Self-Attention Gate-Shift Networks for Video Action Recognition

Daniele Cappuccio
Politecnico di Torino
265554

Daniele Paliotta
Politecnico di Torino
265873

Abstract

Action recognition is a renowned task that copes with the identification of actions from video clips, which can be shallowly seen as a sequence of 2D images. As a matter of fact, action recognition is considered the natural extension of image classification to multiple frames. Nonetheless, this task has not seen the incredible success that image classification had over the last few years or, as it has been put, this field still hasn't reached its "AlexNet moment". In this paper, we tackle the Something-Something V1 video action recognition challenge putting together two approaches, the one featuring a Self Attention mechanism and the one including a Gate-Shift Module, so as to better capture spatio-temporal context across frames with little computational overhead.

1. Introduction

The application of deep learning in computer vision has shown outstanding progress over the last few years, especially in well-known tasks such as object classification and semantic segmentation of images. However, similar tasks applied to video have not known the same success, and proposed models are still far away from being extremely accurate.

The main reasons can be traced back to a series of factors, including, but not limited to: the need to combine spatial, temporal and acoustic information into a coherent stream, the higher computational costs, and the lack of a standard benchmark dataset like ImageNet for images.

Hence, in mid 2017, Twenty Billion Neurons (TwentyBN) proposed a brand new large-scale dataset called "Something-Something V1"[1] to promote advancements in the video action recognition field.

A second version of this dataset has been released recently, and it features more than twice the size of samples, greatly reduced label noise, increased video resolution, and other minor improvements. Nonetheless, we are gonna stick to the first version of the dataset because it still represents a standard benchmark for our task.

2. Related work

Before the advent of deep learning, most of the algorithms proposed for action learning were based on shallow hand-crafted features, combined into a fixed-sized video level description. A classifier, like SVM or RF, was then trained to perform the final prediction.

In 2014, two breakthrough papers changed drastically the way of approaching this problem. They differed for the design choice around dealing with spatio-temporal information.

Karpathy et al.[2] proposed multiple ways to fuse temporal information using 2D pre-trained convolutions. For instance, "Early Fusion" combines information across an entire time window immediately on the pixel level. This is done by adding a temporal extent T to the filters of the first convolutional layer (e.g., $T = 30$ is approximately a second).

Simonyan and Zisserman[6], on the other hand, developed an architecture with two different networks, one for spatial context using standard RGB images, and one for temporal context in the form of stacked optical flow vectors. The two streams were trained separately and combined using SVM.

Recent works try to overcome the limitations and the complexity of these solutions. Temporal Shift Module (TSM), by Lin et al.[3], is a compromise between the low complexity of a 2D-CNN and the performance of a 3D-CNN. TSM shifts part of the channels along the temporal dimension, facilitating the exchange of information among neighboring frames.

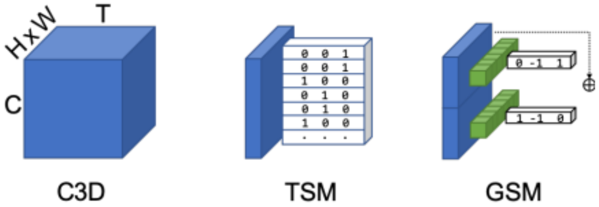

Figure 1. High-level architectural view of C3D, TSM, and GSM.

