With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition – Appendix

Evangelos Kazakos[1]
Evangelos.Kazakos@bristol.ac.uk
Jaesung Huh[2]
jaesung@robots.ox.ac.uk
Arsha Nagrani†[2]
arsha.nagrani@gmail.com
Andrew Zisserman[2]
az@robots.ox.ac.uk
Dima Damen[1]
Dima.Damen@bristol.ac.uk

[1] Dept. of Computer Science
University of Bristol
Bristol, UK

[2] Visual Geometry Group
University of Oxford,
Oxford, UK

A EPIC-KITCHENS-100: Results on the Test Set

In Table 3 of the main paper, we compare to published works on the validation set of EPIC-KITCHENS-100. Unfortunately, most works do not report on the leaderboard test set. In Table 1, we provide results on the test set comparing our model to baselines from [2], as well as Ego-Exo [7] that distills knowledge from a much larger training set. MTCN outperforms all other methods, including the competitive method of [7], showcasing that multimodal temporal context from consecutive actions is more beneficial than pretraining large models (ResNet101) using egocentric signals from third-person datasets.

B EPIC-KITCHENS-55 Results

We also compare our model to works that report on the earlier version of this dataset, namely EPIC-KITCHENS-55 [1] in Table 2. We report results for the Seen split (S1). We opted to include these in the appendix to avoid confusion in the main paper as the results are not comparable across these two dataset versions. We compare MTCN with two audio-visual approaches [16] and [18], as well as [17] which was one of the first works to utilise temporal context. We also report the performance of [7] which evaluates their method on both EPIC-KITCHENS-55 and EPIC-KITCHENS-100. Our MTCN outperforms all approaches.

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

†Now at Google Research.

	Overall						Unseen Participants			Tail-classes		
	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Top-1 Accuracy (%)			Top-1 Accuracy (%)		
Model	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
TSN [15]	59.03	46.78	33.57	87.55	72.10	53.89	53.11	42.02	27.37	26.23	14.73	11.43
TRN [19]	63.28	46.16	35.28	88.33	72.32	55.26	57.54	41.36	29.68	28.17	13.98	12.18
TBN [6]	62.72	47.59	35.48	88.77	73.08	56.34	56.69	43.65	29.27	30.97	19.52	14.10
TSM [9]	65.32	47.80	37.39	89.16	73.95	57.89	59.68	42.51	30.61	30.03	16.96	13.45
SlowFast [3]	63.79	48.55	36.81	88.84	74.49	56.39	57.66	42.55	29.27	29.65	17.11	13.45
Ego-Exo [7]	66.07	51.51	39.98	**89.39**	76.31	60.68	59.83	45.50	32.63	33.92	22.91	16.96
MTCN - v.f. SlowFast [3]	**68.44**	**55.41**	**44.10**	88.74	**78.04**	**61.69**	**61.82**	**47.62**	**34.94**	**34.77**	**28.60**	**20.45**

Table 1: Results on the test set of EPIC-Kitchens-100.

	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
Model	Verb	Noun	Action	Verb	Noun	Action
LFB [17]	60.0	45.0	32.7	88.4	71.8	55.3
G-Blend [16]	66.7	48.5	37.1	88.9	71.4	56.2
AV-SlowFast [18]	65.7	46.4	35.9	89.5	71.7	57.8
Ego-Exo [7]	65.97	47.99	37.09	**90.32**	70.72	56.32
MTCN (Ours)	**69.12**	**51.30**	**40.77**	90.18	**73.53**	**59.15**

Table 2: Comparison with SOTA on the Seen split (S1) of EPIC-KITCHENS-55.

C Language model analysis and baselines

In this section, we assess the statistical significance of our language model and compare the performance of our MTCN to variants using baseline language models. All the experiments in this section as well as in Sections D and E and the visualisations in Section H are performed on the validation set of EPIC-KITCHENS-100 and using the SlowFast visual features.

