Gate-Shift Networks for Video Action Recognition Supplementary Document

Swathikiran Sudhakaran[1], Sergio Escalera[2,3], Oswald Lanz[1]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]Computer Vision Center, Barcelona, Spain
[3]Universitat de Barcelona, Barcelona, Spain
{sudhakaran,lanz}@fbk.eu, sergio@maia.ub.es

This supplementary document includes details about the backbone CNN architecture, additional t-SNE plots, results on Something Something-V1 dataset using ensemble of models, and visualization samples using saliency tubes. We also provide a supplementary video with the visualization samples. Code and models are available at https://github.com/swathikirans/GSM.

1. Architecture Details

We provide the details of the CNN architectures used in our GSM models.

1.1. BN-Inception

Tab. 1 shows the architecture of GSM BN-Inception. The Inception modules used are shown in Fig. 4 of the paper. The table also lists the output size after each layer.

Type	Kernel size/ stride	Output size
Conv	$7 \times 7/2$	$112 \times 112 \times 64$
Max Pool	$3 \times 3/2$	$56 \times 56 \times 64$
Conv	$1 \times 1/1$	$56 \times 56 \times 64$
Conv	$3 \times 3/1$	$56 \times 56 \times 192$
Max Pool	$3 \times 3/2$	$28 \times 28 \times 192$
Inception-GSM 1 (Inc3a)		$28 \times 28 \times 256$
Inception-GSM 1 (Inc3b)		$28 \times 28 \times 320$
Inception-GSM 2 (Inc3c)		$14 \times 14 \times 576$
Inception-GSM 1 (Inc4a)		$14 \times 14 \times 576$
Inception-GSM 1 (Inc4b)		$14 \times 14 \times 576$
Inception-GSM 1 (Inc4c)		$14 \times 14 \times 608$
Inception-GSM 1 (Inc4d)		$14 \times 14 \times 608$
Inception-GSM 2 (Inc4e)		$7 \times 7 \times 1056$
Inception-GSM 1 (Inc5a)		$7 \times 7 \times 1024$
Inception-GSM 1 (Inc5b)		$7 \times 7 \times 1024$
Avg Pool	$7 \times 7/1$	$1 \times 1 \times 1024$
Linear		$1 \times 1 \times C$

Table 1: Gate-Shift BN-Inception Architecture. All convolution layers are followed by Batch Normalization (BN) layer and ReLU non-linearity. C is the number of classes in the dataset.

1.2. InceptionV3

The architecture of GSM InceptionV3 is shown in Tab. 2 along with the size of the outputs after each layer. We apply an input of size 229×229 instead of the standard size of 299×299. This reduces the computational complexity without affecting the performance of the model. The Inception blocks with GSM used in the model are shown in Fig. 1.

Type	Kernel size/ stride	Output size
Conv	$3 \times 3/2$	$114 \times 114 \times 32$
Conv	$3 \times 3/1$	$112 \times 112 \times 32$
Conv	$3 \times 3/1$	$112 \times 112 \times 64$
Max Pool	$3 \times 3/2$	$56 \times 56 \times 64$
Conv	$3 \times 3/1$	$56 \times 56 \times 80$
Conv	$3 \times 3/1$	$54 \times 54 \times 192$
Max Pool	$3 \times 3/2$	$27 \times 27 \times 192$
Inception-GSM (Fig. 1a)		$27 \times 27 \times 256$
Inception-GSM (Fig. 1a)		$27 \times 27 \times 288$
Inception-GSM (Fig. 1c)		$27 \times 27 \times 288$
Inception-GSM (Fig. 1b)		$13 \times 13 \times 768$
Inception-GSM (Fig. 1b)		$13 \times 13 \times 768$
Inception-GSM (Fig. 1b)		$13 \times 13 \times 768$
Inception-GSM (Fig. 1b)		$13 \times 13 \times 768$
Inception-GSM (Fig. 1b)		$13 \times 13 \times 768$
Inception-GSM (Fig. 1d)		$6 \times 6 \times 1280$
Inception-GSM (Fig. 1e)		$6 \times 6 \times 2048$
Inception-GSM (Fig. 1e)		$6 \times 6 \times 2048$
Avg Pool	$6 \times 6/1$	$1 \times 1 \times 2048$
Linear		$1 \times 1 \times C$

Table 2: Gate-Shift InceptionV3 Architecture. All convolution layers are followed by BN layer and ReLU non-linearity. C is the number of classes in the dataset.

