Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings

Michael Wray
University of Bristol

Diane Larlus
Naver Labs Europe

Gabriela Csurka
Naver Labs Europe

Dima Damen
University of Bristol

Abstract

We address the problem of cross-modal fine-grained action retrieval between text and video. Cross-modal retrieval is commonly achieved through learning a shared embedding space, that can indifferently embed modalities. In this paper, we propose to enrich the embedding by disentangling parts-of-speech (PoS) in the accompanying captions. We build a separate multi-modal embedding space for each PoS tag. The outputs of multiple PoS embeddings are then used as input to an integrated multi-modal space, where we perform action retrieval. All embeddings are trained jointly through a combination of PoS-aware and PoS-agnostic losses. Our proposal enables learning specialised embedding spaces that offer multiple views of the same embedded entities.

We report the first retrieval results on fine-grained actions for the large-scale EPIC dataset, in a generalised zero-shot setting. Results show the advantage of our approach for both video-to-text and text-to-video action retrieval. We also demonstrate the benefit of disentangling the PoS for the generic task of cross-modal video retrieval on the MSR-VTT dataset.

1. Introduction

With the onset of the digital age, millions of hours of video are being recorded and searching this data is becoming a monumental task. It is even more tedious when searching shifts from video-level labels, such as 'dancing' or 'skiing', to short action segments like 'cracking eggs' or 'tightening a screw'. In this paper, we focus on the latter and refer to them as fine-grained actions. We thus explore the task of *fine-grained action retrieval* where both queries and retrieved results can be either a video sequence, or a textual caption describing the fine-grained action. Such free-form action descriptions allow for a more subtle characterisation of actions but require going beyond training a classifier on a predefined set of action labels [20, 30].

As is common in cross-modal search tasks [26, 36], we learn a shared embedding space onto which we project both videos and captions. By nature, fine-grained actions can be described by an actor, an act and the list of objects involved in the interaction. We thus propose to learn a separate embedding for each part-of-speech (PoS), such as for instance verbs, nouns or adjectives. This is illustrated in Fig. 1 for two PoS (verbs and nouns). When embedding verbs solely, relevant entities are those that share the same verb/act regardless of the nouns/objects used. Conversely, for a PoS embedding focusing on nouns solely, different actions performed on the same object are considered relevant entities. This enables a PoS-aware embedding, specialised for retrieving a variety of relevant/irrelevant entities, given that PoS. The outputs from the multiple PoS embedding spaces are then combined within an encoding module that produces the final action embedding. We train our approach end-to-end, jointly optimising the multiple PoS embeddings and the final fine-grained action embedding.

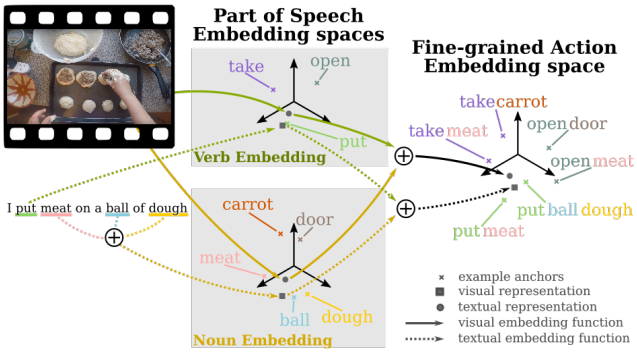

Figure 1. We target fine-grained action retrieval. Action captions are broken using part-of-speech (PoS) parsing. We create separate embedding spaces for the relevant PoS (*e.g.* Noun or Verb) and then combine these embeddings into a shared embedding space for action retrieval (best viewed in colour).

This approach has a number of advantages over training a single embedding space as is standardly done [7, 8, 15, 22, 24]. Firstly, this process builds different embeddings that can be seen as different views of the data, which contribute to the final goal in a collaborative manner. Secondly, it allows to inject, in a principled way, additional information but without requiring additional annotation, as parsing a caption for PoS is done automatically. Finally, when considering a single PoS at a time, for instance verbs, the cor-

responding PoS-embedding learns to generalise across the variety of actions involving each verb (*e.g*. the many ways 'open' can be used). This generalisation is key to tackling more actions including new ones not seen during training.

We present the first retrieval results for the recent large-scale EPIC dataset [6] (Sec 4.1), utilising the released free-form narrations, previously unexplored for this dataset, as our supervision. Additionally, we show that our second contribution, learning PoS-aware embeddings, is also valuable for general video retrieval by reporting results on the MSR-VTT dataset [39] (Sec. 4.2).

2. Related Work

Recently, neural networks trained with a ranking loss considering image pairs [27], triplets [35], quadruplets [5] or beyond [32], have been considered for metric learning [17, 35] and for a broad range of search tasks such as face/person identification [29, 5, 16, 2] or instance retrieval [10, 27]. These learning-to-rank approaches have been generalised to two or more modalities. Standard examples include building a joint embedding for images and text [11, 36], videos and audio [33] and, more related to our work, for videos and action labels [15], videos and text [8, 14, 40] or some of those combined [25, 24, 22].

