applied sciences

MDPI

Article

Multiple Visual-Semantic Embedding for Video Retrieval from Query Sentence

Huy Manh Nguyen, Tomo Miyazaki *, Yoshihiro Sugaya and Shinichiro Omachi

Graduate School of Engineering, Tohoku University, Sendai 9808579, Japan;
nmhuy@iic.ecei.tohoku.ac.jp (H.M.N.); sugaya@iic.ecei.tohoku.ac.jp (Y.S.); machi@ecei.tohoku.ac.jp (S.O.)
* Correspondence: tomo@tohoku.ac.jp

Abstract: Visual-semantic embedding aims to learn a joint embedding space where related video and sentence instances are located close to each other. Most existing methods put instances in a single embedding space. However, they struggle to embed instances due to the difficulty of matching visual dynamics in videos to textual features in sentences. A single space is not enough to accommodate various videos and sentences. In this paper, we propose a novel framework that maps instances into multiple individual embedding spaces so that we can capture multiple relationships between instances, leading to compelling video retrieval. We propose to produce a final similarity between instances by fusing similarities measured in each embedding space using a weighted sum strategy. We determine the weights according to a sentence. Therefore, we can flexibly emphasize an embedding space. We conducted sentence-to-video retrieval experiments on a benchmark dataset. The proposed method achieved superior performance, and the results are competitive to state-of-the-art methods. These experimental results demonstrated the effectiveness of the proposed multiple embedding approach compared to existing methods.

Keywords: video retrieval; visual-semantic embedding; multiple embedding spaces

Citation: Nguyen, H.M.; Miyazaki, T.; Sugaya, Y.; Omachi, S. Multiple Visual-Semantic Embedding for Video Retrieval from Query Sentence. *Appl. Sci.* **2021**, *11*, 3214. https://doi.org/10.3390/app11073214

Academic Editor: Hee-Deok Yang

Received: 3 March 2021
Accepted: 31 March 2021
Published: 3 April 2021

1. Introduction

Video has become an essential source for humans to learn and acquire knowledge. Due to the increased demand for sharing and accumulating information, there is a massive amount of video being produced in the world every day. However, compared to images, videos usually contain much semantic information, and thus it is hard for humans to organize videos. Therefore, it is critical to developing an algorithm that can efficiently perform multimedia analysis and automatically understand the semantic information of videos.

A common approach for video analysis and understanding is to form a joint embedding space of video and sentences using multimodal learning. Similar to the fact that humans experience the world with multiple senses, the goal of multimodal learning is to develop a model that can simultaneously process multiple modalities, such as visual, text, and audio, in an integrated manner by constructing a joint embedding space. Such models can map various modalities into a shared Euclidean space where distances and directions capture useful semantic relationships. This enables us to not only learn the semantic presentations of videos by leveraging other modalities information but also show great potential in tasks such as cross-modal retrieval and visual recognition.

Recent works in learning an embedding space bridge the gap between sentence and visual information by utilizing advancements in image and language understanding [1,2]. Most approaches build the embedding space by connecting visual and textual embedding paths. Generally, a visual path uses a Convolutional Neural Network (CNN) to transform visual appearances into a vector. Likewise, a Recurrent Neural Network (RNN) embeds sentences in a textual path. However, capturing the relationship between video and

sentence remains challenging. Recent works suffer from extracting visual dynamics in a video, such as object movements, actions, scene changes, etc. [3–6].

In this paper, we propose a novel framework equipped with multiple embedding networks so that we can capture various relationships between video and sentence, leading to more compelling video retrieval. Precisely, one network captures the relationship between an overall appearance in the video and a textual feature. Others consider consecutive appearances or action features. Thus, the networks learn their own embedding spaces. We fuse the similarities measured in the multiple spaces using the weighted summing strategy to produce the final similarity between video and sentence.

The main contribution of this paper is a novel approach to measure the similarity between video and sentence by fusing similarities in multiple embedding spaces. Consequently, we can measure the similarity with multiple understandings and relationships of video and sentence. We also emphasize that the proposed method can quickly expand the number of embedding spaces. We demonstrated the effectiveness of the proposed method by the experimental results. We conducted video retrieval experiments using query sentences on the standard benchmark dataset and demonstrated an improvement of our approach compared to existing methods.

2. Related Work

2.1. Vision and Language Understanding

There have been many efforts in understanding vision and language, which focus on making a connection between visual and linguistic information. Various applications need such a connection to realize tagging [7–10], retrieval [11], captioning [12–14], and visual question answering [15–18].

Image tagging is a task to learn a connection between image and tag, a short description of the image in a few words. Generally, tags are assigned by multi-label classification using a binary classifier, e.g., nearest neighbor [7]. There are, in the literature, works on more complex cases. Xu et al. tagged images using practical vocabularies for mobile users on social media [8]. They learned missing and defective tags of the images from the training dataset. Li et al. recovered missing tags from partially tagged images [9]. They used low-rank approximation to partiallyoptimize completed tags from the given tags. Rahman et al. addressed zero-shot image tagging, which assigns tags according to unseen objects [10]. They located objects and perform binary classification using the multiple Instance Learning (MIL) framework. Image captioning is a task for generating description of images [19]. This task needs a connection between sentences and visual information, such as objects, their attributes, and their relationships. A common approach is based on CNNs and RNN [13]. Specifically, CNN extracts a set of feature vectors from an input image. Then, RNN iteratively transforms each feature vector into words, resulting in a caption. Chen et al. use a graph structure to describe object, attribute, and relationship in images [14]. The graph structure facilitates making connections to texts. Visual question answering is a task to generate answer description for a question about an image [15–17]. This task requires further understandings of visual and textual information. A model needs to extract a target from the question, find the target in the image, and generate an answer. Thus, two connections are necessary, question-to-image and image-to-answer.

2.2. Video and Sentence Embedding

Video and sentence embedding is to learn a joint embedding space of visual and textual features. The embedding is essential for video retrieval from query sentences.

