International Journal of Computer Vision (2021) 129:921–941
https://doi.org/10.1007/s11263-020-01392-1

Hierarchical Visual-Textual Knowledge Distillation for Life-Long Correlation Learning

Yuxin Peng[1] · Jinwei Qi[1] · Zhaoda Ye[1] · Yunkan Zhuo[1]

Received: 3 January 2019 / Accepted: 14 October 2020 / Published online: 5 January 2021

Abstract

Correlation learning among different types of multimedia data, such as visual and textual content, faces huge challenges from two important perspectives, namely, *cross modal* and *cross domain*. **Cross modal** means the heterogeneous properties of different types of multimedia data, where the data from different modalities have inconsistent distributions and representations. This situation leads to the first challenge: cross-modal similarity measurement. **Cross domain** means the multisource property of multimedia data from various domains, in which data from new domains arrive continually, leading to the second challenge: model storage and retraining. Therefore, correlation learning requires a cross-modal continual learning approach, in which only the data from the new domains are used for training, but the previously learned correlation capabilities are preserved. To address the above issues, we introduce the idea of life-long learning into visual-textual cross-modal correlation modeling and propose a visual-textual life-long knowledge distillation (VLKD) approach. In this study, we construct a hierarchical recurrent network that can leverage knowledge from both semantic and attention levels through adaptive network expansion to support cross-modal retrieval in life-long scenarios across various domains. The results of extensive experiments performed on multiple cross-modal datasets with different domains verify the effectiveness of the proposed VLKD approach for life-long cross-modal retrieval.

Keywords Cross-modal retrieval · Life-long learning · Hierarchical knowledge distillation · Attention transfer · Adaptive network expansion

1 Introduction

For the past few decades, multimedia data have been increasing explosively. Visual and textual content are two major forms on the Internet that have immensely enlarged the horizons for human beings. Accordingly, there is a huge demand for retrieving massive amounts of multimedia resources in many practical applications, such as intelligent search engines and multimedia digital libraries. Therefore, cross-modal retrieval (Peng et al. 2017) to perform retrieval across multimedia data has become an important research topic. As shown in Fig. 1, the purpose of cross-modal retrieval is to retrieve various types of multimedia information with a query of any modality. Cross-modal retrieval is more flexible and useful than the traditional single-modal retrieval for acquiring more comprehensive information.

The key point of visual-textual cross-modal retrieval is correlation learning, which effectively correlates data from different modalities. Thus, the retrieval process can search for relevant results in any modality. Naturally, semantic correlations exist among multimedia content. For example, images and texts are correlated when they describe the same semantics as an object or event. Therefore, the core issue is to understand the latent semantic correlation among the data of different modalities. However, because large amounts of multimedia data are distributed across diverse domains, correlation learning faces serious challenges from two important perspectives, namely, "cross modal" and "cross domain".

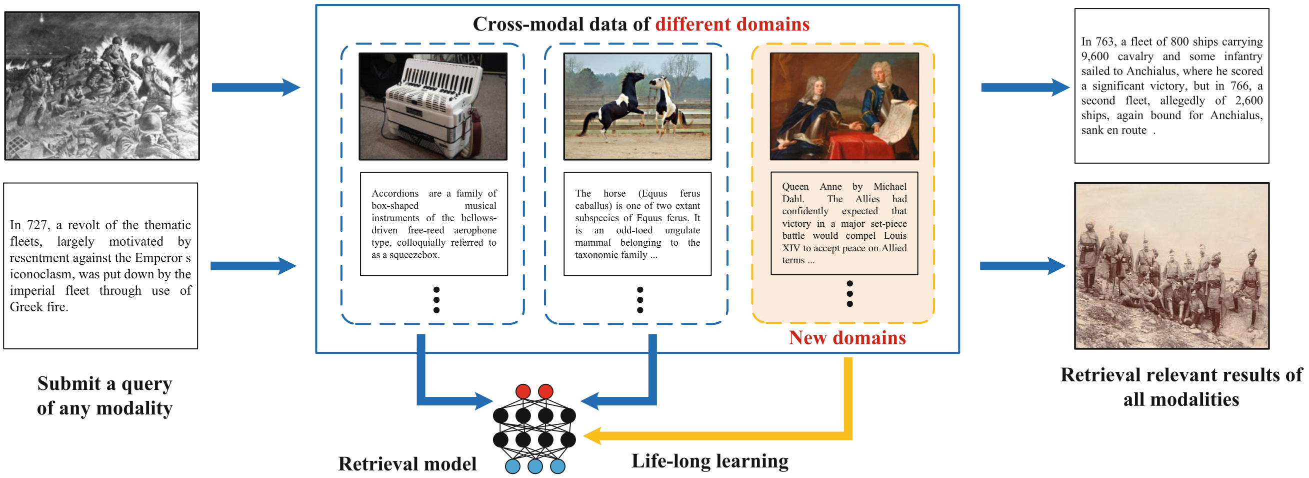
Cross modal, as the first challenge for correlation learning, involves the heterogeneous properties of multimedia data, which leads to inconsistent distributions and repre-

Communicated by Josef Sivic.

This work was supported by the National Natural Science Foundation of China under Grant 61925201 and Grant 61771025.

✉ Yuxin Peng
pengyuxin@pku.edu.cn

[1] Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China


Cross-modal data of different domains
New domains
Submit a query of any modality
Retrieval model
Life-long learning
Retrieval relevant results of all modalities

In 727, a revolt of the thematic fleets, largely motivated by resentment against the Emperor s iconoclasm, was put down by the imperial fleet through use of Greek fire.

Accordions are a family of box-shaped musical instruments of the bellows-driven free-reed aerophone type, colloquially referred to as a squeezebox.

The horse (Equus ferus caballus) is one of two extant subspecies of Equus ferus. It is an odd-toed ungulate mammal belonging to the taxonomic family ...

Queen Anne by Michael Dahl. The Allies had confidently expected that victory in a major set-piece battle would compel Louis XIV to accept peace on Allied terms ...

In 763, a fleet of 800 ships carrying 9,600 cavalry and some infantry sailed to Anchialus, where he scored a significant victory, but in 766, a second fleet, allegedly of 2,600 ships, again bound for Anchialus, sank en route .