2. t-SNE

We first visualize the t-SNE plot of features for the models used in the ablation study, *i.e.*, model with no GSM (Fig. 2a), model with 1 GSM (Fig. 2b), model with 5 GSM (Fig. 2c) and model with 10 GSM (Fig. 2d). All figures plot the features of the 10 action groups presented in [2].

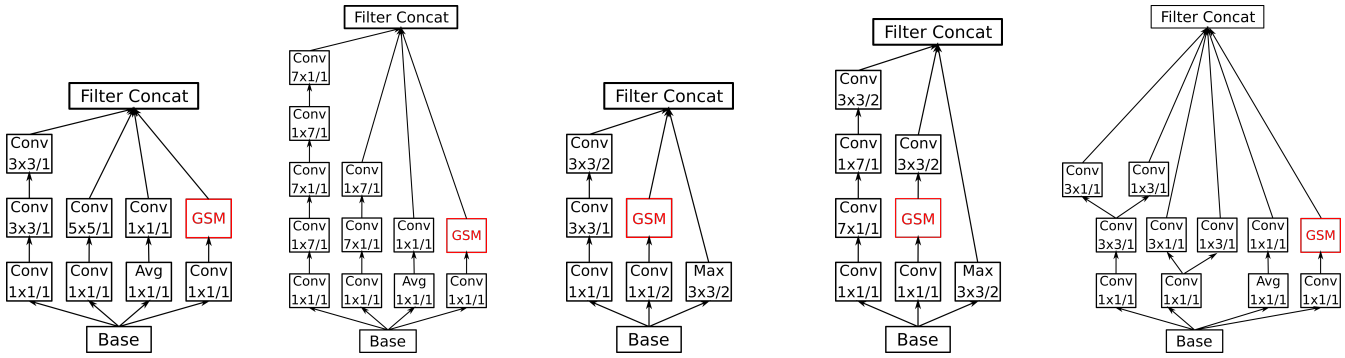


Figure 1: Inception blocks with GSM used in the InceptionV3 architecture.

From the figures, one can see that adding GSM into the CNN results in a reduction of intra-class variability and in an increase of inter-class variability. Fig. 3 shows the t-SNE plot of features from the last four Inception blocks of BN-Inception with 10 GSM. From the figure, we can see that the semantic separation increases as we move towards the top layers of the backbone.

