Improving Causality in Interpretable Video Retrieval

Varsha Devi, Georges Quénot, Philippe Mulhem

Improving Causality in Interpretable Video Retrieval

Varsha Devi, Georges Quénot, Philippe Mulhem
Univ. Grenoble Alpes, CNRS,
Grenoble INP (Institute of Engineering Univ. Grenoble Alpes), LIG,
Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

***Abstract* – This paper focuses on the causal relation between the detection scores of concept (or tag) classifiers and the ranking decisions based on these scores, paving the way for these tags to be used in the visual explanations. We first define a measure for quantifying a causality on a set of tags, typically those involved in visual explanations. We use this measure for evaluating the actual causality in the explanations generated using a recent interpretable video retrieval system (Dong et al. [4]), which we find to be quite low. We then propose and evaluate improvements for significantly increasing this causality without sacrificing the retrieval accuracy of the system.**

***Keywords* –**video retrieval, interpretability, causality

I. Introduction

State-of-the-art approaches for cross-modal video retrieval rely on dual-stream neural architecture that project video and text samples into a common embedding space. These architectures [3] first considered one latent space. They obtained good performances on Text-To-Video (TTV) or Video-To-Text (VTT) retrieval tasks but they operated as black boxes, providing no explanation or justification for their results

To improve the interpretability of these systems, subsequent approaches replaced the latent space with a concept space whose dimensions are aligned with a set of "concepts" or "tags" corresponding to the most frequently used terms in a training set. Such systems implement at the same time a TTV/VTT matching task and a related concept classification task. Such TTV/VTT matching is implemented using *only* the concept classification scores, thereby enforcing a strict *causal* relation between concept classification and TTV/VTT retrieval [30]. The system operation is then interpretable as retrieval decisions (based on similarities) use only classification scores corresponding to tags meaningful to humans.

The retrieval performance of such a concept-based approach is slightly degraded compared to a purely latent one, as the constraints coming from the classification make the retrieval less optimal. Hybrid approaches, combining both a latent and a concept space, achieve higher performances than each of the latent and concept spaces alone [4], [30], at the price of a lower level of interpretability and causality, as the latent branch of the system remains opaque.

Figure 1 (from [4]) illustrates how explanation/justification can be provided to a user using these hybrid approaches: tag clouds show the concepts found to be the most relevant (with sizes related to their estimated importance)for the query and for the 4 top-ranked retrieved documents, . A user can evaluate to what extent these tag clouds are actually relevant to the query and to the documents and to what extent they match. However, it's important to note that these tag-clouds do not provide information about their contribution to the overall retrieval decision. Hence, our work takes place before such displays: we study how to measure concepts detections scores' causal contribution in retrieval decisions, and we propose ways to increase such causality on a state of the art system.

We rely on the dual stream implementation of [4], which uses a *dual space* (i.e. latent and concept) to map the latent and concept features of video and text, and a *dual task* learning approach, where the system simultaneously performs video-text retrieval and video and text classification tasks. [4] achieved state-of-the-art performance on several video retrieval benchmarks. We used the code shared by the authors and the MSR-VTT collection [32] for conducting the experiments. The contributions of this paper are:

1. the proposal of a metric for quantifying the causal contribution of a set of concepts involved in the visual explanation/justification of a retrieval decision;
2. the use of this metric to show that the causality is actually quite low when using the top-10 detected concepts of [4], as the non-displayed concepts contribute much more to the similarity measure than the displayed ones;
3. a method that significantly improves the causality of the explanation of [4] without degrading the retrieval accuracy.

In section 2, we discuss the related work; in section 3, we present a metric for quantifying causality in visual explanations and justifications; in section 4, we describe the two methods that we propose for improving this causality; in section 5, we present and discuss the results of our experiments; and in section 6, we conclude and discuss future work.