Fig. 1 An example of visual-textual cross-modal retrieval with image and text, where the cross-modal data come from various domains. We need to learn a unified retrieval model to support life-long learning when data in the new domain arrive continually

sentations of different modalities. Therefore, measuring the similarities and constructing correlations across different modalities is challenging. Inspired by the human cognitive process, which fuses multiple senses such as vision, audition and language, common-space learning is intuitive for modelling a joint distribution over the multimedia data by projecting them into a common shared space.

Cross domain, as the second challenge for correlation learning, involves the multisource property of multimedia data. Specifically, data of different modalities are both dynamically increasing and come from various domains. Training a separate model for each new domain is unrealistic because multiple models would exist for different domains, resulting in unacceptable storage consumption. Thus, we need a unified sharing model to support continual learning that can discover new correlation capabilities for new domains while simultaneously maintaining the performance on existing domains. Because the existing data in old domains may become unavailable due to storage limits, direct retraining or fine-tuning on data in new domains may decrease the performance in both the new and old domains. On the one hand, the data in various domains are usually largely different and have diverse semantic distributions, which causes a model trained from the old domains to be unfit for new domains. On the other hand, the retraining may cause catastrophic forgetting that has an adverse effect on the model's performance in the old domains, making the newly learned model no longer optimal for the old domains.

To address the above issues, the main goal of this paper is to introduce life-long continual learning into cross-modal correlation modeling. It will not only support cross-modal retrieval across various domains with a unified sharing model, but also further promote retrieval performance through cross-domain knowledge transfer and sharing. Thus, in this paper, we propose the Visual-textual Life-long Knowledge Distillation (VLKD) approach that can leverage knowledge learned from existing data at both the semantic and attention level. We seek to obtain a better performance in the new domain while simultaneously preserving the original correlation capabilities and conducting training only with the data in the new domain. The main contributions of this paper are as follows:

- **Visual-Textual Hierarchical Recurrent Network** is constructed to fully exploit both global and fine-grained contextual information within visual and textual content, which can provide complementary hints and knowledge for correlation learning.
- **Cross-Domain Semantic-Level Knowledge Distillation** is proposed including both intra-domain distribution alignment and inter-domain knowledge distillation. Such knowledge distillation can explore the semantic correlations between visual and textual content, while preserving the previously learned cross-modal correlation knowledge during life-long learning.
- **Cross-Modal Attention-Level Knowledge Transfer** is proposed including intra-modality and inter-modality attention transfer. Thus, we can promote attention learning from both intra-modality and inter-modality perspectives, while also boosting fine-grained correlation learning in new domains under life-long scenarios.
- **Life-Long Adaptive Network Expansion** is proposed to adaptively expand the network capacity, thereby enabling the model to adapt to new domains in life-long learning setting. It can also serve to identify the genuinely essential parameters that preserve cross-modal knowledge across different domains.

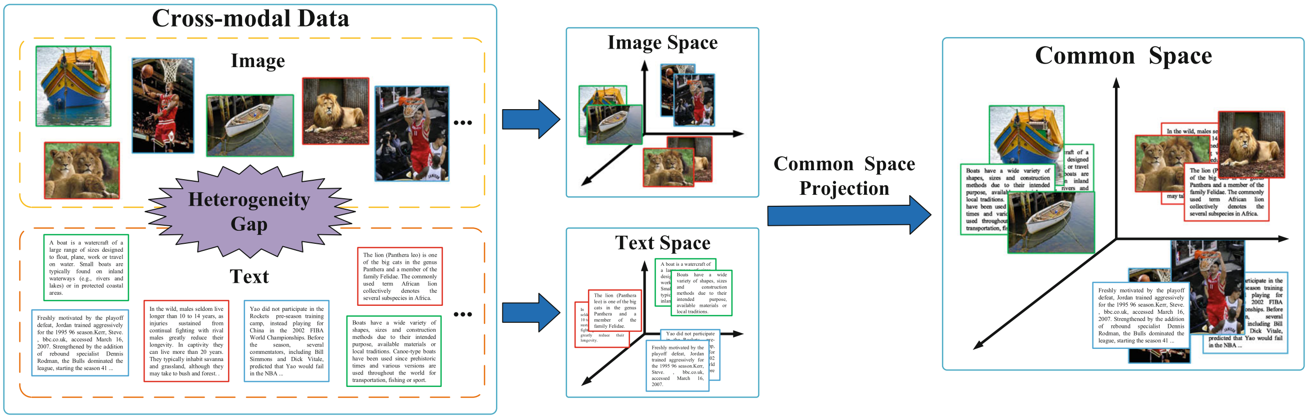


Fig. 2 Illustrations of the mainstream framework for visual-textual cross-modal correlation learning, which aims to project heterogeneous data from the individual feature spaces of different modalities into a single common space. Thus, a similarity measurement can be adopted to directly establish the correlations among cross-modal data

The main differences between this paper and our previous conference paper CmLL (Qi et al. 2018) can be summarized into the following 3 aspects: (1) *Attention-level knowledge transfer* is newly proposed to apply an attention mechanism over the visual and textual fine-grained inputs to highlight the discriminative parts. It can transfer both inter-modality and intra-modality attention knowledge learned from previous domains to promote discriminative correlation learning in the new domain. Our previous CmLL did not consider the attention information provided by the fine-grained parts, which cannot exploit fine-grained alignment effectively. (2) *Adaptive network expansion with parameter selection and pruning* is proposed not only to expand the capacity of a hierarchical recurrent network to absorb new knowledge, but also to adaptively select the important parameters while dropping the irrelevant ones to eliminate redundancy. Our previous CmLL only augments the network parameters by a fixed number simply in the output layer of the hierarchical network, which is not comprehensive and may increase network complexity. (3) *Extensive experiments on two additional new datasets and parameter experiments* were conducted to comprehensively verify the effectiveness of our approach, which is significantly improved compared to our previous CmLL.

2 Related Works

In this section, we briefly review the representative methods of visual-textual cross-modal retrieval and conventional life-long learning. Cross-modal retrieval is the task of our research, while life-long learning is the training scenario, which has realistic acquirements in practical applications.

2.1 Visual-Textual Cross-Modal Retrieval

The mainstream of cross-modal retrieval approaches follow the basic idea of constructing a single common space, as shown in Fig. 2. The goal is to effectively correlate the heterogeneous data by measuring the cross-modal similarities in this common space. The existing methods can be summarized into two categories based on the correlation model: traditional methods and deep learning based methods.