Representing text. Early works in image-to-text cross-modal retrieval [9, 11, 36] used TF-IDF as a weighted bag-of-words model for text representations (either from a word embedding model or one-hot vectors) in order to aggregate variable length text captions into a single fixed sized representation. With the advent of neural networks, works shifted to use RNNs, Gated Recurrent Units (GRU) or Long Short-Term Memory (LSTM) units to extract textual features [8] or to use these models within the embedding network [15, 18, 24, 25, 34] for both modalities.

Action embedding and retrieval. Joint embedding spaces are a standard tool to perform action retrieval. Zhang *et al*. [42] use a Semantic Similarity Embedding (SSE) in order to perform action recognition. Their method, inspired by sparse coding, splits train and test data into a mixture of proportions of already seen classes which then generalises to unseen classes at test time. Mithun *et al*. [24] create two embedding spaces: An activity space using flow and audio along with an object space from RGB. Their method encodes the captions with GRUs and the output vectors from the activity and object spaces are concatenated to rank videos. We instead create Part of Speech embedding spaces which we learn jointly, allowing our method to capture relationships between *e.g*. verbs and nouns.

Hahn *et al*. [15] use two LSTMs to directly project videos into the Word2Vec embedding space. This method is evaluated on higher-level activities, showing that such a visual embedding aligns well with the learned space of Word2Vec to perform zero-shot recognition of these coarser-grained classes. Miech *et al*. [21] found that using NetVLAD [3] results in an increase in accuracy over GRUs or LSTMs for aggregation of both visual and text features. A follow up on this work [22] learns a mixture of experts embedding from multiple modalities such as appearance, motion, audio or face features. It learns a single output embedding which is the weighted similarity between the different implicit visual-text embeddings. Recently, Miech *et al*. [23] propose the HowTo100M dataset: A large dataset collected automatically using generated captions from youtube of 'how to tasks'. They find that fine-tuning on the weakly-paired video clips allows for state-of-the-art performance on a number of different datasets.

Fine-grained action recognition. Recently, several large-scale datasets have been published for the task of fine-grained action recognition [6, 12, 13, 31, 28]. These generally focus on a closed vocabulary of class labels describing short and/or specific actions.

Rohrbach *et al*. [28] investigate hand and pose estimation techniques for fine-grained activity recognition. By compositing separate actions, and treating them as attributes, they can predict unseen activities via novel combinations of seen actions. Mahdisoltani *et al*. [20] train for four different tasks, including both coarse and fine grain action recognition. They conclude that training on fine-grain labels allows for better learning of features for coarse-grain tasks.

In our previous work [38], we explored action retrieval and recognition using multiple verb-only representations, collected via crowd-sourcing. We found that a soft-assigned representation was beneficial for retrieval tasks over using the full verb-noun caption. While the approach enables scaling to a broader vocabulary of action labels, such multi-verb labels are expensive to collect for large datasets.

While focusing on fine-grained actions, we diverge from these works using open vocabulary captions for supervision. As recognition is not suitable, we instead formulate this as a retrieval problem. Up to our knowledge, no prior work attempted cross-modal retrieval on fine-grained actions. Our endeavour has been facilitated by the recent release of open vocabulary narrations on the EPIC dataset [6] which we note is the only fine-grained dataset to do so. While our work is related to both *fine-grained action recognition* and *general action retrieval*, we emphasise that it is neither. We next describe our proposed model.

3. Method

Our aim is to learn representations suitable for cross-modal search where the query modality is different from the target modality. Specifically, we use video sequences with textual captions/descriptions and perform *video-to-text* (vt) or *text-to-video* (tv) retrieval tasks. Additionally, we would like to make sure that classical search (where the query and

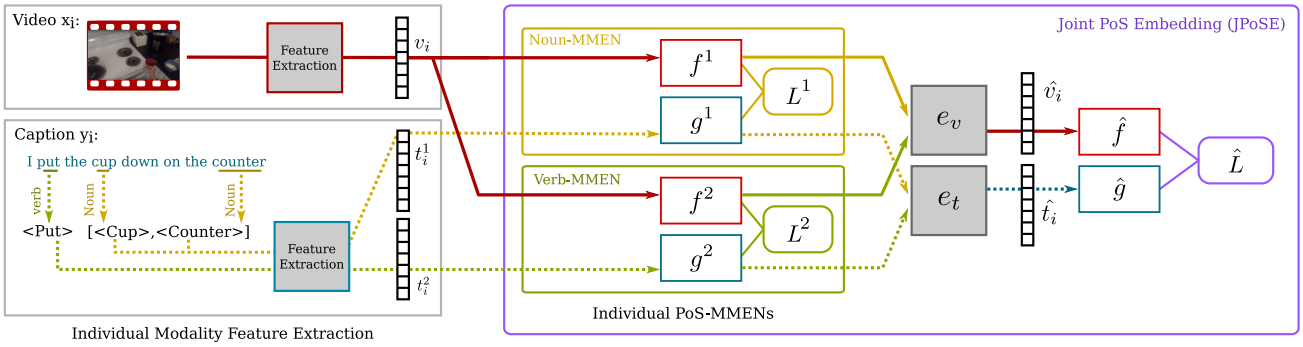

Figure 2. Overview of the JPoSE model. We first disentangle a caption into its parts of speech (PoS) and learn a Multi-Modal Embedding Network (MMEN, Sec. 3.1) for each PoS (Sec. 3.2). The output of these PoS-MMENs are then encoded (e_v, e_t) to get new representations $\hat{v}_i$ and $\hat{t}_i$ on top of which the final embeddings $\hat{f}$ and $\hat{g}$ are learnt. JPoSE learns all of those jointly (Sec. 3.3), using a combination of PoS-aware L^1, L^2, defined in Eq. (5) and PoS-agnostic $\hat{L}$ losses, defined in Eq. (6). Non-trained modules are shown in grey.

