Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings Supplementary Material

Michael Wray
University of Bristol

Diane Larlus
Naver Labs Europe

Gabriela Csurka
Naver Labs Europe

Dima Damen
University of Bristol

EPIC	SEEN			
	vv	vt	tv	tt
Random Baseline	12.6	12.6	12.6	12.6
Features(Word2Vec)	–	–	–	50.0
Features(Video)	21.0	–	–	–
CCA Baseline	21.3	23.3	25.7	37.7
MMEN(Caption)	32.0	53.1	47.2	90.0
MMEN([Verb,Noun])	33.2	55.7	48.9	96.1
MMEN(Caption RNN)	31.2	33.7	49.2	92.6
PoS-MMEN(Verb)	31.1	56.2	48.5	**97.1**
JPoSE	**33.7**	**57.1**	**49.9**	**97.1**

Table 1. Verb retrieval task results on the seen test set of EPIC-Kitchens.

EPIC	SEEN			
	vv	vt	tv	tt
Random Baseline	2.17	2.17	2.17	2.17
Features(Word2Vec)	–	–	–	30.9
Features(Video)	10.6	–	–	–
CCA Baseline	11.9	16.9	19.2	52.2
MMEN(Caption)	**18.7**	26.2	20.7	70.9
MMEN([verb,Noun])	18.3	29.8	23.8	90.1
MMEN(Caption RNN)	17.9	20.3	22.0	74.0
PoS-MMEN(Noun)	17.8	31.5	23.6	**92.6**
JPoSE	18.6	**32.2**	**25.5**	**92.6**

Table 2. Noun retrieval task results on the seen test set of EPIC-Kitchens.

EPIC	SEEN		UNSEEN	
	vt	tv	vt	tv
JPoSE(Verb,Noun)*	18.0	13.4	11.5	8.8
JPoSE(Verb,Noun)	**23.2**	**15.8**	**14.6**	**10.2**

Table 3. Cross-modal retrieval results - compared closed (*) to open vocabulary embedding.

1. Individual Part-of-Speech Retrieval (Sec. 3.3)

In the main manuscript, we report results on the task of *fine-grained action retrieval*. For completion, we here present results on individual Part-of-Speech (PoS) retrieval tasks.

In Table 1, we report results for *fine-grained verb retrieval* (*i.e.* only retrieve the relevant verb/action in the video). We include the standard baselines and we additionally report the results obtained by a PoS-MMEN, that is a single embedding for verbs solely. We compare this to our proposed multi-embedding JPoSE. Using JPoSE produces better (or the same) results for both cross-modal and within-modal searches.

Similarly, in Table 2, we compare results for *fine-grained noun retrieval* (*i.e.* only retrieve the relevant noun/object in the video). We show similar increases in mAP over cross-modal and within-modal searches. This indicates the complementary PoS information, from the other PoS embedding as well as the PoS-aware action embedding, helps to better define the individual embedding space.

2. Closed vs Open Vocabulary Embedding

Table 3 compares to JPoSE* trained using only the closed vocabulary of EPIC. In this setup, closed vocabulary was used for building the embedding, but open vocabulary used for testing. Results show that using the full open vocabulary in training yields a sizeable benefit.

3. Text embedding Using RNN

We provide here the results of replacing the text embedding function, g, with an RNN instead of the two layer perceptron for the MMEN method. The RNN was modelled as a Gated Recurrent Unit (GRU). Captions were capped and zero-padded to a maximum length of 15 words. Adding a layer on top of the GRU proved not to be useful. Results of the RNN in the experiments are given under the name MMEN (Caption RNN). Given the singular verb and low noun count RNNs were not tested for the individual PoS-MMENs.

Cross-Modal and Within-Modal Results can be seen in Tables 4 and 5 respectively. The inclusion of the RNN sees improvements in mAP performance for tv, vv and tt compared to MMEN (caption). However, compared to

EPIC	SEEN		UNSEEN	
	vt	tv	vt	tv
Random Baseline	0.6	0.6	0.9	0.9
CCA Baseline	20.6	7.3	14.3	3.7
MMEN (Caption)	14.0	11.2	10.1	7.7
MMEN (Caption RNN)	10.3	13.8	6.3	9.0
MMEN ([Verb, Noun])	18.7	13.6	13.3	9.5
JPoSE(Verb,Noun)	**23.2**	**15.8**	**14.6**	**10.2**

Table 4. Cross-modal action retrieval results on EPIC including MMEN(Caption RNN).

