Article

Associating Images with Sentences Using Recurrent Canonical Correlation Analysis

Yawen Guo ^{1,*}, Hui Yuan ^{2,3} and Kun Zhang ³¹ Institute of Biophysics, Chinese Academy of Sciences, Beijing 100190, China² The College of Automation and Electronic Information, Xiangtan University, Xiangtan 411105, China; hui.yuan@cripac.ia.ac.cn³ Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; kun.zhang@cripac.ia.ac.cn

* Correspondence: yavinguo@outlook.com

Received: 5 July 2020; Accepted: 5 August 2020; Published: 10 August 2020

Abstract: Associating images with sentences has drawn much attention recently. Existing methods commonly represent an image by indistinctively describing all its contents in an one-time static way, which ignores two facts that (1) the association analysis can only be effective for partial salient contents and the associated sentence, and (2) visual information acquisition is a dynamical rather than static process. To deal with this issue, we propose a recurrent canonical correlation analysis (RCCA) method for associating images with sentences. RCCA includes a contextual attention-based LSTM-RNN which can selectively attend to salient regions of an image at each time step, and then represent all the salient contents within a few steps. Different from existing attention-based models, our model focuses on the modelling of contextual visual attention mechanism for the task of association analysis. RCCA also includes a conventional LSTM-RNN for sentence representation learning. The resulting representations of images and sentences are fed into CCA to maximize linear correlation, where parameters of LSTM-RNNs and CCA are jointly learned. Due to the effective image representation learning, our model can well associate images with sentences with complex contents, and achieve better performance in terms of image annotation and retrieval.

Keywords: image retrieval; image and sentence matching; canonical correlation analysis

1. Introduction

Associating images with sentences plays an important role in many applications, e.g., finding sentences given an image query for image annotation and caption, and retrieving images with a sentence query for image search. Recently, various methods have been proposed to deal with this problem. Most of them first extract representations for images and sentences, and then use either a structured objective [1,2] or a canonical correlation objective [3,4] to learn a latent joint space for association analysis, where matched images and sentences have small distance or high correlation.

Obtaining effective representations for images is crucial for the association analysis. An image usually contains very abundant contents, but its associated sentence can only describe partial salient contents due the limited size. Thus the following association analysis can only be effective for these salient contents and their corresponding textual descriptions since only they share the same semantics. As a result, the rest undescribed contents in the image are redundant, and tend to bring in noisy side effects in the association analysis. However, most existing methods [1,3–8] ignore this fact and represent an image by simultaneously processing its entire contents. Although other methods [2,9,10] consider to associate the salient contents (regions) with the corresponding descriptions (words) in the image and sentence, respectively, they have to pre-establish the salient regions using manual annotations or detection algorithms. In addition, the existing methods usually represent the visual

information in an one-time static process, which is in contrast to the neuroscience fact [11] that visual information acquisition is actually a dynamic rather than static process. Selective attention mechanism plays an essential role in this process. In particular, human visual system does not focus its attention on an entire image for only once. Instead, it selectively attends to salient regions at different locations to collect all important information in a sequential way.

Recently, some attention-based models have been proposed in various vision applications [12–15], but the modelling of attention mechanism in the context of association analysis is still under investigated. In fact, existing models mostly focus on the problems of recognizing salient image regions using dense supervision information, which makes them less generally applicable to the association analysis that aims to fuse the information of all the salient regions to obtain a unified representation. Moreover, they rarely consider the modelling of context [16] which actually has the prominent effect on attention mechanism. Ba et al. [17] consider to exploit the context information only at the first time step during the attention process, but they ignore the fact that context information modulates the attention mechanism during all the time steps [18].

In this work, we propose a recurrent canonical correlation analysis (RCCA) method for associating images with sentences using the visual attention. The proposed RCCA includes a contextual attention-based LSTM-RNN for image representation learning, which can selectively focus on salient regions of an image by referring to context information at each time step, and sequentially aggregate local features corresponding to the regions at all time steps to obtain the final image representation. RCCA also exploits a conventional LSTM-RNN to learn representation for the associated sentence by orderly taking its contained words as sequential inputs. Then the obtained representations of images and sentences are fed into CCA for association analysis, where parameters of LSTM-RNNs and CCA are jointly learned by maximizing linear correlation between images and sentences. To demonstrate the effectiveness of the proposed RCCA, we perform several experiments of image annotation and retrieval on three publicly available datasets, and achieve satisfactory results.

Our main contributions can be summarized as follows. We propose the recurrent canonical correlation analysis for associating images with sentences, which models the contextual visual attention mechanism and learns image representation in a biologically-plausible dynamical process. We systematically study the modelling of context in conjunction with attention mechanism, and demonstrate that the use of dense context information does help much. We achieve better performance in the tasks of image annotation and retrieval.

2. Related Work

2.1. Association between Images and Sentences

Existing methods for associating images with sentences usually include two steps: (1) feature extraction for images and sentences, and (2) association analysis based on the obtained features. Many recent methods [2–4,6–10,19] exploit CNNs [20,21] to obtain image features, i.e., the 4096-dimensional output vector in the last fully-connected layer. While for sentences, since they are sequential data, most methods use LSTM-RNNs [22] to model their long-range temporal dependency, and exploit the hidden states at the last step as their representations. After obtaining features, they use either a structured objective [1,2,7] or a canonical correlation objective [3,4,9] to learn a latent joint space, in which matched image-sentence pairs have small distance or high correlation.