the retrieved results have the same modalities) could still be performed in that representation space. The latter are referred to as *video-to-video* (vv) and *text-to-text* (tt) search tasks. As discussed in the previous section, several possibilities exist, the most common being embedding both modalities in a shared space such that, regardless of the modality, the representation of two relevant entities in that space are close to each other, while the representation of two non-relevant entities are far apart.

We first describe how to build such a joint embedding between two modalities, enforcing both cross-modal and within-modal constraints (Sec. 3.1). Then, based on the knowledge that different parts of the caption encode different aspects of an action, we describe how to leverage this information and build several disentangled Part of Speech embeddings (Sec. 3.2). Finally, we propose a unified representation well-suited for fine-grained action retrieval (Sec. 3.3).

3.1. Multi-Modal Embedding Network (MMEN)

This section describes a Multi-Modal Embedding Network (MMEN) that encodes the video sequence and the text caption into a common descriptor space.

Let $\{(v_i, t_i) | v_i \in V, t_i \in T\}$ be a set of videos with v_i being the visual representation of the i^{th} video sequence and t_i the corresponding textual caption. Our aim is to learn two embedding functions $f : V \rightarrow \Omega$ and $g : T \rightarrow \Omega$, such that $f(v_i)$ and $g(t_i)$ are close in the embedded space Ω. Note that f and g can be linear projection matrices or more complex functions *e.g.* deep neural networks. We denote the parameters of the embedding functions f and g by θ_f and θ_g respectively, and we learn them jointly with a weighted combination of two cross-modal ($L_{v,t}, L_{t,v}$) and two within-modal ($L_{v,v}, L_{t,t}$) triplet losses. Note that other point-wise, pairwise or list-wise losses can also be considered as alternatives to the triplet loss.

The **cross-modal losses** are crucial to the task and ensure that the representations of a query and a relevant item for that query from a different modality are closer than the representations of this query and a non-relevant item. We use cross-modal triplet losses [19, 36]:

$$L_{v,t}(\theta) = \sum_{(i,j,k)\in\mathcal{T}_{v,t}} \max\left(\gamma + d(f_{v_i}, g_{t_j}) - d(f_{v_i}, g_{t_k}),\, 0\right)$$
$$\mathcal{T}_{v,t} = \{(i,j,k)\,|\, v_i \in V, t_j \in T_{i+}, t_k \in T_{i-}\} \quad (1)$$

$$L_{t,v}(\theta) = \sum_{(i,j,k)\in\mathcal{T}_{t,v}} \max\left(\gamma + d(g_{t_i}, f_{v_j}) - d(g_{t_i}, f_{v_k}),\, 0\right)$$
$$\mathcal{T}_{t,v} = \{(i,j,k)\,|\, t_i \in T, v_j \in V_{i+}, v_k \in V_{i-}\} \quad (2)$$

where γ is a constant margin, $\theta = [\theta_f, \theta_g]$, and $d(.)$ is the distance function in the embedded space Ω. T_{i+}, T_{i-} respectively define sets of relevant and non relevant captions and V_{i+}, V_{i-} the sets of relevant and non relevant videos sequences for the multi-modal object (v_i, t_i). To simplify the notation, f_{v_i} denotes $f(v_i) \in \Omega$ and g_{t_j} denotes $g(t_j) \in \Omega$.

Additionally, **within-modal losses**, also called structure preserving losses [19, 36], ensure that the neighbourhood structure within each modality is preserved in the newly built joint embedding space. Formally,

$$L_{v,v}(\theta) = \sum_{(i,j,k)\in\mathcal{T}_{v,v}} \max\left(\gamma + d(f_{v_i}, f_{v_j}) - d(f_{v_i}, f_{v_k}),\, 0\right)$$
$$\mathcal{T}_{v,v} = \{(i,j,k)\,|\, v_i \in V, v_j \in V_{i+}, v_k \in V_{i-}\} \quad (3)$$

$$L_{t,t}(\theta) = \sum_{(i,j,k)\in\mathcal{T}_{t,t}} \max\left(\gamma + d(g_{t_i}, g_{t_j}) - d(g_{t_i}, g_{t_k}),\, 0\right)$$
$$\mathcal{T}_{t,t} = \{(i,j,k)\,|\, t_i \in T, t_j \in T_{i+}, t_k \in T_{i-}\} \quad (4)$$

using the same notation as before. The final loss used for the MMEN network is a weighted combination of these four losses, summed over all triplets in $\mathcal{T}$ defined as follows:

$$L(\theta) = \lambda_{v,v}L_{v,v} + \lambda_{v,t}L_{v,t} + \lambda_{t,v}L_{t,v} + \lambda_{t,t}L_{t,t} \quad (5)$$

where λ is a weighting for each loss term.