II. Related work

Cross-Modal retrieval aims at retrieving a ranked list of relevant items in a modality, for a given query in another modality. We focus here both on the Text-to-Video and Video-to-Text retrieval cases [2]–[4],[8],[30],[34], which have achieved significant progress in recent years. In concept-based approach, the aim is to use pre-defined sets of visual concepts to generate concept-based video-text representation and map them to concept space for similarity computation [5],[10],[12], [17], [22], [26]–[29]. These approaches are interpretable and work better when the right concepts are accurately identified without ambiguity and mapped to video and text. In order to deal with the problems of ambiguity in concept-based

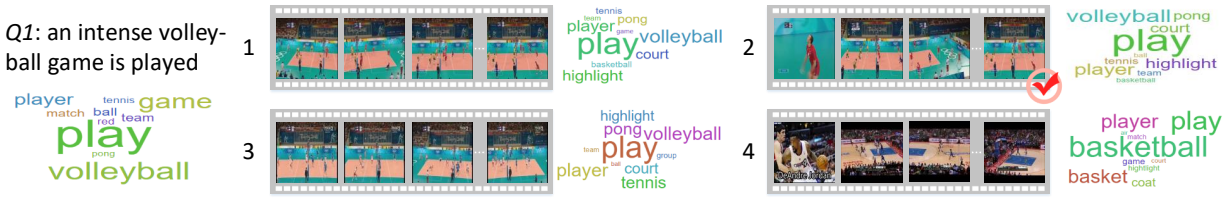

Fig. 1. Tag clouds for justifying the retrieved results for one query (from [4])

.

approach, *concept-free approaches* were proposed, which map encoded videos and text features to high dimensional latent spaces for similarity computation [3],[13],[14],[19],[31],[33]. They are effective in retrieval, but lack interpretability. To combine the advantages of both approaches, hybrid models that fuse video-text similarity in concept space and latent space at a later stage has become the norm [4], [6], [10], [22], [29], [30]. These approaches are effective and achieve some level of interpretability. *Dong et al.* [4] and *Wu et al.* [30] proposed a hybrid model with dual space (latent and concept) and dual task (retrieval and classification) and achieved state of the art performance for video-text retrieval. However, in their case, the displayed tag-cloud-based explanations for retrieved results do not reveal the causal effect of one or several tags on the retrieval decision as a whole.

This paper focuses on evaluating and enhancing the causality in tag-cloud-based explanations of dual encoding model [4] within a hybrid approach, by quantifying and augmenting the causal contribution of visual concept classes, specifically those employed in tag-cloud explanations. By assigning greater weight to fewer relevant concepts, we seek to amplify their causality in the retrieval decision-making process without impacting retrieval accuracy. Similar to our concept weighing approach, a few papers [20], [21] propose methods to enhance the plausibility of attention maps in RNN or transformer-based models, which are commonly used to explain classification model decisions. The approaches aim to provide more reliable explanations using fewer important words/tokens by giving them greater weight, thus addressing the concept of "parsimony" in attention maps. Additionally, *Liao et al.* [15] develop frameworks for automatically learning compact and parsimonious representations by focusing on a small subset of informative features while disregarding irrelevant or redundant ones. This leads to more compact and interpretable representations. Although *Liao et al.*'s work differs in the specific context of classification or retrieval, it aligns with the idea of achieving parsimony and interpretability in models.

Apart from the parsimonious models, researchers have explored causality in machine learning to gain a deeper understanding of the classification models [24],[25],[35],[36]. Our study also differs from *Yang et al.* [35], who propose a causality-inspired framework for Video-Moment Retrieval, employing a structural causal model to analyze the impact of queries and video content on prediction outcomes. However, we aim to quantify the true causal effect of a set of predicted concepts in retrieval decisions, rather than focusing on effect of queries and video content on prediction. To the best of our knowledge, we are the first to provide a quantitative measure of the causal contribution of visual concept classes in retrieval explanations, contributing to a better understanding and interpretation of retrieval models.

III. Analysis of causality

A. Quantifying causality

If we consider only the concept-based part of dual encoding model [4], videos and queries are represented only by the detection scores of the concepts. Videos are then ranked in decreasing order of the similarity of their representations with that of the query content. [4] uses by default the Jaccard similarity function for the concepts between v and s respectively a video sample and a text sample:

$$sim_{con}(v,s) = \frac{\sum_{i=0}^{i=K} \min(g(v)_i, g(s)_i)}{\sum_{i=0}^{i=K} \max(g(v)_i, g(s)_i)} \tag{1}$$

with g being the function projecting v and s into the concept space, and K being the number of dimensions of the concept space. The g function contains a final sigmoid function that normalizes the concept detection scores between 0 and 1 (used also in the binary cross-entropy during the concept classification training).

The cosine similarity function may also be considered, as it is already used by default on the latent space of [4]. Such cosine similarity is defined as:

$$sim_{con}(v,s) = \frac{h(v).h(s)}{\|h(v)\|.\|h(s)\|} = \frac{\sum_{i=0}^{i=K} h(v)_i.h(s)_i}{\|h(v)\|.\|h(s)\|} \tag{2}$$

with h being the function projecting v and s into the concept space without the final sigmoid function.

