

An Evaluation of Action Recognition Models on EPIC-Kitchens

Will Price
 University of Bristol, UK
 will.price@bristol.ac.uk

Dima Damen
 University of Bristol, UK
 dima.damen@bristol.ac.uk

1. Introduction

We benchmark contemporary action recognition models (TSN [12], TRN [14], and TSM [7]) on the recently introduced EPIC-Kitchens dataset [1] and release pre-trained models on GitHub¹ for others to build upon. In contrast to popular action recognition datasets like Kinetics [5], Something-Something [2], UCF101 [10], and HMDB51 [6], EPIC-Kitchens is shot from an egocentric perspective and captures daily actions in-situ. In this report, we aim to understand how well these models can tackle the challenges present in this dataset, such as its long tail class distribution, unseen environment test set, and multiple tasks (verb, noun and, action classification). We discuss the models' shortcomings and avenues for future research.

2. Models

We benchmark 3 models: Temporal Segment Networks (TSN) [12], Temporal Relational Networks (TRN) [14], and Temporal Shift Module (TSM) based networks [7], including a variety of their variants. These models are evaluated under a uniform training and testing regime to ensure the results are directly comparable. TSN is the earliest model of the three and both TRN and TSM can be viewed as evolutionary descendants of TSN, integrating temporal modelling. In the following paragraphs, we provide an explanation of how network inputs are sampled and a brief summary of the design of each network.

Sampling Inputs to the models, *snippets*, are sampled according to the TSN sampling strategy. An action clip is split into n equally sized *segments* and a snippet is sampled at a random position within each of these. For an RGB network, the input is a single frame and for a flow network it is a stack of 5 (u, v) optical flow pairs (proposed in the two-stream CNN [9]).

TSN [12] Temporal Segment Networks propagate each snippet through a 2D CNN backbone and aggregate the class scores across segments through average or max pooling. As a consequence, TSN is unable to learn temporal cor-

relations across segments. TSN is typically trained on RGB and optical flow modalities and combined by late-fusion.

TRN [14] Temporal Relation Networks propagate snippets through a 2D CNN, like in TSN, up to the pre-classification layer. These produce features rather than class confidence scores. In order to support inter-segment temporal modelling, these segment-level features are then processed by a modified relational module [8] sensitive to item ordering. Two variants of the TRN module exists: a single scale version which computes a single n -segment relation, and a multi-scale (M-TRN) variant which computes relations over ordered sets of segment features of size 2 to n . Once the relational features have been computed, they are summed and fed to a classification layer.

TSM [7] These networks functionally operate just like TSN, snippets are sampled per segment, propagated through the backbone, and then averaged. However, unlike TSN, the backbone is modified to support reasoning across segments by shifting a proportion of the filter responses across the temporal dimension. This opens the possibility for subsequent convolutional layers to learn temporal correlations.

3. Experiments

In this section, we examine how a variety of factors impact model performance such as backbone choice, input modality, and temporal support. We analyse model performance across tasks from the perspective of the more defining characteristics of the dataset: the long-tail class distribution, and the domain gap between the seen and unseen kitchen test sets.

3.1. Experimental details

Tasks EPIC-Kitchens has three tasks within the action recognition challenge: classifying the verb, noun, and action (the verb-noun pair) of a given trimmed video. We follow the approach in [1], and replace the classification layer of each model with two output FC layers, one for verbs v and one for nouns n . The models are trained with an averaged softmax cross-entropy loss over each classification layer: $\mathcal{L} = 0.5(\mathcal{L}_n + \mathcal{L}_v)$. We obtain action predictions

¹github.com/epic-kitchens/action-models

from verb and noun predictions assuming the tasks are independent. Later, we examine the impact of integrating action priors computed from the training set for action classification. Performance on these tasks are evaluated on two test sets: seen kitchens (S1) and unseen kitchens (S2). The unseen kitchens test set contains videos from novel environments, whereas the seen kitchens split contains videos from the same environments used in training.

Training We train all models with a batch size of 64 for 80 epochs using an ImageNet pretrained model for initialisation. SGD is used for optimisation with momentum of 0.9. A weight decay of 5×10^{-4} is applied and gradients are clipped at 20. We replace the backbone’s classification layer with a dropout layer, setting $p = 0.7$. We train RGB models with an initial learning rate (LR) of 0.01 for ResNet-50 based models and 0.001 for BN-Inception models. flow models are trained with an LR of 0.001. These LRs were the maximum we could achieve whilst maintaining convergence. The LR is decayed by a factor of 10 at epochs 20 and 40.