Traditional cross-modal retrieval methods usually embed the data from different modalities through linear projections, which are learned by optimizing statistical values. One representative method is canonical correlation analysis (CCA) (Rasiwasia et al. 2010), which learns projection matrices by maximizing the pairwise correlation of heterogeneous data with the same semantics. CCA is a classical solution that inspired many subsequent studies. For instance, Akaho (2006) incorporate a kernel function to explore the nonlinear cross-modal correlation, and Ranjan et al. (2015) extend CCA by considering the high-level semantics in multi-label annotations. Similar to Rasiwasia et al. (2010), Li et al. (2003) learn projection matrices of different modalities and optimize by directly minimizing the Frobenius norm between pairwise data. Wei et al. (2017) apply modality-specific projections for different cross-modal retrieval tasks. In addition, some other traditional methods have focused on graph regularization, in which graphs of data from different modalities are constructed to exploit cross-modal correlations. Zhai et al. (2013) adopt metric learning and graph regularization to learn the project matrices, and they further model the cross-modal correlation using several separate graphs for different modalities (Zhai et al. 2014). Wang et al. (2016a) also adopt graph regularization to preserve the inter-modality and intra-modality correlations simultaneously. Peng et al. (2016b) exploit the fine-grained information of different modalities of data by constructing a unified hypergraph.

Because deep learning has shown its strengths on many single-modal tasks such as image/text classification (Krizhevsky et al. 2012; Zhang et al. 2015) and object recognition (He et al. 2016), deep learning based methods have become a new trend in cross-modal retrieval research. Andrew et al. (2013) combine a deep neural network with CCA to learn better projections to common space. Ngiam et al. (2011) adopt a bimodal autoencoder to represent cross-modal correlations in the shared layer. Feng et al. (2014) jointly model the cross-modal correlation and reconstruction information with a correspondence autoencoder (Corr-AE). Wei et al. (2017) perform a deep semantic matching using deep network to map images and texts into their label vectors. Peng et al. (2016a) propose hierarchical structure with multiple deep networks, intend to preserve intra-modality information as well as inter-modality information. Then, they further employ both coarse-grained instances and fine-grained patches and introduces a multi-task learning strategy to learn the cross-modal correlations more precisely (Peng et al. 2018a).

Karpathy and Li (2015) infer the latent alignments between sentence segments and image regions to associate the two modalities through a common space. Wang et al. (2016b) propose a two-branch neural network to learn joint embeddings for images and text. Reed et al. (2016) extend the structured joint embedding and propose a word-based LSTM model to learn fine-grained image descriptions. Huang et al. (2018a) propose a semantically enhanced image-sentence matching model to learn semantic concepts and the correct semantic order for image representations. Fukui et al. (2016) propose multimodal compact bilinear pooling to combine visual and textual representations. Eisenschtat and Wolf (2017) propose a bidirectional neural network architecture that projects the data from two modalities into common space using Euclidean loss.

Furthermore, an attention mechanism is utilized to exploit the discriminative fine-grained information during cross-modal correlation learning. For example, a modality-specific cross-modal similarity measurement (MCSM) method (Peng et al. 2018b) is proposed to model the imbalanced and complementary relationships between different modalities by using an attention mechanism. Wang et al. (2018) exploit a joint attention mechanism to perform joint global and co-attentive representation learning between images and text. Lee et al. (2018) propose the stacked cross attention network (SCAN), which exploits the latent alignments between local image regions and keywords.

Despite great progress in cross-modal correlation learning, the methods mentioned above cannot be applied directly to implement continual learning. Without preserving learned knowledge, simply retraining the network on the data from the new domain would lead to catastrophic forgetting of previous domains. Therefore, it is imperative to incorporate a mechanism designed for life-long learning to allow knowledge transfer among cross-modal domains.

2.2 Life-Long Learning

Life-long learning remains a long-standing challenge for machine learning and neural networks, because catastrophic forgetting (Goodfellow et al. 2013) occurs when training data from different domains are input to models separately and sequentially. However, it is often infeasible to present all the data at the same time due to the enormous storage consumption of multimedia data. Therefore, it is essential to realize a life-long learning style that has the ability to acquire new knowledge and refine existing knowledge simultaneously based on continual inputs. Following the above idea, Mitchell et al. (2018) define a never-ending learning paradigm and propose a never-ending language learner that continually improves its reading competence over time.

Other attempts to alleviate catastrophic forgetting maintain the network architecture during training, but impose constraints on the neural weights updating. Li and Hoiem (2018) propose the concept of learning without forgetting (LwF), which utilizes knowledge distillation. This approach prevents the output of optimal model's in original tasks from shifting significantly when model is trained on new training data. Kirkpatrick et al. (2016) introduce elastic weight consolidation (EWC), which adopts a quadratic penalty to slow down the learning of certain weights based on how important they are to previous domains. Zenke et al. (2017) propose an approach similar to Kirkpatrick et al. (2016), but differs in the weight importance measurement. It computes synaptic relevance over the entire learning trajectory in an online manner. However, without increasing the network capacity, the above methods may result in sub-optimal performance due to limited neural resources.

The other category of methods explores how to dynamically expand the architecture of the network in response to new information. Rusu et al. (2016) propose progressive neural networks, which allocate additional fix-sized subnetwork when a new domain arrives. Although this approach is immune to forgetting the old domains, the number of parameters in the network expands substantially when faced with a large number of sequential tasks. Yoon et al. (2017) take a step forward to dynamically adjust its network capacity upon the arrival of each task.