3.2. Disentangled Part of Speech Embeddings

The previous section described the generic Multi-Modal Embedding Network (MMEN). In this section, we propose to disentangle different caption components so each component is encoded independently in its own embedding space. To do this, we first break down the text caption into different PoS tags. For example, the caption *"I divided the onion into pieces using wooden spoon"* can be divided into verbs, $[divide, using]$, pronouns, $[I]$, nouns, $[onion, pieces, spoon]$ and adjectives, $[wooden]$. In our experiments, we focus on the most relevant ones for fine-grained action recognition: verbs and nouns, but we explore other types for general video retrieval. We extract all words from a caption for a given PoS tag and train one MMEN to only embed these words and the video representation in the same space. We refer to it as a PoS-MMEN.

To train a PoS-MMEN, we propose to adapt the notion of relevance specifically to the PoS. This has a direct impact on the sets $V_{i+}, V_{i-}, T_{i+}, T_{i-}$ defined in Equations (1)-(4). For example, the caption *'cut tomato'* is disentangled into the verb *'cut'* and the noun *'tomato'*. Consider a PoS-MMEN focusing on verb tags solely. The caption *'cut carrots'* is a relevant caption as the pair share the same verb *'cut'*. In another PoS-MMEN focusing on noun tags solely, the two remain irrelevant. As the relevant/irrelevant sets differ within each PoS-MMEN, these embeddings specialise to that PoS.

It is important to note that, although the same visual features are used as input for all PoS-MMEN, the fact that we build one embedding space per PoS trains multiple visual embedding functions f^k that can be seen as multiple views of the video sequence.

3.3. PoS-Aware Unified Action Embedding

The previous section describes how to extract different PoS from captions and how to build PoS-specific MMENs. These PoS-MMENs can already be used alone for PoS-specific retrieval tasks, for instance a verb-retrieval task (*e.g.* retrieve all videos where "cut" is relevant) or a noun-retrieval task.[1] More importantly, the output of different PoS-MMENs can be combined to perform more complex tasks, including the one we are interested in, namely fine-grained action retrieval.

[1]Our experiments focus on action retrieval but we report on these other tasks in the supplementary material.

Let us denote the k^{th} PoS-MMEN visual and textual embedding functions by $f^k : V \rightarrow \Omega^k$ and $g^k : T \rightarrow \Omega^k$. We define:

$$\begin{aligned} \hat{v}_i &= e_v(f^1_{v_i}, f^2_{v_i}, \ldots, f^K_{v_i}) \\ \hat{t}_i &= e_t(g^1_{t_i}, g^2_{t^2_i}, \ldots, g^K_{t^K_i}) \end{aligned} \quad (6)$$

where e_v and e_t are encoding functions that combine the outputs of the PoS-MMENs. We explore multiple pooling functions for e_v and e_t: *concatenation*, *max*, *average* - the latter two assume all Ω_k share the same dimensionality.

When $\hat{v}_i$, $\hat{t}_i$ have the same dimension, we can perform action retrieval by directly computing the distance between these representations. We instead propose to train a final PoS-agnostic MMEN that unifies the representation, leading to our final JPoSEmodel.

Joint Part of Speech Embedding (JPoSE). Considering the PoS-aware representations $\hat{v}_i$ and $\hat{t}_i$ as input and, still following our learning to rank approach, we learn the parameters $\hat{\theta}_{\hat{f}}$ and $\hat{\theta}_{\hat{g}}$ of the two embedding functions $\hat{f} : \hat{V} \rightarrow \Gamma$ and $\hat{g} : \hat{T} \rightarrow \Gamma$ which project in our final embedding space Γ. We again consider this as the task of building a single MMEN with the inputs $\hat{v}_i$ and $\hat{t}_i$, and follow the process described in Sec. 3.1. In other words, we train using the loss defined in Equation (5), which we denote $\hat{L}$ here, which combines two cross-modal and two within-modal losses using the triplets $\mathcal{T}_{v,t}, \mathcal{T}_{t,v}, \mathcal{T}_{v,v}, \mathcal{T}_{t,t}$ formed using relevance between videos and captions based on the action retrieval task. As relevance here is not PoS-aware, we refer to this loss as PoS-agnostic. This is illustrated in Fig. 2.

We learn the multiple PoS-MMENs and the final MMEN jointly with the following combined loss:

$$L(\hat{\theta}, \theta^1, \ldots \theta^K) = \hat{L}(\hat{\theta}) + \sum_{k=1}^{K} \alpha^k L^k(\theta^k) \quad (7)$$

where α^k are weighting factors, $\hat{L}$ is the PoS-agnostic loss described above and L^k are the PoS-aware losses corresponding to the K PoS-MMENs.