It is believed that obtaining good image representations, such as by CNNs and fisher vector (FV) [4,23], can greatly facilitate the following association analysis. However, these methods indistinctively describe all the contents of an image, which might be suboptimal since only salient contents and their associated textual descriptions can be effective in association analysis. Although other works [2,9,10] consider to associate these salient regions with their descriptions, they have to explicitly use object detectors or manual annotations to pre-establish the salient regions.

Different from them, our model can automatically select and attend to salient regions in the image by modelling the contextual attention mechanism, which is more efficient and biologically-plausible.

The most related work to ours is [24], but it has to perform an one-to-one similarity measurement during testing, which makes it much more time-consuming than ours. Note that we do not make comparisons with recent state-of-the-art technologies [25–27], since they all use advance image representation modules like ResNet [28] and bottom-up features [29].

2.2. Attention-Based Models

There have been several works simulating the attention mechanism. Alex Graves [30] exploits RNNs and differentiable Gaussian filters to simulate the attention mechanism, and applies it to handwriting synthesis. The model is able to predict a word conditioned on an auxiliary annotation sequence. Gregor et al. [12] introduce the deep recurrent attentive writer for image generation, which develops a novel spatial attention mechanism based on 2-dimensional Gaussian filters to mimic the foveation of human eyes. Larochelle and Hinton [31] propose high-order Boltzmann machines to learn how to accumulate information from an image over several fixations. Mnih et al. [32] present a RNN-based model to select a sequence of regions and then only process the selected regions from an image to extract information. Ba et al. [17] present a recurrent attention model for multiple objects recognition. By training with reinforcement learning, the model can attend to some most label-relevant regions of an image. Bahdanau et al. [13] propose a neural machine translator which can search for relevant parts of a source sentence to predict a target word, without explicit alignment. Inspired by it, Xu et al. [15] develop an attention-based caption model which can automatically learn to fix gazes on salient objects in an image and generate the corresponding annotated words. In addition, this work is also related to recent works [33–35] focusing on semantic segmentation based on attention-based networks. To the best of our knowledge, we make the first attempt to consider the attention mechanism in the context of association between images and sentences, and systematically study the modelling of contextual attention mechanism.

3. Association between Images and Sentences via Recurrent Canonical Correlation Analysis

In this section, we will introduce the proposed recurrent canonical correlation analysis (RCCA) for associating images with sentences, by describing its included three modules in details: (1) a contextual attention-based LSTM-RNN for image representation learning, (2) a conventional LSTM-RNN for sentence representation learning, and (3) canonical correlation analysis based on the resulting image and sentence representations for model learning.

3.1. Image Representation Using Contextual Attention-Based LSTM-RNN

The contextual attention-based LSTM-RNN is illustrated in Figure 1 (at the bottom). Given an image, we first use a CNN [36] to extract two kinds of features: (1) multiple local features $\{\mathbf{a}_l | \mathbf{a}_l \in \mathbb{R}^F\}_{l=1,\dots,L}$, i.e., feature maps from the last convolutional layer, where $\mathbf{a}_l$ describes l -th local region in the image and L is the total number of local regions, and (2) a context feature $\mathbf{m} \in \mathbb{R}^D$, i.e., a feature vector from the last fully-connected layer, describing the global context information of the image. Based on these features, we exploit a contextual attention scheme to obtain sequential representations $\{\mathbf{e}_t\}_{t=1,\dots,T}$ for the static image, where $\mathbf{e}_t$ is a combined local feature describing the attended salient regions at the t -th time step. Then we use LSTM-RNN to model the long-rang temporal dependency in the sequential representations, by separately taking them as inputs to hidden states at different time steps $\{\mathbf{h}_t \in \mathbb{R}^{D_x}\}_{t=1,\dots,T}$. As time goes on, all the hidden states propagate their values along the time axis until the end, and thus the hidden state at the last time step $\mathbf{h}_T$ indicates the fused representation of all the attended salient regions.

Figure 1. The proposed recurrent canonical correlation analysis (RCCA).

The details of the contextual attention scheme at the t -th time step can be formulated as follows. As illustrated in Figure 2a, given context feature $\mathbf{m}$, local feature set $\{\mathbf{a}_l | \mathbf{a}_l \in \mathbb{R}^F\}_{l=1, \dots, L}$ and hidden state at previous time step $\mathbf{h}_{t-1}$, we compute $\{p_{t,l} | p_{t,l} \in [0, 1]\}_{l=1, \dots, L}$ for all the local regions, where $p_{t,l}$ indicates the probability that the l -th region will be attended to at the t -th time step:

$$p_{t,l} = \frac{e^{\hat{p}_{t,l}}}{\sum_{l=1}^L e^{\hat{p}_{t,l}}}, \quad \hat{p}_{t,l} = f_{\text{gated}}([\mathbf{h}_{t-1}, \mathbf{m}, \mathbf{a}_l]) \quad (1)$$

where $f_{\text{gated}}(\cdot)$ is a gated fusion unit [37] to fuse all the context information in terms of $\hat{\mathbf{h}}_{t-1}$, $\mathbf{m}$ and $\mathbf{a}_l$ as input, and $\hat{\mathbf{h}}_{t-1}$ is obtained by modulating previous hidden state $\mathbf{h}_{t-1}$ with context information $\mathbf{m}$. Note that the usage of context information is necessary, since it can not only provide the complementary global information about the current attended regions, but also encourage the model to focus on other unexplored regions under the principle of information maximization.