Statistical significance of LM. We train 10 audio-visual transformers and 10 corresponding language models with different random seeds. Table 3 shows the mean and standard deviation top-1 and top-5 accuracy without and with the language model. Utilising the language model improves performance on average with a low std, demonstrating that the improvement from the language model is statistically significant. We further showcase that by conducting T-tests on verb, noun and action top-1 accuracies, obtaining a p-value of $3.6e-2$, $6.0e-4$, $9.7e-4$, respectively.

Baselines comparison. We compare our MTCN that uses a transformer based language model to two baselines, N-gram and Bi-directional LSTM (BiLSTM). For N-gram, we follow a similar procedure to natural language processing. In particular, from all action sequences of length 9 in the training set, we derive the heuristic probability of occurrence of the centre action given the preceding and succeeding actions. We train a BiLSTM with 3 layers and a hidden size of 512. The rest hyperparameters are the same as the transformer encoder.

Results are shown in Table 4. It turns out that only a few preceding-succeeding action sequences in the training set also appear in the validation set, resulting in no difference in performance when N-gram is added comparing to not using a language model. Our transformer-based Masked Language Model (MLM) outperforms both the N-gram and BiLSTM, showcasing that it is beneficial to use a deep neural network language model over a heuristic prior and that MLM with transformers outperforms recurrent architectures in this problem.

	Top-1 Accuracy (%)			Top-5 Accuracy (%)		
LM	Verb	Noun	Action	Verb	Noun	Action
✗	70.26 ± 0.27	55.70 ± 0.22	44.90 ± 0.20	91.12 ± 0.13	**79.03 ± 0.18**	**64.79 ± 0.17**
✓	**70.52 ± 0.25**	**56.08 ± 0.21**	**45.25 ± 0.18**	**91.13 ± 0.13**	**79.03 ± 0.18**	64.58 ± 0.18

Table 3: Mean and standard deviation of multiple runs both w. & w/o language model in the validation set of EPIC-KITCHENS-100.

	Overall						Unseen Participants			Tail-classes		
	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Top-1 Accuracy (%)			Top-1 Accuracy (%)		
Model	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
No LM	70.23	55.82	45.00	91.13	**79.06**	**64.58**	63.29	46.38	35.02	**41.76**	32.26	24.41
N-gram	70.23	55.84	45.02	91.13	**79.06**	64.49	63.29	46.38	35.02	**41.76**	32.26	24.41
BiLSTM	70.57	55.97	45.09	**91.14**	**79.06**	**64.55**	63.29	46.76	35.31	40.68	32.47	24.15
Transformer enc. (proposed)	**70.60**	**56.26**	**45.48**	**91.14**	**79.06**	63.06	**63.76**	**47.14**	**35.87**	41.36	**32.84**	**24.70**

Table 4: Performance of MTCN in the validation set of EPIC-KITCHENS-100 using different language models.

D Online recognition

The focus of this work is to leverage both past and future context to predict an action. In this section however, we explore the performance of our model in online recognition, i.e. using only the preceding actions as context to predict the current action. This approach can be used to recognise actions in an online fashion for streaming videos. For this setting, we train the audio-visual transformer to predict the last action in the sequence. We do not train a new language model for this task; we simply mask and predict the last action in the sequence instead of the centre one.

Results are demonstrated in Table 5 by varying w. Our model can also utilise temporal context in this setting, as performance improves for $w > 1$ with optimal top-1 accuracy at $w = 7$ and optimal accuracy on tail-classes at $w = 9$. Compared to our original proposal that utilises also future context (see Table 1 on main paper), the overall performance degrades, indicating that leveraging future context is beneficial.

E Architecture ablations

In Table 6, we explore different number of layers in MTCN, both without and with (layer-wise) weight sharing, and compare each case with a single layer. Note that we use the same number of layers and sharing strategy for both AV and LM. We use bold to indicate best performance within each group rather than overall. Best results are obtained using four layers in most metrics, both without & with weight sharing. These outperform a single layer, demonstrating that is beneficial to use a multi-layered transformer. Although MTCN without weight sharing performs slightly better, our proposed model has 2.7× less parameters with only a minor drop in performance.