Mallya and Lazebnik (2018) propose Packnet, which iteratively prunes and retrains the network to avoid catastrophic forgetting. Moreover, Xu and Zhu (2018) incorporate a reinforcement learning method to determine the optimal number of nodes to add for each layer when facing new tasks. In addition to appending extra units of neural network for new

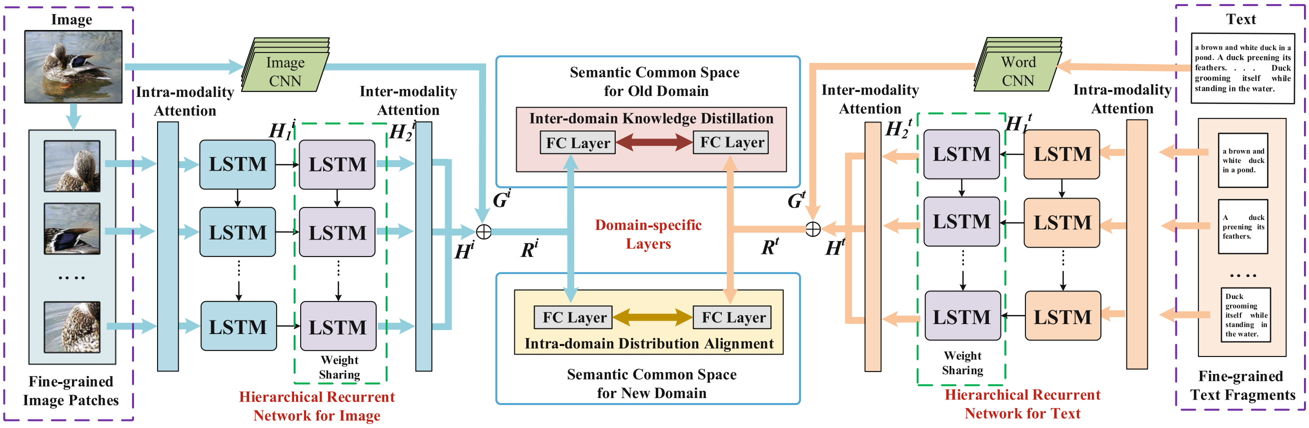


Image
Image CNN
Intra-modality Attention
Inter-modality Attention
H_1^i
H_2^i
LSTM
Weight Sharing
Fine-grained Image Patches
Hierarchical Recurrent Network for Image
Semantic Common Space for Old Domain
Inter-domain Knowledge Distillation
FC Layer
FC Layer
Domain-specific Layers
G^i
H^i
R^i
R^t
G^t
H^t
FC Layer
FC Layer
Intra-domain Distribution Alignment
Semantic Common Space for New Domain
Word CNN
Inter-modality Attention
Intra-modality Attention
H_2^t
H_1^t
Hierarchical Recurrent Network for Text
Text
a brown and white duck in a pond. A duck preening its feathers. . . . Duck grooming itself while standing in the water.
a brown and white duck in a pond.
A duck preening its feathers.
Duck grooming itself while standing in the water.
Fine-grained Text Fragments

Fig. 3 An overview of our VLKD approach. We construct a visual-textual hierarchical recurrent network for cross-modal life-long correlation learning that leverages knowledge from both the semantic and attention levels. The goal of our approach is to preserve previously learned correlations while boosting the performance in the new domain without affecting the learned correlations in the old domains

domains, Triki et al. (2017) train an autoencoder for each domain. It can capture the features with crucial information, while preserving the performance on old domains by preventing reconstructions of those features from changing. Aljundi et al. (2017) add experts trained on individual tasks to the network sequentially, then measure the task relatedness and select the most relevant expert during testing using a autoencoder implemented with a gate mechanism. Inspired by the recent advances in deep generative models, Shin et al. (2017) simply train generators to mimic the data of old domains that were no longer available. The key challenge in these methods lies in adding extra structures precisely to avoid unnecessarily complex structures.

However, the aforementioned methods are all designed to address single-modal scenarios, such as image classification. The existing methods are unsuitable for addressing the inconsistent distributions and complex correlations of different modalities to transfer knowledge across heterogeneous data. Instead, we propose a cross-modal life-long learning strategy that aims to absorb the complex correlation knowledge from previous cross-modal domains of visual and textual contents by adaptive network expansion. Our proposed approach considers both inter-modality and inter-domain characteristics during network training to implement effective life-long learning.

3 Our VLKD Approach

3.1 Notation and Overview

We propose a visual-textual life-long knowledge distillation (VLKD) approach to perform correlation learning under a life-long scenario. The network architecture is shown in Fig. 3. The visual-textual hierarchical recurrent network is constructed to exploit the global and fine-grained context information within each modality and further share the knowledge at a high level of the hierarchical network. We further propose semantic-level knowledge distillation, attention-level knowledge transfer and adaptive network expansion for cross-modal life-long learning. The goal of our approach is to preserve the original correlation ability in old domains while improving the performance in a new domain to implement life-long cross-modal retrieval. In summary, to build the initial cross-modal domain, we first train a hierarchical recurrent network using Eqs. (3) and (6). As new cross-modal domains arrive, the cross-modal life-long correlation is performed through semantic-level knowledge distillation, attention-level knowledge transfer and adaptive network expansion, as shown in Algorithm 1.

Algorithm 1: Cross-modal life-long learning procedure

```
Input: Shared parameters $\theta_S$, domain-specific parameters $\theta_D$ for
       old domains, cross-modal training data and corresponding
       category labels for new domain.
Output: Optimized VLKD model.
Compute the target probability distribution $y_p$ and initial
attention weights $a_k^i, a_k^t, b_k^i, b_k^t$ for the data in the new domain
with $\theta_S$ and $\theta_D$.
Randomly initialize new parameters $\theta_D^*$ in the domain-specific
layers for the new domain, and expand the parameters of the
hierarchical recurrent network.
Update both the shared parameters $\theta_S$ and the domain-specific
parameters $\theta_D^*$ for the new domain using Eqs. (3), (6), (8) and
(10).
Delete the parameters with small $l_2$-regularization values using
Eq. (11).
Fine-tune the both shared parameters $\theta_S$ and the domain-specific
parameters $\theta_D^*$ for the new domain using Eqs. (3), (6), (8) and
(10).
```

Next, we give a formal definition of the cross-modal domain for visual-textual correlation learning, which is denoted as $D = \{I, T\}$ which consists of image (I) and text (T) data. $I = \{i_p, c_p\}_{p=1}^{n}$ indicates that there are n instances of the image data, where i_p is the p-th image instances and has the semantic category label c_p. $T = \{t_p, c_p\}_{p=1}^{n}$ indicates that there are n text instances, which are defined similarly to the image instances. The goal of cross-modal retrieval is to measure the semantic similarity $sim(i_p, t_p)$ between the images and text by modeling the correlation among these multimedia data. The overall goal is to achieve cross-modal retrieval, so that relevant text can be searched by an image query, and vice versa.