In image retrieval, we need to learn a connection between query texts and target images, which focus on building a joint visual-semantic space [11]. Recent works in image-to-text retrieval embed image and sentence into a joint embedding trained with ranking loss. A penalty is applied when an incorrect sentence is ranked higher than the correct sentence [20–24]. Another popular loss function is triplet ranking [1,25–27]. The VSE++

model improves the triplet ranking loss by focusing on the hardest negative samples (the most similar yet incorrect sample in a mini-batch) [28].

Like image-text retrieval approaches, most video-to-text retrieval methods learn a joint embedding space [29–31]. The method [3] incorporates web images searched with a sentence into an embedding process to take into account fine-grained visual concepts. However, the method treats video frames as a set of unrelated images and averages them out to get a final video embedding vector. Thus, it may lead to inefficiency in learning an embedding space since temporal information of the video is lost.

Mithun et al. tackle this issue [32] by learning two different embedding spaces to consider temporal and appearance. Furthermore, they extract audio features from the video for learning space. This approach achieved accurate performance for the sentence-to-video retrieval task. However, this approach puts equal importance on both embedding spaces. Practically, equal importance does not work well. There are cases that one embedding space is more important than the others in capturing semantic similarity between video and sentence. Therefore, we propose a novel mechanism emphasizing a space so that we can know the critical visual cues.

The existing works suffer from extracting visual dynamics in a video, such as object movements, actions, scene changes, etc. [3–6]. The main obstacle is due to the limited number of embedding spaces, i.e., only a single space. An embedding space aligns visual and textual features globally. Thus, it is hard to align textual and local visual features in videos, such as temporal information. Therefore, we propose to use multiple embedding spaces. Specifically, we align global and local features in separate embedding spaces. Additionally, we measure the similarity between video and sentences in each space and merge the similarities using dynamical weights. Consequently, we can align video and sentence globally and locally.

3. the Proposed Method

3.1. Overview

We describe the problem of learning a joint embedding space for video and sentence embedding. Given video and sentence instances, which are sequences of frames and words, the aim is to train embedding functions that map them into a joint embedding space. Formally, we use embedding functions $\mathrm{f} : \mathcal{X} \rightarrow \mathcal{Z}$ and $\mathrm{g} : \mathcal{Y} \rightarrow \mathcal{Z}$, where $\mathcal{X}$ and $\mathcal{Y}$ are video and sentence domains, respectively, and $\mathcal{Z}$ is a joint embedding space. The similarity $\mathrm{s}(\mathrm{f}(x), \mathrm{g}(y))$ between $\mathcal{X}$ and $\mathcal{Y}$ is calculated in $\mathcal{Z}$ by a certain measurement. Therefore, the ultimate goal is learning f and g satisfying the following equation: $\mathrm{s}(\mathrm{f}(x_i), \mathrm{g}(y_i)) > \mathrm{s}(\mathrm{f}(x_i), \mathrm{g}(y_j)), \forall\ i \neq j$. This encourages similarity to increase in a same pair x_i and y_i, whereas it decreases in a different pair x_i and y_j.

As we illustrated in the overview of the proposed framework in Figure 1, there are two networks for embedding videos: global and the sequential visual networks. These networks have their counterparts that are embedding the sentence. Thus, we develop multiple visual and textual networks that are responsible for embedding a video and a sentence, respectively. We form one embedding space by merging two networks (one visual and one textual) so that we can align a visual feature to textual information. Specifically, the global visual network aligns average visual information of the video to the sentence. Likewise, the sequential visual network aligns a temporal context to the sentence. Consequently, the networks receive a video and sentence as inputs and map them into the joint embedding spaces.