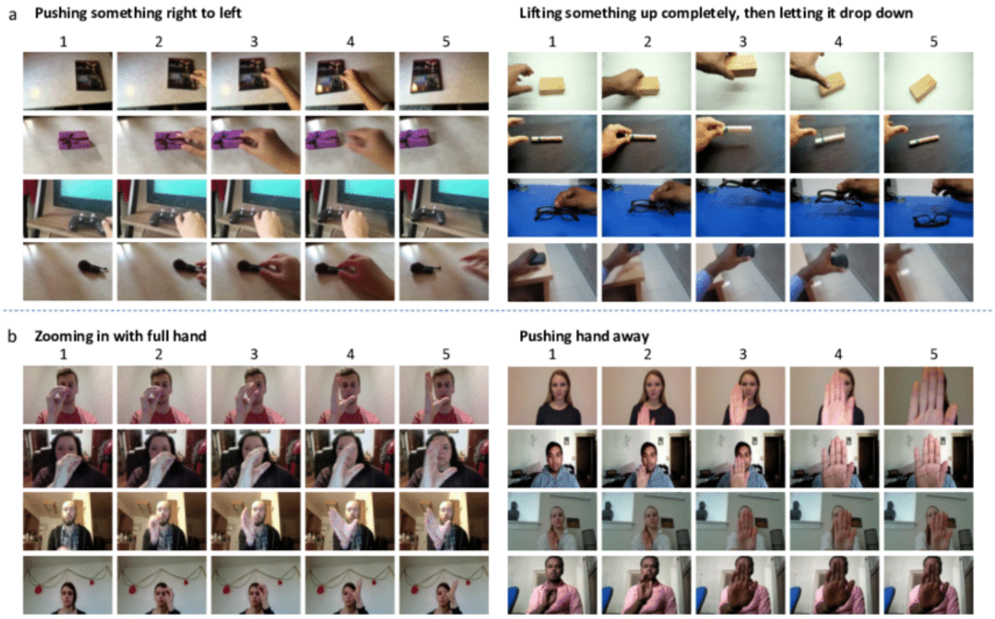
a
Pushing something right to left
1
2
3
4
5
Lifting something up completely, then letting it drop down
1
2
3
4
5
b
Zooming in with full hand
1
2
3
4
5
Pushing hand away
1
2
3
4
5

Figure 2. Something-V1 is an action recognition dataset of realistic action videos. It is one of the largest and most widely used dataset in the research community for benchmarking state-of-the-art video action recognition models.

In general, many recent proposals have been focusing on clever ways of factorizing 3D-Convolutions to overcome its huge computational cost and to avoid the curse of dimensionality, as summarized in [8].

An analoguous idea can be found in the work of Sudhakaran et al.[7], who propose a Gate-Shift Module (GSM) that enables a 2D-CNN to adaptively route features through time and combine them, at almost no additional parameters and computational overhead.

Both TSM and GSM have been tested (and achieved state-of-the-art) on Something-V1. As of Feb. 2020, GSM-based models lead the global ranking on Something-V1, with a top1 accuracy of 55.16% and a top5 accuracy of 82.49% (obtained ensembling 4 different models).

3. Something-Something V1 dataset

The 20BN Something-Something dataset is a collection of more than 100,000 densely-labeled video clips that show humans performing pre-defined basic actions with everyday objects. Videos are annotated with 174 different labels (624 videos per label), which briefly describe the performed action. Tab.1 summarizes the dataset.

Fig.4 shows that data is well distributed with respect to the labels. The top 30 most recurring labels amount for 23.8% of the total samples.

Total no. of videos	108,499
Training set	86,017
Validation set	11,522
Test set (w/o labels)	10,960
Total no. of labels	174
Average no. of videos per label	624
Original video length	Variable
No. of encoded frames per second	12
Image height	100
Image width	Variable

Table 1. Some numbers about Something-V1.

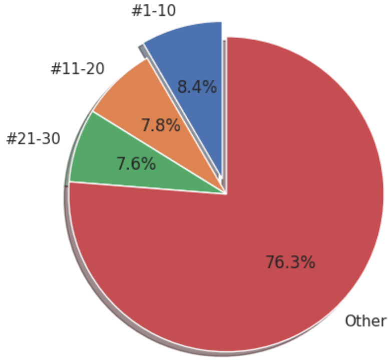

Figure 4. Label distribution in Something-V1. The ten most recurring labels account for 8.4% of the total number of videos.

