applied sciences

MDPI

Article

Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability

Ricardo Kleinlein *, Cristina Luna-Jiménez, David Arias-Cuadrado, Javier Ferreiros and Fernando Fernández-Martínez

Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación,
Universidad Politécnica de Madrid, Avda. Complutense 30, 28040 Madrid, Spain; cristina.lunaj@upm.es (C.L.-J.);
david.ariasc@alumnos.upm.es (D.A.-C.); javier.ferreiros@upm.es (J.F.); fernando.fernandezm@upm.es (F.F.-M.)
* Correspondence: ricardo.kleinlein@upm.es

Abstract: Not every visual media production is equally retained in memory. Recent studies have shown that the elements of an image, as well as their mutual semantic dependencies, provide a strong clue as to whether a video clip will be recalled on a second viewing or not. We believe that short textual descriptions encapsulate most of these relationships among the elements of a video, and thus they represent a rich yet concise source of information to tackle the problem of media memorability prediction. In this paper, we deepen the study of short captions as a means to convey in natural language the visual semantics of a video. We propose to use vector embeddings from a pretrained SBERT topic detection model with no adaptation as input features to a linear regression model, showing that, from such a representation, simpler algorithms can outperform deep visual models. Our results suggest that text descriptions expressed in natural language might be effective in embodying the visual semantics required to model video memorability.

Keywords: BERT; sentence-BERT; transformer; topic detection; video memorability; linear regression; DenseNet-121; PCA

check for updates

Citation: Kleinlein, R.; Luna-Jiménez, C.; Arias-Cuadrado, D.; Ferreiros, J.; Fernández-Martínez, F. Topic-Oriented Text Features Can Match Visual Deep Models of Video Memorability. *Appl. Sci.* **2021**, *11*, 7406. https://doi.org/10.3390/app11167406

Academic Editor: Francesco Bianconi

Received: 28 June 2021
Accepted: 9 August 2021
Published: 12 August 2021

1. Introduction

The human brain has evolved to potentially hold an incredibly large amount of detailed visual memories for long periods of time, even a lifetime. It is thanks to this ability that the vast majority of people can quickly identify other people, places and objects that have been previously seen with little to no effort. Interestingly, and opposed to the traditional belief, human memory is not completely subjective, but rather an intrinsic property of an image [1]. In this way, the existing literature seems to suggest that not all input stimuli are equally well remembered.

The fast pace at which digital media production is growing allows us to enjoy hundreds of hours of video clips from almost anywhere. Although TV was previously the most common multimedia source of information, it is customary nowadays to browse the internet, read the news and consume video streaming services from a variety of different electronic devices. Consequently, we are now exposed to huge amounts of video clips, and companies and institutions alike struggle to catch our attention and make their messages persist in our memory. It is in this context that a system able to automatically predict what videos will remain memorable and which ones will be quickly forgotten would have huge applicability, as well as an obvious scientific interest.

Even though our understanding on the computational modeling of media memorability is still in its early stages, there is a large record of previous work on this topic. From the fields of psychology and cognitive sciences, the interest dates back to the seminal studies of R.N. Shepard (1967) and Standing (1973) [2,3]. Nonetheless, the advent of computational tools and pattern recognition models has opened the gates for further investigation on the properties that make something particularly memorable, starting from the seminal work of Isola et al. [1,4,5].

Even though there is a general agreement now in recognizing semantic features as high-quality predictors of an image or video memorability [6], the way these semantics may be implemented into a computational model are still largely unknown. We reckon natural language as a human-friendly mechanism to describe with any desired degree of detail most of the scenes and experiences in our lives and, therefore, a short description should be able to convey most of the implicit semantics within a video. In this paper, we focus on the potential of such textual captions alone as predictors of media memorability.

The rest of the paper is organized as follows: Section 2 outlines previous studies on the computational modeling of media memorability and its relationship with high-level semantic features. Afterwards, although they have not been developed by us, for the sake of clearness Section 3 introduces the reader to the data sets used in our experiments. The intrinsic semantics of each dataset are studied from an unsupervised point of view in Section 4. Our media memorability predictive models are described in full in Section 5, after which Section 6 delves into the experimentation process carried out. Finally, we include some conclusions and hints on future paths to extend this investigation in Section 7.

2. Related Work

Long thought to be completely subjective, the combination of cognitive sciences and computer science is throwing light upon our knowledge on perceptual appraisals such as the asthetic experience, the interest raised by an image or the memorability of a media clip. In fact, the work of Jaegle et al. [7] shows that there are specific brain regions (such as the inferotemporal cortex) that play a fundamental role in encoding and filtering what information is worth keeping in memory. A fundamental discovery made by them is that these brain areas specialize in different topics over time. Hence, for instance, natural landscapes tend to be rather forgotten, as opposed to faces, which are generally well remembered even in monkeys. Contrary to intuition, emotional traits do not seem to contribute positively to memorability [8,9].

Experiments suggest that memory depends more on the conceptual structure of the input than its perceptual distinctiveness [10]. Moreover, it seems that a major principle in creating new memories comes from brain dealing with scene and object representations at the same level of abstraction [11]. That alone highlights the need for global descriptors of the media content if the goal is to predict its likelihood to be remembered.

As it has been already commented, although details of images and videos are well remembered, most of the process of recalling a previous experience comes from the overall perception (i.e., the global representation) that we make of them. Natural language provides an efficient and robust way to describe most of the world around us. As a matter of fact, language can be used to encapsulate the semantics of a video in a concise, relatively cheap and comprehensible way. Furthermore, recent advances such as the BERT model have greatly contributed to enhance natural language processing [12]. The BERT family of algorithms is based on the Transformer architecture [13], a particular type of neural processing unit that outperforms traditional LSTM cells in many problems dealing with data structured in sequences [14].

Sentence-BERT (or SBERT) is a novel framework that allows computing sentence, paragraph and even image-level embeddings [15]. As similar word and sentence-level encoders, SBERT keeps similar samples close to each other in the embeddings' space. In particular, the SBERT models we use throughout this study are previously trained on topic detection in text documents. Therefore, with this model we are able to compute a numerical representation of the topics depicted in a video, given that a short text describes the video as objectively as possible. We hypothesize that, given that the dataset contains a range large enough of different topics, even pretrained topic detection models can perform comparably to dedicated visual models in terms of video memorability prediction.

3. Datasets

When choosing a set of data to validate our hypothesis, two aspects are of the utmost importance. Firstly, we are interested in video memorability, and therefore we exclude any image-only or audio-only corpora [16–18]. Secondly, we require that every video sample is accompanied by at least one textual description. Each of these captions must be a single sentence, no longer than a few words. Furthermore, every video must also have a memorability score attached indicating its degree of memorability. Amidst the corpora accessible by the research community that meet our criteria, two in particular stand out: VideoMem and Memento10K [19,20]. Although it is true that data are extracted from different sources and their annotation procedures are not exactly equal, in this study we shall consider that both sets of labels represent the same concept closely enough to model them following the very same principles. The next paragraphs discuss the particularities of each dataset in depth.