We propose to quantify the causality of a group of tags (which may be those presented using clouds as in figure 1) in a retrieval decision by the sum of their feature effects, themselves taken as their *relative* overall contribution in the similarity measure used for the ranking of the results. We observe that in the *sim* functions presented above, the numerators are based on a sum of per-tag terms. As we are interested in the relative importance of individual tags or of group of tags, we may get rid of the *sim* denominators and normalize the terms so that the sum of their absolute values is equal to one (all values are positive in the Jaccard case but not necessarily in the cosine one). This gives for the Jaccard and cosine individual tag contributions:

$$w_i(v,s) = \frac{\min(g(v)_i, g(s)_i)}{\sum_{j=0}^{j=K} \min(g(v)_j, g(s)_j)} \quad \text{or} \quad \frac{|h(v)_i.h(s)_i|}{\sum_{j=0}^{j=K} |h(v)_j.h(s)_j|} \tag{3}$$

The causal effect, defined in $[0,1]$, of a set of tags G is defined as:

$$c(G,v,s) = \sum_{i \in G} w_i(v,s) \tag{4}$$

This function measures the causal effect of one or several tags for a single pair (s,v). We define the "causality at k" as the causality defined as in equation (4) with G corresponding to the k tags contributing the most to the computation of the similarity score:

$$c_k(v,s) = \max_{G \subset 1,K,|G|=k} \sum_{i \in G} w_i(v,s) \tag{5}$$

From this measure defined for one pair (v,s), we derive global measures on a whole cross-modal collection by computing statistics such as the mean (equation (6)) and the standard deviation of this value on a set of pairs P.

$$C_k(P) = \frac{1}{|P|} \sum_{(v,s) \in P} c_k(v,s) \tag{6}$$

P may be the set of all possible pairs in the collection or only the set of matching pairs. We can also consider the set of pairs obtained using all of the text queries and, for each of them, the top-n retrieved videos, or the opposite using video queries and retrieved texts.

In our case, causality in explanations/justifications relies only on the detections scores for the displayed tags. This is the case by design for the dimensions in a concept space, but not for the dimensions in a purely latent space as these have no meaning for humans. The causal weight of any element coming from the latent space in the concept-based visual explanation/justification should then be strictly zero. In the latent-space-only approach, no concept detection scores are available anyway for displaying tag clouds. However, such scores are available in hybrid approaches, as the decision is made partly on similarities $sim_{lat}(v,s)$ coming from the latent space and partly on similarities $sim_{con}(v,s)$ coming from the concept space. The overall similarity is a weighted sum (after a global scale normalization) $sim(v,s) = \alpha.sim_{lat}(v,s) + (1-\alpha).sim_{con}(v,s)$. The overall causality should logically be a weighted sum based only on the concept scores multiplied by the $(1-\alpha)$ factor in which, as the causality on the latent part should be zero.

B. Evaluating causality of the target system

We have evaluated the tag-detection-score-to-similarity causality using the pre-trained hybrid model provided by the authors of [4] on the MSR-VTT dataset [32]. In this hybrid model, the concept-based similarity accounts for 40% of the global score, As described above, the causal weight of the concepts is reduced accordingly. Figure 2 shows the mean and standard deviations of the individual $w_k(v,s)$ and cumulative $c_k(v,s)$ contributions of the tags ranked by decreasing contributions for the matched pairs (associated videos and captions). We observe from these curves that:

- Even for similar pairs, the individual and cumulative causalities of the first few tags are very small: less than 0.5% for the first tag and less than 4% accumulated for the first 10. This indicates that the actual causality in visual explanations such as illustrated if figure 1 is of only 4% if we consider a similarity based only on the concept space and even only 1.6% for the whole hybrid approach.

- A large majority of the tags have a significant contribution to the similarity measure and therefore to the ranking decision. We also observed that the retrieval performance is very degraded if we include only the first few tens of tags in the Jaccard distance. Even if the displayed tags seems relevant to both the caption query and the retrieved video, the ranking decision is actually mostly made on the terms beyond the first few tens.

Table I presents the retrieval performance and the causality at 10 and at 30 of the original hybrid approach from [4], as well as of a number of variants aiming at improving the performance and/or the causality values. "2048-d latent" corresponds to a latent-space only version; "1536d+512d hybrid" is the original hybrid (GitHub) version "512-d (hyb. train.)" is the same hybrid system in which only the concept-based part is used for the ranking.

