ConTra: (Con)text (Tra)nsformer for Cross-Modal Video Retrieval

Supplementary Material

Adriano Fragomeni Michael Wray Dima Damen

Department of Computer Science, University of Bristol, UK

In the supplementary material we first detail the tables corresponding to the figures in the main paper in Sec. 1. The supplementary also includes additional ablation experiments in Sec. 2. We then provide an analysis of the complexity of the proposed method in Sec. 3.

1 Tables corresponding to figures in the main paper

Table 1 details the results presented in Fig. 3 and reports Recall at $K = \{1, 5, 10\}$ (R@K) and median rank (MR) for all the datasets when different lengths of local clip context are used.

	YouCook2								ActivityNet CS								EPIC-KITCHENS-100							
	Sentence-To-Clip				Clip-To-Sentence					Sentence-To-Clip				Clip-To-Sentence					Sentence-To-Clip				Clip-To-Sentence				
m	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
0	15.6	39.1	52.0	10	13.3	37.1	50.6	10	207.7	4.3	14.3	21.9	51	4.4	14.6	22.9	50	82.4	17.9	38.6	47.9	12	24.0	46.4	55.7	7	230.5
1	16.5	41.1	54.2	**8**	14.1	38.6	51.7	10	216.2	5.7	17.5	26.0	41	5.7	17.9	27.2	39	100.0	21.4	42.5	51.8	10	27.8	50.5	59.7	**5**	253.7
2	**17.1**	42.0	54.4	**8**	**14.9**	39.8	52.8	**9**	221.0	5.8	17.9	26.8	38	6.2	19.2	28.3	37	104.2	21.6	42.6	53.0	9	28.0	51.0	60.7	**5**	256.9
3	16.7	**42.1**	**55.2**	**8**	14.8	40.5	53.9	**9**	**223.2**	**5.9**	18.4	**27.6**	**38**	**6.4**	19.3	**28.5**	37	**106.1**	21.6	43.2	**53.6**	**8**	**28.6**	51.6	**61.3**	**5**	259.9
4	16.7	41.8	54.9	**8**	14.7	40.3	**54.5**	**9**	222.9	**5.9**	**18.5**	27.2	**38**	**6.4**	**19.4**	28.4	**36**	105.8	**22.2**	43.4	53.4	9	28.2	**52.0**	61.1	**5**	**260.3**
5	16.5	**42.1**	54.4	**8**	14.8	**40.7**	53.7	**9**	222.2	5.8	18.1	27.2	**38**	6.3	19.1	28.4	**36**	104.9	21.2	**43.7**	53.5	**8**	28.2	51.6	61.2	**5**	259.4

Table 1: Analysis of the importance of temporal clip context (CC), reporting Recall ↑ and Median Rank ↓.

Table 2 and Table 3 are the extensions of Fig. 7 and Fig. 8 from the main paper, respectively, and report sentence-to-clip results for all the datasets when different lengths of local clip context are used.

	YouCook2 (S2C)					ActivityNet CS (S2C)					EPIC-KITCHENS-100 (S2C)				
m	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	RSum
0	16.0	39.7	52.5	9	108.1	4.3	14.3	21.9	51	40.5	17.9	38.6	47.9	12	104.4
1	17.5	42.7	56.0	8	116.2	6.0	19.5	29.2	30	54.7	22.0	46.6	56.3	7	124.9
2	**17.7**	**43.9**	**56.9**	**7**	**118.5**	6.7	21.1	31.3	27	59.1	24.1	50.1	60.4	**5**	134.6
3	17.3	42.7	56.8	8	116.8	6.9	21.7	**32.0**	**26**	60.6	25.6	51.2	61.7	**5**	138.5
4	17.1	42.8	55.8	8	115.7	**7.0**	**21.8**	**32.0**	**26**	**60.8**	**26.2**	52.6	62.7	**5**	141.5
5	16.6	41.2	55.4	8	113.2	**7.0**	21.5	31.9	**26**	60.4	25.8	**53.2**	**63.6**	**5**	**142.6**

Table 2: Analysis of temporal context in text only for sentence-to-clip.