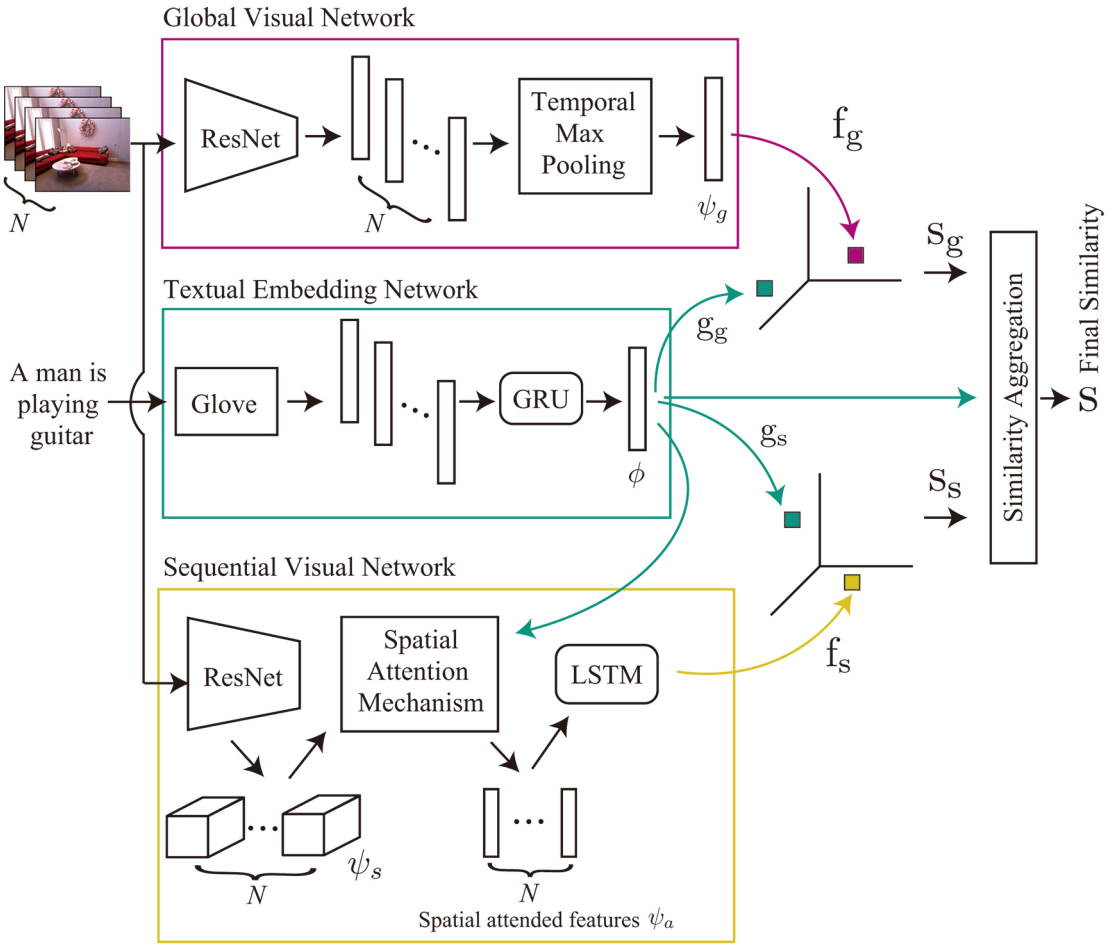

Figure 1. Overview of the proposed framework. f_g, f_s, g_g, and g_s are embedding functions.

We propose to fuse similarities measured in the embedding spaces to produce the final similarity. One embedding space has the global visual network extracting a visual object feature from the video. Thus, this embedding space describes a relationship between global visual appearances and textual features. The other embedding space captures the relationship between sequential appearances and textual features. The similarity scores of the two embedding spaces are then combined using the weighted sum to put more emphasis on one embedding space than the other space.

Our motivation is twofold. Firstly, we aim to develop a more robust similarity measurement by combining multiple embedding spaces in a weighted manner instead of using a hard-coded average. Videos and sentences require flexible attention, such as attention to global and local visual features. For example, the words in "a girl is dancing" need different visual feature perspectives. "Girl" needs to attend global visual features. In contrast, "dancing" needs to attend local visual features to capture temporal information. Secondly, in order to highlight spatial and temporal information of videos according to a sentence, we need to develop a mechanism that utilizes textual information to emphasize spatial and temporal features. Overall, we can dynamically attend to global and local visual features by varying their weights. Consequently, we can learn appropriate correspondences between visual and textual features.

3.2. Textual Embedding Network

We decompose the sentence to variable-length sequences of tokens. Then, we transform each token into a 300-dimensional vector representation by using the pre-trained GloVe model [33]. The length of the tokens depends on the sentence. Therefore, in order to obtain a fixed-length meaningful representation, we encode the GloVe vectors using the

Gated Recurrent Unit (GRU) [34] with H hidden states, resulting in a vector $\phi(y) \in \mathbb{R}^H$. We set $H = 512$. This embedded vector $\phi(y)$ goes to four processing modules: global and sequential visual networks, spatial attention mechanism, and similarity aggregation. We further transform $\phi(y)$ in each processing module. H is the dimension of a joint embedding space for textual and visual features. We determined the dimension $H = 512$ by considering two facts. Firstly, the visual feature needs higher dimensions than the textual feature. Images contain richer information than texts. Secondly, the dimension of the initial textual feature is 300 as extracted by GloVe. Considering these two facts, we determined H as 512.

3.3. Global Visual Network

The global visual network aims to obtain a vector representation that is a general visual feature over the video frames. We divide the input video frames into N chunks, and one frame is sampled from each chunk by random sampling. We set $N = 20$. We heuristically determined the chunks as $N = 20$. Although there are no specific reasons, the experimental results show that the N worked well. We extract visual features from the sampled frames using the ResNet-152 pre-trained on the ImageNet dataset [35]. Specifically, we resize the sampled frames 320×240 to 224×224, and then the ResNet encodes them, resulting in 2048-dimensional N vectors. Note that we extract the vectors directly from the last fully connected layer of the ResNet. Subsequently, we apply average-pooling to the N vectors to merge them. Consequently, we obtain a feature vector $\psi_g(x) \in \mathbb{R}^{2048}$ containing a global visual feature in the video.

We learn a joint embedding space of the global visual and textual networks. As defined in Equations (1) and (2), we use embedding functions to embed $\psi_g(x)$ and $\phi(y)$ into a D-dimensional space.

$$\mathrm{f_g}(x) = W_g \psi_g(x) + b_g \tag{1}$$

$$\mathrm{g_g}(y) = \hat{W}_g \phi(y) + \hat{b}_g \tag{2}$$

There are learnable parameters $W_g \in \mathbb{R}^{2048 \times D}$, $\hat{W}_g \in \mathbb{R}^{H \times D}$, and $b_g, \hat{b}_g \in \mathbb{R}^D$. We set $D = 512$ in this paper.

We use cosine similarity to measure the similarity $\mathrm{s_g}(y, x)$ between the video and sentence in the joint embedding space.

$$\mathrm{s_g}(y, x) = \frac{\mathrm{f_g}(x) \cdot \mathrm{g_g}(y)}{\| \mathrm{f_g}(x) \| \| \mathrm{g_g}(y) \|} \tag{3}$$

3.4. Sequential Visual Network

Similar to the global visual network, we divide the input video frames into N chunks and take the first frames of each chunk as the input of the sequential visual network. Thus, we use 20 discontinuous frames. Then, we use the ResNet-152 to extract a $7 \times 7 \times 2048$-dimensional vector $\psi_s(x)$ from each frame at the last convolution layer of the ResNet. The $\psi_s(x)$ contains visual features in spatial regions. Considering spatial regions where we should pay further attention may change by sentences, we need to explore relationships between spatial and textual features. Therefore, we incorporate a spatial attention mechanism into the sequential visual network. We apply the attention mechanism to $\psi_s(x)$ to emphasize spatial regions. Finally, we capture the sequential information of the video using a single layer Long-Short Term Memory (LSTM) [36] with an H-dimensional hidden state. We denote the vector embedded by the sequential visual network as $\mathrm{f_s}(x) \in \mathbb{R}^D$.