3.1. VideoMem

Introduced in 2018 by R. Cohendet et al., VideoMem is composed of 10,000 videos, with memorability scores attached as ground-truth labels. It represents the first attempt to build a large-scale corpus regarding video visual memory, which were typically limited to a few hundred of samples in previous works. Added to this, videos were not collected from TRECVID or Hollywood-like films [21,22], but extracted from raw footage originally designed to serve as generic material for professional editing of films or advertisements. Clips are of high quality (HD) and are provided as short soundless videos of 7 s in *.webm* format, with a bitrate of 3000 kps for 24 fps. The rationale behind this is to have a diverse dataset in terms of topics while minimizing side-effects derived from having more than one semantic unit within the same video. All of them are briefly described by a caption that sums up the visual semantics of the scene.

Its authors also adapted an annotation protocol previously developed to image memorability to the task of video memorability [1,4]. They gathered human participants in sessions during which they were presented a series of videos, and researchers asked them to hit a button every time they remembered having watched any video previously projected within the same session. Participants carried out this procedure twice. Whereas in the first one they had to respond to videos within the same session, therefore providing a label for the short-term memory (understood as memorability after a few minutes), in the second session the task referred to the videos projected during the first session. This second phase was carried out between 24 and 72 h after the first one, and it can be thought of as a long-term memorability score. Not all participants went through this second phase. This caused an inconsistency in the number of annotations per video between the short-term and the long-term cases. On average, every video was seen as a repeated target by 38 people to estimate the short-term labels whereas only 13 did this for the long-term ones. In both cases the final memorability score of a video is computed as the average percentage of people that successfully recognised having seen a video. Videos are never presented in the same fixed order, but randomly. Because it has been observed that video memorability depends in a linear way on the number of videos between two occurrences [19,20], in order to homogenize memorability labels among videos a linear correction (firstly introduced in [16]) is applied to the raw hit scores, thus obtaining the final set of labels.

However, as can be noticed from Figure 1, there is not a clear relationship between short and long term memorability scores. A video that is on average well remembered in the short-run does not necessarily remain in memory a couple of days after with the same intensity. Likewise, clips that may go unnoticed in a first stage can instead trigger a quick and clear memory during the long-term session. Because of this apparent lack of correlation between them, in this study we tackle them as two distinct problems.

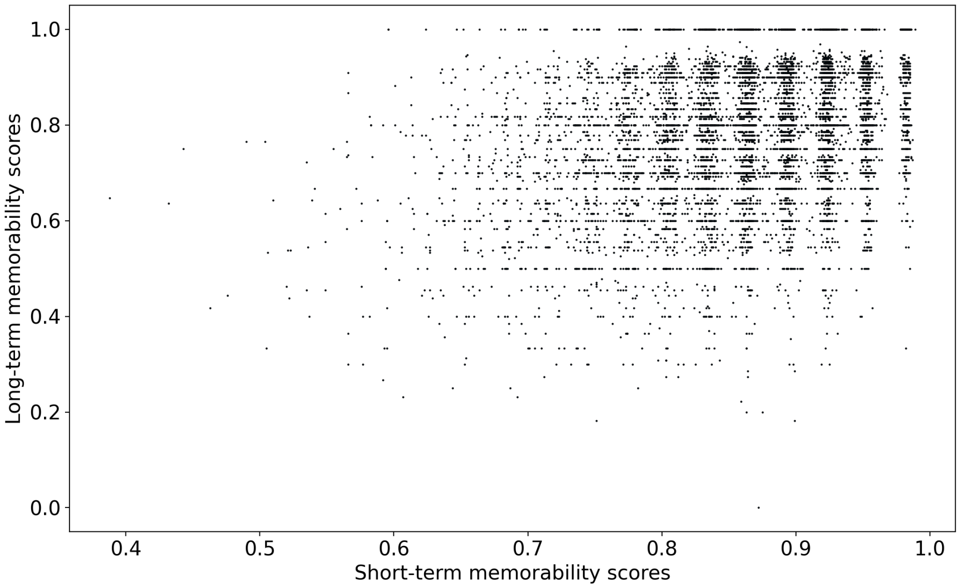


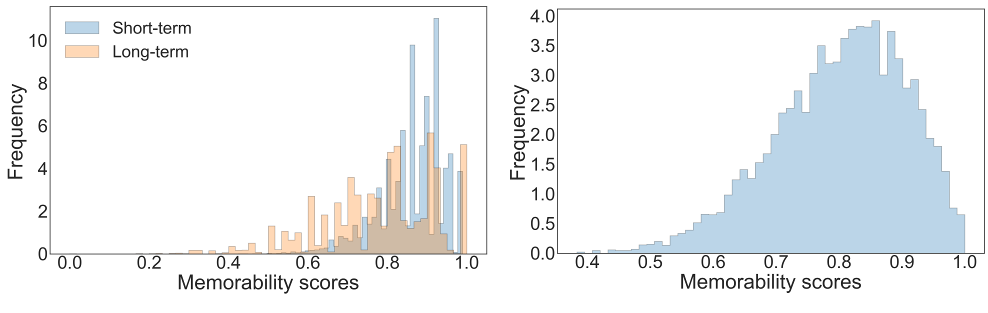
Figure 1. VideoMem's short-term scores against long-term ones. There does not seem to exist any correlation between them.

3.2. *Memento10K*

Memento10K is a corpus of 10,000 short videos released in 2020, centered around video visual memory. The authors collected a database aimed at modeling video memory decay as a function of the interval between the repetitions of a video during a visual memory game. Footage was collected scrapping from the Internet, keeping exclusively those videos that could be considered "home videos" by human workers. Hence, as a rule, the quality of the clips is considerably worse than VideoMem's. Interestingly, Memento10K's content places more attention on human actions and motion. This is particularly relevant given that only one semantic unit is intended to be depicted in each video (which are around 3 s long on average), and therefore the clips present more sudden changes in image and a higher degree of optical flow.

Because the goal of their study included predicting the memorability decay of a video (and not only its raw memorability score), the visual memory game previously explained was modified accordingly. In their experiment, crowdworkers from Amazon's Mechanical Turk (AMT) were presented a series of three-second video clips, asking them to press the space bar every time they saw a repeated video. The key difference is that they only carried out a single session, changing the lag between repetitions of the target clips. They found that annotations at different lag times are not comparable in a straightforward manner, but that the participants' success rate recognizing a video decays linearly as a function of the number of projected videos between two repetitions. Thus, in order to build a consistent set of labels, they performed a post-processing of the scores. As a result, target labels represent the likelihood a human will remember a video after a lag of 80 videos of similar length. Analogously to VideoMem, visual semantics is described by a set of captions written by human annotators. However, instead of a single one, Memento10K's clips attach five such texts each. These descriptions do not present emotional content and are limited to objective summaries.

The distributions of labels in VideoMem (both short and long-term) and Memento10K are displayed in Figure 2. In general, memorability tends to be fairly high in all cases, with barely any videos being completely forgotten. Although it is true that Memento10K presents a smoother range of scores, more homogeneously distributed despite having approximately the same number of samples, distributions display a comparable trend, and that motivates us to consider them three separate and independent problems focused on the same phenomenon, the visual memorability of videos.