3. Ensemble Results

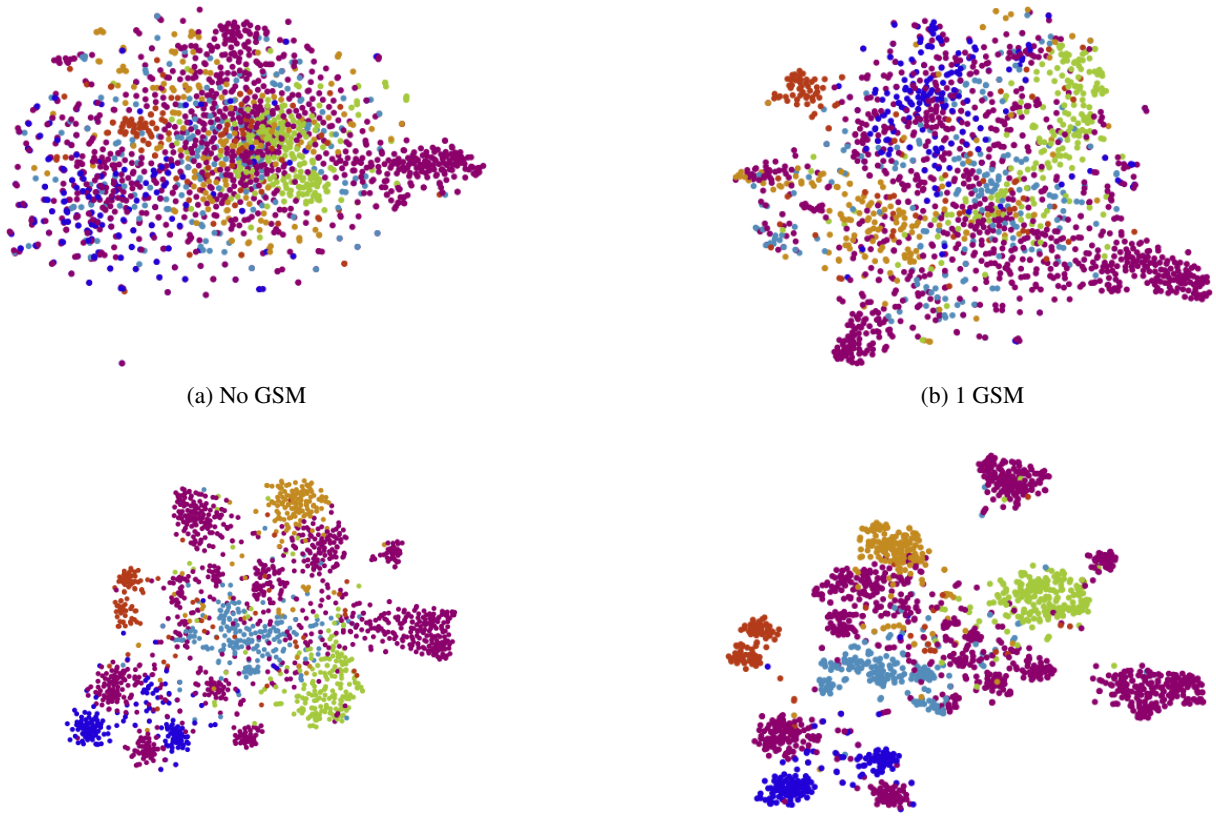
Tab. 3 lists the action recognition accuracy, number of parameters and FLOPS obtained by ensembling the models presented in this work on Something Something-V1 dataset. The first and second blocks in the table list the accuracy obtained with individual models when evaluated using 1 and 2 clips, respectively. The third block shows the recognition accuracy with different ensemble models. Ensembling is done by combining GSM InceptionV3 models that are trained with different number of input frames. We average the prediction scores obtained from individual models to compute the performance of the ensemble. From the table, it can be seen that the accuracy is increasing as more models are being added. Using models with different number of input frames enables the ensemble to encode the video with different temporal resolutions. Such an ensemble has some analogy with SlowFast [1]. With an ensemble of models trained on 8, 12, 16 and 24 frames, we achieve a state-of-the-art recognition accuracy of 55.16%. We include the parameter and complexity trade-off in Fig. 4. From the figure, we can see that the ensemble of GSM family achieves the state-of-the-art recognition performance with fewer parameters than previous state-of-the-art [3].

4. Visualization

We show 'visual explanations' for the decisions made by GSM. We use the approach of saliency tubes [4] for generating the visualizations. In this approach, the frames and their corresponding regions that are used by the model for making a decision are visualized in a form of saliency map. Figs. 5 and 6 compare the saliency tubes generated by the TSN baseline and the proposed GSM approach on sample videos from the validation set of Something Something-V1 dataset. We use the models with BNInception backbone trained using 16 frames for generating the visualizations. Each column in the figures show the 16 frames that are applied as input to the respective networks with the saliency tubes overlaid on top. We show TSN on the left side and GSM on the right side. The classes that improved the most by plugging in GSM on TSN are chosen for visualization. These classes require strong temporal reasoning for understanding the action. From the figures, we can see that TSN focuses on the objects present in the video irrespective of where and when the action takes place, while *GSM enables temporal reasoning by focusing on the active object(s) where and when an action is taking place*. For example, in Fig. 5a, an example from the class `putting something in front of something`, TSN focuses on the object that is present in the scene, the pen in the first few frames and the cup in the later frames. On the other hand, GSM makes the decision from the frames where the cup is introduced into the video. Similarly, in the example from the class `taking one of many similar things on the table` shown in Fig. 5d, TSN is focusing on the object, the matchbox, in all the frames while GSM makes the decision based on those frames where the action is taking place.