In Table 7, we compare the effect of different types of positional encodings. Particularly, we replace our chosen absolute learnt positional encoding with relative positional encodings [12] and Fourier feature positional encodings [5]. Fourier feature positional encodings replace our learnable absolute positional encodings with non-learnable ones represented as a vector of log-linearly spaced frequency bands up to a maximum frequency. Relative positional encodings replace our absolute positional encodings of the inputs, with positional

	Overall						Unseen Participants			Tail-classes		
	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Top-1 Accuracy (%)			Top-1 Accuracy (%)		
w	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
1	67.93	52.29	41.30	90.53	76.47	61.52	61.13	44.60	32.58	42.05	27.42	21.48
3	68.42	54.15	42.59	91.20	78.52	61.10	**61.69**	**44.41**	**32.11**	40.11	31.26	22.58
5	68.58	54.04	42.75	**90.96**	78.27	62.04	59.81	43.94	**32.11**	39.49	31.11	22.48
7	**68.88**	**54.31**	**42.96**	90.89	77.87	62.39	61.41	43.38	32.02	40.51	32.00	23.61
9	68.77	54.28	42.77	90.66	77.72	**62.44**	60.38	45.07	31.83	**40.80**	**32.68**	**23.86**
11	67.83	54.04	42.13	90.63	**78.85**	62.10	57.46	43.94	31.46	36.88	30.74	21.96

Table 5: Online action recognition results by varying temporal context length in the validation set of EPIC-KITCHENS-100.

		Overall						Unseen Participants			Tail-classes		
		Top-1 Accuracy (%)			Top-5 Accuracy (%)			Top-1 Accuracy (%)			Top-1 Accuracy (%)		
Layers	Shared	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
1	-	69.58	55.04	43.71	91.27	79.02	63.96	61.03	46.01	33.33	42.10	32.42	24.09
2	✗	69.49	55.41	43.94	**91.13**	78.96	63.86	**62.72**	46.57	34.37	**41.42**	**32.32**	23.90
4	✗	**71.01**	**56.55**	**46.04**	90.98	**79.28**	**63.97**	62.35	**47.61**	**35.96**	39.94	31.89	**24.22**
6	✗	69.58	55.68	44.89	90.28	78.17	63.37	61.31	45.63	34.46	38.75	32.21	23.90
2	✓	69.82	55.37	43.81	91.09	78.97	**64.26**	61.50	44.32	32.68	**42.05**	32.58	23.74
4	✓	**70.60**	**56.26**	**45.48**	**91.14**	**79.06**	63.06	**63.76**	**47.14**	**35.87**	41.36	**32.84**	**24.70**
6	✓	69.58	55.68	44.89	90.28	78.17	63.37	61.31	45.63	34.46	38.75	32.21	23.90

Table 6: Analysis of performance using different number of layers, both w. and w/o weight sharing. Results are shown in the validation set of EPIC-KITCHENS-100.

	Overall						Unseen Participants			Tail-classes		
	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Top-1 Accuracy (%)			Top-1 Accuracy (%)		
Pos. enc.	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action	Verb	Noun	Action
Fourier PE	69.60	56.13	44.63	90.65	78.86	63.31	63.29	45.63	35.12	38.86	32.53	23.09
Relative PE	70.32	**56.30**	45.37	91.01	**79.35**	**64.04**	61.41	46.67	33.99	**41.42**	**33.74**	**24.96**
Absolute PE (Proposed)	**70.60**	56.26	**45.48**	**91.14**	79.06	63.06	**63.76**	**47.14**	**35.87**	41.36	32.84	24.70

Table 7: Comparison of different positional encodings (PE) using the validation set of EPIC-KITCHENS-100.

encodings representing distances between tokens and placed within the self-attention layers. As shown in the table our proposed absolute learnable positional encodings outperform Fourier feature positional encodings in all metrics (except top-5 action accuracy). Comparing to relative positional encodings, our positional encodings are slightly better in top-1 verbs and actions, as well as in unseen participants, while relative positional encodings perform slightly better in top-5 accuracy and tail classes. Overall, there are no notable differences between the different choices of positional encodings.