EPIC	SEEN		UNSEEN	
	vv	tt	vv	tt
Random Baseline	0.6	0.6	0.9	0.9
CCA Baseline	13.8	62.2	18.9	68.5
Features(Word2Vec)	–	62.5	–	71.3
Features(Video)	13.6	–	21.0	–
MMEN (Caption)	17.2	63.8	20.7	69.6
MMEN (Caption RNN)	17.6	73.5	22.1	76.1
MMEN ([Verb, Noun])	17.6	83.5	22.5	84.7
JPoSE(Verb,Noun)	**18.8**	**87.7**	**23.2**	**87.7**

Table 5. Within-modal action retrieval results on EPIC including MMEN(Caption RNN).

MMEN ([Verb,Noun]) or JPoSE (Verb,Noun) using the entire caption still leads to worse results for both cross and within modal retrieval.

4. Additional MSR-VTT Experiments (Sec. 4.2)

Table 6 of this supplementary is an expanded version of Table 7 in the main paper testing a variety of different combinations for PoS. For each row, an average of 10 runs is reported. This experiment also includes the removal of the NetVLAD layer in the MMEN, substituting it with mean pooling which we label as AVG.

Results show that, on their own, Determinants, Adjectives and Adpositions achieve very poor results. We also report three JPoSE disentanglement options: (Verb, Noun), (Caption\Verb, Verb) and the one in the main paper (Capiton\Noun, Noun). The table shows that the best results are achieved when nouns are disentangled from the rest of the caption.

References

[1] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 3

MSR-VTT Retrieval	Video-to-text R@1	R@5	R@10	MR	Text-to-Video R@1	R@5	R@10	MR
Mixture of Experts [1]*	–	–	–	–	12.9	36.4	51.8	10
Random Baseline	0.3	0.7	1.1	502	0.3	0.7	1.1	502
CCA Baseline	2.8	5.6	8.2	283	7.0	14.4	18.7	100
MMEN(DET AVG)	0.0	0.2	0.5	214	0.3	1.0	2.2	264
MMEN(ADJ AVG)	0.0	0.3	0.7	216	0.1	1.1	2.6	260
MMEN(ADP AVG)	0.1	0.6	1.5	172	0.7	2.8	5.0	185
MMEN(Verb AVG)	1.1	5.4	11.1	57	3.2	10.9	17.4	57
MMEN(Noun AVG)	10.0	28.0	40.0	16	10.7	29.7	43.5	15
MMEN(DET NetVLAD)	0.0	0.1	0.3	241	0.1	1.1	2.4	255
MMEN(ADJ NetVLAD)	0.0	0.0	0.1	232	0.2	1.2	2.0	262
MMEN(ADP NetVLAD)	0.1	0.7	1.5	174	0.6	2.9	4.9	190
MMEN(Verb NetVLAD)	0.7	4.0	8.3	70	2.9	7.9	13.9	63
MMEN(Noun NetVLAD)	10.8	31.3	42.7	14	10.8	30.7	44.5	13
MMEN([V, N, DET] AVG)	9.0	28.4	41.0	15	7.7	24.2	36.0	20
MMEN([Verb,Noun] AVG)	12.9	34.0	46.7	12	12.6	32.6	46.3	12
MMEN([V, N, ADP] AVG)	13.0	33.0	46.0	13	12.2	33.0	46.0	13
MMEN([V, N, ADJ] AVG)	12.4	32.9	45.3	13	11.0	31.2	44.3	13
MMEN([V, N, ADJ, ADP] AVG)	13.0	32.3	45.9	12	11.1	31.5	44.3	13
MMEN([V, N, DET] NetVLAD)	14.8	38.3	52.5	9.1	12.4	33.6	46.3	13
MMEN([Verb,Noun] NetVLAD)	15.6	39.4	55.1	9.0	13.6	36.8	51.7	10
MMEN([V, N, ADP] NetVLAD)	15.8	40.3	55.1	8.5	13.8	36.7	51.0	10
MMEN([V, N, ADJ] NetVLAD)	16.3	40.1	54.1	8.9	14.0	36.2	50.9	10
MMEN([V, N, ADJ, ADP] NetVLAD)	16.1	39.7	53.8	8.9	13.4	36.2	51.3	10
MMEN(Caption AVG)	12.4	32.8	45.6	12	11.4	31.2	43.8	14
MMEN(Caption NetVLAD)	15.8	40.2	53.6	9	13.8	36.7	50.7	10.3
JPoSE(Verb, Noun)	15.5	39.3	53.8	9	13.7	37.6	52.2	9.6
JPoSE(Caption\Verb,Verb)	15.9	39.2	**55.5**	**8**	13.4	36.8	52.0	10
JPoSE(Caption\Noun,Noun)	**16.4**	**41.3**	54.4	8.7	**14.3**	**38.1**	**53.0**	**9**

Table 6. MSR-VTT Video-Caption Retrieval results using recall@k (R@k, higher is better) and median Rank (MR, lower is better). For each row, an average of 10 runs is reported. *We include results from [1], only available for Text-to-Video retrieval.