	YouCook2 (S2C)					ActivityNet CS (S2C)					EPIC-KITCHENS-100 (S2C)				
m	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	RSum	R@1	R@5	R@10	MR	RSum
0	16.0	39.7	52.5	9	108.1	4.3	14.3	21.9	51	40.5	17.9	38.6	47.9	12	104.4
1	40.1	68.5	78.8	**2**	187.4	16.4	37.3	49.4	11	103.1	48.0	72.5	79.8	2	200.3
2	45.5	73.7	82.2	**2**	201.4	22.7	48.1	60.6	6	131.4	60.6	81.3	86.9	**1**	228.8
3	**47.0**	**75.4**	84.7	**2**	**207.1**	25.3	52.6	64.8	**5**	141.9	67.0	85.9	90.8	**1**	243.7
4	46.4	75.3	**84.9**	**2**	206.6	**26.3**	**53.2**	**65.3**	**5**	**144.8**	72.2	89.4	92.9	**1**	254.5
5	46.3	74.7	83.9	**2**	204.9	25.7	**53.2**	64.7	**5**	143.6	**73.8**	**90.4**	**93.7**	**1**	**257.9**

Table 3: Analysis of temporal context in both text and video for sentence-to-clip.

#Layers		#Heads		Sentence-to-Clip				Clip-to-Sentence				
V	T	V	T	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
1	1	1	1	**17.1**	42.0	**55.2**	**8**	14.4	39.3	53.5	**9**	221.5
1	1	2	2	16.7	**42.1**	**55.2**	**8**	14.8	**40.5**	**53.9**	**9**	**223.2**
1	1	4	4	16.5	41.7	**55.2**	**8**	14.6	39.9	53.8	**9**	221.7
1	1	8	8	16.7	41.3	**55.2**	**8**	**15.2**	39.9	53.6	**9**	221.9
1	1	16	16	16.4	41.6	55.0	**8**	14.8	40.0	53.5	**9**	221.3
2	1	2	2	16.1	41.1	54.9	**8**	14.5	39.3	52.9	**9**	218.8
1	2	2	2	16.1	41.1	53.8	9	14.7	39.8	52.9	**9**	218.4
2	2	2	2	16.2	40.3	54.2	9	14.5	39.4	52.3	**9**	216.9

Table 4: Different number of layers and heads for Video (V) and Text (T) transformer encoders (YC2).

2 Extra Ablation Studies

In this section, we present some additional experiments on all datasets. Additional ablation studies on YouCook2 are in Sec. 2.1, ActivityNet Clip-Sentence in Sec. 2.2, EPIC-KITCHENS-100 in Sec. 2.3, a further analysis on the neighbouring loss, L_{NEI} is in Sec. 2.4 followed by context vs clip length in Sec. 2.5.

2.1 Ablations on YouCook2 (YC2)

Number of Heads and Layers. We study how the performance changes by varying the number of stacked encoder layers in clip and text transformers, and the number of heads per layer in Table 4. We first fix the number of layers and change the number of heads in both encoders. The best result is achieved when using only 2 heads. Overall, we can assert that the method is robust to the number of heads, as performance varies only marginally when the number of heads is adjusted. Next, we fix the number of heads to 2 – our best result from above, and change the number of layers in our encoders, for both clip transformer and text transformer. The best performance is achieved using 1 layer. Note that YouCook2 is the smallest dataset. We ablate the number of heads and layers also in Sec 2.3 for larger datasets.

Temperature Parameter. Several works [3–5] proposed to set the temperature parameter $\tau = 0.07$. We thus follow this in all our experiments. Here, we test empirically this choice by varying τ as shown in Table 5. Increasing the value of τ, drops the performance of our model in both retrieval tasks. While higher, but comparable, results are achieved with $\tau = 0.05$, we keep the standard τ to remain directly comparable to other works that keep τ at 0.07.