We also considered variants trained only with the concept space (no latent space) and with two different vocabulary sizes: "512-d Jaccard" and "256-d Jaccard". Finally, we also tried a concept-only training with a cosine similarity as used in the case of the latent space instead of the Jaccard one, also with two different vocabulary sizes: "512-d cosine" and "256-d cosine".

Regarding causality, we chose the causalities at 10 and 30 as they correspond to practically useful values in the sense that 10 is the number of tags that a user can grasp simultaneously, e.g.; [18] mentions that human are unable to process more than a few, typically 7 ± 2, stimulus at one time, and 30 is a reasonable bound on the number of tags that could be validly assigned to a given caption or video. Explanations involving more than these numbers are unlikely to be causally correct and the components beyond them would likely be used just as latent dimension in a quite opaque way.

The "512-d (hyb. train.)" case corresponds to the curves displayed in figure 2; the causality for "2048-d latent" would be 0 (or rather n/a); and the causality for "1536d+512d hybrid" is in between. The causality for "512-d cosine" is significantly higher because the decreasing of the sorted component values happens to be much faster in this case. For both the Jaccard and cosine versions there is a significant increase in the causality when the vocabulary size is decreased, which is expected as the relative weight values automatically increase when their count decreases.

In our study, we are currently examining the dual encoding model proposed by *Dong et al.* [4] in 2021 for causality analysis. However, there are other interpretable models, *e.g., Wu et al.* [30], that use similar methods. So, it's likely that these alternative models will show similar behavior. Our approach, focused on causality, is generalized and can be applied to retrieval models based on classification tasks, offering a broader understanding of their behavior and performance.

TABLE I

Comparison on the MSR-VTT task [32] for the original hybrid approach [4] and for some selected variant. mAP (3^{rd} last column) represents average of the TTV and VTT mAPs and last two "C@n" columns for the causality at n on the matched pairs. Metrics are same as in [4] except "C@n", and described in Section V.

Method	**Text-to-Video Retrieval**					**Video-to-Text Retrieval**					**SumR**	**mAP**	**C@10**	**C@30**
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP				
2048-d latent [4]	11.0	29.2	39.8	19	20.2	18.8	42.7	56.2	8	9.3	197.7	14.0	n/a	n/a
1536d+512d hybrid	11.8	30.6	41.8	17	21.4	21.6	45.9	58.5	7	10.3	210.2	15.8	1.6	4.0
512-d (hyb. train.)	9.7	26.2	36.2	25	18.2	19.3	43.8	56.0	8	9.2	191.1	13.7	3.9	10.0
512-d Jaccard	10.4	28.2	39.1	20	19.5	19.8	42.4	55.0	8	9.4	194.8	14.5	4.0	10.0
256-d Jaccard	10.9	29.2	40.2	18	20.3	18.2	41.3	53.4	9	9.1	193.2	14.7	8.2	19.6
512-d cosine	10.6	28.8	39.2	20	19.9	20.3	44.0	56.6	7	9.9	199.6	14.9	10.4	23.5
256-d cosine	11.0	29.5	40.4	18	20.5	19.8	44.2	56.2	8	9.8	201.0	15.1	17.4	37.8

IV. Improving causality

We have seen above that the causality in the actual visual explanations is very low because, instead of having the causal weights mostly distributed on only a few tags as it would be expected if only relevant tags were detected with significant scores, we have quite the opposite with most tags being detected with similar and non-negligible scores, as can be seen in figure 2. This is a bit less pronounced when the cosine similarity is used but still significant enough for the causalities at 10 and at 30 to remain quite low. Using only 256 tags instead of 512 significantly improves these causalities with no significant loss or even with a slight gain in performance but the causalities still remain low.

The causal weights are so much spread over all the available tags because all the tags are always detected to some degree with an average "probability" of about 0.4. This means that, in average, about 200 tags out of 512 are detected, which is not what is expected and is much larger than the average tag frequency in the training data. This is likely due to the fact that the detections scores depend on two loss functions: one for the classification task and one for the retrieval task. The latter probably shifts the equilibrium of the former leading to over-detection. Also, likely, the detection scores become partly an estimated tag probability and partly a latent component not related to the classification task, adding noise to the shift. We observed that several detectors are very poor.