The sequential visual network uses LSTM to capture a meaningful sequential representation of the video. However, the LSTM transforms all spatial details of a video into a flat representation, resulting in losing its spatial reasoning with the sentence. Therefore, we employ the spatial attention mechanism in order to obtain a spatial relationship between video and sentence. Inspired by the work [18], we develop the spatial attention mechanism to learn with regions in a frame to attend to each word in the sentence.

We illustrate the spatial attention mechanism in Figure 2. The mechanism associates the visual feature vector $\psi_s(x)$ with the textual vector $\phi(y)$ to produce a spatial attention map $a \in \mathbb{R}^{7\times7}$. Then, we combine $\psi_s(x)$ with a to produce the spatial attended feature $\psi_a(x)$. Formally, $\psi_a(x)$ and a are defined in Equation (5) and Equation (6), respectively.

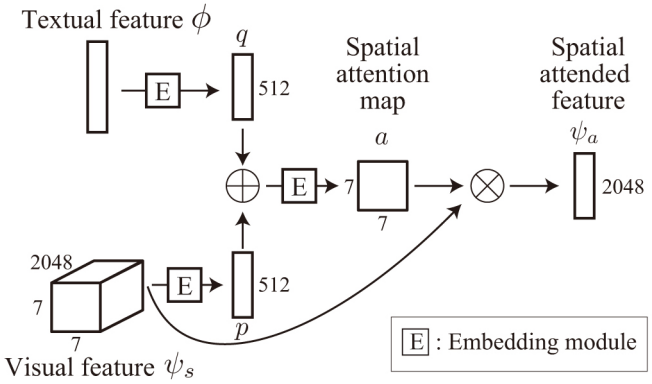

Figure 2. The spatial attention mechanism used in the sequential visual network.

Specifically, Equation (4) defines an embedding function of the sequential visual network. γ represents the LSTM, and $\otimes$ is the element-wise product. The p and q mean intermediate outputs, which are 512-dimensional vectors. We have learnable parameters $W_a \in \mathbb{R}^{512\times(7\times7)}$, $W_p \in \mathbb{R}^{(7\times7\times2048)\times512}$, $\hat{W}_q \in \mathbb{R}^{H\times512}$, $b_a \in \mathbb{R}^{7\times7}$, and $b_p, \hat{b}_q \in \mathbb{R}^{512}$.

$$\mathrm{f_s}(x) = \gamma(\psi_a(x)) \tag{4}$$

$$\psi_a(x) = \psi_s(x) \otimes a \tag{5}$$

$$a = \mathrm{softmax}(\tanh(W_a(p+q) + b_a)) \tag{6}$$

$$p = \tanh(W_p\psi_s(x) + b_p) \tag{7}$$

$$q = \tanh(\hat{W}_q\phi(y) + \hat{b}_q) \tag{8}$$

The joint embedding space of the sequential visual and textual networks is formed by $\mathrm{f_s}(x)$ and $\mathrm{g_s}(y)$. We measure the similarity $\mathrm{s_s}(x,y)$ in this joint embedding space. The formulations are described below, where $\hat{W}_s \in \mathbb{R}^{H\times D}$ and $\hat{b}_s \in \mathbb{R}^D$ are learnable parameters.

$$\mathrm{g_s}(y) = \hat{W}_s\phi(y) + \hat{b}_s \tag{9}$$

$$\mathrm{s_s}(y,x) = \frac{\mathrm{f_s}(x) \cdot \mathrm{g_s}(y)}{\| \mathrm{f_s}(x) \| \| \mathrm{g_s}(y) \|} \tag{10}$$

3.5. Similarity Aggregation

There are many approaches to aggregate similarities. An average is a straightforward approach. Some cases work well with an average. However, the average may cause unexpected behaviors if an inaccurate similarity is considerably high or low. Therefore, we adopt the Mixture of Experts fusion strategy [37] for aggregating similarities with weights that changes according to the input sentence. Consequently, we can emphasize one embedding space using the weights for merging the multiple similarities.

We propose to aggregate the similarities measured in multiple embedding spaces so that we can produce the final similarity $\mathrm{s}(x,y)$ with various understandings of videos and sentences. We illustrate the proposed similarity aggregation in Figure 3. Specifically, we merge the similarities using the weight $W_m \in \mathbb{R}^2$ generated by considering the textual feature. $\hat{W}_t \in \mathbb{R}^{D\times2}$ is a learnable parameter. concat() is a function that concatenates given scalar values.

$$\mathrm{s}(x,y) = W_m\big(\mathrm{concat}(\mathrm{s_g}(x,y), \mathrm{s_s}(x,y))\big) \tag{11}$$

$$W_m = \mathrm{softmax}(\hat{W}_t\phi(y)) \tag{12}$$

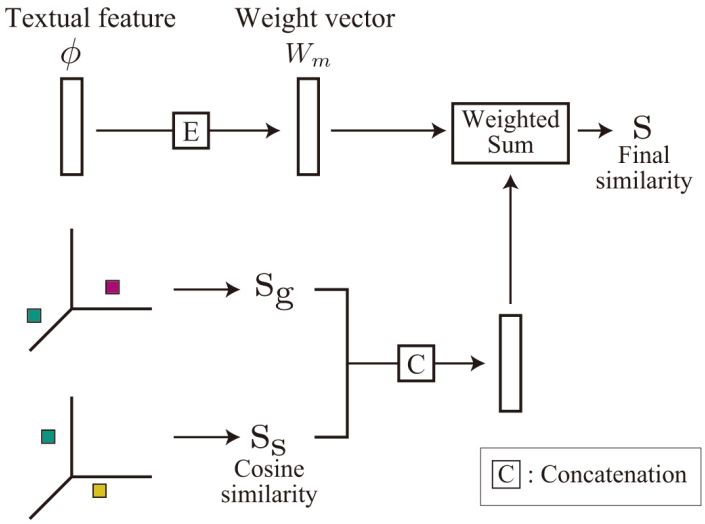

Figure 3. Similarity aggregation.