	Sentence-to-Clip				Clip-to-Sentence				
τ	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
0.05	**17.2**	41.9	**55.6**	**8**	**15.1**	40.4	**54.7**	**8**	**224.9**
0.07	16.7	**42.1**	55.2	**8**	14.8	**40.5**	53.9	9	223.2
0.1	16.2	40.3	53.9	9	14.5	39.6	52.5	9	217.0
0.7	12.5	33.2	46.0	13	10.9	31.8	43.1	15	177.5
1.0	13.0	34.9	47.2	12	12.5	33.3	45.2	13	186.1

Table 5: Analysis of the performance varying τ in L_{CML} and L_{NEI} (YC2).

Loss weights. In the main paper, our objective function is defined as follows:

$$L = \lambda_{CML} L_{CML} + \lambda_{NEI} L_{NEI} + \lambda_{UNI} L_{UNI} \tag{1}$$

where $\lambda_{CML} = \lambda_{UNI} = \lambda_{NEI} = 1$ showcasing that we outperform other methods without hyperparameter tuning. Table 6 shows the performance of ConTra on YouCook2 when different combination of weights are used. By applying a Grid Search approach we are able to find the best combination of weights of the multiple loss empirically and improve the RSum of 7.6 points when $\lambda_{CML} = 1$, $\lambda_{UNI} = 12$ and $\lambda_{NEI} = 2$. The uniformity loss's weight has the largest value, i.e. $\lambda_{UNI} = 12$, but this high value is similar to other works [1].

Weights Loss			Sentence-to-Clip				Clip-to-Sentence				
λ_{CML}	λ_{UNI}	λ_{NEI}	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
1	1	1	16.7	42.1	55.2	8	14.8	40.5	53.9	9	223.2
2	1	1	16.6	40.8	55.1	8	14.5	39.5	53.5	9	220.0
1	1	2	16.9	42.7	55.6	8	14.8	39.9	54.2	9	224.1
1	1	5	16.1	40.4	54.3	9	14.6	38.7	52.4	9	216.5
1	2	1	**17.3**	42.1	55.3	8	15.1	40.4	54.3	9	224.5
1	5	1	17.1	42.8	55.7	8	15.5	41.1	54.8	**8**	227.0
1	10	1	16.7	43.2	56.2	8	15.7	41.6	55.9	**8**	229.3
1	12	1	17.0	**43.4**	56.1	8	15.6	**42.2**	55.9	**8**	230.2
1	15	1	17.1	42.9	56.0	8	15.6	41.8	55.6	**8**	229.0
1	12	2	16.9	43.3	**56.8**	**7**	**15.9**	41.7	**56.2**	**8**	**230.8**

Table 6: Ablation of loss weights in clip context scenario (YC2).

Batch Size. We vary the batch size in Table 7. The overall performance of the model increases with the batch size, showing the importance of having varied negatives per batch when using the NCE loss. Given memory limits, we were unable to increase the batch size further, but we anticipate diminishing returns above 512.

	Sentence-to-Clip				Clip-to-Sentence				
B	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
64	15.8	41.2	54.6	**8**	13.8	39.5	52.4	**9**	217.3
128	16.4	41.7	55.1	**8**	14.2	40.0	53.1	**9**	220.5
256	**16.8**	41.9	54.7	**8**	14.4	39.5	52.9	**9**	220.2
512	16.7	**42.1**	**55.2**	**8**	**14.8**	**40.5**	**53.9**	**9**	**223.2**

Table 7: Analysis of the performance varying the batch size, B, (YC2).