3.2 Visual-Textual Hierarchical Recurrent Network for Correlation Learning

As shown in Fig. 3, we construct two levels of recurrent networks to capture the fine-grained context information from the visual and textual content. The network weights at the second level are shared among the different modalities, which is intended to share the knowledge from the visual and textual contents. We also add two levels of attention layers at the bottom and at the top of the hierarchical network. The first level aims to model the intra-modality attention information within each modality, while the second level aims to model the inter-modality information in the weight-sharing network. In addition, we construct convolutional networks to explore the global visual and textual information. Finally, domain-specific layers are added to maintain model diversity to ensure the adaptivity and extensibility of the new domain.

3.2.1 Cross-Modal Inputs

The hierarchical recurrent network adopts sequential input to exploit the fine-grained context information for each modality. For each image instance i_p, the separate features from different image regions are extracted from the last pooling layer (pool5) of the 19-layer VGGNet (Simonyan and Zisserman 2014). We organize the image features as a sequence denoted as $S_i = \{v_1^i, \ldots, v_{m_i}^i\}$ for a total of m_i regions. For each text instance t_p, we first extract a k-dimensional vector for each word using Word2Vec model. Then, we divide the text instance into several fragments (by paragraph or sentence) and utilize Word-CNN (Kim 2014) to extract separate features from these different text fragments. Each fragment consisting of w words is represented as a $w \times k$ matrix. The textual sequence is denoted as $S_t = \{v_1^t, \ldots, v_{m_t}^t\}$, which contains m_t fragments organized in their original order. Furthermore, we also generate a global visual representation G^i for each complete image using the 19-layer VGGNet, as well as a global textual representation G^t for the full text using Word-CNN.

3.2.2 Network Architecture

The sequential features for image and text contain rich fine-grained contextual information. These features are then input into the visual-textual hierarchical recurrent network with the two levels of attention layers. First, considering that the importances of the visual regions in images and text in textual fragments are different, we adopt the first level of intra-modality attention layer to model the sequential inputs for image and text. Specifically, for image sequential features S_i, we adjust the scale of features using the intra-modality attention mechanism, which is intended to highlight the discriminative fine-grained information in the visual context and is calculated as follows:

$$a_k^i = \frac{e^{F_{intra}^i(v_k^i)}}{\sum_{p=1}^{m_i} e^{F_{intra}^i(v_p^i)}}, \quad \hat{v}_k^i = a_k^i v_k^i \tag{1}$$

where F_{intra}^i is a fully-connected layer that outputs a one-dimensional value for each image feature. The above equation represents the intra-modality attention, which is depicted at the left side of Fig. 3. As mentioned, it is intended to highlight the discriminative fine-grained information in the visual context. We calculate the attention weight a_k^i for each image region, and obtain an input sequence for the LSTM denoted as $\hat{S}_i = \{\hat{v}_1^i, \ldots, \hat{v}_{m_i}^i\}$. The textual sequence is adjusted similarly, to $\hat{S}_t = \{\hat{v}_1^t, \ldots, \hat{v}_{m_t}^t\}$ with F_{intra}^t, as shown at the right of Fig. 3.

For the hierarchical recurrent network, we adopt a long short term memory (LSTM) unit to model the fine-grained context information using two levels. The first level of the recurrent network takes the visual and textual sequences $\hat{S}_i$ and $\hat{S}_t$ as inputs respectively. The outputs from hidden units are denoted as $H_1^i = \{h_1^i, \ldots, h_{m_i}^i\}$ for images and $H_1^t = \{h_1^t, \ldots, h_{m_t}^t\}$ for text, and they contain rich fine-grained modality-specific context information. Next, H_1^i and H_1^t form the inputs to the second level of the recurrent network, which follows a process similar to that of the first level mentioned above: it takes H_1^i and H_1^t as inputs and outputs H_2^i and H_2^t for image and text, respectively. However, in contrast to the first level, the weights in the second level are shared among the visual and textual sequences. The idea behind sharing the weights is to share the high-level contextual knowledge learned from both modalities, allowing inter-modality context correlation to be fully exploited.

Furthermore, we place the inter-modality attention layers after the second level of the weight-sharing recurrent network. The layers are also shared across the different modalities to model the joint attention information. Specifically, the inter-modality attention weights for the visual sequence are calculated as follows:

$$b_k^i = \frac{e^{F_{inter}(h_k^i)}}{\sum_{p=1}^{m_i} e^{F_{inter}(h_p^i)}}, \quad H^i = \sum_{k=1}^{m_i} b_k^i h_k^i \tag{2}$$

where F_{inter} is a fully-connected layer that generates a one-dimensional value. Thus, the final visual fine-grained context representation is represented as H^i, which is further concatenated with its corresponding global representation G^i to obtain a visual common representation denoted as $R^i = \{r_1^i, \ldots, r_n^i\}$. Similarly, the text features H^t and G^t are concatenated into textual common representation $R^t = \{r_1^t, \ldots, r_n^t\}$, which contains both fine-grained context and global information.

We then input the visual and textual common representations into the domain-specific layer to ensure the adaptivity and extensibility of the model to a new domain. The domain-specific layer is a single fully-connected layer for each modality followed by a softmax layer that transforms the concatenated visual and textual representations into semantic common representations. These layers have domain-independent parameters and obtain the classification probabilities $Y^i = \{y_1^i, \ldots, y_n^i\}$ and $Y^t = \{y_1^t, \ldots, y_n^t\}$ for life-long learning process.

3.2.3 Common Space Learning

We first propose *intra-domain distribution alignment* to enhance the learned common space by transferring semantic knowledge between the different modalities within each domain.

Specifically, we design a joint embedding loss to preserve the relative similarity between image and text during common space learning. Thus, the ranking information in common space can be effectively modeled. In addition, we utilize the maximum mean discrepancy (MMD) criterion (Gretton et al. 2012) to match the distributions of the image and text representations in common space. This approach further enhances the alignment between different modalities. The objective function is defined as follows:

$$\mathcal{L}_{intra} = f_{mmd}(R^i, R^t) + f_{je}(r^{i+}, r^{i-}, r^{t+}, r^{t-}) \tag{3}$$

$$f_{mmd} = \left\| E_I(r_p^i) - E_T(r_p^t) \right\|_{\mathcal{H}}^2 \tag{4}$$

$$\begin{aligned} f_{je} = \max(0, &[d(r^{i+}, r^{t+}) - d(r^{i+}, r^{t-})] \\ &+ [d(r^{i+}, r^{t+}) - d(r^{t+}, r^{i-})] + \alpha) \end{aligned} \tag{5}$$

where the squared formulation of MMD in the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ is adopted. In the MMD function, $E_I(f(x)) = dis(f(x), \mu(I))$ for $f \in \mathcal{H}$ is the kernel function and $\mu(I)$ denotes the mean embedding of modality I in the reproducing kernel Hilbert space $\mathcal{H}$. We regard r_p in our loss function as $f(x)$ above and adopt cosine distance as dis. In the cross-modal joint embedding loss f_{je}, $d(r^{i+}, r^{t+})$ denotes the similarity of matched image/text pairs, while the similarities of mismatched pairs are denoted as $d(r^{i+}, r^{t-})$ and $d(r^{t+}, r^{i-})$. α is the margin parameter. Therefore, we can align the distributions of the different modalities using the above objective function, and further preserve the relative similarity ranking information to boost cross-modal correlation learning.