4. Experiments

We first tackle fine-grained action retrieval on the EPIC dataset [6] (Sec. 4.1) and then the general video retrieval task on the MSR-VTT dataset [39] (Sec. 4.2). This allows us to explore two different tasks using the proposed multi-modal embeddings.

The large English spaCy parser [1] was used to find the Part Of Speech (PoS) tags and disentangle them in the captions of both datasets. Statistics on the most frequent PoS tags are shown in Table 1. As these statistics show, EPIC contains mainly nouns and verbs, while MSR-VTT

has longer captions and more nouns. This will have an impact of the PoS chosen for each dataset when building the JPoSE model.

4.1. Fine-Grained Action Retrieval on EPIC

Dataset. The EPIC dataset [6] is an egocentric dataset with 32 participants cooking in their own kitchens who then narrated the actions in their native language. The narrations were translated to English but maintain the open vocabulary selected by the participants. We employ the released free-form narrations to use this dataset for fine-grained action retrieval. We follow the provided train/test splits. Note that by construction there are two *Seen* and *Unseen* test sets, referring to whether the kitchen has been seen in the training set. We follow the terminology from [6], and note that this terminology should not be confused with the zero-shot literature which distinguishes seen/unseen classes. The actual sequences are strictly disjoint between all sets.

Additionally, we train only on the many-shot examples from EPIC excluding all examples of the few shot classes from the training set. This ensures each action has more than 100 relevant videos during training and increases the number of zero-shot examples in both test sets.

Building relevance sets for retrieval. The EPIC dataset offers an opportunity for fine-grained action retrieval, as the open vocabulary has been grouped into semantically relevant verb and noun classes for the action recognition challenge. For example, *'put'*, *'place'* and *'put-down'* are grouped into one class. As far as we are aware, this paper presents the first attempt to use the open vocabulary narrations released to the community.

We determine retrieval relevance scores from these semantically grouped verb and noun classes[2], defined in [6]. These indicate which videos and captions should be considered related to each other. Following these semantic groups, a query *'put mug'* and a video with *'place cup'* in its caption are considered relevant as *'place'* and *'put'* share the same verb class and *'mug'* and *'cup'* share the same noun class. Subsequently, we define the triplets $\mathcal{T}_{v,t}, \mathcal{T}_{t,v}, \mathcal{T}_{v,v}, \mathcal{T}_{t,t}$ used to train the MMEN models and to compute the loss $\hat{L}$ in JPoSE.

When training a PoS-MMEN, two videos are considered relevant only within that PoS. Accordingly, *'put onion'* and *'put mug'* are relevant for verb retrieval, whereas, *'put cup'* and *'take mug'* are for noun retrievals. The corresponding PoS-based relevances define the triplets $\mathcal{T}^k$ for L^k.

4.1.1 Experimental Details

Video features. We extract flow and appearance features using the TSN BNInception model [37] pre-trained on Kinetics and fine-tuned on our training set. TSN averages the features from 25 uniformly sampled snippets within the video. We then concatenate appearance and flow features to create a 2048 dimensional vector (v_i) per action segment.

[2]We use the verb and noun classes purely to establish relevance scores, the training is done with the original open vocabulary captions.

Parts of Speech	EPIC count	EPIC avg/caption	MSR-VTT count	MSR-VTT avg/caption
Noun	34,546	1.21	418,557	3.33
Verb	30,279	1.06	245,177	1.95
Determiner	6,149	0.22	213,065	1.69
Adposition	5,048	0.18	151,310	1.20
Adjective	2,271	0.08	79,417	0.63

Table 1. Statistics of the 5 most common PoS tags in the training sets of both datasets: total counts and average counts per caption.

Text features. We map each lemmatised word to its feature vector using a 100-dimension Word2Vec model, trained on the Wikipedia corpus. Multiple word vectors with the same part of speech were aggregated by averaging. We also experimented with the pre-trained 300-dimension Glove model, and found the results to be similar.

Architecture details. We implement f^k and g^k in each MMEN as a 2 layer perceptron (fully connected layers) with ReLU. Additionally, the input vectors and output vectors are L2 normalised. In all cases, we set the dimension of the embedding space to 256, a dimension we found to be suitable across all settings. We use a single layer perceptron with shared weights for $\hat{f}$ and $\hat{g}$ that we initialised with PCA.

Training details. The triplet weighting parameters are set to $\lambda_{v,v} = \lambda_{t,t} = 0.1$ and $\lambda_{v,t} = \lambda_{t,v} = 1.0$ and the loss weightings α^k are set to 1. The embedding models were implemented in Python using the Tensorflow library. We trained the models with an Adam solver and a learning rate of $1e^{-5}$, considering batch sizes of 256, where for each query we sample 100 random triplets from the corresponding $\mathcal{T}_{v,t}, \mathcal{T}_{t,v}, \mathcal{T}_{v,v}, \mathcal{T}_{t,t}$ sets. The training in general converges after a few thousand iterations, we report all results after 4000 iterations.

Evaluation metrics. We report mean average precision (mAP), *i.e.* for each query we consider the average precision over all relevant elements and take the mean over all queries. We consider each element in the test set as a query in turns. In within-modal retrieval, the query item (video or caption) is removed from the test set for that query.