After obtaining the probabilities, we exploit the “soft attention” scheme [13] to compute the attended local feature $\mathbf{e}_t$. We perform element-wise multiplication between each local feature $\mathbf{a}_l$ and its corresponding probability $p_{t,l}$, and sum all the products together:

$$\mathbf{e}_t = \sum_{l=1}^L p_{t,l} \mathbf{a}_l \quad (2)$$

Then, we can use LSTM-RNN [38] to model the temporal dependency in the sequential representations $\{\mathbf{e}_t\}_{t=1, \dots, T}$ as shown in Figure 2b, where (initial) memory state ($\mathbf{c}_0$) $\mathbf{c}_t$, (initial) hidden state ($\mathbf{h}_0$) $\mathbf{h}_t$, input gate $\mathbf{i}_t$, forget gate $\mathbf{f}_t$ and output gate $\mathbf{o}_t$ can be computed as follows:

$$\mathbf{c}_0 = f_{MLP,c}(\frac{1}{L} \sum_{l=1}^L \mathbf{a}_l), \quad \mathbf{h}_0 = f_{MLP,h}(\frac{1}{L} \sum_{l=1}^L \mathbf{a}_l) \quad (3)$$

$$\mathbf{i}_t = \sigma(W_{ei}\mathbf{e}_t + W_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i), \quad \mathbf{f}_t = \sigma(W_{ef}\mathbf{e}_t + W_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (4)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(W_{\mathbf{ec}}\mathbf{e}_t + W_{\mathbf{hc}}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (5)$$

$$\mathbf{o}_t = \sigma(W_{\mathbf{eo}}\mathbf{e}_t + W_{\mathbf{ho}}\mathbf{h}_{t-1} + \mathbf{b}_o), \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (6)$$

where $f_{MLP,c}(\cdot)$ and $f_{MLP,h}(\cdot)$ are two multilayer perceptrons associated with $\mathbf{c}$ and $\mathbf{h}$, respectively, σ denotes the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

Figure 2. Illustrations of contextual attention scheme and LSTM-RNN at the t -th time step.

3.2. Sentence Representation Using Conventional LSTM-RNN

Considering that sentences are naturally sequential data consisting of ordered words, and conventional LSTM-RNN can well model long-range contextual information, so we directly exploit the conventional LSTM-RNN to obtain sentence representations. As shown in Figure 1 (at the top), conventional LSTM-RNN orderly takes a word as the input to the corresponding hidden state at each time step, and all the hidden states propagate their values along the time axis. The hidden state at the last time step collects all the information in the sentence, which can thus be regarded as the representation of the entire sentence.

We denote $[\mathbf{w}_1, \dots, \mathbf{w}_t, \dots, \mathbf{w}_{T'}]$ as a sentence associated with the image, $\mathbf{w}_t \in \{0, 1\}^M$ as the t -th word in the sentence, T' as the number of words in the sentence (also the number of time steps in the conventional LSTM-RNN), M as the size of word vocabulary, and $\{\mathbf{h}_t \in \mathbb{R}^{D_y}\}_{t=1, \dots, T'}$ as the hidden states at all time steps. Since the original representations of words are high-dimensional “one-hot” vectors, we embed them into low-dimensional real-valued vectors using an embedding matrix $E \in \mathbb{R}^{U \times M}$. The details of the conventional LSTM-RNN can be formulated as follows:

$$\mathbf{i}_t = \sigma(W_{\mathbf{wi}}E\mathbf{w}_t + W_{\mathbf{hi}}\mathbf{h}_{t-1} + \mathbf{b}_i), \mathbf{f}_t = \sigma(W_{\mathbf{wf}}E\mathbf{w}_t + W_{\mathbf{hf}}\mathbf{h}_{t-1} + \mathbf{b}_f) \quad (7)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(W_{\mathbf{wc}}E\mathbf{w}_t + W_{\mathbf{hc}}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (8)$$

$$\mathbf{o}_t = \sigma(W_{\mathbf{wo}}E\mathbf{w}_t + W_{\mathbf{ho}}\mathbf{h}_{t-1} + \mathbf{b}_o), \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (9)$$

Different from the contextual attention-based LSTM-RNN, here we do not include any explicit definition of the initial memory state and the hidden state, similar to [38].

3.3. Model Learning with Canonical Correlation Analysis

We regard the last hidden states of two modality-specific LSTM-RNNs as the representations of images and sentences, denoted by $X \in \mathbb{R}^{D_x \times N}$ and $Y \in \mathbb{R}^{D_y \times N}$, respectively, where N is the number of given image-sentence pairs, and D_x and D_y are the dimensions of obtained representations of images and sentences, respectively. We feed these representations to canonical correlation analysis (CCA) to learn two linear projections W_x and W_y for X and Y , respectively, so that their correlation is high:

$$(W_x^*, W_y^*) = \arg \max_{W_x, W_y} \text{corr}(X, Y) = \arg \max_{W_x, W_y} \frac{W_x^T \Sigma_{xy} W_y}{\sqrt{W_x^T \Sigma_{xx} W_x W_y^T \Sigma_{yy} W_y}} \quad (10)$$

The covariances $\sum_{xx}$, $\sum_{yy}$ and $\sum_{xy}$ are estimated as: $\sum_{xx} = \frac{1}{N-1} \bar{X} \bar{X}^T + \lambda_x I$, $\sum_{yy} = \frac{1}{N-1} \bar{Y} \bar{Y}^T + \lambda_y I$ and $\sum_{xy} = \frac{1}{N-1} \bar{X} \bar{Y}^T$, respectively, where $\bar{X}$ and $\bar{Y}$ are centered representations, and $\lambda_x I$ and $\lambda_y I$ are regularization terms to ensure the positive definiteness of $\sum_{xx}$ and $\sum_{yy}$.