3.6. Optimization

As described in Section 3.1, we optimize the proposed architecture by enforcing similarity of a video x_i, and its counterpart sentence y_i will be greater than similarities of the video x_i and other sentence y_j, such as $\mathrm{s}(x_i, y_i) \geq \mathrm{s}(x_i, y_j)$ or $\mathrm{s}(x_j, y_i)$. We achieve this by using the triplet ranking loss [38–40], where α is a margin.

$$\mathcal{L}_{\mathrm{s}}(x_i, y_i, y_j) = \max\{0, \alpha - \mathrm{s}(x_i, y_i) + \mathrm{s}(x_i, y_j)\} \tag{13}$$

$$\mathcal{L}_{\mathrm{v}}(x_i, y_i, x_j) = \max\{0, \alpha - \mathrm{s}(x_i, y_i) + \mathrm{s}(x_j, y_i)\} \tag{14}$$

Given a dataset $\mathcal{D} = (x_i, y_i)_{i=1}^{N}$, with N pairs, we optimize the following equation by stochastic gradient descent [41,42].

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\mathcal{L}_{\mathrm{s}}(x_i, y_i, y_j) + \mathcal{L}_{\mathrm{v}}(x_i, y_i, x_j)\right) \tag{15}$$

The training time was less than 3 h using a GTX 1080 Ti GPU. We fixed the parameters of ResNet and GloVe during training. Thus, training time was relatively short. We note that the proposed method is worse than the other methods in the computational complexity due to the number of embedding spaces. Addintionally, the complexity of the proposed method increases linearly according to the number of spaces.

4. Extendability

The proposed architecture has extendability of new embedding networks. The steps of the extension are straightforward. We build visual and textual networks and then merge them to form a joint embedding space. In this paper, we add embedding networks using the Inflated 3D Convolutional Neural Network (I3D) model [43] so that the networks can capture video activities. We utilize the pre-trained RGB-I3D model to extract embedding vectors from 16 continuous frames of video. Consequently, a 1024-dimensional embedding vector $\mathrm{f_e}(x)$ is produced for each video.

The joint space is learned by using $\mathrm{f_e}(x)$ and the textual embedding vector $\mathrm{g_e}(y)$, and the similarity $\mathrm{s_e}(x, y)$ is measured in this joint embedding space. We transform $\phi(y)$ using $\hat{W}_e \in \mathbb{R}^{D \times 1024}$ and $\hat{b}_e \in \mathbb{R}^{1024}$. The similarity aggregation is also straightforward to extend.

Specifically, we concatenate all similarities and merge them with a weight $W'_m \in \mathbb{R}^M$, where M represents a number of embedding spaces. $M = 3$ in this case.

$$\mathrm{g_e}(y) = \hat{W}_e \phi(y) + \hat{b}_e \tag{16}$$

$$\mathrm{s_e}(x, y) = \frac{\mathrm{f_e}(x) \cdot \mathrm{g_e}(y)}{\| \mathrm{f_e}(x) \| \| \mathrm{g_e}(y) \|} \tag{17}$$

We stress that the extendability is vital since this enables us to incorporate other feature extraction models into the framework quickly. There are abundant approaches to extract features from videos [44–53]. Various aspects and approaches are necessary for the understanding of videos and sentences.

5. Experiments

We carried out the sentence-to-video retrieval task on the benchmark dataset to evaluate the proposed method. The task is retrieving the video associated with the query sentence from the test videos. We calculated similarities over the test videos with the query sentence, and then we picked up videos according to the similarities in descending order.

We reported the experimental results using rank-based performance metrics, i.e., Recall@k and Median rank. The Recall@k calculates the percentage of the correct video in top-k retrieved results. In this paper, we set $k = 1, 5, 10$. Median rank calculates the median of the ground-truth results in the ranking. For Recall@k, the bigger value indicates better performance. When Median Rank is a lower value, retrieved results are closer to the ground-truth items. Therefore, a lower median rank means better retrieval performance.

Following the convention in sentence-to-video retrieval, we used the Microsoft Research Video to Text dataset (MSR-VTT) [54], which is a large-scale video benchmark for video understanding. The MSR-VTT contains 10,000 video clips from YouTube with 41.2 h in total. The dataset provides videos in 20 categories, e.g., music, people, and sport. Each video is associated with 20 different description sentences. The dataset consists of 6513 videos for training, 497 videos for validation, and 2990 videos for testing.

We evaluated four variants of the proposed method: single-space, sequential or I3D dual-space, and triple-space models. Firstly, the single-space model represents the proposed framework composed of the global visual and textual embedding networks. These two networks form a single embedding space. Then, we measured the final similarity in the single embedding space. Secondly, the sequential dual-space model (dual-S) is the proposed framework using two embedding spaces learned by the global and sequential visual networks, and textual embedding networks. We measured the final similarity by merging two similarities, $\mathrm{s_g}$ and $\mathrm{s_s}$, as described in Equation (11). Thirdly, the I3D dual-space model (dual-I) has global visual and I3D embedding networks. Lastly, the triple-space model added the I3D and textual embedding networks into the dual-space model. We produced the final similarity by merging $\mathrm{s_g}$, $\mathrm{s_s}$, and $\mathrm{s_e}$ with the similarity aggregation.

5.1. Sentence-to-Video Retrieval Results

We summarize the results of the sentence-to-video retrieval task on the MSR-VTT dataset in Table 1. We compared the proposed method to the existing methods [28,32,38,55]. The proposed method obtained 7.1 (dual) at R@1, 21.2 (triple) at R@5, 32.4 at R@10 (triple), and 29 at MR (triple). These are competitive with [32,55], which are the state of the art.

Table 1. The results of sentence-to-video retrieval task on the Microsoft Research Video to Text dataset (MSR-VTT) dataset. The bold and underlined results represent the first- and the second-best, respectively.