F EGTEA Implementation Details

Visual features. For EGTEA, we train SlowFast [3] using the EPIC-KITCHENS pre-trained model, by sampling a clip of 2s from an action segment similar to EPIC-KITCHENS. We use a learning rate of 0.001, no warm-up, and we keep the batch normalisation layers frozen. All unspecified hyperparameters remain unchanged. For feature extraction, we follow the same procedure as EPIC-KITCHENS, except that we use clips of 2s rather than 1s.

	Visual		Visual + LM	
w	Top-1(%)	Mean Class (%)	Top-1(%)	Mean Class (%)
1	72.26	64.98	72.26	64.98
3	72.55	64.86	**73.59**	65.87
5	**73.10**	**65.42**	73.49	65.57
7	72.26	64.38	73.19	65.31
9	72.55	64.86	73.44	**66.02**

Table 8: Ablation of temporal context extent and language model in the first test split of EGTEA.

Train/Val Details. Here, we discuss differences in the architecture for training/evaluating EGTEA. Remember that for EGTEA we train only vision and language as EGTEA does not contain audio. First, as there is no audio input to the transformer, we do not use modality encodings either. Second, following previous methods [4, 8, 10, 11, 13, 14] that train using a single head for actions and report only action accuracy, we use a single summary embedding for actions, rather than verb/noun embeddings. Accordingly, the language model utilises a single word-embedding for actions, with a dimension of 512. For training the visual-only transformer, we use a learning rate of 0.001, train the model for 50 epochs and decay the learning rate at epochs 25 and 38, while keeping all other hyperparameters unchanged. We use same hyperparameters for the language model. For evaluation, differently than EPIC-KITCHENS, we average the predictions of the 10 clips per action, rather than feeding all 10 clips in the transformer.

G Ablation of temporal context and language model in EGTEA

We study the effect of the temporal context length both with and without the language model on the first test split of EGTEA. Results are shown in Table 8. For the visual only model, top-1 accuracy increases when we increase the length of temporal context from $w = 1$ to 5, and optimal results for both top-1 and mean class accuracy are obtained for $w = 5$. When the language model is incorporated top-1 accuracy increases from $w = 1$ to 3 and then decreases while best mean class accuracy is obtained at $w = 9$. These findings showcase that our model successfully utilises context in this dataset as well. The language model is helpful for EGTEA, and provides a bigger boost in performance than EPIC-KITCHENS, possibly due to the absence of audio modality. Finally, it is worth noting that after the addition of the language model best performance is obtained at a shorter temporal context, showing that shorter sequences of actions provide a stronger prior in this dataset.

H Attention Visualisation

In Fig. 1, we show additional qualitative examples, similar to Fig. 4 in the main paper. These demonstrate how our model attends to temporal context. In the first three examples, the model predicts the centre action correctly, while in the last one it gives incorrect predictions. In the 'wash chopping board' example, the model particularly attends to actions containing