(**a**) (**b**)

Figure 2. Memorability ground-truth label distribution for the three datasets used in our study. All of them are skewed towards high-memorability scores. (**a**) VideoMem's distribution of scores; (**b**) Memento10K's distribution of scores.

4. Unsupervised Topic Discovery from Video Captions

There is a general agreement that semantics play a fundamental role in understanding and remembering a video. In particular, there seems to be specific topics that greatly determine whether a video or image will be memorable or not [1,16]. Our work focuses on this idea; we aim at extracting embeddings from the captions and use them as efficient descriptors of the video's visual semantics. For this to work, however, two requisites must be fulfilled: on the one hand, our pretrained model must be able to capture the semantics of our set of sentences and embody it in a numerical embedding. Otherwise the automatically extracted embeddings would barely make sense from the point of view of language processing. On the other hand, the semantic units we are able to extract must show some degree of alignment with the memorability scores.

4.1. Out-of-Vocabulary Words

The vocabulary observed at training time by the model in charge of encoding text features must coincide as much as possible with the vocabulary available in our captions. In order to encode texts into a numerical form, we compute their sentence embeddings using a pretrained transformer-based network Sentence-BERT [15]. This model is a variation over the widely popular BERT language model [12]. In particular, we make use of the *nli-bert-base* pretrained version, which is trained on a combination of the SNLI and the Multi-Genre NLI datasets [23,24]. These text corpora account for a million of sentence pairs labeled as *contradiction*, *entailment* or *neutral*. The sources these sentences were extracted from are rich in topics, coming from fiction novels, travel guides or photo captions, to name a few. By means of this model, we are able to compute a single 768-dimensional embedding for every sentence.

Still, there is a gap between the vocabulary observed by the pretrained model and that of our datasets. This is a fundamental issue to investigate, given that memorability datasets span several different topics and our SBERT model should be able to extract meaningful features for any of them. In fact, although SBERT successfully builds sentence-level embeddings from the sequence of words in a sentence, it tags out-of-vocabulary (OOV) words as *unkown*. Whereas in VideoMem, 21% of the words are labeled so, in Memento10K, this rate rises to 30%. Amidst these OOV words we mostly observe terms used in casual conversations such as *xmas*, *veggie* or *piggy*. These words contribute with a fixed value to the sentence embedding. It is fundamental to highlight that no adaptation at all is performed over these embeddings or the Sentence-BERT model used to compute them, thus these words remain as unknown to the model, introducing some degree of noise to the set of embeddings.

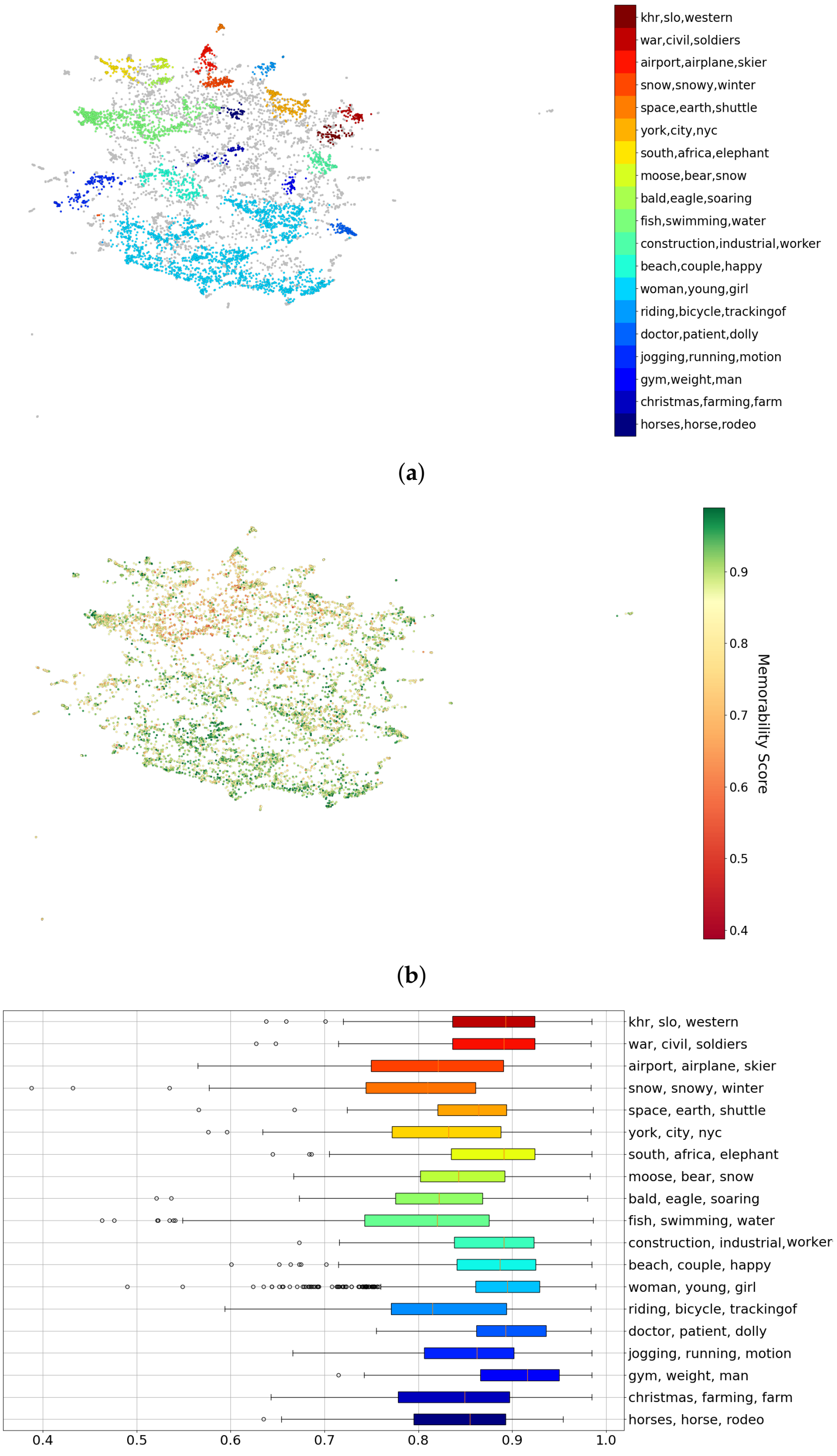
4.2. Relationship between Topics and Memorability

The second of the conditions required in a topic-oriented modeling of memorability is that there is an effective relationship between those topics and their overall degree of memorability. We thus aim at finding the set of topics available in our corpora, trying at the same time to understand the way in which they relate with memorability scores. However, we have no direct access to the semantic content of the videos, which need to be derived from the video captions in an unsupervised way.