5. Video

The visualization samples for Something Something-V1 dataset discussed in previous section (Figs. 5 and 6), comparing saliency tubes generated by TSN and GSM, are displayed in the supplementary video `GSM-CVPR20.mp4`.



(a) No GSM (b) 1 GSM

(c) 5 GSM (d) 10 GSM

Figure 2: t-SNE visualization of features from networks that use (a) No GSM, (b) 1 GSM, (c) 5 GSMs and (d) 10 GSMs.

Model	#Frames	Params. (M)	FLOPs (G)	Accuracy (%)
GSM InceptionV3	8	22.21	26.85	49.01
GSM InceptionV3	12	22.21	40.26	51.58
GSM InceptionV3	16	22.21	53.7	50.63
GSM InceptionV3	24	22.21	80.55	49.63
GSM InceptionV3	8×2	22.21	53.7	50.43
GSM InceptionV3	12×2	22.21	80.55	51.98
GSM InceptionV3	16×2	22.21	107.4	51.68
GSM InceptionV3	24×2	22.21	161.1	50.35
GSM InceptionV3 En1	8+12	44.42	67.13	52.57
GSM InceptionV3 En2	8+12+16	66.63	120.83	54.04
GSM InceptionV3 En3	8+12+16+24	88.84	201.38	54.88
GSM InceptionV3 En3	8×2+12×2+16+24	88.84	268.47	55.16

Table 3: Recognition Accuracy obtained on Something Something-V1 dataset by ensembling different models.