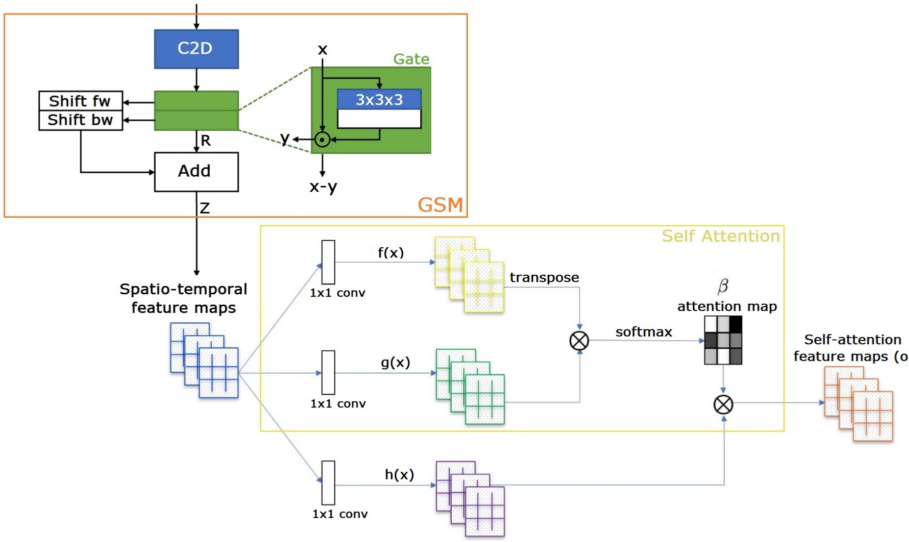

Figure 3. The building block of Self-Attention Gate-Shift Networks. The GSM module is in charge of implementing group spatial gating and forward-backward temporal shift. The spatio-temporal features are fed into the Attention module, which takes care of modelling and scoring relationships between them.

4. Self-Attention Gate-Shift Networks

In this section, we present Attention-based Gate-Shift Networks, which combine spatio-temporal feature extraction of Gate Shift Networks with the long-range dependency modelling of attention mechanisms for action recognition. We describe its building block, the Self-Attention Gate-Shift Module (**SA-GSM**), and how this can be integrated into well-known base models, such as Inceptionv3.

4.1. SA-GSM

GSM is used to turn a Convolutional 2D model (C2D) into a lightweight spatio-temporal feature extractor. It first applies spatial convolution on the input - inherited from the CNN base model. The gate (Figure 3, in green), composed of a single 3D convolution kernel with tanh calibration, separates the output of the convolution into group-gated features and residuals. The features are shifted forward and backward in time, and added to the residual to be propagated to the attention module.

If X of shape $C \times T \times W \times H$ is the input of the GSM, $X = [X_1, X_2]$ is the group=2 split of X along the channel dimension C, and $W = [W_1, W_2]$ are the two gating kernels, the output $Z = [Z_1, Z_2]$ can be computed as:

$$Y_1 = \tanh(W_1 * X_1) \odot X_1$$

$$Y_2 = \tanh(W_2 * X_2) \odot X_2$$

$$R_1 = X_1 - Y_1$$

$$R_2 = X_2 - Y_2$$

$$Z_1 = shift\_fw(Y_1) + R_1$$

$$Z_2 = shift\_bw(Y_2) + R_2$$

The Self-Attention module follows the same principles as the one found in SAGANs[10]. The features extracted from the GSM layer are first transformed into two feature spaces $f(x) = W_f(x)$, $g(x) = W_g(x)$ to compute the attention. Matrix multiplication is performed between $f(x)$ and $g(x)$:

$$s_{ij} = f(x_i)^T g(x_j) \tag{1}$$

A softmax classifier computes the values $\beta_{j,i}$ which indicate the extent to which the model attends to the i-th location when synthesizing the j-th region.

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})} \tag{2}$$

We now define two additional learned weight matrices - namely W_h and W_v - implemented as 1×1 convolutions, so that the output of the attention layer $o = (o_1, o_2, ..., o_N)$ corresponds to:

$$h(x_i) = W_h x_i$$

$$v(x_i) = W_v x_i$$

$$o_j = v(\sum_{i=1}^{N} \beta_{j,i} h(x_i)) \tag{3}$$

The output of the attention layer is scaled by a learnable parameter γ initialized to 0 and added to the input feature map. This γ trick allows the model to learn the easy task first, relying on local neighborhood, and progressively increase the complexity by broading the focus on non-local evidence.

$$y_j = \gamma o_j + x_j \tag{4}$$

With respect to a vanilla GSM, SA-GSM allows the spatio-temporal features - the inputs to the Attention - to interact with each other and find out who they should pay more attention to. The outputs are aggregates of these interactions and their respective scores. In other terms, this allows the model to selectively focus patches of the video frames in which most of the ”action” happens over time.