To that aim, we carry out two consecutive transformations over the data. First, the 768-dimensional SBERT embeddings' dimensionality is reduced to 5 dimensions through UMAP [25] using the cosine distance over the input data. This technique is commonly used since it performs a non-linear dimensionality reduction while preserving most of the spatial distribution of the embeddings in the original space. Afterwards, a HDBSCAN clustering method on the euclidean distance is applied [26]. One of the main advantages of the DBSCAN family of algorithms is that it is not a requisite to know the number of classes a priori. Additionally, it allows samples to be considered noise when they are not sufficiently close to a densely-connected set of points, thus leaving us with a set of topics whose data are close enough to be sure they share a great amount of semantic information.

This way, Figure 3a displays the UMAP projection of VideoMem's SBERT embeddings over the plane, colored according by the topics found in the clustering phase. Samples labeled as noise are displayed in gray color. Notice that we do not have a fixed set of tags to represent each topic but, instead, we define a topic by its top three most representative words according to a TF-iDF word scoring method [27]. In this technique, all the captions from a cluster are grouped together as sentences of a unique document. The trade-off between the occurrence of a word within a cluster, altogether with the distinctiveness it provides against other clusters, yields the TF-iDF word scoring for every document observed. Comparing Figure 3a,b, we can see that there seems to be some sort of link between particularly memorable topics and others that are overall easily forgotten in the short-term. Figure 3c offers further information regarding this fact. In spite of not finding topics particularly forgettable or memorable, topics do have on average different degrees of memorability each, encouraging us to use this knowledge to build our models. The exact same phenomena are observed in Figure 4 but for the long-term memorability. In this case, however, the difference between topics is not that evident, something we attribute to the relative lack of annotations per sample of this problem, which may be preventing topics to clearly show the properties seen in the short-term, something that may be only hinted at here.

Contrarily to clips in VideoMem, Memento10K's videos are described by not one, but five different captions. As alternatives to exploit the semantic information within them, we evaluate two strategies: a single average BERT embedding computed at an early stage from the sentence-level initial embeddings (early AVG), and taking each sentence independently and averaging over their posteriors at prediction time (late AVG). Figure 5 shows the relationship between topics and memorability found in the early AVG scenario. From this figure it seems clear that this dataset is centered around people and human actions, whereas the distribution of scores per topic shows great variability among different subjects found. Nonetheless, videos related to *{people, water, man}* seem to stand out as particularly memorable. Still, averaging over embeddings leaves us with less information available about the clusters, given people is the main topic in the corpus and words such as *man*, *person* or *people* appear in almost all of them. Figure 6 presents the same results but for the late AVG method. In this case, it is indeed reasonable to say that there are specific topics that show an overall higher memorability rate than others. In particular, the clusters found to relate to *{woman, girl, camera}*, *{food, spoon, baby}* and *{gun, man, shooting}*. Subjects that are the closest related to nature and landscapes generally exhibit worse average memorability ratios, in line with the existing literature.

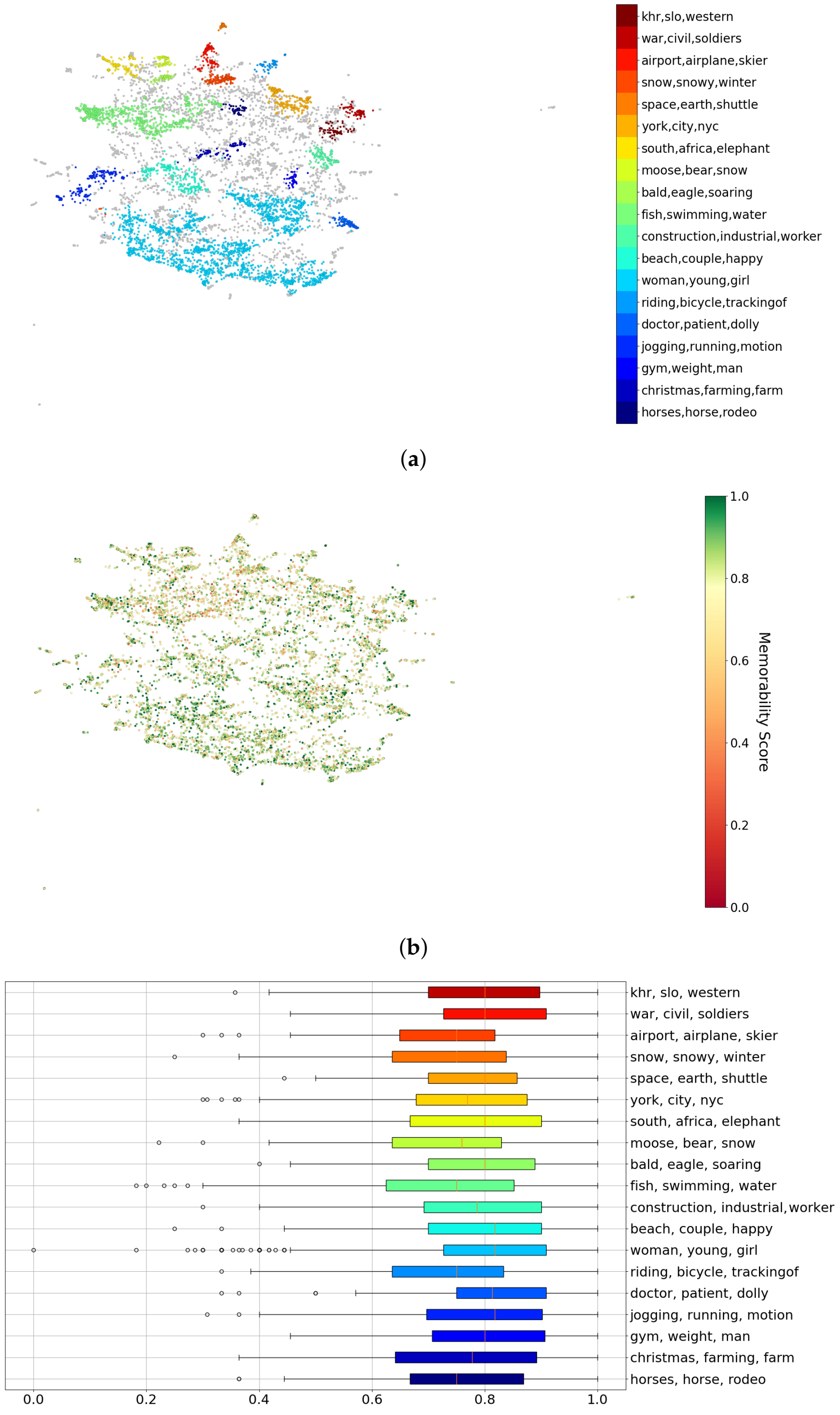


(**a**)

(**b**)

(**c**)

Figure 3. Automatic topic discovery from Sentence-BERT vector embeddings in the short-term memorability problem. From an unsupervised point of view, there seems to be some topics that tend to produce stronger memories. In particular, those related to people. (**a**) VideoMem's topics. Each one is defined by each cluster's top 3 most representative words. Gray samples cannot be assigned to any topic. (**b**) Samples shown in color according to their memorability score. Low memorability regions can be spotted. (**c**) Distribution of memorability scores within each detected topic.