Since the objective is invariant to the scaling of W_x and W_y , the projections can be constrained to have unit variance. Then by assembling the top k projection vectors into the columns of projection matrices, denoted by $\hat{W}_x$ and $\hat{W}_y$, the objective can be written as:

$$\max_{\hat{W}_x, \hat{W}_y} \text{tr}(\hat{W}_x^T \sum_{xy} \hat{W}_y), \text{ s.t., } \hat{W}_x^T \sum_{xx} \hat{W}_x = \hat{W}_y^T \sum_{yy} \hat{W}_y = 1 \quad (11)$$

Then according to [39], the optimum of this objective can be attained at:

$$(\hat{W}_x^*, \hat{W}_y^*) = (\sum_{xx}^{-1/2} U_k, \sum_{yy}^{-1/2} V_k) \quad (12)$$

where U_k and V_k are matrices of the first k left- and right- singular vectors of $T = \sum_{xx}^{-1/2} \sum_{xy} \sum_{yy}^{-1/2}$, respectively. As a result, the objective is equivalent to the sum of the top k singular values of T , i.e., the trace norm:

$$\text{corr}(X, Y) = \|T\|_{\text{tr}} = \text{tr}((T^T T)^{1/2}) \quad (13)$$

It should be noted that the parameters of two modality-specific LSTM-RNNs and CCA are jointly learned, where the parameters of LSTM-RNNs are adaptively trained to optimize the CCA objective using stochastic gradient descent. The gradients of $\text{corr}(X, Y)$ with respect to parameters of LSTM-RNNs can be computed by first obtaining the gradients with respect to X and Y as follows, and then performing backpropagation.

$$\begin{aligned} \frac{\partial \text{corr}(X, Y)}{\partial X} &= \frac{1}{N-1} (2\nabla_{xx} \bar{X} + \nabla_{xy} \bar{Y}), \quad \frac{\partial \text{corr}(X, Y)}{\partial Y} = \frac{1}{N-1} (2\nabla_{yy} \bar{Y} + \nabla_{yx} \bar{X}) \\ \nabla_{xx} &= -\frac{1}{2} \sum_{xx}^{-1/2} U D U^T \sum_{xx}^{-1/2}, \quad \nabla_{xy} = \sum_{xx}^{-1/2} U V^T \sum_{yy}^{-1/2} \\ \nabla_{yy} &= -\frac{1}{2} \sum_{yy}^{-1/2} V D V^T \sum_{yy}^{-1/2}, \quad \nabla_{yx} = \sum_{yy}^{-1/2} V U^T \sum_{xx}^{-1/2} \end{aligned}$$

where $U D V^T$ is the singular value decomposition (SVD) of T .

Similar to [40,41], although the objective of CCA is a function of the entire training set, we perform mini-batch based optimization due to the high computational cost. Such implementations often lead to the fact that, directly optimizing parameters of LSTM-RNNs from random initialization cannot obtain good results. To provide a better initialization for the parameters, we exploit a structured objective for parameter pretraining, which has been widely used in cross-modal learning [2,7,10,41]:

$$\begin{aligned} W_{LSTM-RNN}^* &= \arg \min_{W_{LSTM-RNN}} \sum_{i=1}^N \sum_{k=1}^K \max \{0, m - d(\mathbf{x}_i, \mathbf{y}_i) + d(\mathbf{x}_i, \mathbf{y}_k)\} \\ &\quad + \max \{0, m - d(\mathbf{x}_i, \mathbf{y}_i) + d(\mathbf{x}_k, \mathbf{y}_i)\} \end{aligned} \quad (14)$$

where $W_{LSTM-RNN}$ denotes all the parameters of two LSTM-RNNs, $(\mathbf{x}_i, \mathbf{y}_i)$ are the obtained representations of the i -th matched images and sentences, $\mathbf{x}_k$ is a mismatched (contrastive) image representation for $\mathbf{y}_i$, and vice-versa with $\mathbf{x}_k$, K is the total number of contrastive samples, m is a tuning parameter and $d(\cdot, \cdot)$ is a cosine similarity measurement. This structured objective encourages representations of matched images and sentences to be more similar than those of mismatched ones. Such initial representations are helpful for the following CCA optimization, since maximizing the correlation between two more similar representations is much easier.

In addition, we add a doubly stochastic regularization [15] to the objective, which forces the contextual attention-based LSTM-RNN to pay equal attention to every local region of the image during the dynamical attention process:

$$R = \lambda \sum_{l=1}^L \left(1 - \sum_{t=1}^T p_{t,l} \right) \quad (15)$$

where λ is a balancing parameter. In our experiments, we find that using this regularization can further improve the performance.

4. Experimental Results

To demonstrate the effectiveness of the proposed RCCA, we perform experiments of image annotation and retrieval on three publicly available datasets.

4.1. Datasets and Protocols

The three evaluation datasets and their corresponding experimental protocols are described as follows. (1) **Flickr8k** [42] consists of 8000 images collected from the Flickr website, each of which may contain people or animal performing actions, and is accompanied with five sentences describing the content of the image. This dataset provides the standard training, validation and testing splits, with 6000, 1000 and 1000 images, respectively. (2) **Flickr30k** [43] is an extension of Flickr8K, which consists of 31,783 images collected from the Flickr website. Each image is also accompanied with five sentences, and the sentences are annotated in a similar way as those in Flickr8K. We use the public training, validation and testing splits, which contains 28,000, 1000 and 1000 images, respectively. (3) **Microsoft COCO** [44] consists of 82,783 training and 40,504 validation images, each of which is associated with five sentences. We use the public training, validation and testing splits [45], with 82,783, 4000 and 1000 images, respectively.