4.1.2 Results

First, we consider cross-modal and within-modal fine-grained action retrieval. Then, we present an ablation study as well as qualitative results to get more insights. Finally we show that our approach is well-suited for zero-shot settings.

Compared approaches. Across a wide of range of experiments, we compare the proposed **JPoSE** (Sec. 3.3) with

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**

Table 2. Cross-modal action retrieval on EPIC.

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 (Verb)	15.2	11.7	20.1	15.8
MMEN (Noun)	16.8	30.1	21.2	34.1
MMEN (Caption)	17.2	63.8	20.7	69.6
MMEN ([Verb, Noun])	17.6	83.5	22.5	84.7
JPoSE (Verb, Noun)	**18.8**	**87.7**	**23.2**	**87.7**

Table 3. Within-modal action retrieval on EPIC.

some simpler variants based on **MMEN** (Sec. 3.1).

For the captions, we use 1) all the words together without distinction, denoted as 'Caption', 2) only one PoS such as 'Verb' or 'Noun', 3) the concatenation of their respective representations, denoted as '[Verb, Noun]'.

These models are also compared to standard baselines. **Random Baseline** randomly ranks all the database items, providing a lower bound on the mAP scores. The **CCA-baseline** applies Canonical Correlation Analysis to both modalities v_i and t_i to find a joint embedding space for cross-modal retrieval [9]. Finally, **Features (Word2Vec)** and **Features (Video)**, which are only defined for within-modal retrieval (*i.e.* vv and tt), show the performance when we directly use the video representation v_i or the averaged Word2Vec caption representation t_i.

Cross-modal retrieval. Table 2 presents cross-modal results for fine-grained action retrieval. The main observation is that the proposed JPoSE outperforms all the MMEN variants and the baselines for both video-to-text (vt) and text-to-video retrieval (tv), on both test sets. We also note that MMEN ([Verb, Noun]) outperforms other MMEN variants, showing the benefit of learning specialised embeddings. Yet the full JPoSE is crucial to get the best results.

Within-modal retrieval. Table 3 shows the within-modal retrieval results for both text-to-text (tt) and video-to-video (vv) retrieval. Again, JPoSE outperforms all the flavours of MMEN on both test sets. This shows that by learning a cross-modal embedding we inject information from the other modality that helps to better disambiguate and hence to improve the search.

EPIC				SEEN			
Learn	$\hat{L}$	(e_v, e_t)	$(\hat{f}, \hat{g})$	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**

Table 4. Ablation study for JPoSE showing the effects of different encodings, training PoS-MMENs, independently or jointly with $\hat{f}$ and $\hat{g}$ being the identity function Id or being learnt.

Ablation study. We evaluated the role of the components of the proposed JPoSE model, for both cross-modal and within-modal retrieval. Table 4 reports results comparing different options for the encoding functions e_v and e_t in addition to learning the model jointly both with and without learned functions $\hat{f}$ and $\hat{g}$. This confirms that the proposed approach is the best option. In the supplementary material, we also compare the performance when using the closed vocabulary classes from EPIC to learn the embedding. Results demonstrate the benefit of utilising the open vocabulary in training.

Zero-shot experiments. The usage of the open vocabulary in EPIC lends itself well to zero-shot settings. These are the cases for which the open vocabulary verb or noun in the test set is not present in the training set. Accordingly, all previous results can be seen as a Generalised Zero-Shot Learning (GZSL) [4] set-up: there exists both actions in the test sets that have been seen in training and actions that have not. Table 5 shows the zero-shot (ZS) counts in both test sets. In total 12% of the videos in both test sets are zero-shot instances. We separate cases where the noun is present in the training set but the verb is not, denoted by ZSV (zero-shot verb), from ZSN (zero-shot noun) where the verb is present but not the noun. Cross-modal ZS retrieval results for this interesting setting are shown in Table 6. We compare JPoSE to MMEN (Caption) and baselines.

Results show that the proposed JPoSE model clearly improves over these zero-shot settings, thanks to the different views captured by the multiple PoS embeddings, specialised to acts and objects.

Qualitative results. Fig. 3 illustrates both video-to-text and text-to-video retrieval. For several queries, it shows the relevance of the top-50 retrieved items (relevant in green, non-relevant in grey).

Fig. 4 illustrates our motivation that disentangling PoS embeddings would learn different visual functions. It