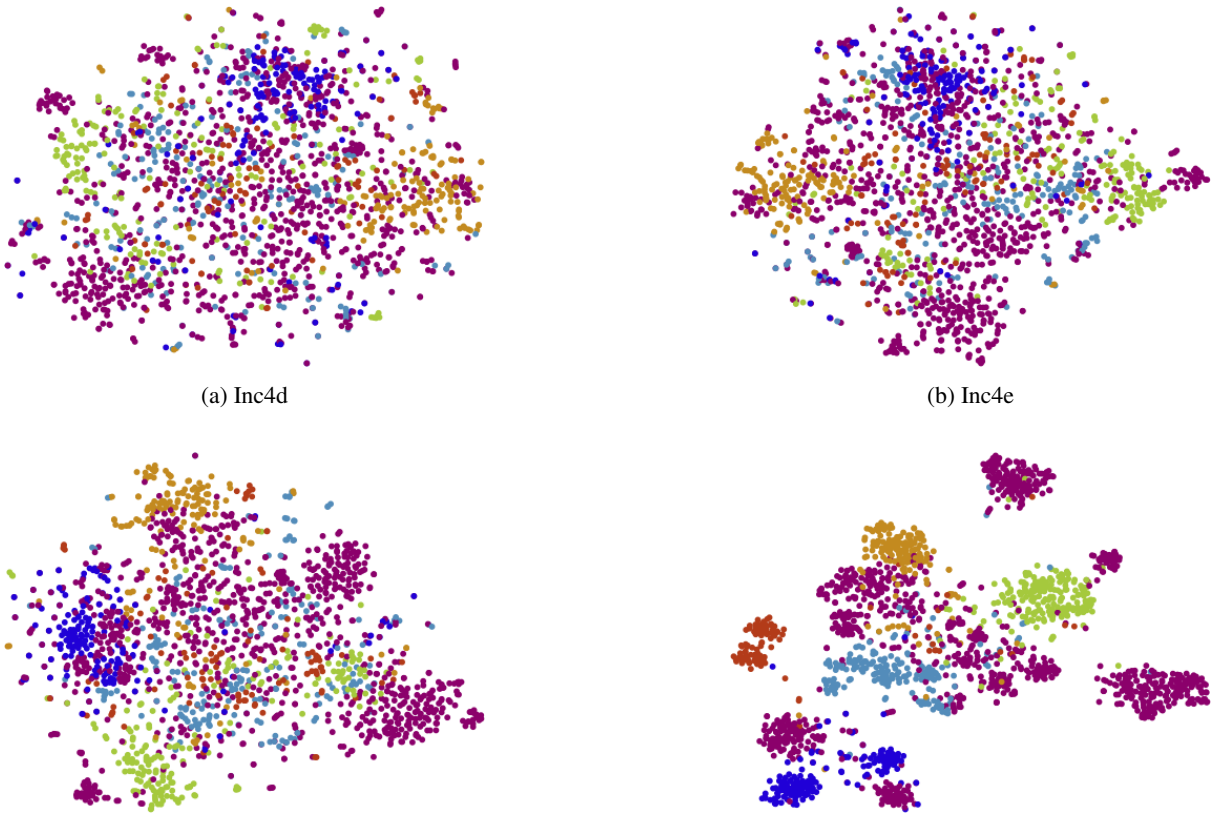
References

[1] C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast Networks for Video Recognition. In *Proc. ICCV*, 2019. 2

[2] R. Goyal, S.E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proc. ICCV*, 2017. 1

[3] B. Martinez, D. Modolo, Y. Xiong, and J. Tighe. Action recognition with spatial-temporal discriminative filter banks. In *Proc. ICCV*, 2019. 2

[4] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe. Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions. In *Proc. ICIP*, 2019. 2



(a) Inc4d (b) Inc4e

(c) Inc5a (d) Inc5b

Figure 3: t-SNE visualization of features from intermediate layer of GSM BN-Inception.

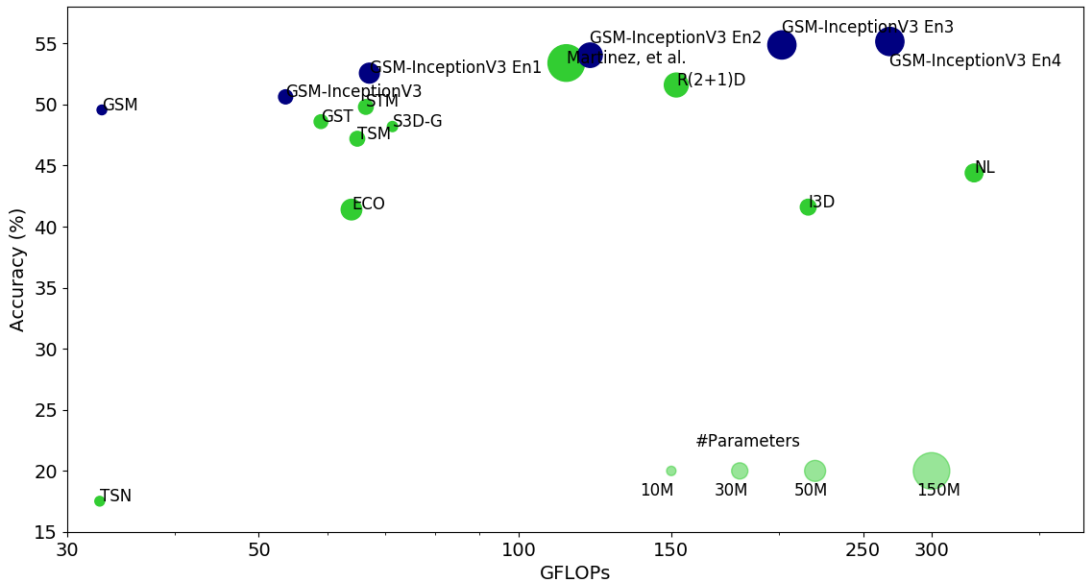


Figure 4: Accuracy-vs-complexity of state-of-the-art on Something-V1. Size indicates number of parameters (M, in millions).