4.2. Experimental Details

During model learning, we use the Stochastic Gradient Descent (SGD) to optimize the parameters, in which the learning rate is 0.005, batch size is 128 and gradient clipping is $[-10, 10]$. The model is trained for 30 epochs to guarantee its convergence, based on the Pytorch platform [46] and accelerated with a NVIDIA Titan X GPU.

The task of image annotation and retrieval can be jointly formulated as cross-modal retrieval, i.e., given an image, the goal is to retrieve highly matched sentences for annotation, and vice versa. The commonly used evaluation criteria are “R@1”, “R@5” and “R@10”, i.e., recall rate at the top 1, 5 and 10 results. Another one is “Med r” which is the median rank of the first ground truth result. We also compute an additional criterion “Sum” to evaluate the overall performance for both image annotation and retrieval.

4.3. Image and Sentence Matching

To systematically study the modelling of contextual attention mechanism, as shown in Figure 3, we propose four variants of RCCA: (a) RCCA-nc does not use any context information at all the time steps in the contextual attention-based LSTM-RNN, (b) RCCA-fc uses the context information only at the first time step, (c) RCCA-lc uses the context information only at the last time step, and (d) RCCA-na does not use the attention mechanism but only the context information. We denote RCCA-ac as the original model in Figure 1 which uses the context information at all the time steps. We also develop an ensemble model RCCA* in a similar way as [8], by summing five cross-modal similarity matrices generated from the five RCCAs together,

For all the RCCAs, we use the 19-layer VGG Network [21] to extract 512 feature maps (with a size of 14×14) in “conv5-4” layer as multiple local features, and a feature vector in “fc7” layer as the context feature. Therefore, the dimensions of local and context features are $F = 512$ and $D = 4096$, respectively,

and the total number of local regions is $L = 196$. Some hyperparameters of RCCAs are illustrated as follows: $D_x = 1024$, $D_y = 1024$, $\lambda_x = 0.1$, $\lambda_y = 0.1$, $\lambda = 10$, $K = 100$, $U = 300$ and $m = 0.2$. For the pretraining, we find that using 3 pretraining epoches can always achieve satisfactory performance. The number of time steps in the contextual attention-based LSTM-RNN is $T=3$ (more details in Section 4.4).

Figure 3. Different usages of context information in the four RCCA variants.

We compare RCCAs with several previous state-of-the-arts on the three datasets in Tables 1–3, respectively. All the compared methods can be classified into two classes: (1) global matching and (2) local matching. The global matching methods [1,3,4,6,7,45,47,48] directly learn global representations for image and sentence, and then associate them by computing the global similarity in an one-to-one manner. While the local matching methods [2,5,8–10] first detect local objects and words in the image and sentence, and then summarize their local similarities to finally obtain the global similarity.

Table 1. Comparison results of image annotation and retrieval on the Flickr8K dataset (* indicates ensemble or multi-model methods).

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r	
DeViSE [1]	4.8	16.5	27.3	28.0	5.9	20.1	29.6	29	104.2
SDT-RNN [5]	6.0	22.7	34.0	23.0	6.6	21.6	31.7	25	122.6
Deep Fragment [2]	12.6	32.9	44.0	14	9.7	29.6	42.5	15	171.3
RVP (T+I) [6]	11.7	34.8	48.6	11.2	11.4	32.0	46.2	11	184.7
m-RNN [45]	14.5	37.2	48.5	11	11.5	31.0	42.4	15	185.0
DCCA [3]	17.9	40.3	51.9	9	12.7	31.2	44.1	13	197.9
DVSA (BRNN) [10]	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4	199.9
MNLM [7]	18.0	40.9	55.0	8	12.5	37.0	51.5	10	214.9
NIC [47]	20.0	-	61.0	6	19.0	-	64.0	5	-
m-CNN-st [8]	18.1	44.1	57.9	7	14.6	38.5	53.5	9	226.7
m-CNN* [8]	24.8	53.7	67.1	5	20.3	47.6	61.7	5	275.2
FV(HGLMM) [4]	28.5	58.4	71.7	4	20.6	49.4	64.0	6	292.6
FV* [4]	31.0	59.3	73.7	4	21.3	50.0	64.8	5	300.1
Ours:									
RCCA-nc	19.5	44.8	58.2	7	14.7	39.9	53.0	9	230.1
RCCA-na	22.8	48.9	63.3	6	16.6	41.1	54.0	9	246.7
RCCA-fc	23.3	49.4	64.0	6	17.0	41.2	53.5	8	248.4
RCCA-lc	25.6	51.9	65.4	5	17.7	41.2	53.7	9	255.5
RCCA-ac	26.5	55.0	67.9	4	18.2	45.3	58.2	7	271.1
RCCA*-f30k	28.1	56.1	66.3	4	18.5	43.8	59.6	7	272.4
RCCA*-coco	29.0	57.7	68.2	4	19.7	45.3	61.1	6	281.0
RCCA*	30.3	59.6	69.7	3	20.6	47.9	62.1	6	290.2

From Table 1, we can see that the pretraining results can already perform better than most compared methods. Then after CCA optimization, the performance of all the RCCAs can be further improved. Note that our best RCCA* still performs a little worse than FV*. It is mainly attributed to the factor that our model has a large number of learning parameters which cannot be well fitted using only the limited number of training data on the Flickr8k dataset.