Testing Models are evaluated using 10 crops (center and corner crops as well as their horizontal flips) for each clip. The scores from these are averaged pre-softmax to produce a single clip-level score. Fusion results are obtained by averaging the softmaxed scores obtained for each modality.

3.2. Results

Backbone choice To choose a high performing backbone, we compare BN-Inception [4, 11] to ResNet-50 [3] across the 3 models, training and testing with 8 segments. We did not test TSM with BN-Inception as the authors state that the shift module is harmful unless placed in a residual branch [7]. The top-1/5 accuracy across tasks is reported in Table 1 where the results show ResNet-50 to be superior to BN-Inception in 14/18 cases when examining top-1 action accuracy across both test sets.

Aggregate performance We now compare models with ResNet-50 backbones across tasks in Table 1 using top-1/5 accuracy. On the verb task, an intrinsically more temporal problem than classifying nouns, both M-TRN and TSM out perform TSN, especially when operating on RGB frames instead of flow. This can be explained by TSN’s inability to learn inter-segment correlations as only average or max pooling is used in aggregating class scores across segments. TSN flow models outperform their RGB counterparts; this can be attributed to the network being passed temporal information in the form of stacked optical flow frames. The 2D convolutions inside the network can learn temporal relations within the stack. Both (M-)TRN and TSM flow models outperform TSN flow showing that inter-segment reasoning is complimentary to intra-segment reasoning.

Unlike verb classification, noun classification does not rely on temporal modelling *as much* since objects can be

recognised from a single frame. TSM and TSN perform best on this task, with TRN models lagging 2–3% points behind. A possible explanation for the observed drop is that the relation module within TRN places heavy emphasis on extracting temporal relational information, which is of little relevance in recognising objects. Noun performance drops considerably across models when switching from RGB to flow as the former is a much better modality for recognising objects. Unexpectedly, TSM improves top-1 noun accuracy by 1% point over TSN. Additionally, we find all fusion models improve over the RGB models alone. We hypothesise that the temporal information here is helping disambiguate the action relevant object from those that are simply present in the environment.

Classifying actions, the joint task of classifying both verb and noun, is clearly very challenging, with the best top-1 accuracy on actions being 29.9% and 17.9% for the seen and unseen test set respectively. Even at top-5, the best results are 49.8% and 32.8%. Despite flow’s superior results on verb classification on the unseen test set, the inferior noun performance drags flow models below RGB models on both test sets.

An enduring approach, pioneered by the 2SCNN [9], has been to ensemble networks trained on different modalities through late fusion at test-time. Averaged across all model variants, fusing both modalities results in a 2.9%, 5.8%, and 9.7% relative improvement over the best performing single modality model for verb, noun, and action classification respectively. The best model on the seen test set is TSM fusion, followed by M-TRN fusion. On the unseen test set, the trend is reversed with M-TRN out-performing TSM.

Novel environment robustness It is interesting to examine the relative drop in model performance from the seen to unseen test set to determine the models’ ability to generalise to new environments. Table 1 shows that flow models are more robust to the domain gap between the seen kitchens and unseen kitchens test sets only suffering an average 22% relative drop in top-1 action accuracy compared to a 44% drop for RGB models, and 39% for fused models. The domain gap on fused models suggests that the RGB model’s predictions dominates those of the flow model. We find that flow models consistently outperform RGB models for verb classification on the unseen test set. We hypothesis this is due to the absence of appearance information in optical flow, forcing flow models to focus on motion. Motion is more environment-invariant and salient to the classification of verbs than the visual cues the RGB models will use.

Class performance analysis To further understand the differences between models, we look at confusion amongst the top-20 most frequent classes in training in Fig. 1.