In order to improve the causality from the first few tags, we propose to modify the detection scores by applying a function to them so that the causal weight becomes more concentrated on the first few tags. There are several ways to do this. First, considering the tag probabilities used in the Jaccard similarity (equation (1)), simply applying a power transformation with an exponent p greater than 1 automatically increases the relative weights of the first terms. Second, the tag probabilities $g(v)$ or $g(s)$ are obtained by applying a sigmoid function to "raw" detection scores $h(v)$ or $h(s)$; we can then apply a bias b (shift) and/or a gain a (scale) to these raw scores before applying the sigmoid function, performing a kind of Platt normalization [23], possibly correcting the influence of the retrieval loss in the classification calibration. Combining transformations, we replace $g(x)_i = \sigma(h(x)_i)$ by:

$$\left(g_{(a,b,p)}(x)\right)_i = (\sigma(a(h(x)_i - b)))^p \tag{7}$$

with σ being the sigmoid (expit) function and x being either a video sample v or a text sample s. The original function corresponds to $(a,b,p) = (1,0,1)$. Similarly, in order to improve the causality from the first few tags with the cosine similarity (equation 2), we replace $h(x)_i$ by:

$$\left(\left(h_{(a,b,p)}(x)\right)\right)_i = (a(h(x)_i - b))^p \tag{8}$$

the main difference being that the sigmoid transform is not used with the cosine similarity. Again, the original function corresponds to $(a,b,p) = (1,0,1)$ but it can be noted that, as a scale factor, the a parameter has no effect in the cosine similarity, which is related to an angle between vectors. We will then keep $a = 1$ in this case.

For appropriate values of the a, b and p parameters, the transformations described in equations (7) and (8) increase the contrast between the values used for the similarity computation and therefore the causality over the first few most contributing tags. Indeed, these transformations do impact the retrieval performance of the system as well, sometimes positively and sometimes negatively, depending upon the choice of the a, b and p parameters. These parameters should then be chosen in order to obtain the best compromise between causality and accuracy. This is done by giving preference first to the accuracy −as we generally do not want to sacrifice it to causality− and second to the causality as long as this does not hurt accuracy. The corresponding optimal a, b and p parameters are obtained by direct search on the validation set, one at a time, and iteratively.

V. Experiments

Data. We conducted experiments on the MSR-VTT dataset [32] that contains in total 10K video clips with 20 captions provided for each video, i.e. 200K captions. We used the official split of MSR-VTT dataset, containing 6,513 video clips for training, 497 for validation and 2,990 video clips for evaluating causality, and retrieval accuracy.

Implementation details. We used PyTorch code[1] provided by the authors of [4]. In order to measure causality and retrieval performance of system, the concept space is trained and evaluated in 3 different settings: (i) *Concept-Hybrid*:

[1] https://github.com/danieljf24/hybrid_space

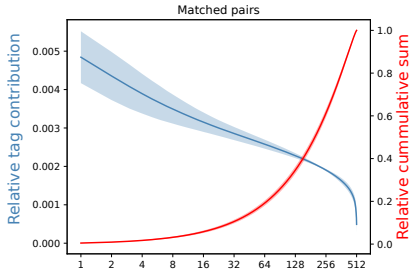

Fig. 2. Individual and cumulative contribution (mean $\pm$ standard deviation) for matched (v,s) pairs, of the tags ranked by decreasing contributions.

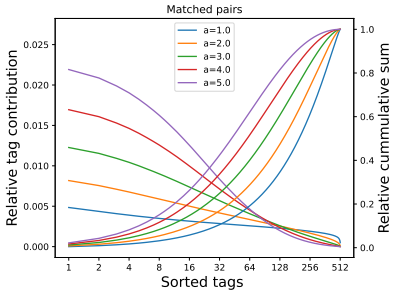

Fig. 3. Per-tag (decreasing curves) and cumulative (increasing curves) causality for different values of scale a.

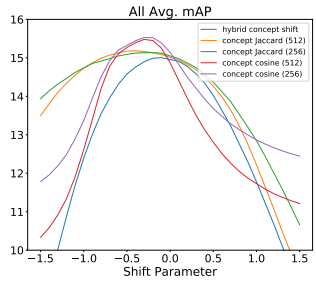

Fig. 4. Global mAP evolution for the shift (for optimal scale) parameters for the five considered system variants.

where the concept space is trained in hybrid mode (latent and concept both), for testing of causality and retrieval, only concept space part is used, (ii) *Concept-Jaccard*: In this setting the dual encoding model is trained and tested on concept space part only with Jaccard coefficient as a similarity metric, and (iii) *Concept-cosine*: As we have seen in Table I (row-8), there is slight improvement in accuracy when using cosine similarity in concept space so we also reported results for "concept-cosine" setting.