Many existing works, such as Wei et al. (2017) and Wang et al. (2017), use category labels to preserve the semantic information. Instead, we map the encoded visual and textual representations r_p^i and r_p^t into a semantic common space through domain-specific layer if global-level category label is available, and obtain the semantic common representations y_p^i and y_p^t guided by the category information. It can model the cross-modal correlation while simultaneously preserving the semantic information. Specifically, we adopt the following semantic loss function:

$$\mathcal{L}_{se} = \sum_{p=1}^{N} (f_{sm}(y_p^i, c_p) + f_{sm}(y_p^t, c_p)) \tag{6}$$

where $f_{sm}(y_p, c_p)$ is the cross entropy loss function, defined as:

$$f_{sm}(y_p, c_p) = -\sum_{q=1}^{L} 1\{q = c_p\} \log(y_p) \tag{7}$$

where c_p is the corresponding category label. There are a total of L categories in this domain. N indicates the number of image-text pairs in one batch. In addition, $1\{q = c_p\}$ equals 1 when $q = c_p$, and 0 otherwise.

This semantic loss functions to make the distances between images and texts within the same category as close as possible in the common space, which can preserve the semantic consistency of common representations.

3.3 Cross-Modal Life-Long Learning

We perform cross-modal life-long correlation learning by semantic-level knowledge distillation and attention-level knowledge transfer. The goal is to preserve the original correlation ability for existing domains while improving the performance in the new domain by transferring knowledge from both the semantic and attention aspects. Furthermore, we propose an adaptive network expansion strategy to increase model's capacity to absorb knowledge from the new domain.

3.3.1 Semantic-Level Knowledge Distillation

Because we have obtained the semantic common spaces for different domains, we next perform semantic-level knowledge distillation to transfer the knowledge from both intra-domain and inter-domain perspectives. We propose *inter-domain knowledge distillation* to maintain the semantic distribution from the old domains during the life-long learning process, which requires the outputs from the domain-specific layers for the old domains to remain close to their initial outputs. Inspired by Hinton et al. (2015), the objective function for inter-domain knowledge distillation is defined as follows:

$$\mathcal{L}_{inter} = -\sum_{p=1}^{N} y_p \log(\hat{y}_p) \tag{8}$$

where the predicted probability distribution in the semantic common space for the old domain is denoted as $\hat{y}_p$, while y_p is the target probability distribution recorded from the original network.

We aim to guide the predicted probability distribution $\hat{y}_p$ to remain close to the target y_p. To provide a concrete example, we have built a common space for the data of the old domain in which the data of different categories are gathered into different centers. When the data comes from the new domain, we first project the new data into the common space of the old domain, and record the relative distance between the new data and the original centers. Then, during the life-long training, we try to keep the relative distances unchanged to preserve the positions of the original centers, which effectively preserves the original semantic knowledge and transfers it to the new domain to achieve knowledge distillation.

3.3.2 Attention-Level Knowledge Transfer

To preserve the knowledge learned from the discriminative fine-grained information in the old domains, we perform attention-level knowledge transfer for cross-modal life-long learning from both intra-modality and inter-modality perspectives. Zagoruyko and Komodakis (2016) have demonstrated that attention information can be transfered as knowledge. Thus, we define the objective function of intra-modality attention-level knowledge transfer for visual content as follows:

$$\mathcal{AL}^i_{intra} = \frac{1}{m_i}\sum_{k=1}^{m_i}(a^i_k - \hat{a}^i_k)^2 \tag{9}$$

where n is the total number of image regions in the sequence. $\hat{a}^i_k$ is the initial recorded attention weight, while a^i_k is generated during the current process as defined in Eq. (1). Therefore, complete objective function for the attention-level knowledge transfer is defined as:

$$\mathcal{AL} = \mathcal{AL}^i_{intra} + \mathcal{AL}^t_{intra} + \mathcal{AL}^i_{inter} + \mathcal{AL}^t_{inter} \tag{10}$$

where $\mathcal{AL}^t_{intra}$ is the intra-modality attention transfer for text calculated with a^t_k, which has a definition similar to $\mathcal{AL}^i_{intra}$. Then $\mathcal{AL}^i_{inter}$ and $\mathcal{AL}^t_{inter}$ are calculated with the inter-modality attention b^i_k and b^t_k in Eq. (2). Through this attention-level knowledge transfer, the learned attention knowledge from the discriminative fine-grained information in the old domains can be fully exploited to boost attention learning in the new domain, while simultaneously maintaining the performance of the old domains .

3.3.3 Adaptive Network Expansion Strategy

Considering that network capacity is not infinite, as increasingly amounts of cross-modal data arrive from new domains continually during life-long learning, the performance would be eventually limited by the network capacity. To address this issue, as shown in Fig. 4, we attempt to increase the network capacity by adding new network parameters to expand the original network.

When we add a new cross-modal domain to the life-long learning process, the parameters of the hierarchical recurrent network are increased to allow the network to absorb knowledge from the new domain. Specifically, we preserve the input and output dimension of hierarchical network, while adding the dimensions to the hidden state from D_{old} to D_{new}. Thus, the expansion ratio is defined as $\beta = \frac{D_{new}-D_{old}}{D_{old}}$.