The verb classification results show the top-3 verbs (accounting for 53% of the actions in training) dominate predictions due to the dataset imbalance, with this effect being

BB	Model	Modality	Verb				Noun				Action			
			Top-1		Top-5		Top-1		Top-5		Top-1		Top-5	
			S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
BN-Inception	TSN	RGB	47.97	36.46	87.03	74.36	38.85	22.64	65.54	46.94	22.39	11.30	44.75	26.32
		Flow	51.68	47.35	84.63	76.95	26.82	21.20	50.64	42.47	16.76	13.49	33.75	27.52
		Fusion	54.70	46.06	87.24	76.65	40.11	24.27	65.81	49.27	25.43	14.78	45.69	29.81
	TRN	RGB	58.26	47.29	87.14	76.54	36.32	22.91	63.30	44.73	25.46	15.06	45.66	28.99
		Flow	55.20	50.32	84.04	77.67	23.95	19.02	47.02	40.25	16.03	12.77	32.92	27.62
		Fusion	61.04	51.83	87.46	79.11	37.90	24.75	63.69	47.35	26.54	16.59	46.37	31.14
	M-TRN	RGB	57.66	45.41	86.91	76.34	37.94	23.90	63.78	46.33	26.62	15.57	46.39	29.57
		Flow	55.92	51.38	84.44	77.74	24.88	20.69	48.37	40.83	16.78	14.00	34.09	28.75
		Fusion	61.12	51.62	87.71	78.42	39.28	26.02	64.36	48.99	27.86	17.34	47.56	32.57
ResNet-50	TSN	RGB	49.71	36.70	87.19	73.64	39.85	23.11	65.93	44.73	23.97	12.77	46.14	26.08
		Flow	53.14	47.56	84.88	76.89	27.76	20.28	51.29	42.23	18.03	13.11	35.18	27.83
		Fusion	55.50	45.75	87.85	77.40	41.28	25.13	66.53	48.11	26.89	15.40	47.35	30.01
	TRN	RGB	58.82	47.32	86.60	76.92	37.27	23.69	62.96	46.02	26.62	15.71	46.09	30.01
		Flow	55.16	50.39	83.87	77.71	23.19	18.50	47.33	40.70	15.77	12.02	33.08	27.42
		Fusion	61.60	52.27	87.20	79.55	38.41	25.74	63.37	47.87	27.58	17.79	46.44	32.20
	M-TRN	RGB	60.16	46.94	87.18	75.21	38.36	24.41	64.67	46.71	28.23	16.32	47.89	29.74
		Flow	56.79	50.36	84.91	77.67	25.00	20.28	48.70	41.45	17.24	13.42	34.80	29.02
		Fusion	62.68	52.03	87.96	78.90	39.82	25.88	64.94	49.03	29.41	17.86	48.91	32.54
	TSM	RGB	57.88	43.50	87.14	73.85	40.84	23.32	66.10	46.02	28.22	14.99	49.12	28.06
		Flow	58.08	52.68	85.88	79.11	27.49	20.83	50.27	43.70	19.14	14.27	36.90	29.60
		Fusion	62.37	51.96	88.55	79.21	41.88	25.61	66.43	49.47	29.90	17.38	49.81	32.67

Table 1: Backbone (BB) comparison using 8 segments in both training and testing evaluating top-1/5 accuracy across tasks. S1 denotes the seen test set, and S2 the unseen test set. Cells are coloured on a per column basis: low high

especially pronounced in the unseen test set. Classes outside the top-20 are rarely correctly classified and instead are classified into one of the majority classes. The fine-grained nature of the verbs seems to pose challenges, particularly in the unseen test set, with similar classes being confused, such as ‘move’ with ‘put’/‘take’, ‘turn’ with ‘mix’, and ‘insert’ with ‘put’. TSN shows increased confusion between classes that differ primarily in their temporal aspects (e.g. ‘put’ vs ‘take’), compared to TSM and TRN. The generic class ‘move’ is hardest to classify, for all models.

For noun classification, the confusion matrices show the models don’t struggle as much to classify less frequent classes compared to verb classification. This is likely as a result of the models benefiting from pretraining on the large-scale ImageNet dataset. However, when fine-tuned, some overfitting to seen environments is observed, as the unseen test set matrices demonstrate that the models generalise less well to new objects. Like the verb results, the fine-grained classes pose a challenge with confusion between similar objects like ‘fork’ with ‘spoon’, and ‘bowl’ with ‘plate’ occurring. Another interesting contrast between the verb and noun tasks is that the top-20 verbs almost never get misclassified into any of the classes outside the top-20, whereas for nouns, there are more misclassifications of top-20 nouns into the long-tail.