Performance Metrics In order to evaluate the accuracy of TTV and VTT retrieval system, we used the proportion of the queries for which at-least one correct document is retrieved among top-K results (R@K, with $K = 1,5,10$), the median rank of first relevant item (Med R), the mean Average Precision (mAP) and sum of R@K for TTV and VTT (SumR). Higher R@K and lower median rank represents better performance of system. To evaluate the causality, we proposed the formula for calculating the averaged causal effect of a group of concepts G over the top-n results of all the queries of the dataset (refer to Section B).

We now explore the impact of our score modifications on the causality of different variants of the system. Then, we check the impact of the proposed modifications on the accuracy. We finally discuss the trade-off between these two criteria.

Impact on causality For each combination of the a, b and p parameters, it is possible to compute the modified tag "probabilities" or scores and to compute from them similarity values, causalities as displayed in figure 2 and performance metrics as displayed in table I. Figure 3 shows how per-tag and cumulative causality curves evolve according to the values of the scale parameters for the "512-d (hyb. train.)" system. As expected, the causality always increases with the value of the p (power) parameter. We also observed that it increases with the values of the a (scale) and b (shift) parameters. This remains for all combinations of these parameters that we tried and is also the same for the other systems using a Jaccard similarity ("512-d Jaccard" and "256-d Jaccard"). Regarding the systems using a cosine similarity ("512-d cosine" and "256-d cosine"), the same behavior is observed for the p and b parameters and, as expected, the a parameter has no effect. As we are interested in values as high as possible for the causality at a few tens of tags, for all the systems, we should use values as high as possible for the p, a (if applicable) and b parameters.

Impact on accuracy Choosing values as high as possible for the p, a and b parameters is likely to have a negative impact on retrieval accuracy. Figure 4 shows the evolution of the global mAP according to the parameter b (shift) of equations 7 and 8. The baseline value is of 0.0 for those parameters We observe that, except (as expected) for the scale parameter with cosine similarity, there is an optimum value for each parameter for the global mAP. The optimum value generally gives a slight performance improvement over the baseline, sometimes significant. Regarding the p and a (when applicable) parameters, the optimum value is significantly higher than the baseline, indicating that it is possible to have a gain simultaneously on the causality and on the accuracy. On the opposite, the optimum value for the accuracy for the b parameter corresponds to a value lower than the baseline so that we loose on one criterion if we optimize on the other.

Joint optimization As previously mentioned, we favor accuracy over causality as users generally do not want to sacrifice the former to the latter. Here, we even try to further improve the accuracy even if we improve less on the causality. This means that we choose the optimum values obtained from the functions displayed in the curves of figure 4 except where the curve is rather flat and the optimum value is close the baseline one, in which case we keep the latter, which is better for the causality. Also, when relevant, we optimize jointly the b parameter and the p or the a parameter. We don't jointly optimize the p and the a parameters as they have a similar effect and keep the other to the baseline value. The optimization is done on the validation set and causalities and accuracies are measured on the test set. We have also checked that the optimal values are quite close on the validation set and on the test set. Table III shows the optimum value combinations found on the validation set for the five system variants considered and Table II compares the original and improved accuracy values for these cases and the original hybrid version.

Discussion We found out that there are many ways to improve the actual causality in visual explanations: by using only a concept space for the retrieval, either with a hybrid training or with a concept-only training, by using a cosine

TABLE II

Causality and performance with and without our improvements for five training conditions. C@10 and C@30 are the causality respectively for the top-10 and top-30 contributing tags. mAP is the mean of the TTV and VTT mAPs. SumR is as defined in [4]. All values are in percentages.

Training	inference	C@10	C@30	mAP	SumR
1536d+512d hyb.	original	1.6	4.0	15.8	210.2
512-d (hyb. tr.)	original	3.9	10.0	13.7	191.1
	improved	10.9	25.5	15.0	203.0
512-d Jaccard	original	4.0	10.0	14.5	194.8
	improved	16.0	29.7	15.0	198.7
256-d Jaccard	original	8.2	19.6	14.7	193.2
	improved	32.0	51.8	15.3	200.8
512-d cosine	original	10.4	23.5	14.9	199.5
	improved	15.6	31.3	15.5	207.0
256-d cosine	original	17.4	37.8	15.1	201.0
	improved	22.3	44.1	15.5	206.7

similarity instead of a Jaccard one, by using a smaller tag vocabulary size, and finally by using a transformation on the tag probabilities or scores with optimized parameters. All of them may lead to a significant improvement in the causality on the first few tags or tens of tags without sacrificing on the retrieval accuracy or with even a slight increase in accuracy too, except in the first considered step which is to drop the use of the purely latent space in the retrieval step.