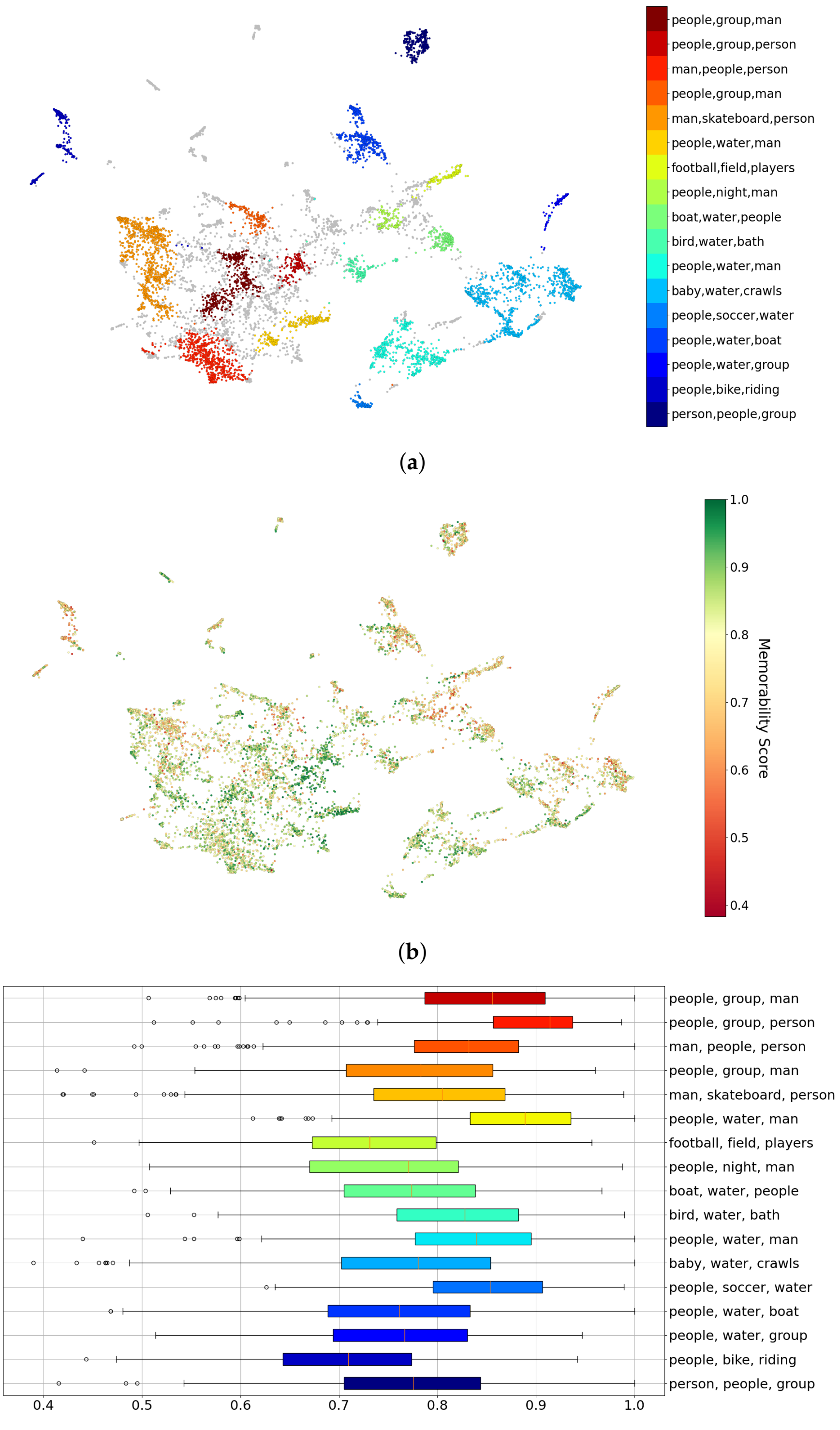


(**a**)

(**b**)

(**c**)

Figure 4. Automatic topic discovery from Sentence-BERT vector embeddings in the long-term memorability problem. Here the separability by topic is not particularly evident, perhaps due to the relatively low amount of annotations. (**a**) VideoMem's topics. Each one is defined by each cluster's top 3 most representative words. (**b**) Samples shown in color according to their memorability score. Again, low memorability regions can be spotted. (**c**) Distribution of memorability scores within each detected topic.