5. Experiments

The CNN backbone of our network is BN-Inception (Inception with added Batch Normalization), and SA-GSM is added inside each Inception block (for a total of 10). Stochastic Gradient Descent with Warm Restarts[4] with a learning rate of 0.01 and momentum 0.9 is used to train the entire network end-to-end. We use a cosine learning rate schedule as described in SGDR. The model is initialized with ImageNet pre-trained weights. In order to augment the data during training, random scaling, flipping and cropping are applied to the training set. We use a batch size of 12.

5.1. State-of-the-art comparison

Something-V1. We first evaluate SA-GSM on the Something-V1 dataset using top1 accuracy as our main metric and top5 accuracy as a side metric. Top5 accuracy refers to the ground truth label being in the 5 most probable predicted labels.

We were able to run our model only for 15 epochs, due to computational resources constraints. This is not enough to train the model appropriately and achieve state-of-the-art results. For instance, the authors of GSM trained their network on Something-V1 for 60 epochs.

Nonetheless, we collected training data using Tensorboard and notice a slight increase in accuracy (both top1 and top5, corresponding to a minor training and validation loss) w.r.t. the GSM-enabled architecture, at least for the first 5 epochs. In terms of net parameters, this is achieved by means of a 4% increase (from 10.5×10^6 to 11.0×10^6), assuming the same BN-Inception backbone.

As shown in Fig.5, our model resembles the trends of GSM on the training set. Most of the plots perfectly overlap, and we only achieve a 0.5% gain in top1 accuracy at the end of the 15th epoch.

However, the difference is more pronounced on the validation set, in which we achieve a 14.0% top1 accuracy compared to the 11.8% of GSM (see Fig.6). Also the top5 accuracy shows a non-negligible increase, from 30.2% to 37.6%. The validation has been performed at the end of each epoch for both models. It is important to note that these results still refer to the end of the 15th epoch. Both models have been trained with the same set of hyperparameters - i.e., the ones described in [7].

After training, we went on to test our models using different types of test-time augmentation. Quite surprisingly, we found that byb adding some specific flavours of transformations such as **group scaling**, **group normalization**, and **group center cropping**, we manage to reach a 32% top1 accuracy, and a 62% top5 accuracy, with an average class accuracy of 29.23%.

Diving48. Diving48 is a comparatively small fine-grained dataset of competitive diving. It consists of 18k trimmed video clips of 48 unambiguous dive sequences. We tested SA-GSM on this dataset, achieving worse results w.r.t. a vanilla GSM. This might be due to the fact that the increase in the number of parameters requires more training time in order to converge in this setting. It is also worth noting that different runs achieved visibly better results using the SA-GSM model, suggesting that model initialization might play a central role in improving performance.

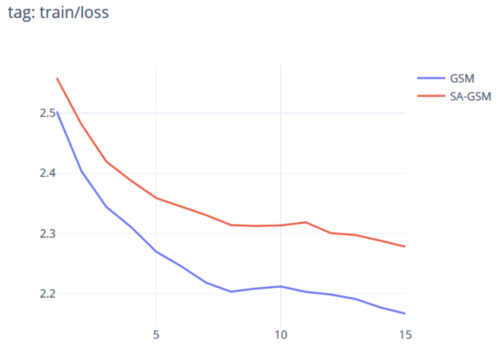

Figure 7. Train loss on Diving48.

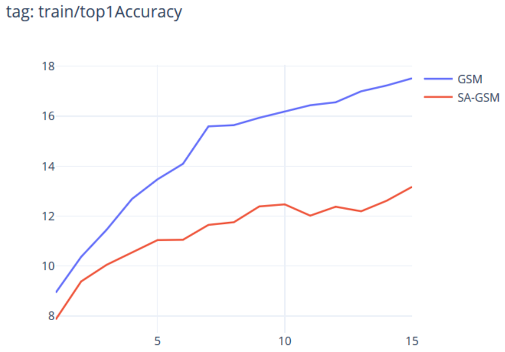

Figure 8. Train top1 accuracy on Diving48.