TABLE III

Optimal values for the p (power), a (scale) and b (shift) parameters on the validation set for five system variants.

Training	p	a	b
512-d (hyb. tr.)	1.00	2.7	0.0
512-d Jaccard	1.00	2.9	0.0
256-d Jaccard	1.00	1.8	0.0
512-d cosine	1.07	n/a	-0.25
256-d cosine	0.98	n/a	-0.24

Regarding the transformations, we found that a scale-only transformation was the best for systems using the Jaccard similarity and that a transformation based on both shift and power was best for systems using the cosine similarity. The use of cosine similarity may lead to better accuracy for the improved version but with a slightly lower improvement in causality. The accuracy of the improved cosine versions using a concept space only is closed to that of the original full hybrid version.

Regarding the size of the tag vocabulary, the accuracy is comparable for 512-tag and 256-tag versions while the causality is greatly improved for the latter. We tried to reduce further the tag vocabulary size but the accuracy begins to drop significantly for sizes going below about 200 tags [1].

One might question whether the modified tag probabilities or scores still represent well the detection scores from the tag classifiers. Both the Jaccard- and cosine-specific transformations are actually doing a *re-calibration* of these. In fact, the original "tag probabilities" are unlikely to be well calibrated because they correspond to an average detection of 40% of the tags (i.e. ~200 concepts), which is much larger than the actual average tag annotation in the training data, and because the calibration is biased due to the fact that the tag probabilities are subject to two different and competing loss functions (for classification and for retrieval). By reducing the average detection rate of the tags, it is likely that the proposed transformations actually leads to a *better* calibration of the detection scores and to more meaningful "tag probabilities".

VI. Conclusion and Perspectives

In this paper, we have proposed an evaluation measure for quantifying the causality in ranking for retrieval of human readable tag used in visual explanations. Then, we extended a video retrieval state of the art approach, [4], in a way to enforce a higher causality, without negatively impacting the performance of the system. Our proposal relies on a modification of the tag scores computation, through a generalization of the original sigmoid function (for the Jaccard similarity case and an equivalent for the cosine similarity case), in order to increase the relative effect of the top tags. In such case, the major part of the matching function is supported by a few tens of dimensions, which is much more suitable for causality explanation. We show that our proposal increases our causality measure by up to an order of magnitude without loosing significantly on the accuracy. This study has been conducted in the case of the system proposed by [4] but both the observations and the improvements should be generalizable to other interpretable systems for multimedia retrieval that similarly rely on a similarity in a conceptual space.

This preliminary work shows that, though it is possible to significantly improve the causality in visual explanations without sacrificing performance, a 100% causality in such visual justifications / explanations is still far away. Other experiments that we conducted show that it is possible to

strictly enforce a 100% causality, but with a very significant penalty on the accuracy, typically halving the global mAP value. Any compromise in between is also likely to be achievable but, in general, users will not want to trade away accuracy for causality. We believe that it is possible to further improve the causality in visual justifications / explanations by introducing modifications in the system beyond the simple detection score transformations introduced by equations (7) and (8). This may involve inserting such transformations as layers with learnable parameters and/or modifying the classification and retrieval loss functions for a better cooperation between the two tasks and a better probability calibration.

The quantification of the causality that we proposed may also be improved. As it is defined, it makes sense and gives reasonable insights of what is going on, but it does not capture everything. In the case of the Jaccard similarity, it takes into account only minimum (numerator) part the per-tag probabilities while their maximum or their difference also has some effect. This is less a problem for the cosine similarity if the $h(x)$ vectors are L^2-normalized. Also and relatedly, the current version considers only a single (v,s) pair and is adequate for explaining why a document is retrieved or not but it is possibly less adequate for explaining why a document is ranked before another one. We believe that these aspects are at least partially indirectly taken into account by the proposed quantification of the causality but it could adapted to address them directly.

REFERENCES

[1] Varsha Devi, Philippe Mulhem, and Georges Quénot. Analysis of the Complementarity of Latent and Concept Spaces for Cross-Modal Video Search. In *CBMI 2022: International Conference on Content-based Multimedia Indexing*, pages 84–90, Graz, Austria, September 2022. ACM.