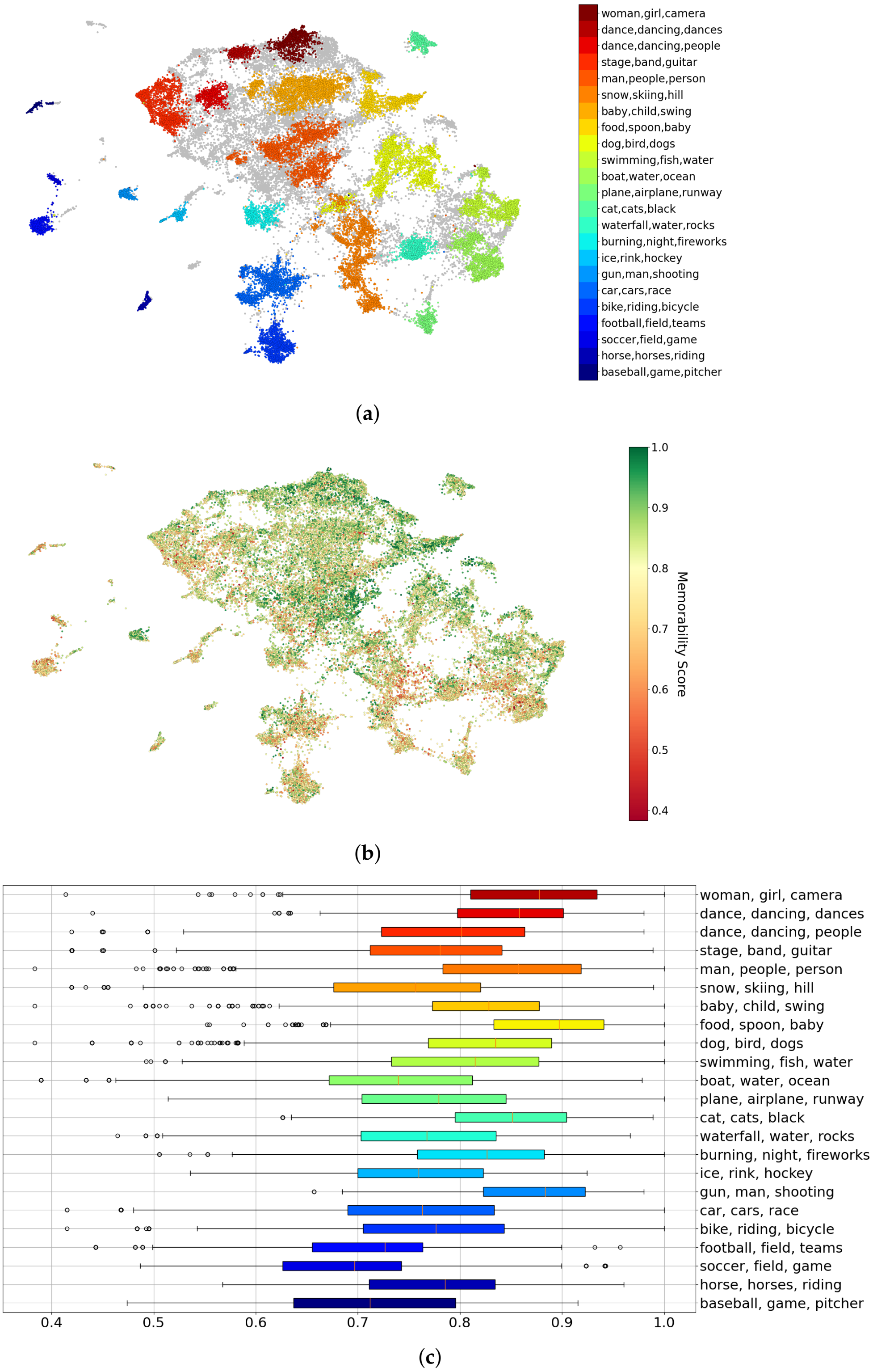


(**a**)

(**b**)

(**c**)

Figure 5. Automatic topic discovery from Sentence-BERT vector embeddings in the Memento10K dataset under an early AVG strategy. Topics found appear mainly targeted at human actions. (**a**) Memento10K's topics detected from the early AVG scheme. (**b**) Samples shown in color according to their memorability score. (**c**) Distribution of memorability scores within each detected topic.



(**a**)

(**b**)

(**c**)

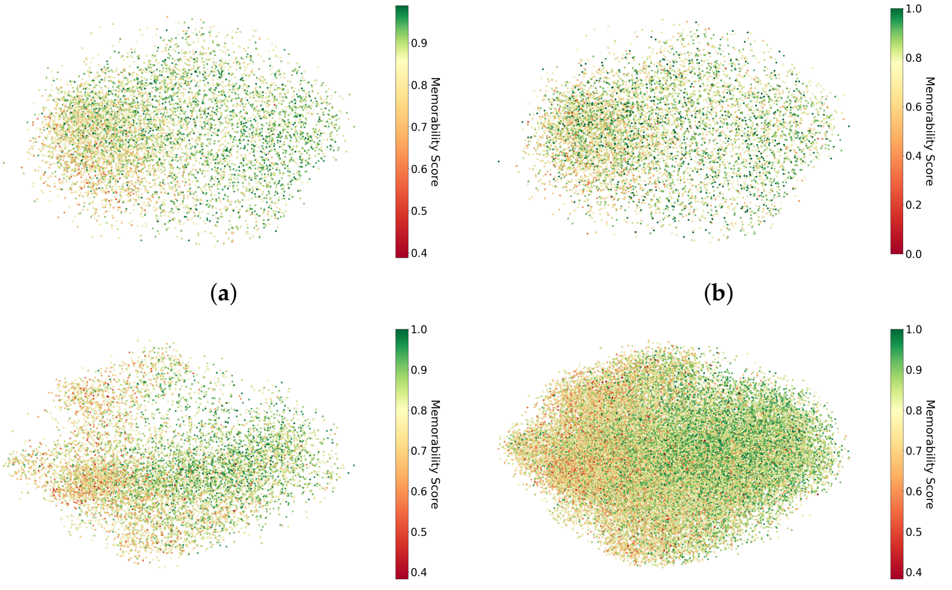
Figure 6. Automatic topic discovery from Sentence-BERT vector embeddings in the Memento10K dataset under a late AVG strategy. A finer degree of detail can be found here about the content of the topics. (**a**) Memento10K's topics detected from a late AVG scheme. (**b**) Samples shown in color according to their memorability score. (**c**) Distribution of memorability scores within each detected topic.

5. Predictive Models of Video Memorability

The aim of our study is to evaluate whether simple features extracted from short descriptive texts of the visual semantics of a video can be competitive against purely visual features fed to deep neural models. Consequently, predictive models are deliberately kept as simple as possible, leaving most of the weight of the prediction on the shoulders of the text features.

5.1. Textual Branch: SBERT-Based Models

Although UMAP provides us with features that apparently capture the semantics of our captions, this method applies non-linear transformations to the input data, potentially causing undesirable side-effects. In fact, after reducing the dimensionality of the original SBERT embeddings under a *Principal Components Analysis* (PCA), the relationship between said vector representation and their memorability labels seems to emerge in a direct way (Figure 7). This approach applies only linear operations over the input space, thus keeping a representation space closer to the original one. It can be seen that the least memorable samples tend to cluster together in all the cases, whereas highly memorable ones display a more diverse spatial distribution. Because even in a very low dimensionality space, samples display such alignment with the task of predicting video memorability, we use a PCA-projected version of the SBERT embeddings as input features to our models.



(**a**) (**b**)

(**c**) (**d**)

Figure 7. 2-dimensional PCA projection of SBERT embeddings by their memorability score. (**a**) VideoMem (short-term). (**b**) VideoMem (long-term). (**c**) Memento10K (early AVG). (**d**) Memento10K (late AVG).

The lower branch of the diagram in Figure 8 shows that linear regression models are used when modeling memorability from text encodings. The input to these models is an embedding from a pretrained Sentence-BERT, one vector at a time. We also explore the effect of reducing the size of the original embeddings (768-dimensional) via a PCA decomposition. We reckon there is some degree of redundancy in the original captions describing the visual semantics, partially manageable by extracting the most relevant features within data. We do not carry out any data augmentation procedure or text processing over the original sentences or their representation.

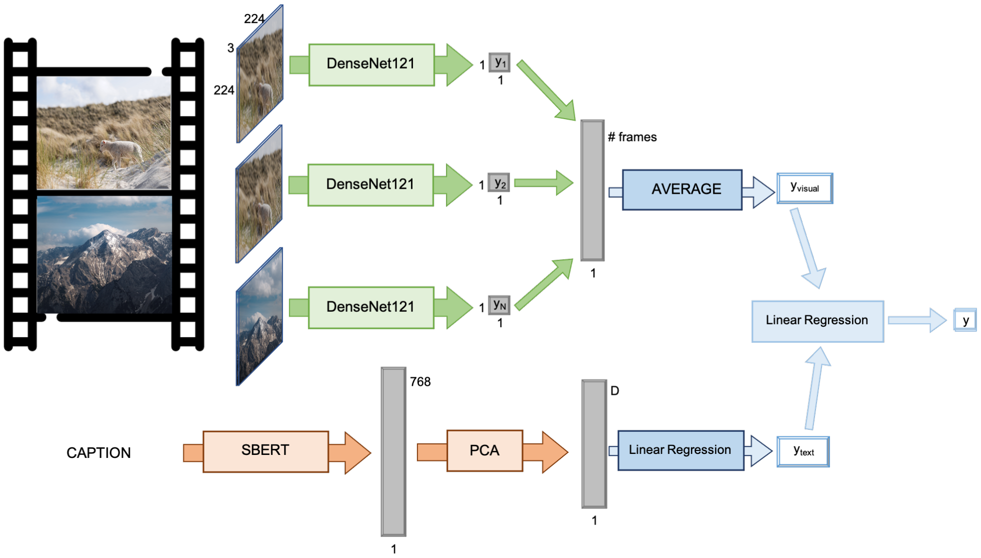


Figure 8. Predictive models used in this study. Visual and text-based branches are combined in a late fusion scheme.