When exploiting large-scale training data on the Flickr30k (an extension of Flickr8k) and Microsoft COCO datasets in Tables 2 and 3, RCCA* can achieve much better performance than all the compared methods. Our best single model RCCA-ac outperforms both the best single model FV (HGLMM) on the Flickr30k dataset, and the best single model m-CNN-st on the Microsoft COCO dataset. These observations demonstrate that and dynamically learning representations for images is more suitable for association analysis.

Table 2. Comparison results of image annotation and retrieval on the Flickr30K dataset (* indicates ensemble or multi-model methods, and [†] indicates using manual annotations).

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med <i>r</i>	R@1	R@5	R@10	Med <i>r</i>	
DeViSE [1]	4.5	18.1	29.2	26	6.7	21.9	32.7	25	113.1
SDT-RNN [5]	9.6	29.8	41.1	16	8.9	29.8	41.1	16	160.3
RVP (T+I) [6]	12.1	27.8	47.8	11	12.7	33.1	44.9	12.5	178.4
Deep Fragment [2]	14.2	37.7	51.3	10	10.2	30.8	44.2	14	188.4
DCCA [3]	16.7	39.3	52.9	8	12.6	31.0	43.0	15	195.5
NIC [47]	17.0	-	56.0	7	17.0	-	57.0	7	-
DVSA (BRNN) [10]	22.2	48.2	61.4	4.8	15.2	37.7	50.5	9.2	235.2
MNLM [7]	23.0	50.7	62.9	5	16.8	42.0	56.5	8	251.9
LRCN [48]	-	-	-	-	17.5	40.3	50.8	9	-
m-RNN [45]	35.4	63.8	73.7	3	22.8	50.7	63.1	5	309.5
FV (HGLMM) [4]	34.4	61.0	72.3	3	24.4	52.1	65.6	5	309.8
FV* [4]	35.0	62.0	73.8	3	25.0	52.7	66.0	5	314.5
m-CNN-st [8]	27.0	56.4	70.1	4	19.7	48.4	62.3	6	283.9
m-CNN* [8]	33.6	64.1	74.9	3	26.2	56.3	69.6	4	324.7
RTP [†] * [9]	37.4	63.1	74.3	-	26.0	56.0	69.3	-	326.1
Ours:									
RCCA-nc	27.5	53.3	66.9	5	20.9	46.7	58.8	7	274.1
RCCA-na	34.4	61.6	71.2	3	23.9	51.1	61.8	5	304.0
RCCA-fc	32.2	60.4	72.3	3	23.7	53.1	65.0	5	306.7
RCCA-lc	32.2	61.2	72.4	3	23.9	53.4	65.8	5	308.9
RCCA-ac	36.0	65.8	75.6	3	25.8	53.9	65.7	5	322.8
RCCA*	39.3	68.7	78.2	2	28.7	57.2	69.8	4	341.9

When comparing among all the RCCAs, we can conclude as follows. (1) Using context information to modulate attention mechanism is very necessary, since RCCA-fc, RCCA-lc and RCCA-ac all perform much better than RCCA-nc. (2) Exploiting only context information without the attention mechanism, RCCA-na achieve worse results than RCCA-fc, RCCA-lc and RCCA-ac, which verifies the effectiveness of the attention mechanism. (3) RCCA-ac performs the best among all the RCCA variants, which indicates that the best way of modelling the context is using dense context information at all the time steps. (4) The ensemble of the five RCCA variants RCCA* can greatly improve the performance.

To evaluate the cross-dataset generalization of the proposed model, we use the train splits of Flickr30K and Microsoft COCO datasets to train two models: RCCA*-f30k and RCCA*-coco, respectively, and then perform test on the test split of Flickr8K dataset. We can see that either RCCA*-f30k or RCCA*-coco achieves worse performance than the original RCCA*. It mainly results from the fact that the image and sentence contents are quite different between the two datasets. RCCA*-coco performs much better than RCCA*-f30k, because the train split of Microsoft COCO has a larger size than that of Flickr30K.

Table 3. Comparison results of image annotation and retrieval on the Microsoft COCO dataset (* indicates ensemble or multi-model methods).

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med <i>r</i>	R@1	R@5	R@10	Med <i>r</i>	
STD (bi-skip) [49]	32.7	67.3	79.6	3	24.2	57.1	73.2	4	334.1
STD * [49]	33.8	67.7	82.1	3	25.9	60.0	74.6	4	344.1
m-RNN [45]	41.0	73.0	83.5	2	29.0	42.2	77.0	3	345.7
FV (HGLMM) [4]	37.7	66.6	79.1	3	24.9	58.8	76.5	4	343.6
FV * [4]	39.4	67.9	80.9	2	25.1	59.8	76.6	4	349.7
DVSA [10]	38.4	69.9	80.5	1	27.4	60.2	74.8	3	351.2
MNLM [7]	43.4	75.7	85.8	2	31.0	66.7	79.9	3	382.5
m-CNN-st [8]	38.3	69.6	81.0	2	27.4	63.4	79.5	3	359.2
m-CNN * [8]	42.8	73.1	84.1	2	32.6	68.6	82.8	3	384.0
Ours:									
RCCA-nc	37.4	70.3	81.5	2	29.7	65.3	79.8	3	364.0
RCCA-na	40.7	71.0	84.6	2	32.9	68.8	81.0	3	379.0
RCCA-fc	41.9	73.5	84.1	2	33.4	68.1	81.8	3	382.8
RCCA-lc	42.3	77.8	87.9	2	34.3	68.4	81.0	3	391.7
RCCA-ac	44.9	79.6	87.7	2	35.8	71.2	83.3	2	402.5
RCCA *	49.4	80.1	89.5	2	37.9	73.5	84.9	2	415.3

4.4. Analysis of the Number of Time Steps

For a sequential sentence, the number of time steps in conventional LSTM-RNN is naturally the length of the sentence. While for a static image, we have to manually set the number of time steps T . Ideally, T should roughly equal to the number of salient objects appeared in the image, so that the model can separately attend to these objects within T steps to collect all information. In the following, we gradually increase T from 1 to 9, and testify the impact of different numbers of time steps on the performance of RCCA-ac and RCCA-nc.