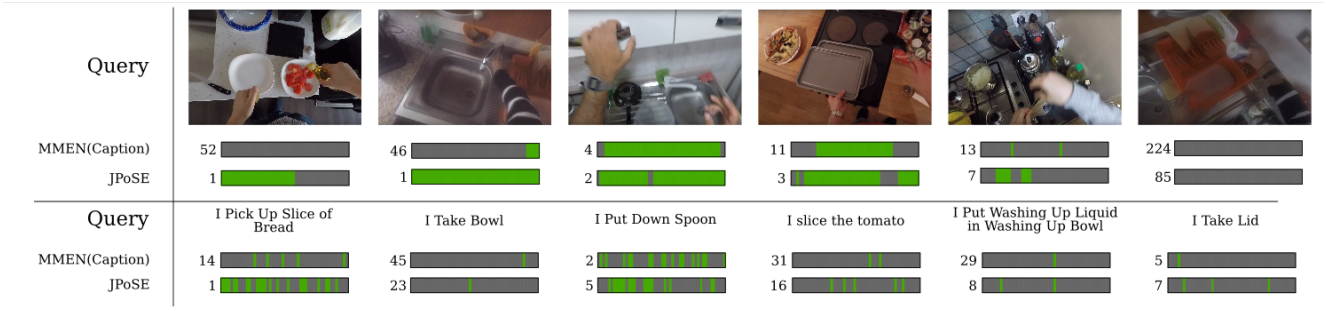

Figure 3. Qualitative results for video-to-text (top) and text-to-video (bottom) action retrieval on EPIC. For several query videos (resp. captions) we show the quality of the top 50 captions (resp. videos) retrieved with green/grey representing relevant/irrelevant. The number in front of the colour-coded bar shows the rank of the first relevant retrieval (lower rank is better).

Figure 4. Maximum activation examples for visual embedding in the noun (left) and the verb (right) PoS-MMEN. Examples of similar objects over different actions are shown in the noun embedding (left) [chopping board vs cutlery] while the same action is shown over different objects in the verb embedding (right) [open/close vs put/take].

EPIC	All			ZSV		ZSN	
	Videos	Verbs	Nouns	Videos	Verbs	Videos	Nouns
Train	26,710	192	1005	–	–	–	–
Seen	8,047	232	530	52	119	367	80
Unseen	2,929	136	243	57	63	275	127

Table 5. Number of videos, and number of open-vocabulary verbs and nouns in the captions, for the three splits of EPIC. For both test sets we also report zero-shot (ZS) instances, showing the number of verbs and nouns that were not seen in the training set, as well as the corresponding number of videos.

EPIC	ZSV		ZSN	
	vt	tv	vt	tv
Random Baseline	1.57	1.57	1.64	1.64
CCA Baseline	2.92	2.96	4.36	3.25
MMEN (Caption)	5.77	5.51	4.17	3.32
JPoSE	**7.50**	**6.47**	**7.68**	**6.17**

Table 6. Zero shot experiments on EPIC.

presents maximum activation examples on chosen neurons within f_i for both verb and noun embeddings. Each cluster represents the 9 videos that respond maximally to one of these neurons. We can remark that noun activations indeed correspond to objects of shared appearance occurring in different actions (in the figure, chopping boards in one and cutlery in the second), while verb embedding neuron activations reflect the same action applied to different objects (open/close *vs*. put/take).

4.2. General Video Retrieval on MSR-VTT

Dataset. We select **MSR-VTT [39]** as a public dataset for general video retrieval. Originally used for video captioning, this large-scale video understanding dataset is increasingly evaluated for video-to-text and text-to-video retrieval [8, 22, 24, 41, 23]. We follow the code and setup of [22] using the same train/test split that includes 7,656 training videos each with 20 different captions describing the scene and 1000 test videos with one caption per video. We also follow the evaluation protocol in [22] and compute recall@k (R@K) and median rank (MR).

In contrast to the EPIC dataset, there is no semantic groupings of the captions in MSR-VTT. Each caption is considered relevant only for a single video, and two captions describing different videos are considered irrelevant even if they share semantic similarities. Furthermore, disentangling captions yields further semantic similarities. For example, "*A cooking tutorial*" and "*A person is cooking*", for a verb-MMEN, will be considered irrelevant as they belong to different videos even though they share the same single verb 'cook'.

Consequently, we can not directly apply JPoSE as proposed in Sec. 3.3. Instead, we adapt JPoSE to this problem as follows. We use the Mixture-of-Expert Embeddings (MEE) model from [22], as our core MMEN network. In fact, MEE is a form of multi-modal embedding network in that it embeds videos and captions into the same space. We

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 [22]*	–	–	–	–	12.9	36.4	51.8	10.0
Random Baseline	0.3	0.7	1.1	502.0	0.3	0.7	1.1	502.0
CCA Baseline	2.8	5.6	8.2	283.0	7.0	14.4	18.7	100.0
MMEN (Verb)	0.7	4.0	8.3	70.0	2.9	7.9	13.9	63.0
MMEN (Caption\Noun)	5.7	18.7	28.2	31.1	5.3	17.0	26.1	33.3
MMEN (Noun)	10.8	31.3	42.7	14.0	10.8	30.7	44.5	13.0
MMEN ([Verb, Noun])	15.6	39.4	**55.1**	9.0	13.6	36.8	51.7	10.0
MMEN (Caption)	15.8	40.2	53.6	9.0	13.8	36.7	50.7	10.3
JPoSE (Caption\Noun, Noun)	**16.4**	**41.3**	54.4	**8.7**	**14.3**	**38.1**	**53.0**	**9.0**

Table 7. MSR-VTT Video-Caption Retrieval results. *We include results from [22], only available for Text-to-Video retrieval.