5.2. Visual Branch: Baseline

We seek a model able to provide features strongly aligned to the problem of video memorability estimation, and most importantly, coming exclusively from the visual perspective. Inspired by [20], we mimic their visual baseline, depicted schematically in the upper branch in Figure 8. Basically, an ImageNet-pretrained DenseNet-121 architecture [28,29] is further finetuned on the LaMem dataset, an image memorability corpus [16]. Finally, the model is retrained on our memorability datasets. This model goes in order through the frames of a video, computing a memorability score independently for every frame and then averaging over these predictions to compute a video-level final estimation. When training the model, we assign all the frames from a video the same memorability score.

Nonetheless, in order to alleviate some computational burden, video frames are extracted at 3 fps, and resized to 224 × 224 pixels, in RGB color format. Given the strong dependency between the perception of scenes and its visual properties, no data augmentation is applied at any stage at training time.

5.3. Mixture of Modalities

It is natural to wonder whether visual and textual representations can help each other at modeling video memorability and, if so, to what degree. Hence, we also experiment with a basic multimodal approach in which input data to a linear regressor come as the late fusion of the posterior probabilities emitted by both the visual system and the best text-based model.

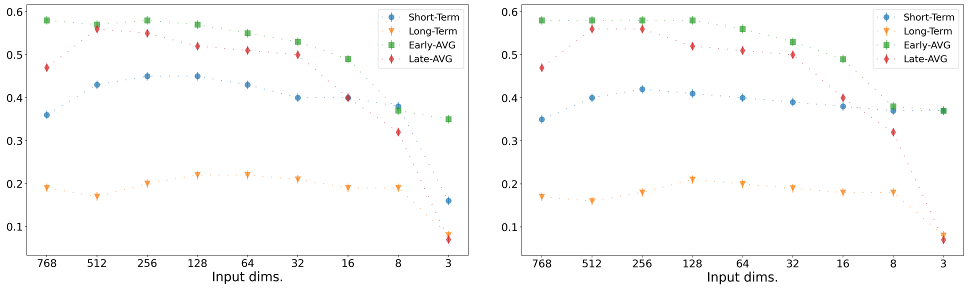
6. Experimentation and Discussion

Memorability scores are defined here as probability rates, thus the task of predicting them corresponds to a regression one. To measure the closeness between the ground truth labels and the estimated ones, we report on the Peason's correlation and the Spearman's rank correlation indices. Whereas the Pearson's coefficient measures the degree to which two variables are linearly correlated, the Spearman's correlation gives insights into how well a more general monotonic function can describe the relationship between those variables, allowing a bit more flexibility.

The whole pipeline and experiments were implemented in Pytorch [30]. Optimization of the models at training time is performed using an Adam optimizer [31], with the initial learning rate fixed at 0.001. We use the mean squared error as our loss function. In spite of having at our disposal official training, validation and test data partitions, the labels of the latter have not been released to the public to this date for neither of the datasets, making it impossible to compare our works in a straightforward way. As a consequence, all our results are obtained following a five-fold cross-validation strategy over the development sets, using at each step three folds to train, one to validate and the last one to test. When building the folds, the integrity of the videos is maintained. This is trivial in the case of VideoMem since there is only a single caption per video, but it is of the utmost importance in the case of Memento10K, given that each clip is characterized by five sentences. We evaluate our systems' performance every epoch, and training is stopped after five consecutive epochs with no improvement on the Spearman's coefficient over the validation set.

6.1. VideoMem

Table 1 shows the results obtained when exploring different input representations in the VideoMem's data, both for the short-term and long-term labels. In the short-term case, the most noticeable aspect is that the visual baseline is outperformed by the textual models. Interestingly, PCA dimensionality reduction seems to enhance the performance of the original embeddings, confirming that there is some redundancy in the original representation. From Figure 9 it is noticeable that although input lengths smaller than 128 begin to worsen, it is not until a drastic dimensionality reduction is carried out that these embeddings fall behind the visual model. The multimodal approach seems to be capturing the best of both modalities, improving upon the best single-modality option in a statistically significant way.



(**a**) (**b**)

Figure 9. The performance of linear regression models is enhanced when the original SBERT input representation is reduced via PCA, although low dimensionality spaces seem to progressively worsen. (**a**) Pearson's coefficient. (**b**) Spearman's coefficient.

Table 1. Memorability prediction on the VideoMem dataset. Reduced versions of the SBERT text embeddings easily improve upon the visual baseline.

Model	Input Dims.	Pearson's Coeff. (±0.01)		Spearman's Coeff. (±0.01)	
		Short-Term	Long-Term	Short-Term	Long-Term
Visual Baseline	224 × 224 × 3	0.38	0.11	0.39	0.12
SBERT (LR)	256	0.45	0.2	0.42	0.18
	128	0.45	**0.22**	0.41	**0.21**
Multimodal	2	**0.48**	0.2	**0.45**	0.19

We can observe similar results for the problem of predicting long-term memorability. Analogously to the previous case, BERT embeddings display robustness to PCA transforms, indeed reaching a peak performance when input is reduced to 128 or 64 dimensions. Except in the case of a radical diminution of the vector dimensionality, it is also the case that text-based models outperform the visual baseline by a significant margin. Another interesting issue is that the late fusion of posteriors does not benefit from the different predictions with respect to the textual model, whose posteriors conform the input to a multimodal system, and which predicts memorability at a comparable level. We reckon that labels from this case are noisier due to the smaller number of annotations per sample, hence posing an additional challenge when a linear regression tries to learn a pattern exclusively from the visual and textual posteriors.

6.2. Memento10K

It can be seen from Table 2 that our approach based on PCA-projections of SBERT embeddings successfully improves upon the visual baseline, too. Analogously to the case of VideoMem, Figure 9 shows how the embeddings are robust to dimensionality reduction, and only at very low dimensionalities the proposed system fails to deliver over the visual model. Likewise to the previous case, the multimodal pipeline seems to further improve the rates achieved at predicting media memorability, yet we cannot speak of statistical significance against text-based approaches. It is worth mentioning that averaging over the raw SBERT embeddings at an early stage apparently yields an overall slightly better result than averaging over posteriors, yet that difference is not statistically significant in the 256-dimensional vector space.

Table 2. Memorability prediction on Memento10K dataset using as input the average of the captions' BERT embeddings against the visual baseline.