Method	R@1	R@5	R@10	MR
VSE [38]	5.0	16.4	24.6	47
VSE++ [28]	5.7	17.1	24.8	65
Multi-Cues [32]	7.0	20.9	29.7	38
Cat-Feats [55]	**7.7**	**22**	<u>31.8</u>	32
Ours (single)	5.6	18.4	28.3	41
Ours (dual-S)	<u>7.1</u>	19.8	31	<u>30</u>
Ours (triple)	6.7	<u>21.2</u>	**32.4**	**29**

VSE and VSE++ adopt a triplet ranking loss, and VSE++ incorporates hard-negative samples into the loss to facilitate practical training [56]. We adopted this strategy. The results show that VSE++ performs significantly better than VSE at every R@*k*. Although the single-space model and VSE++ adopted similar loss functions, we found slight improvements in performance. However, the dual-space model achieves much better performance compared to VSE++. The results demonstrate the importance of using the sequential visual information of videos for learning an embedding space.

Multi-Cues [32] calculates two similarities in separated embedding spaces and then averages them to produce a final similarity. The proposed method has higher performance compared to Multi-Cues. The similarity aggregation strategy is the main difference between the proposed method and Multi-Cues. Note that the average suffers from aligning videos and sentences due to their variations. Thus, some videos need global visual features, and some need sequential visual features. The proposed method has a flexible strategy for merging similarities. The experimental results show that the proposed strategy is more useful for measuring similarity than a naive merging approach with equal importance to each embedding space, such as average.

The Cat-Feats [55] embeds videos and sentences by concatenating feature vectors extracted by three embedding modules: CNN, bi-directional GRU, and max pooling. Cat-Feats is slightly better than the proposed method at R@1 and R@5, e.g., 7.7 and 7.1 for Cat-Feats and dual-space at R@1, respectively. However, the proposed method (triple-space) outperforms Cat-Feat at R@10 and median rank, such as 32 and 29 by Cat-Feat and the triple, respectively. These results imply that the proposed method and Cat-Feats can assist each other. There is a possibility to improve performance by incorporating the feature concatenation mechanism used in Cat-Feat into the proposed method.

Finally, the proposed method with triple-space achieves better results than single- and dual-space at three metrics: R@5, 10, and median rank. Therefore, The results show that integrating multiple similarities can lead to a better, reliable retrieval performance.

6. Ablation Study

We carried out ablation studies to evaluate the components of the proposed method. We conducted the following three experiments.

6.1. Embedding Spaces

We changed the number of embedding spaces and developed the four variants of the proposed architecture. The experimental results are shown in Table 2. There are certain improvements from single to multiple spaces at all the metrics. Therefore, we can verify the effectiveness of the proposed method that integrates multiple spaces. Subsequently, we compared the two duals: dual-S is better than the dual-I at R@1 and R@10, whereas dual-I is superior at R@5. Thus, dual-S and dual-I can complement each other. The triple contains the embedding spaces of both of the duals, and it outperforms the single and the duals. Therefore, we can confirm the effectiveness of the combination of

multiple embedding spaces, which is the key insight of the proposed method for video and sentence understanding.

Table 2. Evaluation on combinations of embedding spaces.

	Global	Sequential	I3D	R@1	R@5	R@10	MR
single	√			5.6	18.4	28.3	41
dual-S	√	√		7.1	19.8	31	30
dual-I	√		√	6.8	20.2	30.6	30
triple	√	√	√	6.7	21.2	32.4	29

6.2. Spatial Attention Mechanism

We conducted experiments using dual-S with or without the spatial attention mechanism. The dual-S without the attention encodes each frame using ResNet into a 2048-dimensional vector. Then, the LSTM processed the sequences of the vectors. Table 3 shows the results. The dual-S with the attention achieves better results than without attention at all metrics, R@1, R@5, R@10, and median rank. Therefore, we can confirm that the proposed spatial attention mechanism improves performance significantly.

We also observed that the performance of the dual-S with attention is almost competitive with dual-I. However, dual-S without the attention is worse than the dual-I. Considering that the I3D model extracts useful spatial and temporal features for action recognition, the dual-S without attention could not extract sequential features effectively. However, the dual-S with attention obtained such features. We stress that this is another supportive evidence showing the effectiveness of the proposed attention mechanism.

Table 3. Effectiveness of the spatial attention.

	R@1	R@5	R@10	MR
w/o attention	5.8	19.8	28.6	34
w/ attention	7.1	19.8	31	30

6.3. Similarity Aggregation

We investigated the impacts of similarity aggregation using the average or the proposed weighted sum. We used dual-S and dual-I for this investigating experiment. The experimental results are shown in Table 4. There are improvements from the weighted sum at R@1, R@10, and MR in both of the dual-S and dual-I. Therefore, we confirmed the effectiveness of the proposed similarity aggregation module.

Table 4. Performance comparison on similarity aggregation using the average or the proposed weighted sum.

		R@1	R@5	R@10	MR
dual-S	average	6.6	19.9	29.9	31
	weighted	7.1	19.9	31	30
dual-I	average	6.7	20.2	30.4	30
	weighted	6.8	20.2	30.6	30

Furthermore, we performed an analysis of the weights in the similarity aggregation. As described in Equation (12), the weights are flexibly determined according to the given sentence. In other words, the weight represents the importance of embedding spaces. We attempted to go further by analyzing the weights. For simplicity, we used the dual-S model in this analysis. Therefore, the analysis is on the importance of global and sequential visual features. We summarize the statistics of the weights in Table 5. The statistics show that the proposed method assigns larger weights to the global features.

Table 5. Statistics of the weights in similarity aggregation for the global and sequential embedding spaces.

	Average	Min	Max
Global	0.52	0.399	0.61
Sequential	0.48	0.393	0.60

We show the accumulative step histogram for the global weight in Figure 4. The ratio reached 0.75 at the weight 0.5. Thus, 0.75 total instances received larger weights for the global feature. In contrast, the weights of the sequential feature are larger only in the remaining 0.25 instances. Therefore, the global feature is more critical than the sequential feature. Thus, dual-S aggressively used global features.