From Table 4 we can observe that both RCCA-ac and RCCA-nc achieve their best performance when the number of time steps is 3. It indicates that they can get all the important information in an image by iteratively visiting the image three times. Intuitively, most images usually contain at most three objects, which is in consistent with the optimal number of time steps. Note that when T becomes larger than 3, the performance of RCCA-ac and RCCA-nc slightly drops, which results from the fact that an overly complex network architecture could lead to overfitting. When comparing RCCA-ac with RCCA-nc, we find that the usage of context information can greatly improve the performance in all cases, regardless of the number of time steps. Especially for image annotation, with the aid of context information, RCCA-ac seems to exploit more time steps to attend to more contents of the image so that it can produce more accurate results.

Table 4. Different numbers of time steps on the Microsoft COCO dataset. T : number of time steps in contextual attention-based LSTM-RNN.

Method	Image Annotation				Image Retrieval				Sum
	R@1	R@5	R@10	Med <i>r</i>	R@1	R@5	R@10	Med <i>r</i>	
RCCA-ac:									
$T = 1$	42.5	74.9	86.0	2	32.1	67.7	81.4	3	384.6
$T = 3$	44.9	79.6	87.7	2	35.8	71.2	83.3	2	402.5
$T = 5$	44.0	78.0	86.8	2	35.4	71.0	83.1	2	398.3
$T = 7$	44.1	76.9	86.7	2	35.5	71.1	83.2	2	397.5
$T = 9$	43.2	76.5	87.1	2	34.7	71.0	83.1	2	395.6
RCCA-nc:									
$T = 1$	38.5	70.1	80.9	2	28.0	63.2	76.1	3	356.8
$T = 3$	37.4	70.3	81.5	2	29.7	65.3	79.8	3	364.0
$T = 5$	34.9	67.1	78.9	3	28.8	64.5	77.5	3	351.7
$T = 7$	36.2	65.6	77.9	3	28.9	64.0	77.2	3	350.9
$T = 9$	32.6	64.7	76.6	3	27.1	62.6	75.7	3	339.2

4.5. Visualization of Dynamical Attention Maps

To verify whether the proposed model can attend to salient regions of an image at different time steps, we visualize the predicted dynamical attention maps on the contextual attention-based LSTM-RNN. For the predicted probabilities at the t -th time step $\{p_{t,l}\}_{l=1,\dots,196}$, we first reshape it to a probabilistic map, with the size of 14×14 . Its layout is equivalent to that of feature maps extracted in “conv5-4” of the VGG Network. Then we resize the probabilistic map to the same size as its corresponding original image, so that each probability in the resized map measures the importance of an image pixel at the same location. We thus perform element-wise multiplication between the resized probabilistic map and its corresponding image to obtain an attention map, where lighter areas indicate attended regions.

Dynamical attention maps at three successive time steps by RCCA-ac and RCCA-nc are shown in Table 5. We can see that RCCA-ac attends to different objects in the images at different time steps. Note that without the aid of context information, RCCA-nc cannot produce accurate dynamical attention maps as those of RCCA-ac. Particularly, it cannot well attend to the “cow” and “clock” in the two images, respectively. In fact, RCCA-nc always finishes attending to salient objects within the first two steps, and does not focus on meaningful regions at the third step any more. Different from it, RCCA-ac uses more steps to focus on more salient objects, which is also inconsistent with the observation in Section 4.4. These evidences again demonstrate that the modelling of context can greatly facilitate the attention mechanism.

Table 5. Visualization of dynamical attention maps on the Microsoft COCO dataset.

Input Image	RCCA-ac			RCCA-nc		
	1st Step	2nd Step	3rd Step	1st Step	2nd Step	3rd Step

4.6. Error Analysis

By analyzing the results by our proposed model, we find it cannot well generalize to those images containing very complex content. To illustrate this, we present the results of image retrieval by sentence queries in Figure 4, where the numbers in the top left corner are returned ranks (the smaller, the better) of groundtruth images. We can see that the ranks of left three examples are very high and the corresponding image contents are relatively simple. While the ranks of right three examples are very high, which indicates that our model cannot find matched images. According to a recent state-of-the-art [50], it mainly because the complex image contents include many few-shot objects and attributes, which cannot be well associated with few-shot words in the sentences.

Figure 4. Example results of sentence-based image retrieval. The top left numbers are returned ranks of groundtruth matched images.

5. Conclusions and Future Work

In this paper, we have proposed the recurrent canonical correlation analysis (RCCA) method for associating images with sentences. Our main contribution is the modelling of contextual attention mechanism for dynamically learning image representations. We have applied our model to image annotation and retrieval, and achieved better results.