However, simply adding network parameters may cause redundancy during network expansion. Thus, we need to decide which parameters are truly necessary and should be expanded for the new domain and which can be discarded. Therefore, we propose another strategy to adaptively select the parameters that have the most important impact on performance and delete the less relevant and irrelevant ones. To preserve the input and output of dimensions of the hierarchical network, we delete the parameters in the weight matrices whose rows and columns are in the first and second level of hierarchical recurrent network, respectively. Specifically, we determine the number of parameters guided by l_2-regularization, which is defined as follows:

$$\begin{aligned} S1 &= \sum_{k=1}^{h} 1\{\left\| \hat{w}_1(k*)\right\|_{l_2} < \theta\} \\ S2 &= \sum_{k=1}^{h} 1\{\left\| \hat{w}_2(*k)\right\|_{l_2} < \theta\} \\ D_{fin} &= D_{new} - min(S1, S2) \end{aligned} \tag{11}$$

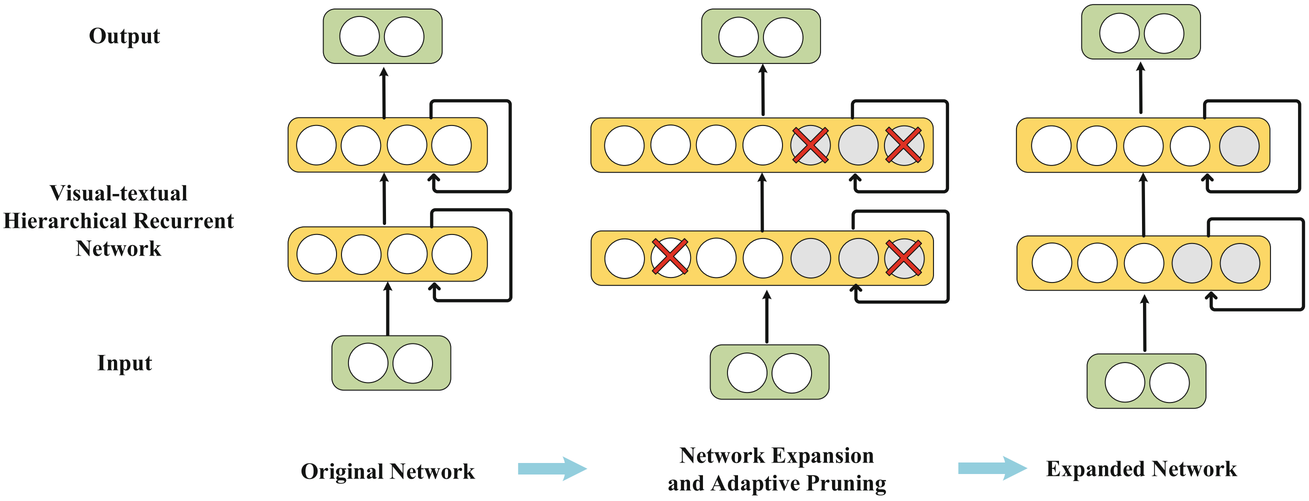
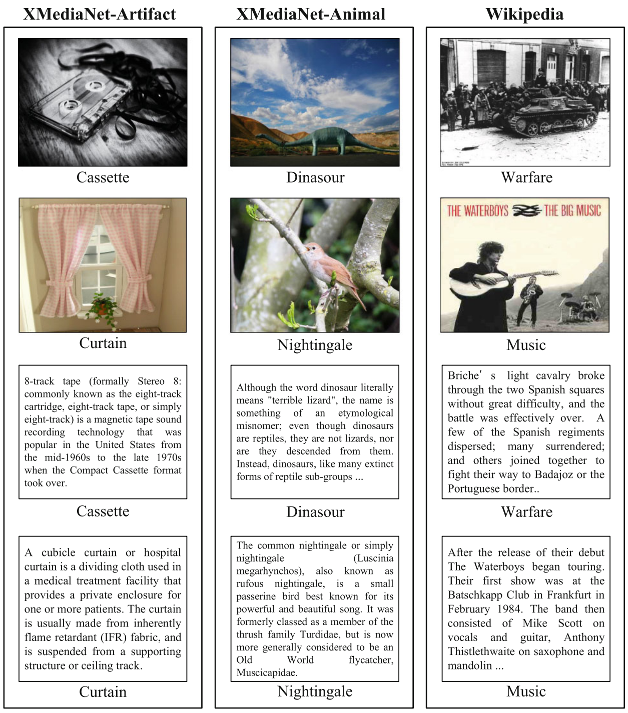

Output
Visual-textual Hierarchical Recurrent Network
Input
Original Network
Network Expansion and Adaptive Pruning
Expanded Network

Fig. 4 Illustration of adaptive network expansion, which first augments the network parameters to absorb knowledge from new domain, then adaptively selects the parameters with important impact on the performance and drops those irrelevant ones to eliminate redundancy

where $\hat{w}_1$ and $\hat{w}_2$ represent the weight matrices in the first and second levels of the hierarchical recurrent network, while $\hat{w}_1(k*)$ and $\hat{w}_2(*k)$ represent the k-th rows and columns of matrix $\hat{w}_1$ and $\hat{w}_2$. $S1$ and $S2$ are the number of rows and columns whose l_2-regularization are smaller than the threshold θ, and D_{fin} is the hidden state dimension in the hierarchical recurrent network after the deletion.

Note that we treat network weights with small magnitudes as redundancy because they contribute very little to the network calculation. Thus, they can be eliminated, while those with large magnitudes are preserved. In addition, we eliminate redundant weights based on a fixed threshold rather than on a pairwise comparison, mainly in consideration of the computational cost during network training. Using this method, we can adaptively expand the parameters to increase the network's capacity to absorb knowledge from the new domain while eliminating redundancy for effective life-long learning.

In summary, we leverage the knowledge learned from existing data at both the semantic level and attention level, which promotes the correlation learning performance in the new domain and preserves the original correlation capabilities of the old domains. In this way, cross-modal retrieval can be performed under life-long scenarios among different domains simultaneously.



8-track tape (formally Stereo 8: commonly known as the eight-track cartridge, eight-track tape, or simply eight-track) is a magnetic tape sound recording technology that was popular in the United States from the mid-1960s to the late 1970s when the Compact Cassette format took over.

Cassette

Although the word dinosaur literally means "terrible lizard", the name is something of an etymological misnomer; even though dinosaurs are reptiles, they are not lizards, nor are they descended from them. Instead, dinosaurs, like many extinct forms of reptile sub-groups ...

Dinasour

Briche′ s light cavalry broke through the two Spanish squares without great difficulty, and the battle was effectively over. A few of the Spanish regiments dispersed; many surrendered; and others joined together to fight their way to Badajoz or the Portuguese border..