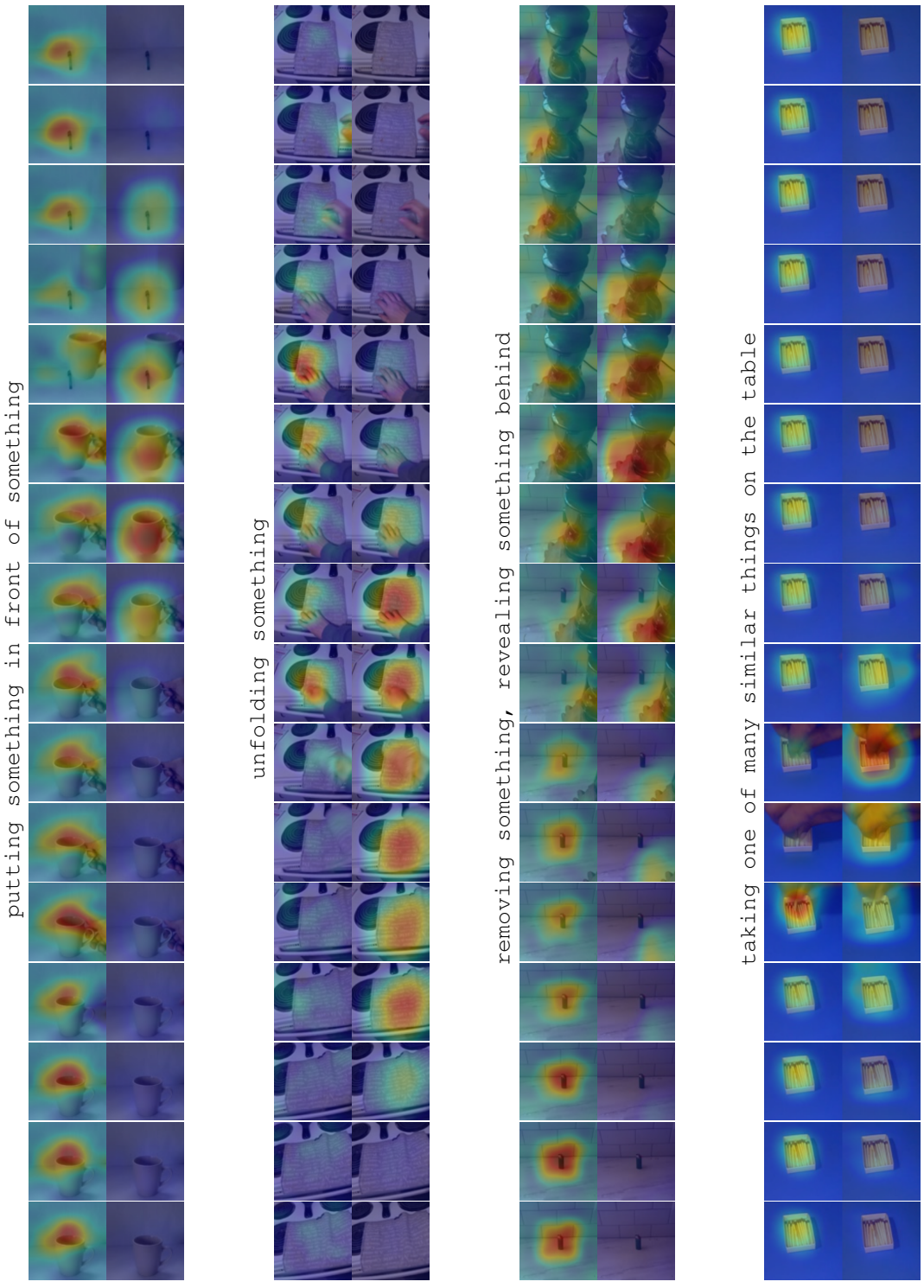


Figure 5: Saliency tubes generated by TSN (left) and GSM (right) on sample videos taken from the validation set of Something Something-V1 dataset. Action labels are shown as text on columns. Please watch video GSM-CVPR20.mp4

