*). Models with large ratio *# single sub-action windows / # total windows*, like Video BagNet-9, are less sensitive to order and achieve good performance. Models with very large temporal RF sizes, like 3D ResNet-50, always see sub-action co-occurrences rather than single sub-actions. Therefore, their ratio *# single sub-action windows / # total windows* is low and their order sensitivity is high, thus performing poorly on the test set with permutations.

Model	Temporal RF	mAP
Single-frame CNN [26]	1	25.4
MultiLSTM [30]	15	29.7
3D ResNet-50 [8]	>100	22.45
Video BagNet-33	33	26.37
Video BagNet-17	17	28.97
Video BagNet-9	9	30.21
Video BagNet-1	1	12.60

Table 2. Classification accuracies of models with small and large temporal RF on the MultiTHUMOS dataset. We compared our evaluated models (bottom rows) to the baselines proposed in [30] (top rows). Despite being trained from scratch, our Video BagNet models with temporal RF 9, 17 and 33 perform comparably to the ImageNet [20] pretrained baselines. Models with smaller temporal RF, e.g. Video BagNet-9, recognize sub-action occurrences and ignore temporal order, achieving the best performance. Video BagNet-1 cannot model motion by seeing just single frames, which has the lowest mean average precision.

ble to current state-of-the-art [32]. Nevertheless, employing this fixed experimental setup for all the analyzed models allows us to fairly compare different temporal RF sizes.

The results in Table 2 show that models with small temporal RF size outperform models with large temporal RF size on this dataset. The highest accuracy is obtained with Video BagNet-9. These results suggest that encoding long-term information, including sub-action order, is hurting the classification of MultiTHUMOS. This long-term information could correspond to the precise order of sub-actions or to the varying durations of different sub-actions. This is sensible: the multi-label classification problem of MultiTHUMOS consists in recognizing all the single sub-actions occurring in a video. Sub-action classification can be achieved by looking at short temporal extents that contain the sub-action. Because of the high variation in the temporal composition of sports videos, overemphasizing long-term information is not necessary or even decreases the sub-action recognition accuracy. On the other hand, for Video BagNet-1 it shows that if the model encodes neither long-term nor short-term information, the accuracy decreases. The results indicate that the short-term information captured by small temporal RF seems essential for good classification performance.

We find that our results are comparable to the baseline models proposed in [30], as illustrated in Table 2. It is worth noting that the single-frame CNN [26], which cannot model temporal information by design, has the advantage of being pre-trained on ImageNet [20], thus explaining the superior performance compared to Video BagNet-1. Similarly, the MultiLSTM model [26] uses pre-trained image features. Despite the lack of pre-training, Video BagNet-9 and 17 achieve 28.97% and 30.21% mAP, which is similar to mAP of 29.7% mAP obtained by Video MultiLSTM.

5. Conclusions

In this paper, we investigate whether spatio-temporal models for long-term action recognition encode sub-action order through their temporal RF. Our experiments reveal that when the temporal RF size is larger than the sub-action duration, the models are sensitive to the sub-action order. We provide a measure for the sensitivity to the sub-action order by a ratio of the number of convolutional windows that contain single sub-actions over the total number of convolutional windows. A higher ratio makes the models less sensitive to the sub-action order.

Sensitivity to sub-action order causes misclassification when the order of sub-actions are different during training and test time. This might occur in long-term action recognition, since it is difficult to collect training samples containing all the sub-action permutations that exist in natural videos. We show that small temporal RFs are robust to permutations of sub-actions, which is beneficial when limited sub-action orders are available at training time. Our study is conducted on 3D convolutional networks. Nevertheless, the conclusions could be generalizable to other spatio-temporal models that use the RF to encode temporal dependencies.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *CoRR*, abs/2102.05095, 2021. 2

[2] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2018. 1, 2, 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[4] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 4

[5] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Learning representations by predicting bags of visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6928–6938, 2020. 2

[6] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017. 2

[7] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017. 2

[8] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 1, 3, 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[10] De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. 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*, pages 7366–7375, 2018. 2

[11] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 1, 2

[12] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Videograph: Recognizing minutes-long human activities in videos. In *ICCV Workshop on Scene Graph Representation and Learning*, 2019. 2

[13] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Pic: Permutation invariant convolution for recognizing long-range activities. *arXiv preprint arXiv:2003.08275*, 2020. 2

[14] Noureldien Hussein, Mihir Jain, and Babak Ehteshami Bejnordi. Timegate: Conditional gating of segments in long-range activities. *arXiv preprint arXiv:2004.01808*, 2020. 1

[15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 3

[16] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014. 2

[17] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 3

[18] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers, 2021. 2

[19] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12056–12065, 2019. 1

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[21] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor. An image is worth 16x16 words, what is a video worth? *CoRR*, abs/2103.13915, 2021. 2

[22] Tao Shen, Yuyu Huang, and Zhijun Tong. Facebagnet: Bag-of-local-features model for multi-modal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[23] Gunnar A Sigurdsson, Olga Russakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In *Proceedings of the IEEE international conference on computer vision*, pages 2137–2146, 2017. 2

[24] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 1

[25] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with

3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1, 2

[29] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. *CoRR*, abs/1812.05038, 2018. 1

[30] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 2017. 1, 2, 4, 6

[31] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1

[32] Jiaming Zhou, Kun-Yu Lin, Haoxin Li, and Wei-Shi Zheng. Graph-based high-order relation modeling for long-term action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8993, 2021. 6