Shared Weights	Positional Encoding	Sentence-to-Clip R@1	R@5	R@10	MR	RSum
✓	shared	47.2	74.8	83.7	**2**	205.7
✓	distinct	47.1	74.9	84.2	**2**	206.2
×	shared	**47.3**	74.8	83.5	**2**	205.6
×	×	46.1	74.5	83.9	**2**	204.5
×	distinct	47.0	**75.4**	**84.7**	**2**	**207.1**

Table 8: Ablation of modality weights and pos. encoding (YC2).

Shared Weights and Positional Encoding. We present additional results on YouCook2 for sentence-to-clip when utilising context in both modalities. We ablate the choice of weights and position encodings in both clip and text transformers. Table 8 compares these results with shared/distinct weights and positional encodings. We also evaluate removing the positional encoding. Using different weights and distinct encodings between modalities achieves the best performance.

2.2 Ablations on ActivityNet CS

Loss Function. In Table 9 we show the improvement given by all the terms of our objective function. We obtain similar results to those we report in Table 6 in the main paper, where the our neighbouring loss L_{NEI} helps the model to better distinguish clips compared to the standard hard mining approach proposed in [2].

	Sentence-to-Clip				Clip-to-Sentence				
Loss	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
L_{NEI}	2.6	9.3	14.8	96	2.5	9.3	15.0	96	53.5
L_{CML}	5.0	16.3	24.9	43	5.6	17.3	26.2	42	95.3
L_{CML}+$L_{HardMining}$	5.1	16.5	25.0	43	5.7	17.5	26.4	41	96.2
L_{CML}+L_{NEI}	**5.9**	18.1	27.2	**38**	6.1	19.0	28.1	**37**	104.4
L_{CML}+L_{NEI}+L_{UNI}	**5.9**	**18.4**	**27.6**	**38**	**6.4**	**19.3**	**28.5**	**37**	**106.1**

Table 9: Ablation of loss function (ActivityNet CS).

2.3 Experiments on EPIC-KITCHENS-100

Loss Function. Similarly, Table 10 illustrates the performance of our model on EPIC-KITCHENS-100 when changing the objective function. As with the other datasets, the neighbouring loss L_{NEI} works better than $L_{HardMining}$ and justifies all the terms of our objective function, i.e. L_{NEI} and L_{UNI}.

Batch Size. Similar to the results obtained in Table 7 on YouCook2, increasing the size of the batch boosts the performance of our model also on EPIC-KITCHENS-100 as highlighted in Table 11.

Number of Heads and Layers. Table 12 shows how the performance changes by varying the number of stacked encoder layers and the number of heads per layer. We

	Sentence-to-Clip				Clip-to-Sentence				
Loss	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
L_{NEI}	5.6	15.8	23.3	61	9.4	28.1	37.8	23	120.0
L_{CML}	21.7	42.4	51.6	**9**	28.1	50.8	60.2	**5**	254.8
L_{CML}+$L_{HardMining}$	20.9	42.4	51.6	**9**	27.8	51.4	60.8	**5**	254.9
L_{CML}+L_{NEI}	21.1	43.0	52.9	**9**	27.3	51.3	60.9	**5**	256.5
L_{CML}+L_{NEI}+L_{UNI}	**22.2**	**43.4**	**53.4**	**9**	**28.2**	**52.0**	**61.1**	**5**	**260.3**

Table 10: Ablation of loss function (EPIC-KITCHENS-100).

	Sentence-to-Clip				Clip-to-Sentence				
B	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
64	19.0	41.0	50.6	10	25.7	48.9	58.6	6	243.8
128	20.7	42.4	52.1	**9**	27.3	51.0	60.7	**5**	254.2
256	21.6	43.1	52.8	**9**	**28.4**	**52.0**	**61.6**	**5**	259.5
512	**22.2**	**43.4**	**53.4**	**9**	28.2	**52.0**	61.1	**5**	**260.3**

Table 11: Analysis of the performance varying the batch size, B (EPIC-KITCHENS-100).

