Fine-Grained Retrieval Through Multiple Parts-of-Speech Embeddings

University of BRISTOL

NAVER LABS EUROPE

ICCV 2019 Seoul, Korea

Michael Wray **Diane Larlus** **Gabriela Csurka** **Dima Damen**

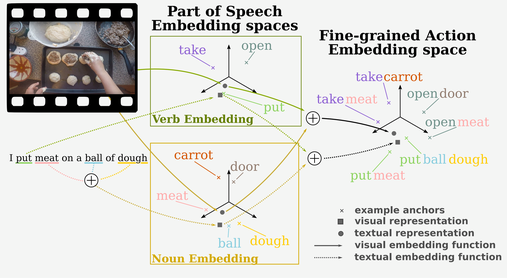
Motivation

Fine-Grained Action Retrieval

I put meat on a ball of dough
I place meat
I distribute minced meat
I put down meat
I move meat

This problem is related, but differs from both fine-grained action recognition and general video retrieval.

Set-Up

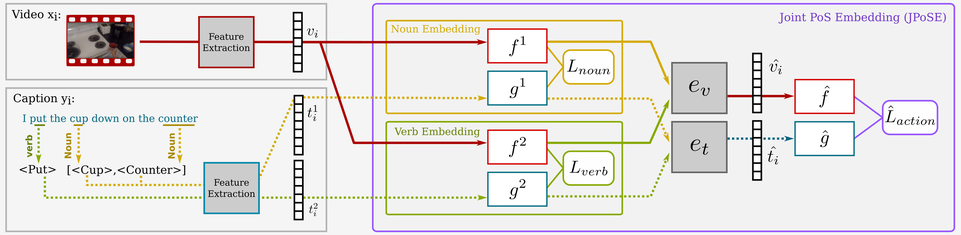


Results

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 (Verb)	3.6	4.0	3.9	4.2
MMEN (Noun)	9.9	9.2	7.9	6.1
MMEN (Caption)	14.0	11.2	10.1	7.7
MMEN ([Verb, Noun])	18.7	13.6	13.3	9.5
JPoSE (Verb, Noun)	**23.2**	**15.8**	**14.6**	**10.2**

Creating explicit embeddings for each PoS and combining them allows for a better learning of the long-tail of word occurences in EPIC-Kitchens and hence to be more discriminative than direct caption or part of caption embedding (MMEN).

Method

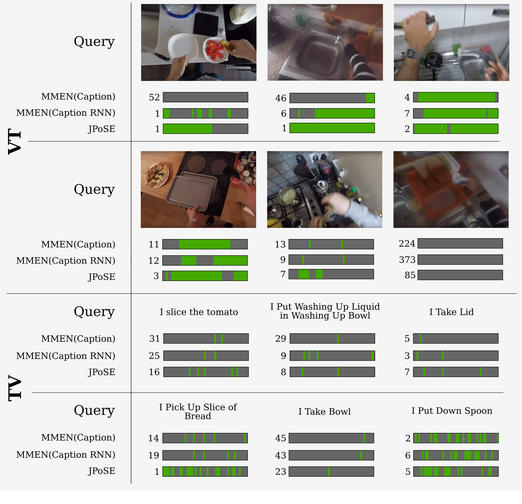


The caption is parsed and split into its constituent Parts of Speech and a corresponding embedding is learned for each. Note that the same visual features are used for both.

To move from Part of Speech Embeddings to the action embedding, an encoding function combines the output and a shared weights layer embeds the final representation.

Qualitative Results

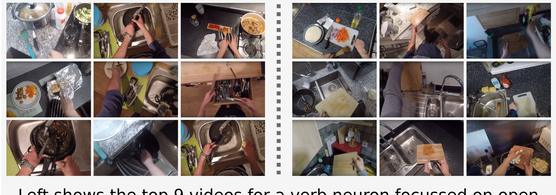
Using JPoSE results in lower ranks of the first correct retrieval (numbered) and more relevant retrievals within the top 50 (relevant shown in green).



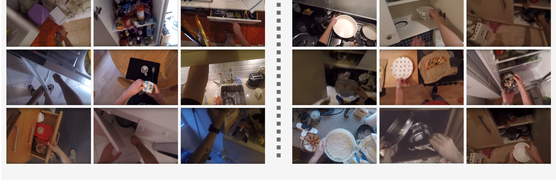
Neuron Activations

The visual embedding functions can be inspected using maximal neuron activations and show the different embeddings they are learning.

Left shows the top 9 videos for a noun neuron focussed on cutlery, right shows a noun neuron for a neuron focussed on cutting boards.



Left shows the top 9 videos for a verb neuron focussed on open and close, right shows a verb neuron for a neuron focussed on take and put.



Ablation Study

EPIC Learn	$\hat{L}$	(e_v,e_t)	$(\hat{f},\hat{g})$	SEEN vv	vt	tv	tt
indep	×	Sum	×	17.4	20.7	13.3	86.5
indep	×	Max	×	17.5	21.2	13.3	86.5
indep	×	Conc.	×	18.3	21.5	14.6	87.1
joint	✓	Sum	(Id, Id)	18.1	21.0	14.3	87.3
joint	✓	Max	(Id, Id)	18.1	22.4	14.8	87.5
joint	✓	Conc.	(Id, Id)	18.3	22.7	15.4	87.6
joint	✓	Conc.	$(\hat{\theta}_{\hat{f}}, \hat{\theta}_{\hat{g}})$	**18.8**	**23.2**	**15.8**	**87.7**

The ablation study on the seen test set of EPIC-Kitchens shows the benefit of the joint learning and the final embedding function θ. Concatenating the Part of Speech Embeddings leads to best performance.

Zero-Shot Results

Zero-shot action retrieval with combinations of either verbs (ZSV) or nouns not seen during training can be performed.

Verbs/nouns unseen during training and seen nouns/verbs make up the ZSV/ZSN test sets.

EPIC	ZSV		ZSN	
	vt	tv	vt	tv
MMEN (Caption)	5.77	5.51	4.17	3.32
JPoSE	**7.50**	**6.47**	**7.68**	**6.17**

By disentangling the caption and learning embedding spaces for verbs and nouns, JPoSE is able to better infer the "missing" part than the classical caption embedding model.

Conclusion

Learning distinct embeddings for each Part-Of-Speech allows for the creation of a space suitable for fine-grained action retrieval.

When tested on EPIC-Kitchens using the open vocabulary labels, the proposed JPoSE method is able to outperform embeddings trained on the caption alone for both many and zero-shot tasks.

We have shown that the notion of disentangling can also be used with success for general video-retrieval task (see experiments with MSR-VTT in the paper)