For action classification, the models perform well on fre-

quent actions, but suffer more misclassifications into the long tail than nouns (as evidenced by the confusion into ‘other’ classes). Confusion within the top-20 actions highlight an issue not visible from the verb and noun matrices: semantically identical classes like ‘turn-on tap’ are confused with ‘open tap’. Whilst these are different classes in the dataset, they refer to the same action. This highlights an issue with the open vocabulary annotation process employed by the dataset: annotators may use different phrases for describing the same action.

We provide qualitative examples in Fig. 2 where TSM and M-TRN correctly classify the actions, but TSN fails. In the top example TSN confuses ‘put’ and ‘take’ as a result of averaging the scores across segments, and thus discarding temporal ordering. M-TRN and TSM show a much larger disparity between the scores of these classes indicating they have better learnt the difference. In the bottom example, TSN again struggles to correctly classify the action. The temporal bounds are quite wide and capture frames just after someone has picked up a bowl, they then open the cupboard and are about to place the bowl. M-TRN and TSM, through their ability to draw correlations across segments, are able to disambiguate the correct class from the action which came before and comes after.

Temporal support How many frames/optical flow snippets does the network need to see before performance sat-

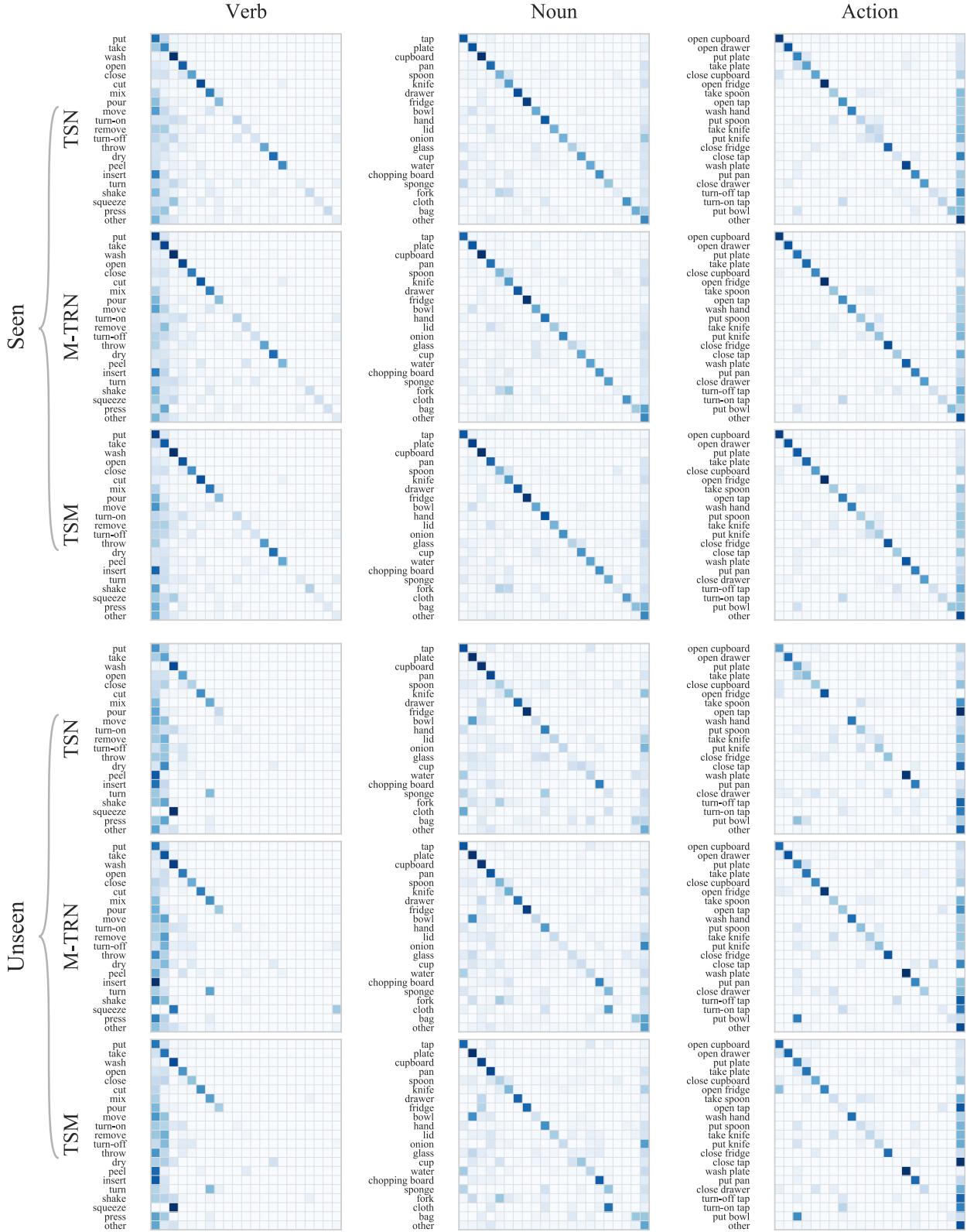


Figure 1: Fusion models' performance on top-20 most frequent classes in training. Classes are ordered from top to bottom in descending order of frequency and any classes outside the top-20 are grouped into a super-class labelled 'other'. [Best viewed on screen]

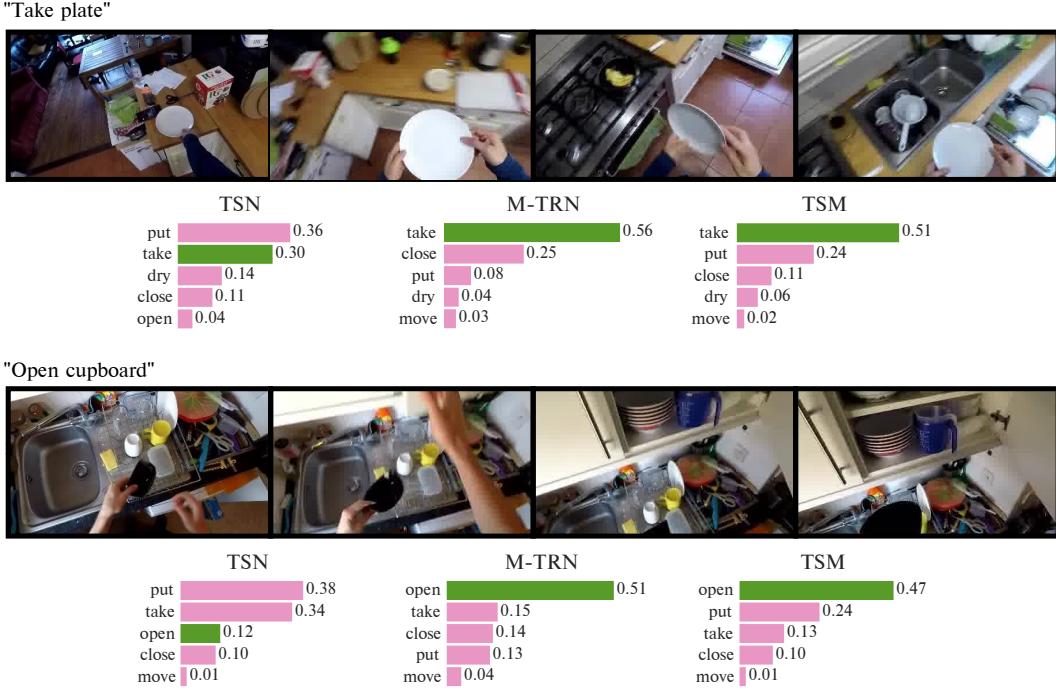


Figure 2: Two examples demonstrating where models capable of temporal reasoning, TRN and TSM, improve over TSN. The bar charts show the model’s scores on the above example with the correct class’ score shown in green.

urates? We examine the answer to this question by training models with different numbers of segments, presenting results in Fig. 3. Overall, flow models benefit more from increasing temporal support, showing monotonically increasing performance, unlike RGB models whose performance saturates at 8 frames, even dropping for the action task when using 16 frames. Curiously, the RGB TSM model is severely harmed by using 16 segments instead of 8, unlike its flow counterpart whose performance improves moving from 8 to 16 segments. This is in contrast to the authors results on Kinetics and Something-something which show an improvement in using 16 frames over 8. This drop was consistently observed across varying LRs suggesting this is not due to a suboptimal learning rate.