Figure 5 shows examples of videos and sentences with assigned weights to the global feature. Videos containing clear scenes tend to have larger weights on the global visual feature since objects in the videos are relatively easy for detection. On the other hand, videos containing unclear scenes tend to assign larger weights to the sequential visual feature.

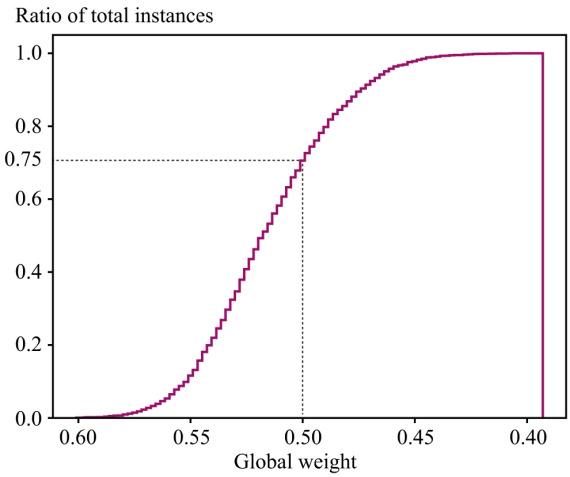

Figure 4. Accumulative step histogram of the weights of the global visual features used in similarity aggregation.

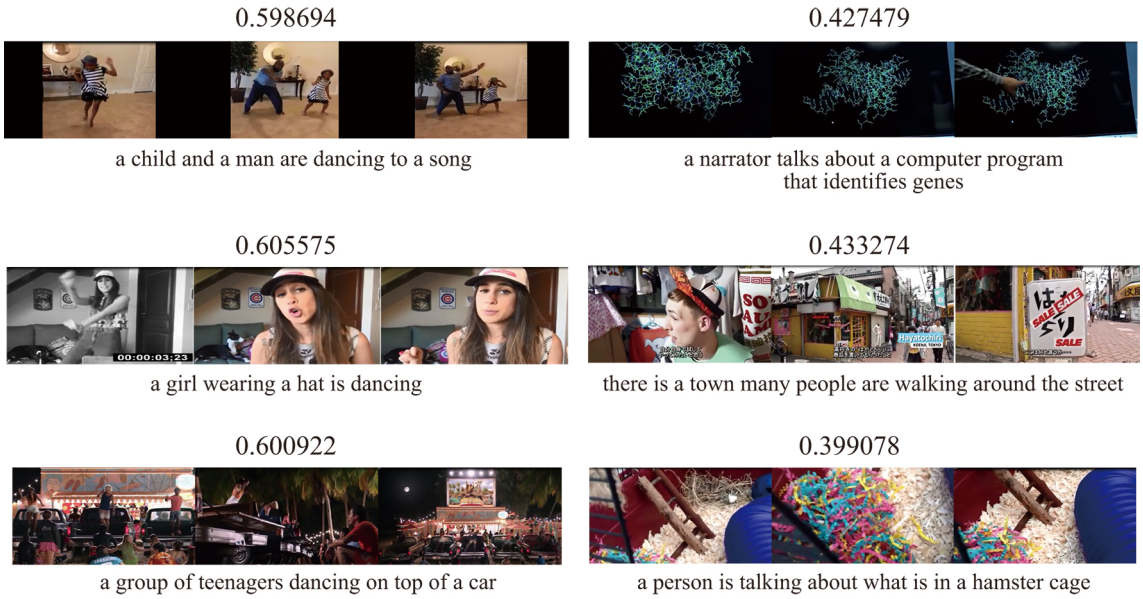

Figure 5. Examples of video and sentence with assigned weight to the global feature. The numbers and the sentences represent the weight for the global feature and query sentences, respectively.

7. Conclusions

In this paper, we presented a novel framework for embedding videos and sentences into multiple embedding spaces. The proposed method uses distinct embedding networks to capture various relationships between visual and textual features, such as global appearance, sequential visual, and action features. We produce the final similarity between a video and a sentence by merging similarities measured in the embedding spaces with the weighted sum. The proposed method can flexibly determine the weights according to a given sentence. Hence, the final similarity can incorporate an essential relationship between video and sentence. We carried out sentence-to-video retrieval experiments on the MSR-VTT dataset to demonstrate that the proposed framework significantly improved the performance when the number of embedding spaces increased. The proposed method achieved competitive results were comparable to the state-of-the-art methods [32,55]. Furthermore, we verified all the critical components in the proposed method through the ablation experiments. Even though the components are individually useful, their cooperation can generate significant improvements. Finally, we stress that there is room for improving the performance of the proposed method. For further improvement, we can extend the number of embedding spaces. As we described in Section 4, the proposed method is readily extendable. Therefore, we can embed sentences and various inputs in individual spaces. Specifically, promising inputs are audio, objects in images, and their relationships.

Author Contributions: Conceptualization, H.M.N. and T.M.; methodology, H.M.N.; software, H.M.N.; validation, H.M.N. and T.M.; formal analysis, H.M.N.; investigation, H.M.N. and T.M.; resources, H.M.N.; data curation, H.M.N.; writing—original draft preparation, H.M.N. and T.M.; writing—review and editing, Y.S. and S.O.; visualization, H.M.N.; supervision, Y.S. and S.O.; project administration, S.O.; funding acquisition, S.O. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by JSPS KAKENHI Grant Number 20H04201, 19K11848, and Yotta Informatics Project by MEXT, Japan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available at http://ms-multimedia-challenge.com/2016/dataset (accessed on 3 March 2021).