[2] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

[3] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9346–9355, Long Beach, CA, USA, 2019. Computer Vision Foundation / IEEE.

[4] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2022.

[5] Markatopoulou Foteini, Moumtzidou Anastasia, Galanopoulos Damianos, Mironidis Theodoros, Kaltsa Vagia, Ioannidou Anastasia, and Spyridon Symeonidis. Iti-certh participation in trecvid 2016. In *TRECVID 2016 Workshop*, 2016.

[6] Danny Francis, Phuong Anh Nguyen, Benoit Huet, and Chong-Wah Ngo. Eurecom at trecvid avs 2019. In *TRECVID*, 2019.

[7] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, 2014.

[8] Richang Hong, Yang Yang, Meng Wang, and Xian-Sheng Hua. Learning visual semantic relationships for efficient visual retrieval. *IEEE Transactions on Big Data*, 1(4):152–161, 2015.

[9] Zhichao Hu and Marilyn A Walker. Inferring narrative causality between event pairs in films. *arXiv preprint arXiv:1708.09496*, 2017.

[10] Po-Yao Huang, Junwei Liang, Vaibhav Vaibhav, Xiaojun Chang, and Alexander Hauptmann. Informedia@ trecvid 2018: Ad-hoc video search with discrete and continuous representations. In *TRECVID Proceedings*, volume 70, 2018.

[11] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 27–34, 2015.

[12] Duy-Dinh Le, Sang Phan, Vinh-Tiep Nguyen, Benjamin Renoust, Tuan A Nguyen, Van-Nam Hoang, Thanh Duc Ngo, Minh-Triet Tran, Yuki Watanabe, Martin Klinkigt, et al. Nii-hitachi-uit at trecvid 2016. In *TRECVID*, volume 25, 2016.

[13] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019.

[14] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 23:4351–4362, 2020.

[15] Renjie Liao, Alex Schwing, Richard Zemel, and Raquel Urtasun. Learning deep parsimonious representations. *Advances in neural information processing systems*, 29, 2016.

[16] Chen Change Loy, Tao Xiang, and Shaogang Gong. Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In *2009 IEEE 12th International Conference on Computer Vision*, pages 120–127. IEEE, 2009.

[17] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. Event detection with zero example: Select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 127–134, 2016.

[18] George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, March 1956.

[19] 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 *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[20] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. A study of the plausibility of attention between rnn encoders in natural language inference. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1623–1629. IEEE, 2021.

[21] Duc Hau Nguyen, Guillaume Gravier, and Pascale Sébillot. Filtrage et régularisation pour améliorer la plausibilité des poids d'attention dans la tâche d'inférence en langue naturelle. In *TALN 2022-Traitement Automatique des Langues Naturelles*, pages 95–103. ATALA, 2022.

[22] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. Vireo@ trecvid 2017: Video-to-text, ad-hoc video search, and video hyperlinking. In *TRECVID*, 2017.

[23] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3), 1999.

[24] Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 519–528. Springer, 2021.

[26] Cees GM Snoek, Xirong Li, Chaoxi Xu, and Dennis C Koelma. University of amsterdam and renmin university at trecvid 2017: Searching video, detecting events and describing video. In *TRECVID*, 2017.

[27] Cees GM Snoek and Marcel Worring. *Concept-based video retrieval*. Now Publishers Inc, 2009.

[28] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. Waseda_meisei at trecvid 2017: Ad-hoc video search. In *TRECVID*, 2017.

[29] Kazuya Ueki, Takayuki Hori, and Tetsunori Kobayashi. Waseda_meisei_softbank at trecvid 2019: Ad-hoc video search. In *TRECVID*, 2019.

[30] Jiaxin Wu and Chong-Wah Ngo. Interpretable embedding for ad-hoc video search. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 3357–3366, Seattle, WA, USA, 2020. ACM.

[31] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4592–4601, 2019.

[32] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, Las Vegas, Nevada, USA, June 2016. Computer Vision Foundation / IEEE.

[33] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[34] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1339–1348, 2020.

[35] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 2021.

[36] Liting Zhou, Jianquan Liu, Shoji Nishimura, Joseph Antony, and Cathal Gurrin. Causality inspired retrieval of human-object interactions from video. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2019.

[37] Yue Zhou, Shuicheng Yan, and Thomas S Huang. Pair-activity classification by bi-trajectories analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.