In the future, we will exploit the similar attention-based scheme to learn good representations for sentences, and further verify its effectiveness on more datasets. In addition, we will try to combine the two tasks of image and sentence matching and image phrase localization into a joint framework to exploit the complementary advantages of global and local matching. To pursue higher performance, we will also consider to exploit the more detailed annotations in Flickr30k Entity [9] between objects and words for fine-grained image and sentence matching.

Author Contributions: Data curation, H.Y.; Investigation, Y.G., H.Y. and K.Z.; Methodology, Y.G.; Project administration, Y.G.; Validation, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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 *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Lake Tahoe, NV, USA, 2013.
2. Karpathy, A.; Joulin, A.; Li, F.F. Deep fragment embeddings for bidirectional image sentence mapping. In *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Montreal, QC, Canada, 2014.
3. Mikolajczyk, F.Y.K. Deep Correlation for Matching Images and Text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015.
4. Klein, B.; Lev, G.; Sadeh, G.; Wolf, L. Associating Neural Word Embeddings with Deep Image Representations using Fisher Vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 7–12 June 2015.
5. Socher, R.; Karpathy, A.; Le, Q.V.; Manning, C.D.; Ng, A.Y. Grounded compositional semantics for finding and describing images with sentences. In *Transactions of the Association for Computational Linguistics (TACL)*; MIT Press: Cambridge, MA, USA, 2014.
6. Chen, X.; Zitnick, C.L. Learning a recurrent visual representation for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Columbus, OH, USA, 24–27 June 2014.
7. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
8. Ma, L.; Lu, Z.; Shang, L.; Li, H. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015.
9. Plummer, B.; Wang, L.; Cervantes, C.; Caicedo, J.; Hockenmaier, J.; Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015.
10. Karpathy, A.; Li, F.F. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Santiago, Chile, 7–13 December 2015.
11. Rensink, R.A. The dynamic representation of scenes. *Vis. Cogn.* **2000**, *7*, 17–42. [[CrossRef](#)]
12. Gregor, K.; Danihelka, I.; Graves, A.; Wierstra, D. DRAW: A recurrent neural network for image generation. *arXiv* **2015**, arXiv:1502.04623.
13. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.

14. Sharma, S.; Kiros, R.; Salakhutdinov, R. Action Recognition using Visual Attention. *arXiv* **2015**, arXiv:1511.04119.
15. Xu, K.; Ba, J.; Kiros, R.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. *arXiv* **2015**, arXiv:1502.03044.
16. Albright, T.D.; Stoner, G.R. Contextual influences on visual processing. *Ann. Rev. Neurosci.* **2002**, *25*, 339–379. [[CrossRef](#)] [[PubMed](#)]
17. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv* **2014**, arXiv:1412.7755.
18. Wang, W.; Chen, C.; Wang, Y.; Jiang, T.; Fang, F.; Yao, Y. Simulating human saccadic scanpaths on natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
19. Jo, Y.; Wi, J.; Kim, M.; Lee, J.Y. Flexible Fashion Product Retrieval Using Multimodality-Based Deep Learning. *Appl. Sci.* **2020**, *10*, 1569. [[CrossRef](#)]
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Lake Tahoe, NV, USA, 2012; pp. 1106–1114.
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
22. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
23. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 17–22 June 2007.
24. Huang, Y.; Wang, W.; Wang, L. Instance-aware image and sentence matching with selective multimodal lstm. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2310–2318.
25. Huang, Y.; Wu, Q.; Wang, W.; Wang, L. Image and sentence matching via semantic concepts and order learning. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2020**, *42*, 636–650. [[CrossRef](#)] [[PubMed](#)]
26. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image-text matching. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 4654–4662.
27. Nguyen, D.K.; Okatani, T. Multi-task learning of hierarchical vision-language representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10492–10501.
28. 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, 27–30 June 2016; pp. 770–778.
29. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6077–6086.
30. Graves, A. Generating sequences with recurrent neural networks. *arXiv* **2013**, arXiv:1308.0850.
31. Larochelle, H.; Hinton, G.E. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Vancouver, BC, Canada, 2010.
32. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Montreal, QC, Canada, 2014.
33. Hu, X.; Yang, K.; Fei, L.; Wang, K. Acnet: Attention based network to exploit complementary features for rgb-d semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Chicago, IL, USA, 22–25 September 2019; pp. 1440–1444.
34. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
35. Yang, K.; Hu, X.; Chen, H.; Xiang, K.; Wang, K.; Stiefelhagen, R. Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swiftnet for surrounding sensing. *arXiv* **2019**, arXiv:1909.07721.

36. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [[CrossRef](#)]
37. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992.
38. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent neural network regularization. *arXiv* **2014**, arXiv:1409.2329.
39. Davis, S.B.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process. (TASSP)* **1980**, *28*, 357–366. [[CrossRef](#)]
40. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013.
41. Yan, F.; Mikolajczyk, K. Deep correlation for matching images and text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
42. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [[CrossRef](#)]
43. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist. (TACL)* **2014**, *2*, 67–78. [[CrossRef](#)]
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*; Springer: Zurich, Switzerland, 2014.
45. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Yuille, A.L. Explain images with multimodal recurrent neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
46. 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 *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Vancouver, BC, Canada, 2019; pp. 8026–8037.
47. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 24–27 June 2014.
48. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
49. Kiros, R.; Zhu, Y.; Salakhutdinov, R.R.; Zemel, R.; Urtasun, R.; Torralba, A.; Fidler, S. Skip-thought vectors. In *Neural Information Processing Systems (NeurIPS)*; NeurIPS Foundation: Montreal, QC, Canada, 2015.
50. Huang, Y.; Wang, L. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 5774–5783.