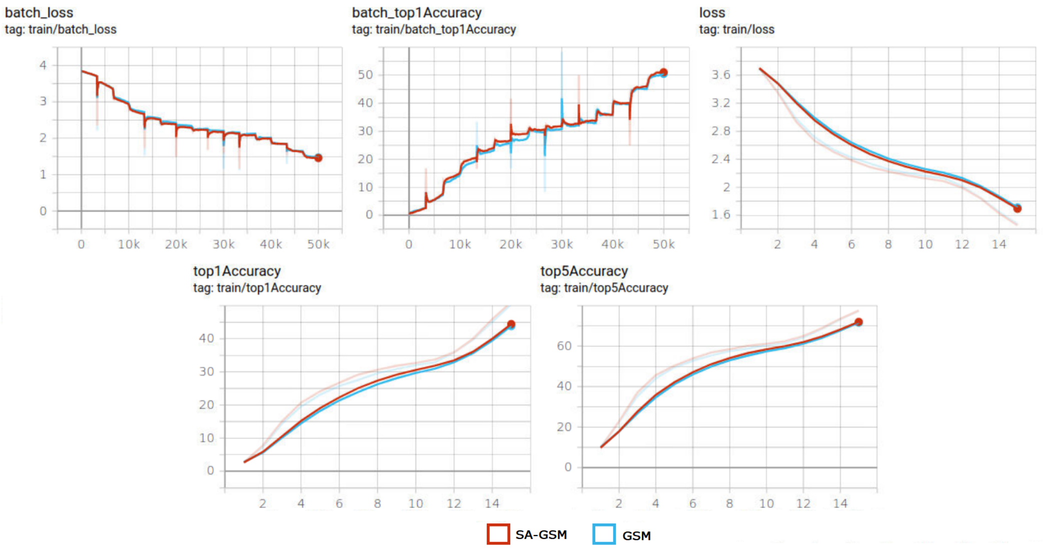

Figure 5. Train loss and accuracy (top1 and top5) on Something-V1. Here, batch loss is computed as the average loss over the batches and its value is reset at the beginning of each epoch. Same goes for batch top1 accuracy.

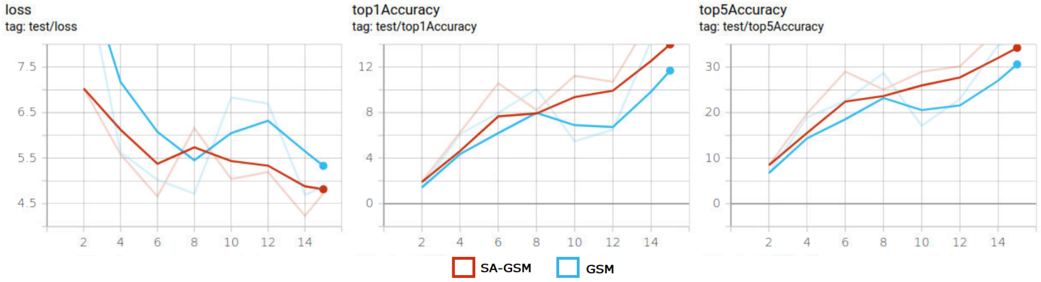

Figure 6. Validation loss and accuracy (top1 and top5) on Something-V1.

5.2. t-SNE

We use t-SNE[5] for the visualization of our high-dimensional dataset. We did not notice a distinct reduction of intra-class variability or an increase of inter-class variability when using SA-GSM. Fig.9 shows the t-SNE plots on Something-V1 (a) and Diving48 (b). The latter clearly shows a better separation among the classes. Obviously, 15 epochs are not enough for our model to get a sharp separation on a huge dataset like Something-V1.

6. Conclusions and future work

Our work introduced the Self-Attention Gate-Shift Module (SA-GSM), a sperimental block that enables CNNs to selectively focus the spatio-temporal features in which most of the "action" happens. We performed the evaluation by analyzing its effectiveness on Something-V1 and Diving48.