Action priors In the previous sections, action predictions have been computed assuming independence between verbs and nouns

$$P(A = (v, n)) = P(V = v)P(N = n), \quad (1)$$

however this is naïve as verb-noun combinations aren’t all as equally likely. For example, it is much more probable to observe ‘cut onion’ than ‘cut chopping board’. In Long-term Feature Banks [13], the authors propose leveraging the prior knowledge of verb-noun co-occurrence in the training set $\mu(v, n)$ to weight the action prediction, *i.e.*

$$P(A = (v, n)) \propto \mu(v, n)P(V = v)P(N = n). \quad (2)$$

Model	Modality	Top-1		Top-5	
		S1	S2	S1	S2
TRN	RGB	+0.05	+1.33	+0.14	+1.43
	Flow	+0.01	+1.43	-0.50	+0.75
M-TRN	RGB	-0.14	+0.99	+0.70	+2.80
	Flow	-0.25	+0.68	-0.61	+0.24
TSM	RGB	+0.02	+0.82	+0.24	+2.42
	Flow	-0.25	+0.89	-0.83	+0.44

Table 2: Percentage point improvement on action task when using action prior across 8-segment ResNet-50 models.

The method in Eq. 2 does not allow zero-shot learning of unseen verb-noun combinations. To remedy this, we apply Laplace smoothing to μ to avoid eliminating the possibility of recognising unseen actions. We evaluate the relative benefit of using action priors in Table 2, finding it provides little benefit on the seen test set, but improves performance on the unseen test set by $\sim 1\%$ point for top-1 accuracy.

4. Released Models

All models required to reproduce the results in Table 1 are made available. We release both RGB and flow models whose predictions can be combined to produce fusion results. To reproduce or compare to these results, the test set predictions should be submitted to the EPIC-Kitchens

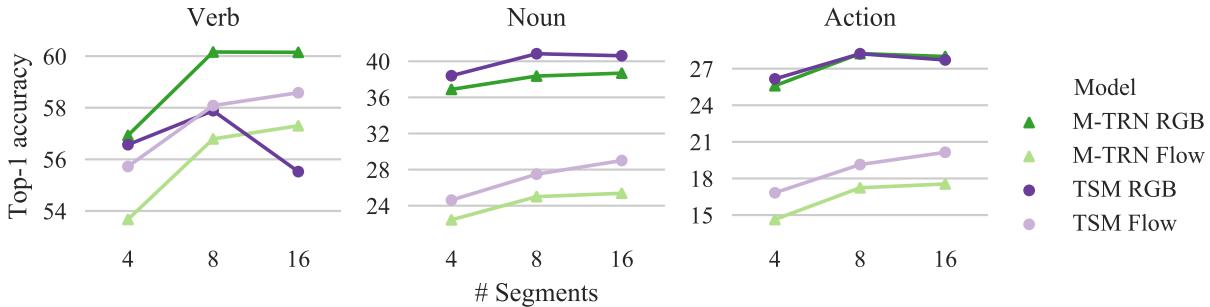


Figure 3: Top-1 accuracy on the seen test set when varying number of segments (during both training/testing) for M-TRN and TSM.

Model	GFLOP/s		Params (M)	
	RGB	Flow	RGB	Flow
TSN	33.12	35.33	24.48	24.51
TRN	33.12	35.32	25.33	25.35
M-TRN	33.12	35.33	27.18	27.21
TSM	33.12	35.33	24.48	24.51

Table 3: Model parameter and FLOP/s count using a ResNet-50 backbone with 8 segments for a single video.

leaderboard² to calculate the performance.

The complexity of the models using ResNet-50 backbone is compared in Table 3,

5. Conclusion

We have benchmarked 3 contemporary models for action recognition and analysed their performance, highlighting areas of good and poor performance. TSM is competitive with M-TRN, and both outperform TSN. These results highlight the necessity for temporal reasoning to recognise actions in EPIC-Kitchens. Yet, the relatively low scores for top-1 accuracy show the challenge is far from solved. Particular issues common to all models are the long-tailed nature of the dataset, fine-grained classes, and difficulty in generalising to unseen environments where we observe a significant drop across all metrics.

References

- [1] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018.
- [2] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, 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. In *ICCV*, 2017.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.
- [6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.
- [7] J. Lin, Chuang. Gan, and S. Han. Temporal shift module for efficient video understanding. *arXiv*, 2018.
- [8] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017.
- [9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.
- [10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012.
- [11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [12] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. *TPAMI*, 2019.
- [13] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019.
- [14] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018.

²<https://epic-kitchens.github.io/2019#challenges>