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

References

1. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.A.; Mikolov, T. DeViSE: A Deep Visual-Semantic Embedding Model. In Proceedings of the Advances in Neural Information Processing Systems, Stateline, NV, USA, 5–10 December 2013; Volume 26.
2. Engilberge, M.; Chevallier, L.; Perez, P.; Cord, M. Finding Beans in Burgers: Deep Semantic-Visual Embedding with Localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3984–3993.
3. Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; Yokoya, N. Learning Joint Representations of Videos and Sentences with Web Image Search. In *European Conference on Computer Vision (ECCV) Workshops*; Springer International Publishing: Amsterdam, The Netherlands, 2016; pp. 651–667.
4. Hendricks, L.A.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; Russell, B. Localizing Moments in Video with Natural Language. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5804–5813.
5. Gao, J.; Sun, C.; Yang, Z.; Nevatia, R. TALL: Temporal Activity Localization via Language Query. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5277–5285.
6. Xu, H.; He, K.; Plummer, B.A.; Sigal, L.; Sclaroff, S.; Saenko, K. *Multilevel Language and Vision Integration for Text-to-Clip Retrieval*; AAAI: Honolulu, HI, USA, 2019; pp. 9062–9069.
7. Verma, Y.; Jawahar, C.V. Image Annotation Using Metric Learning in Semantic Neighbourhoods. In *Computer Vision—ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 836–849.
8. Xu, X.; He, L.; Lu, H.; Shimada, A.; Taniguchi, R. Non-Linear Matrix Completion for Social Image Tagging. *IEEE Access* **2017**, *5*, 6688–6696. [CrossRef]

9. Li, X.; Shen, B.; Liu, B.; Zhang, Y. A Locality Sensitive Low-Rank Model for Image Tag Completion. *IEEE Trans. Multimed.* **2016**, *18*, 474–483. [CrossRef]
10. Rahman, S.; Khan, S.; Barnes, N. Deep0Tag: Deep Multiple Instance Learning for Zero-Shot Image Tagging. *IEEE Trans. Multimed.* **2020**, *22*, 242–255. [CrossRef]
11. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676. [CrossRef] [PubMed]
12. Gong, Y.; Ke, Q.; Isard, M.; Lazebnik, S. A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics. *Int. J. Comput. Vis.* **2014**, *106*, 210–233. [CrossRef]
13. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Sydney, Australia, 6–11 August 2015; Volume 37, pp. 2048–2057.
14. Chen, S.; Jin, Q.; Wang, P.; Wu, Q. Say As You Wish: Fine-Grained Control of Image Caption Generation With Abstract Scene Graphs. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 9959–9968. [CrossRef]
15. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15, Santiago, Chile, 11–18 December 2015; pp. 2425–2433. [CrossRef]
16. Zhang, P.; Goyal, Y.; Summers-Stay, D.; Batra, D.; Parikh, D. Yin and Yang: Balancing and Answering Binary Visual Questions. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5014–5022. [CrossRef]
17. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6325–6334. [CrossRef]
18. Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; Kim, G. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1359–1367.
19. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* **2019**, *51*. [CrossRef]
20. Wang, K.; Yin, Q.; Wang, W.; Wu, S.; Wang, L. A Comprehensive Survey on Cross-modal Retrieval. *arXiv* **2016**, arXiv:1607.06215.
21. Ji, Z.; Wang, H.; Han, J.; Pang, Y. Saliency-Guided Attention Network for Image-Sentence Matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5753–5762.
22. Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; Ma, W. Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 6602–6611.
23. Gu, J.; Cai, J.; Joty, S.; Niu, L.; Wang, G. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7181–7189.
24. Huang, Y.; Wu, Q.; Wang, W.; Wang, L. Image and Sentence Matching via Semantic Concepts and Order Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 636–650. [CrossRef] [PubMed]
25. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv* **2014**, arXiv:1411.2539.
26. Wang, L.; Li, Y.; Lazebnik, S. Learning Deep Structure-Preserving Image-Text Embeddings. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5005–5013.
27. Ye, K.; Kovashka, A. ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
28. Faghri, F.; Fleet, D.J.; Kiros, J.R.; Fidler, S. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; p. 12.
29. Zhang, D.; Dai, X.; Wang, X.; Wang, Y.; Davis, L.S. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1247–1257.
30. Tsai, Y.H.; Divvala, S.; Morency, L.; Salakhutdinov, R.; Farhadi, A. Video Relationship Reasoning Using Gated Spatio-Temporal Energy Graph. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10416–10425.
31. Song, Y.; Soleymani, M. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1979–1988.
32. Mithun, N.C.; Li, J.; Metze, F.; Roy-Chowdhury, A.K. Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval. In Proceedings of the ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, 11–14 June 2018; pp. 19–27.

33. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
34. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In Proceedings of the NIPS 2014 Workshop on Deep Learning, Montreal, QC, Canada, 8–13 December 2014.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
36. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
37. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive Mixtures of Local Experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
38. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal Neural Language Models. In Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research, PMLR, Beijing, China, 21–26 June 2014; Volume 32, pp. 595–603.
39. Chechik, G.; Sharma, V.; Shalit, U.; Bengio, S. Large Scale Online Learning of Image Similarity Through Ranking. *J. Mach. Learn. Res.* **2010**, *11*, 1109–1135.
40. Frome, A.; Singer, Y.; Sha, F.; Malik, J. Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
41. Socher, R.; Karpathy, A.; Le, Q.V.; Manning, C.D.; Ng, A.Y. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 207–218. [CrossRef]
42. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 19–27.
43. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
44. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5534–5542.
45. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
46. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 20–36.
47. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 4489–4497.
48. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6201–6210.
49. Zhang, S.; Guo, S.; Huang, W.; Scott, M.R.; Wang, L. V4D: 4D Convolutional Neural Networks for Video-level Representation Learning. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26 April–1 May 2020.
50. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
51. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
52. Dalal, N.; Triggs, B.; Schmid, C. Human Detection Using Oriented Histograms of Flow and Appearance. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 428–441.
53. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
54. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5288–5296.
55. Dong, J.; Li, X.; Xu, C.; Ji, S.; He, Y.; Yang, G.; Wang, X. Dual Encoding for Zero-Example Video Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 9338–9347.
56. Shrivastava, A.; Gupta, A.; Girshick, R. Training Region-Based Object Detectors with Online Hard Example Mining. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 761–769.