On Semantic Similarity in Video Retrieval - Supplementary Material

Michael Wray Hazel Doughty* Dima Damen
Department of Computer Science, University of Bristol, UK

Here we provide a perceptual study that correlates the proxy measures to human annotators in Section A. Next, we provide information of the correlation between the semantic proxies in Section B, then details on three other proxy measures showcasing their unsuitability to the three datasets in Section C. Finally, we show a tabular version of Figure 7 from the main paper for numerical comparison in future works in Section D.

A. Proxy Measures Human Agreement Study

One might wonder how do the proposed proxies in Section 4.4 correlate to human annotators. To answer this question, we conduct a small-scale human study.

Requesting a human to assign a score relating a video and a caption is challenging and potentially subjective, however, ranking a small number of captions for their relevance to a given video can be achieved. We randomly select 100 videos from both the YouCook2 and MSR-VTT datasets (we focus on these two datasets as they include the most varied captions). For each proposed proxy, we rank the corresponding captions by their similarity to a given video, then select the most/least relevant captions as well as the captions at the 1st, 2nd and 3rd quartiles. This gives us 5 captions that are semantically distinct for the video.

We then asked 3 annotators (out of 6 total annotators) to order these 5 captions by their similarity to the given video. We remove annotation noise by only considering consistently ordered pairs of captions—that is when all 3 annotators agree that caption A is more relevant than B. We then report the percentage of correctly ordered pairs by the proxy, out of all consistently annotated pairs, as the 'Human-to-Proxy' agreement.

Table 1 shows the results of this human study. We note the % of consistent pairs of captions in each case. Results demonstrate that the four proxies correlate well with human rankings, with SYN and BoW giving the best Human-to-Proxy agreement on YouCook2 and MSR-VTT respectively. MET has lower agreements than the other proxy measures due to it penalizing different word orders as discussed in Sec. 5.1 of the main paper.

*Now at University of Amsterdam.

	BoW	PoS	SYN	MET
% Consistent Pairs YouCook2	86.5	78.0	76.3	77.3
% Consistent Pairs MSR-VTT	73.1	78.8	75.6	69.2
Human Agreement YouCook2	91.2	88.8	92.1	85.6
Human Agreement MSR-VTT	93.7	84.8	89.7	87.5

Table 1. Human Study reporting % of caption pairs with agreement between human and proxy on YouCook2 and MSR-VTT. Note: chance is 50%.

B. Correlation Between Semantic Proxies

To determine how similar the four proposed semantic proxies are, we calculate the Pearson correlation coefficient between pairs of semantic proxies for each video in YouCook2, MSR-VTT and EPIC-KITCHENS.

Figure 1 shows this correlation averaged over the videos within a dataset. All proposed semantic proxies have positive correlations, ranging between moderate (0.5-0.7) and high (> 0.7) correlations. We find the agreement between semantic proxies to be stronger at the lower end of the rank with the different methods consistently agreeing on which captions are irrelevant. At the higher end of the rank there tends to be some disagreements between proxies, with SYN and METEOR having the lowest correlation while BoW and PoS having the highest correlation. Importantly, the trend is consistent across the three datasets.

C. Proxies from Learnt Models

C.1. Definition

We compare our proposed proxies (Sec 4 in the main paper) to three other proxies which use learnt features from visual or textual models. Each proxy is defined as the cosine similarity between two vectors:

$$S'(y_i, y_j) = \frac{a(y_i) \cdot a(y_j)}{||a(y_i)|| \times ||a(y_j)||} \quad (1)$$

where $a(\cdot)$ is a trained model.

Textual Similarity We use two language models common in the literature to get representations: Word2Vec [3] and BERT [1]. For Word2Vec, the word vectors are av-

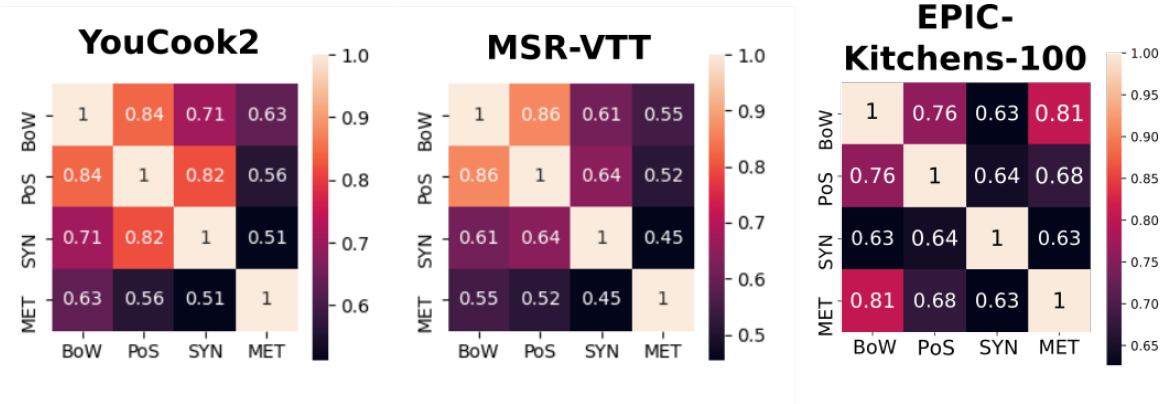

Figure 1. The average Pearson's correlation coefficient between pairs of proposed semantic proxies for YouCook2, MSR-VTT and EPIC-KITCHENS.