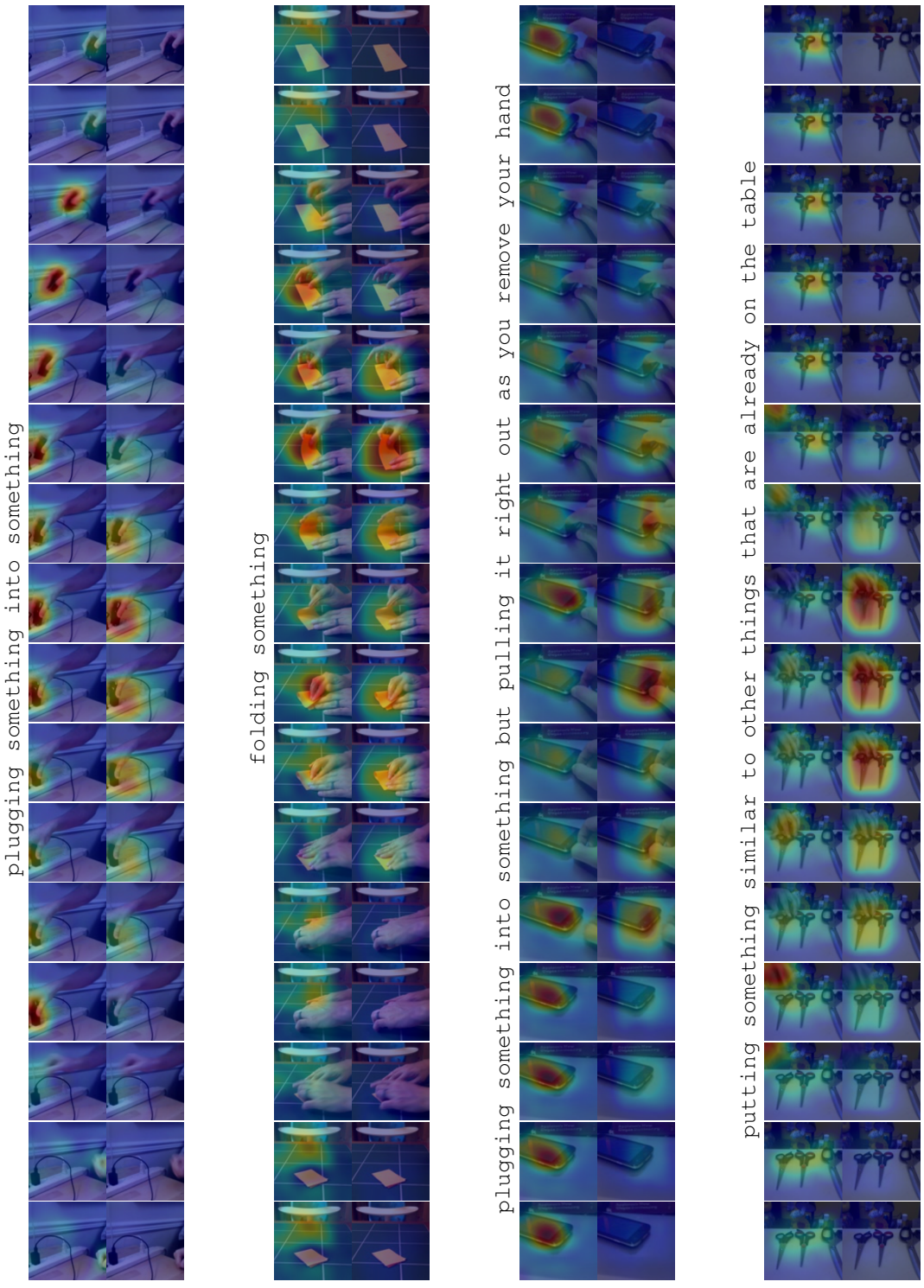


Figure 6: Saliency tubes generated by TSN (left) and GSM (right) on sample videos taken from the validation set of Something Something-V1 dataset. Action labels are shown as text on columns. Please watch video GSM-CVPR20.mp4