The main obstacle throughout our work was the lack of computational power on our side. To perform our experiments, we rented a remote machine equipped with a *Quadro P4000* GPU, but this was definitely not enough to reproduce state-of-the-art results in the given time. Still, we managed to make sensible comparison with the state-of-the-art architecture and our proposal by working with a more manageable amount of data, and by simply training for fewer epochs.

In light of our results, we feel confident to say that adding some kind of attention mechanism in action recognition models might improve the learning process, but more experimentation has to be done. More specifically, one could experiment with different kinds of attention, as well as finding the optimal position of the self-attention layers and their optimal quantity.

Moreover, this new layer might help researchers interpret and find novel insights into the inner workings and decision-making strategies of the network, by carefully inspecting the attention mechanism.

Moreover, just like in the original GSM paper, one could

Method	Backbone	Pre-training	No. of frames	Accuracy (%)
TSN[9] (ECCV '16)	BN-Inception	ImageNet	16	17.52
Multiscale TRN[11] (ECCV '18)	BN-Inception	ImageNet	8	34.44
TSM[3] (ICCV '19)	ResNet-50	Kinetics	16	47.20
GSM[7]	BN-Inception	ImageNet	16	49.56
GSM	InceptionV3	ImageNet	16	50.63
SA-GSM	BN-Inception	ImageNet	8	14.2*

Table 2. Comparison to state-of-the-art on Something-V1. Model ensembles are not included.

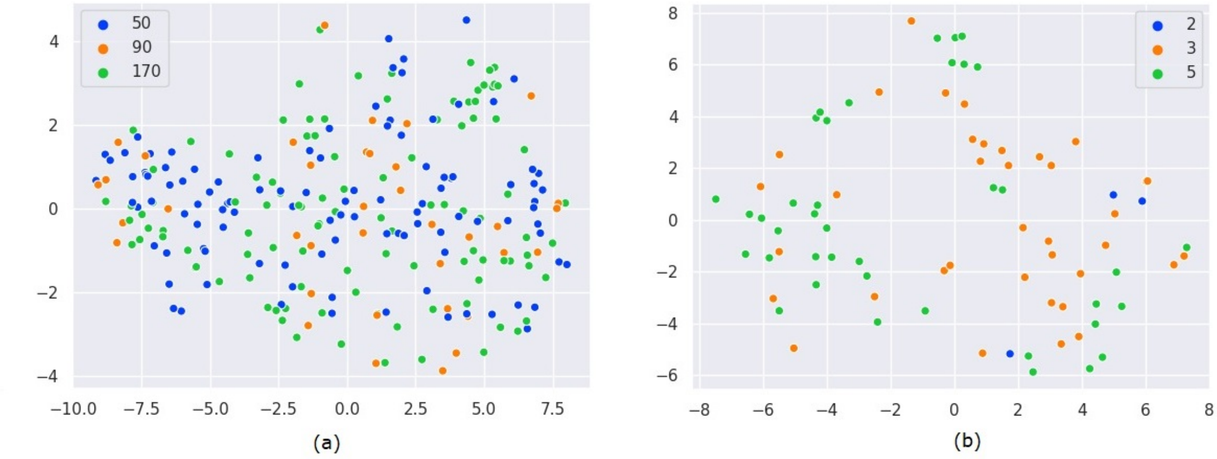

Figure 9. t-SNE visualization of features from the last convolutional layer of SA-GSN (base model: BN-Inception) after 15 epochs of training on (a) Something-V1 and (b) Diving48.

also try to use model ensembles, varying the number of frames for each video, or employing different convolutional architectures, as this usually increases accuracy of several points.

The code we used to train and evaluate our models is available at https://github.com/dpstart/SA-GSM.

References

[1] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261, 2017. 1

[2] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1

[3] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *CoRR*, abs/1811.08383, 2018. 1, 6

[4] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. 4

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 5

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014. 1

[7] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition, 2019. 2, 4, 6

[8] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition, 2017. 2

[9] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. *CoRR*, abs/1608.00859, 2016. 6

[10] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks, 2018. 3

[11] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *CoRR*, abs/1711.08496, 2017. 6