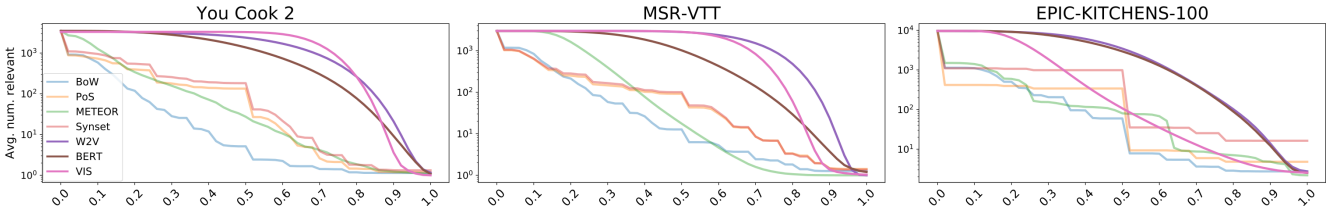

Figure 2. Average number of relevant captions for a video with a given threshold over each dataset and proxy measure including the Word2Vec (W2V), BERT and Visual (VIS).

eraged for a sentence-level representation[1]. When using BERT, we extracted a sentence-level representation using the DistilBERT model from [4].

Visual Similarity For the visual embedding proxy, we use the video features extracted from the pre-trained model. This changes Eq. 5 in the main paper to the following:

$$S_S(x_i, y_j) = \begin{cases} 1 & i == j \\ S''(x_i, x_j) & otherwise \end{cases} \quad (2)$$

Note that a video and a caption are related here purely on the similarity between the video features, making the assumption that the visual contents of video x_j offer a sufficient description of the caption y_j, and that the pre-trained video features offer sufficient discrimination between the videos.

C.2. Proxy Measure Comparisons

We show an extended version of Figure 4 from the main paper, adding the three proxy measures from learnt models in Figure 2. We compare these for the three datasets YouCook2, MSR-VTT and EPIC-KITCHENS.

We find the average number of relevant captions per video from the three learned proxies is much higher than the proposed proxies across almost all thresholds. With lots of captions being considered relevant, this has the effect of inflating nDCG scores.

When analysing the visual proxy, we find that the similarity is not semantic in nature. The visual proxy has high similarities between segments from the same video, further highlighting its unsuitability. Accordingly, using visual similarity from pre-trained models is not suitable as a proxy for semantic similarity.

The BERT and Word2Vec proxies similarly do not produce reasonable proxies of semantic similarities for these three datasets. From Figure 2, both methods produce significantly more relevant captions than proposed metrics. When analysing the results, we note that BERT and Word2Vec relate captions via their context, because of their training which relates words by the co-occurrence rather than their semantic relevance. For example, 'open' and 'close' are often used in the same context of objects, but represent opposite actions. Both Word2Vec and BERT would give much higher similarity to these two, despite being antonyms.

D. Table of Figure 7

Table 2 shows the performance of the different baseline models on all three datasets and proxy measures. See Section 5.3 in the main paper for the discussion of results.

[1]We also tried using the Word Mover's Distance [2] but achieved slightly worse results.

	Proxy	Instance	BoW	PoS	Syn	Met
	Metric	GMR	nDCG			
YouCook2	Random	0.1	23.1	22.1	27.7	66.2
	MEE	7.5	42.1	40.3	45.3	73.3
	MoEE	9.8	41.5	39.1	44.0	73.0
	CE	9.7	41.8	39.3	44.1	73.0
MSR-VTT	Random	0.2	34.0	30.0	11.6	80.4
	MEE	15.7	51.6	48.5	33.5	83.3
	MoEE	22.7	53.9	50.8	36.8	83.9
	CE	22.4	54.0	50.9	36.7	84.0
EPIC	Random	0.0	11.7	4.5	10.7	13.0
	MEE	18.8	39.3	29.2	41.8	41.0
	JPoSE	9.4	39.5	30.2	49.0	44.5

Table 2. Tabular version of Figure 7 from the main paper. Results of evaluating the baseline methods on the different proxy measures for semantic similarity. (GMR=Geometric Mean Recall)

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 1

[2] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015. 2

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 1

[4] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 2