Model	**Input Dims.**	**Pearson's Coeff. (±0.01)**		**Spearman's Coeff. (±0.01)**	
		Early AVG	**Late AVG**	**Early AVG**	**Late AVG**
Visual Baseline	224 × 224 × 3	0.43	0.43	0.42	0.43
SBERT (LR)	256	0.58	0.55	0.58	0.56
Multimodal	2	**0.59**	**0.57**	**0.60**	**0.58**

7. Conclusions and Future Work

The problem of the computational modeling of video memorability has attracted a lot of attention from the research community in recent years due to the growth in computational power and data resources. Even though it seems clear now that high-level semantic information plays a major role in the way our brain stores memories, it remains unclear how we can encapsulate such information in an efficient way that may be later used to automatically predict the memorability of a video. In this work, we propose to use natural language text captions to encode the richness of both elements of a video and their mutual interactions. Humans naturally communicate with each other in this way, hence it stands as a strong alternative to pure visual approaches.

In particular, we study the relationship between the topic depicted in a short video, purposely containing a single semantic unit, and the probability that a person will remember it at different intervals of time. To model topics, we convert the text captions that accompany every video into precomputed sentence-level, BERT-based embeddings. These vectors, which are not adapted to our tasks, were originally trained to perform topic comparison between pairs of sentences. From this representation we extracted a set of topics for each dataset in an unsupervised fashion. We observed that, in tight agreement with previous studies, as a rule subjects related to people and actions are better kept in memory than scenes of nature. However, our current setup leaves more than 20% of the words in our captions as OOV words, therefore worsening their corresponding embedding representation. Further work should study the effect of fine-tuning to video memorability corpora so the sentence embeddings may capture the finest details of the descriptions provided.

Despite this lack of adaptation, simple linear regression models based on these naïve embeddings outperform a deep neural visual model specifically trained to the task. Furthermore, PCA proved to be an appropriate technique to decrease redundancy in the original vector space while keeping apart samples with significantly different memorability scores. In fact, PCA-reduced versions of these embeddings further improved the quality of our estimations over the original ones. Only extremely small dimensionality spaces seem to distort the embeddings space as much as to see their performance below that of the visual alternative.

Inspired by the success of our multimodal approaches, based on a late fusion of the predictions made by two fully independent models, we reckon as another promising line of research the development of novel strategies to merge the features extracted from different sources. Although memorability is eminently a visual aspect of human cognition, models would likely benefit from combining visual descriptors with textual or acoustic information.

Author Contributions: Conceptualization, R.K. and F.F.-M.; data curation, R.K.; formal analysis, R.K., C.L.-J., J.F. and F.F.-M.; funding acquisition, J.F. and F.F.-M.; investigation, R.K.; methodology, R.K., C.L.-J., D.A.-C. and F.F.-M.; software, R.K. and D.A.-C.; supervision, F.F.-M.; visualization, R.K.; writing—original draft, R.K. and D.A.-C.; writing—review and editing, R.K., C.L.-J., J.F. and F.F.-M. All authors have read and agreed to the published version of the manuscript.

Funding: The work leading to these results has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the CAVIAR (MINECO, TEC2017-84593-C2-1-R) and AMIC (MINECO, TIN2017-85854-C4-4-R) projects (AEI/FEDER, UE). Ricardo Kleinlein's research was supported by the Spanish Ministry of Education (FPI grant PRE2018-083225).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in this study are publicly available after sending a request form to the original authors.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for part of this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Isola, P.; Xiao, J.; Parikh, D.; Torralba, A.; Oliva, A. What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1469–1482. [CrossRef] [PubMed]
2. Shepard, R.N. Recognition memory for words, sentences, and pictures. *J. Verbal Learn. Verbal Behav.* **1967**, *6*, 156–163. [CrossRef]
3. Standing, L. Learning 10000 pictures. *Q. J. Exp. Psychol.* **1973**, *25*, 207–222. [CrossRef] [PubMed]
4. Isola, P.; Xiao, J.; Torralba, A.; Oliva, A. What makes an image memorable? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 145–152.
5. Isola, P.; Parikh, D.; Torralba, A.; Oliva, A. Understanding the Intrinsic Memorability of Images. In *Advances in Neural Information Processing Systems*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Granada, Spain, 2011; Volume 24.
6. Bylinskii, Z.; Goetschalckx, L.; Newman, A.; Oliva, A. Memorability: An image-computable measure of information utility. *arXiv* **2021**, arXiv:2104.00805.
7. Jaegle, A.; Mehrpour, V.; Mohsenzadeh, Y.; Meyer, T.; Oliva, A.; Rust, N. Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* **2019**, *8*, e47596. [CrossRef] [PubMed]
8. Bainbridge, W.A.; Isola, P.; Oliva, A. The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **2013**, *142*, 1323–1334. [CrossRef] [PubMed]
9. Baveye, Y.; Cohendet, R.; Perreira Da Silva, M.; Le Callet, P. Deep Learning for Image Memorability Prediction: The Emotional Bias. In Proceedings of the ACM Multimedia 2016, Amsterdam, The Netherlands, 15–19 October 2016; pp. 491–495. [CrossRef]
10. Konkle, T.; Brady, T.F.; Alvarez, G.; Oliva, A. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* **2010**, *139*, 558–578. [CrossRef] [PubMed]
11. Konkle, T.; Brady, T.F.; Alvarez, G.A.; Oliva, A. Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychol. Sci.* **2010**, *21*, 1551–1556. [CrossRef] [PubMed]
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 3–5 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]

13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: California, CA, USA, 2017; Volume 30.
14. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
15. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, 3–7 November 2019.
16. Khosla, A.; Raju, A.S.; Torralba, A.; Oliva, A. Understanding and Predicting Image Memorability at a Large Scale. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
17. Ananthabhotla, I.; Ramsay, D.B.; Paradiso, J.A. HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 920–924.
18. Ramsay, D.; Ananthabhotla, I.; Paradiso, J. The Intrinsic Memorability of Everyday Sounds. In Proceedings of the Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27–29 March 2019; Audio Engineering Society: New York, NY, USA, 2019.
19. Cohendet, R.; Demarty, C.H.; Duong, N.; Engilberge, M. VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 2531–2540. [CrossRef]
20. Newman, A.; Fosco, C.; Casser, V.; Lee, A.; McNamara, B.; Oliva, A. Multimodal Memorability: Modeling Effects of Semantics and Decay on Video Memorability. *arXiv* **2020**, arXiv:2009.02568.
21. Shekhar, S.; Singal, D.; Singh, H.; Kedia, M.; Shetty, A. Show and Recall: Learning What Makes Videos Memorable. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2730–2739. [CrossRef]
22. Cohendet, R.; Yadati, K.; Duong, N.Q.K.; Demarty, C.H. Annotating, Understanding, and Predicting Long-term Video Memorability. In Proceedings of the ICMR'18: 2018 International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018. [CrossRef]
23. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 632–642. [CrossRef]
24. Williams, A.; Nangia, N.; Bowman, S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1112–1122. [CrossRef]
25. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, arXiv:1802.03426.
26. Campello, R.J.G.B.; Moulavi, D.; Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*; Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
27. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed.; Prentice Hall PTR: Upper Saddle River, NJ, USA, 2000.
28. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
29. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
30. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2019; Volume 32.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.