#Layers		#Heads		Sentence-to-Clip				Clip-to-Sentence				
V	T	V	T	R@1	R@5	R@10	MR	R@1	R@5	R@10	MR	RSum
2	2	1	1	20.7	42.1	52.4	**9**	27.3	51.1	60.4	**5**	254.0
2	2	2	2	21.3	43.0	52.9	**9**	27.7	51.1	60.4	**5**	256.4
2	2	4	4	21.5	43.0	53.1	**9**	28.3	51.8	60.8	**5**	258.5
2	2	8	8	**22.2**	**43.4**	**53.4**	**9**	28.2	**52.0**	61.1	**5**	**260.3**
2	2	16	16	21.4	**43.4**	53.1	**9**	28.2	**52.0**	**61.6**	**5**	259.7
1	1	8	8	20.4	41.0	50.9	10	26.1	50.2	60.2	**5**	249.1
2	1	8	8	21.4	41.8	52.3	**9**	**28.6**	51.5	61.3	**5**	256.9
1	2	8	8	20.8	41.0	51.6	10	26.0	50.0	60.2	6	249.6

Table 12: Different number of heads and layers (EPIC-KITCHENS-100) for Video (V) and Text (T) transformer encoders.

first fix the number of layers and change the number of heads in both encoders. We achieve the best result when using 8 heads. Overall, we can see that the method performs better when increasing the number of heads. This can be explained by the larger size of EPIC-KITCHENS-100 compared to YouCook2. Then, we fix the number of heads to 8 and change the number of stacked encoder layers. The best performance remains using 2 layer. We use the same number of layers and heads for ActivityNet Clip-Sentence.

2.4 Neighbouring Loss per Dataset

We provide an expanded version of Fig. 6 from the main paper in Fig. 1 broken down per dataset. Overall, the individual plots follow the results of Fig. 6 in the main paper. In each dataset we can see that L_{NEI} has an important role, the similarity between neighbouring clips $j+1$ and the sentence j tends to decrease for clips and sentences close in space (i.e. higher cosine similarity on the x-axis).

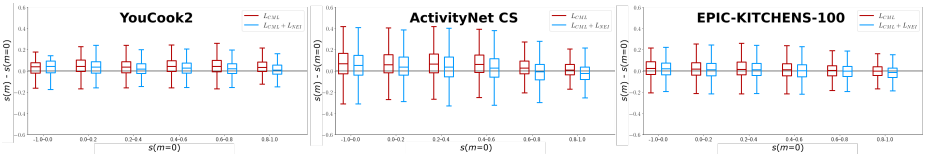

Fig. 1: Expanded version of Fig. 6 from the main paper showing the comparison between similarities to neighbouring clips, $s(m) - s(m = 0)$, with and without using L_{NEI}.

2.5 Context Length w.r.t. Clip Length

We study the relation between context length and clip length by analysing the average improvement of the rank position of all the clips of a certain length, as shown in Fig. 2. We bin all the clips in the test set based on their length considering small intervals for dense datasets. We then calculate the average difference between the rank position when $m = 0$ and $m > 0$. In all the datasets, short clips benefits the most from local clip context when $m > 1$. Long clips demonstrate a different behaviour. In EPIC-KITCHENS-100 and ActivityNet CS, long context helps the most. An interesting behaviour can be highlighted for ActivityNet CS. Although this dataset has the smallest number of clips per video on average, we see that the largest improvement of rank position when $m = 5$. A possible explanation can be that in ActivityNet CS, some clips are as long as the entire video so using context adds fine-grained information that can help the model to retrieve these clips.