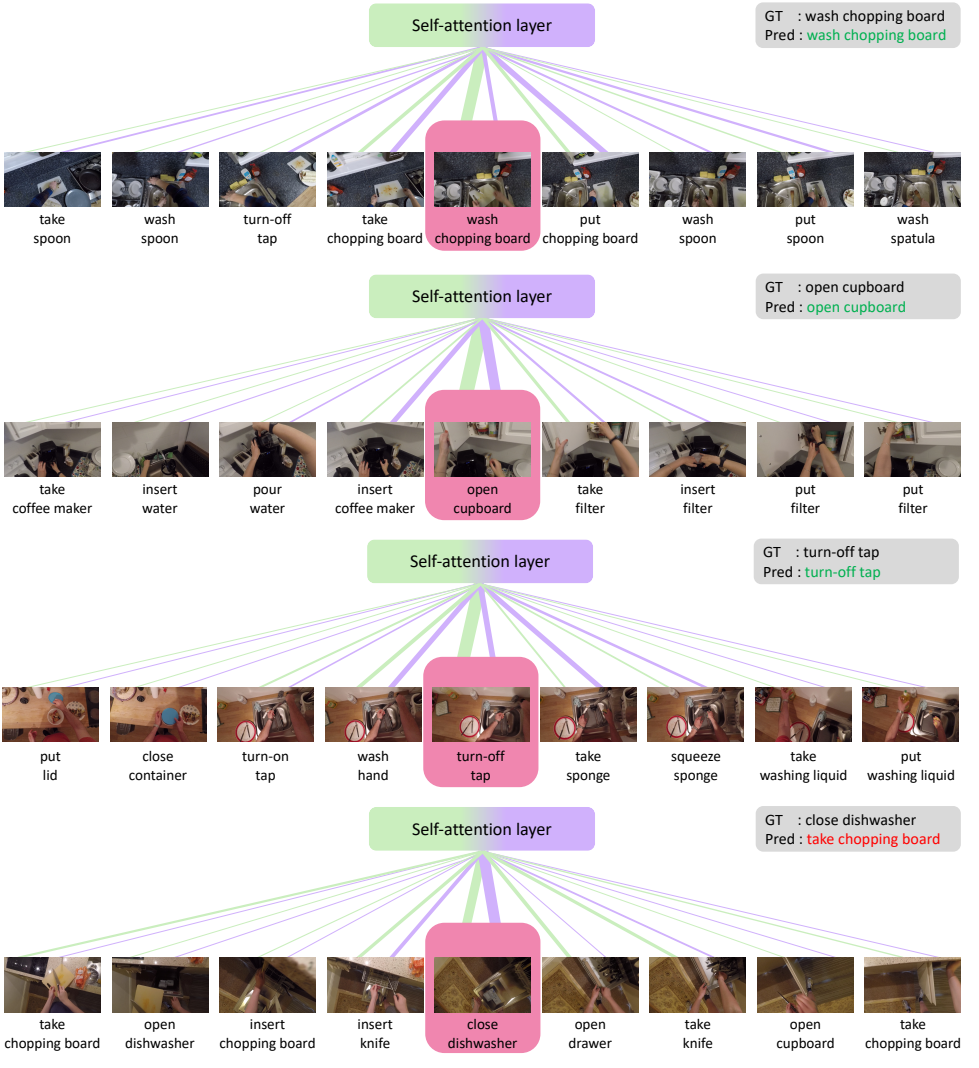

Figure 1: Additional qualitative results of attention weights along with the predictions of our model. Green and purple edges represent attention to visual and auditory tokens, respectively, from the noun summary embedding. Thickness indicates attention weights magnitude to centre (bordered) and temporal context actions.

the chopping board. For 'open cupboard', the model has high audio-visual attention to the centre action, and high attention to the audio of the previous action ('insert coffee maker'), showing that at times audio provides useful temporal context. The importance of audio is also apparent in the third example. A source of error in the model results from confusing the centre action with another action in the sequence; in the fourth example 'close dishwasher' is predicted as 'take chopping board' which corresponds to the first and last actions in the temporal context.

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 *European Conference on Computer Vision (ECCV)*, 2018.

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *IJCV*, 2021.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[4] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020.

[5] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4651–4664, 2021.

[6] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[7] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6943–6953, June 2021.

[8] Yin Li, Miao Liu, and James M. Rehg. 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)*, September 2018.

[9] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *ICCV*, 2019.

[10] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[11] Kyle Min and Jason J. Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1069–1078, January 2021.

[12] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[13] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *British Machine Vision Conference (BMVC)*, 2018.

[14] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[15] 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. In *ECCV*, 2016.

[16] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[17] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[19] 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.