Figure 5. Qualitative results of text-to-video action retrieval on MSR-VTT. $A \leftarrow B$ shows the rank B of the retrieved video from using the full caption MMEN (caption) and the rank A when disentangling the caption JPoSE (Caption\Noun, Noun).

instead focus on assessing whether disentangling PoS and learning multiple PoS-aware embeddings produce better results. In this adapted JPoSE we encode the output of the disentangled PoS-MMENs with e_v and e_t (*i.e.* concatenated) and use NetVLAD [3] to aggregate Word2Vec representations. Instead of the combined loss in Equation (7), we use the pair loss, used also in [22]:

$$L(\theta) = \frac{1}{B}\sum_{i}^{B}\sum_{j\neq i}\max\left(\gamma + d(f_{v_i}, g_{t_i}) - d(f_{v_i}, g_{t_j}),\, 0\right) + \max\left(\gamma + d(f_{v_i}, g_{t_i}) - d(f_{v_j}, g_{t_i}),\, 0\right) \quad (8)$$

This same loss is used when we train different MMENs.

Visual and text features. We use appearance, flow, audio and facial pre-extracted visual features provided from [22]. For the captions, we extract the encodings ourselves[3] using the same Word2Vec model as for EPIC.

Results. We report on video-to-text and text-to-video retrieval on MSR-VTT in Table 7 for the standard baselines and several MMEN variants. Comparing MMENs, we note that nouns are much more informative than verbs for this retrieval task. MMEN results with other PoS tags (shown in the supplementary) are even lower, indicating that they are not informative alone. Building on these findings, we report results of a JPoSE combining two MMENs, one for nouns, and one for the remainder of the caption (Caption\Noun). Our adapted JPoSE model consistently outperforms full-caption single embedding for both video-to-text and text-to-video retrieval. We report other PoS disentanglement results in supplementary material.

Qualitative results. Figure 5 shows qualitative results comparing using the full caption and JPoSE noting the disentangled model's ability to commonly rank videos closer to their corresponding captions.

5. Conclusion

We have proposed a method for fine-grained action retrieval. By learning distinct embeddings for each PoS, our model is able to combine these in a principal manner and to create a space suitable for action retrieval, outperforming approaches which learn such a space through captions alone. We tested our method on a fine-grained action retrieval dataset, EPIC, using the open vocabulary labels. Our results demonstrate the ability for the method to generalise to zero-shot cases. Additionally, we show the applicability of the notion of disentangling the caption for the general video-retrieval task on MSR-VTT.

Acknowledgement Research supported by EPSRC LOCATE (EP/N033779/1) and EPSRC Doctoral Training Partnerships (DTP). We use publicly available datasets. Part of this work was carried out during Michael's internship at Naver Labs Europe.

[3]Note that this explains the difference between the results reported in [22] (shown in the first row of the Table 7) and MMEN (Caption).

References

[1] English spaCy parser https://spacy.io/. 4

[2] Jon Almazán, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *CoRR*, abs/1801.05339, 2018. 2

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 2, 8

[4] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, 2016. 6

[5] Weihua Chen, Xiaotang Chen, Ianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, 2017. 2

[6] 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, 4, 5

[7] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Trans. Multimed.*, 2018. 1

[8] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, and Xun Wang. Dual dense encoding for zero-example video retrieval. *CoRR*, arXiv:1809.06181, 2018. 1, 2, 7

[9] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 2, 6

[10] Albert Gordo, Jon Almazán, Jérome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, 2016. 2

[11] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *CVPR*, 2017. 2

[12] 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. 2

[13] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2

[14] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013. 2

[15] Meera Hahn, Andrew Silva, and James M. Rehg. Action2vec: A crossmodal embedding approach to action learning. In *BMVC*, 2018. 1, 2

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *CoRR*, arXiv:1703.07737, 2017. 2

[17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *ICLR*, 2015. 2

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 2

[19] Wang Liwei, Li Yin, Huang Jing, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *TPAMI*, 2019. 3

[20] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *CoRR*, arXiv:1804.09235, 2018. 1, 2

[21] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *CoRR*, arXiv:1706.06905, 2017. 2

[22] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *CoRR*, arXiv:1804.02516, 2018. 1, 2, 7, 8

[23] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 2, 7

[24] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, 2018. 1, 2, 7

[25] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *ECCV*, 2016. 2

[26] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *CVPR*, 2017. 1

[27] Filip Radenović, Giorgos Tolias, and Ondřej Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*, 2016. 2

[28] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *IJCV*, 2016. 2

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[30] Jie Shao, Kai Hu, Yixin Bao, Yining Lin, and Xiangyang Xue. High order neural networks for video classification. *CoRR*, arXiv:1811.07519, 2018. 1

[31] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[32] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016. 2

[33] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giróo-i Nieto. Cross modal embeddings for video and audio retrieval. In *ECCVW*, 2018. 2

[34] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *CoRR*, arXiv:1609.08124, 2016. 2

[35] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014. 2

[36] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016. 1, 2, 3

[37] 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. 5

[38] Michael Wray and Dima Damen. Learning visual actions using multiple verb-only labels. In *BMVC*, 2019. 2

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 4, 7

[40] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, 2015. 2

[41] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 7

[42] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 2