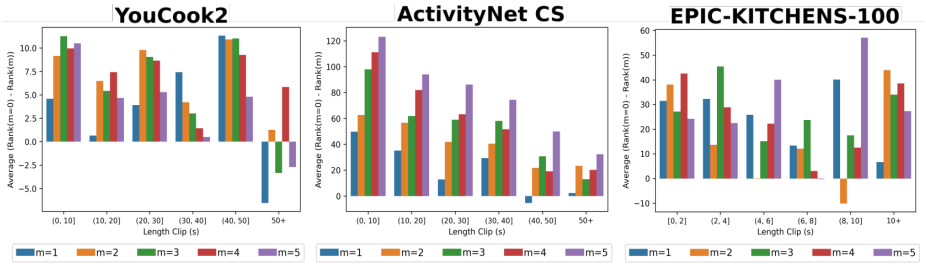

Fig. 2: Analysis of the improvements in the rank position w.r.t the length of the clip

3 Complexity Analysis

We present a computational complexity analysis of ConTra in Table 13, Table 14 and Table 15 for local clip context, text context and when clip and text contexts are used simultaneously, respectively.

For each dataset and each value of m, we report the number of trainable parameters, the FLOPS in training, and the performance of the model, i.e. total RSum in Table 13 and RSum_{S2C} in Table 14 and Table 15. As noted in the tables, increased context only increases the number of parameters slightly, but require more GFlops. As expected,

	YouCook2			ActivityNet CS			EPIC-KITCHENS-100		
m	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum
0	9.549	0.04	207.7	8.675	0.02	82.4	16.903	0.06	230.5
1	9.550	0.08	216.2	8.676	0.06	100.0	16.904	0.12	253.7
2	9.551	0.10	221.0	8.677	0.10	104.2	16.905	0.20	256.9
3	9.552	0.14	223.2	8.677	0.14	106.1	16.906	0.26	259.9
4	9.553	0.16	222.9	8.678	0.16	105.8	16.907	0.32	260.3
5	9.554	0.20	222.2	8.679	0.20	104.9	16.908	0.38	259.4

Table 13: Analysis of Complexity (clip context).

	YouCook2 (S2C)			ActivityNet CS (S2C)			EPIC-KITCHENS-100 (S2C)		
m	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum
0	9.549	0.04	108.1	8.675	0.02	40.5	16.903	0.06	104.4
1	9.550	0.18	116.2	8.676	0.06	54.7	16,904	0.22	124.9
2	9.551	0.28	118.5	8.677	0.10	59.1	16,905	0.36	134.6
3	9.552	0.40	116.8	8.677	0.14	60.6	16,906	0.50	138.5
4	9,553	0.52	115.7	8.678	0.16	60.8	16,907	0.64	141.5
5	9,554	0.62	113.2	8.679	0.20	60.4	16,908	0.78	142.6

Table 14: Analysis of Complexity text context.

	YouCook2 (S2C)			ActivityNet CS (S2C)			EPIC-KITCHENS-100 (S2C)		
m	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum	#Param.(M)	Flops(G)	RSum
0	9.549	0.04	108.1	8.675	0.02	40.5	16.903	0.06	104.4
1	9.551	0.22	187.4	8.677	0.10	103.1	16.905	0.30	200.3
2	9.553	0.36	201.4	8.678	0.18	131.4	16.907	0.50	228.8
3	9.555	0.50	207.1	8.680	0.24	141.9	16.909	0.70	243.7
4	9.558	0.64	206.6	8.681	0.32	144.8	16.911	0.90	254.5
5	9.560	0.78	204.9	8.683	0.38	143.6	16.913	1.10	257.9

Table 15: Analysis of Complexity both context.

using context in both modalities is the most computationally expensive setting as shown in Table 15.

In general, the performance tends to be the comparable or drops marginally for $m > 3$ over all datasets/scenarios. We can argue that using a $1 \leq m \leq 3$ is a good trade-off between performance and computational complexity.

References

1. Chun, S., Oh, S.J., de Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
2. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives. In: British Machine Vision Conference (BMVC) (2018)
3. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
4. Patrick, M., Asano, Y.M., Kuznetsova, P., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. In: International Conference on Computer Vision (ICCV) (2021)
5. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)