Warfare

A cubicle curtain or hospital curtain is a dividing cloth used in a medical treatment facility that provides a private enclosure for one or more patients. The curtain is usually made from inherently flame retardant (IFR) fabric, and is suspended from a supporting structure or ceiling track.

Curtain

The common nightingale or simply nightingale (Luscinia megarhynchos), also known as rufous nightingale, is a small passerine bird best known for its powerful and beautiful song. It was formerly classed as a member of the thrush family Turdidae, but is now more generally considered to be an Old World flycatcher, Muscicapidae.

Nightingale

After the release of their debut The Waterboys began touring. Their first show was at the Batschkapp Club in Frankfurt in February 1984. The band then consisted of Mike Scott on vocals and guitar, Anthony Thistlethwaite on saxophone and mandolin ...

Music

Fig. 5 Image and text examples in 3 cross-modal domains. The data in these domains is largely different that arises huge challenges for cross-modal life-long correlation learning

4 Experiments

We conduct the experiments under life-long scenario on 5 cross-modal domains that are designed to evaluate the effectiveness of our proposed approach. In this section, we first briefly introduce the datasets and experimental configuration under life-long scenario. Then, we present the experimental results and compare with 12 state-of-the-art cross-modal retrieval methods. In addition, we present the comprehensive experimental analyses, including parameter analyses and comparisons with the fine-tuning strategy, as well as ablation studies of each component in our proposed approach.

4.1 Dataset Configuration for Life-Long Scenario

The experiments involve 4 datasets with 5 different cross-modal domains, each of which is split into 3 subsets: a training set, a testing set and a validation set.

PKU XMediaNet dataset (Peng et al. 2017) is a large-scale cross-modal dataset with 5 modalities. We use 2 of the modalities (image and text) for experiment, constituting 40,000 image/text pairs. The images are all collected from Flickr, while the text paragraphs are all obtained from Wikipedia website with relevant topics. Up to 200 independent categories are selected from the WordNet, which consists of 47 animal species and 153 artifact species. Thus, it can be divided into the following two domains, namely, XMediaNet-artifact and XMediaNet-animal.

– **XMediaNet-artifact** contains **153** artifact types, such as violin, airplane, shotgun and camera. There are 24,480 pairs in the training set, 3060 pairs in the testing set and 3060 pairs in the validation set.
– **XMediaNet-animal** consists of **47** species of animals, such as elephant, owl, bee and frog, etc. There are 7520 pairs in the training set, 940 pairs in the testing set and 940 pairs in the validation set.

Wikipedia dataset (Rasiwasia et al. 2010) is the most widely-used dataset for cross-modal retrieval, which is constructed from Wikipedia "featured articles". It has 2,866 image/text pairs of 10 most populated categories on high-level semantics, such as history, music, warfare and so on. We follow the dataset partition strategy of Peng et al. (2018a); Feng et al. (2014) to divide the dataset into 3 subsets, namely 2,173 pairs for training, 231 pairs and 462 pairs for validation and testing respectively.

MS-COCO dataset (Lin et al. 2014) contains 123,287 images, and each image has 5 annotated sentences generated by crowdsourcing via Amazon Mechanical Turk. Following Peng et al. (2018a), 5000 image-text pairs are selected for testing, and 5000 pair for validation, while the rest are for training.

Im2p dataset (Krause et al. 2017) contains 19,561 images from the Visual Genome project, which was originally constructed for generating paragraphs form images. Each image is annotated with a one-paragraph description. It has 14,575 image-paragraph pairs for training, 2487 pairs and 2489 pairs for validation and testing respectively.

Figure 5 provides some examples of these domains. We can observe that there is no overlap between the domains, and these domains are distinctly different from each other. As a result, it is quite challenging to conduct cross-modal life-long learning on these domains. Note that in life-long scenario, the model is trained on 4 cross-modal domains sequentially as follows: "**XMediaNet-artifact** → **XMediaNet-animal** → **Wikipedia** → **MS-COCO**"(VLKD-MSCOCO) and "**XMediaNet-artifact** → **XMediaNet-animal** → **Wikipedia** → **Im2p**"(VLKD-Im2p). The data from the old domains are not available when training on a new domain. We place MS-COCO and Im2p dataset last due to their characteristics for the following reasons. First, the image-sentence matching task in MS-COCO and Im2P datasets is quite different from the cross-modal retrieval task in other domains because its goal is to find exact matches between an image and a sentence. In addition, the category labels are not available for the image-sentence matching task in MS-COCO and Im2p dataset. Thus, we cannot apply semantic-level knowledge distillation to support continual life-long learning.

4.2 Implementation Details

We implemented the architecture of the proposed network in PyTorch. For images, the global representations are extracted from the last fully-connected layer of VGGNet (Simonyan and Zisserman 2014) with 4,096 dimensions. The fine-grained representations are generated from the last pooling layer of VGGNet. Specifically, we use the feature map from pool5 layer of VGGNet with a 7×7 mapping over the image. Thus, there are 49 patches for each image, organized from left to right and from top to bottom, a sequence that mimcs the movement of the human eye when perusing an image. For text, the global representations are obtained by Word-CNN (Kim 2014) with 300 dimensions. The strategy of splitting the text depends on the form of the text instances in the dataset. In Wikipedia and PKU XMediaNet, each text instance is a long article containing multiple paragraphs, and each text instance is split by the paragraph where each paragraph contains relevant content. The fine-grained representations are extracted from text fragments by the same Word-CNN. For MS-COCO dataset, each text instance contains 5 sentences, which is relatively short. Thus, we follow the setting of Yan and Mikolajczyk (2015) that combines all sentences into a single text. Then, we consider each word as a fragment for the fine-grained modeling process. The LSTM network at each level of the hierarchical recurrent architecture has two units in series whose output dimensions are originally set to 512. After each continual learning process on a new domain, we first perform dynamic expansion, increasing the hidden size of the LSTM network by 20%, and then pruning with threshold θ (set as 0.001) as shown in Eq. (11). In addition, the global representations are converted to 1024 dimensions by fully-connected layer. As a result, the final common representations of image and text are the concatenation of the global and fine-grained representations, which have 1,536 dimensions and are fed into domain-specific layers. In addition, we set the margins α in loss function (5) to 0.2, and β is 0.2, which are tuned on the validation set of Wikipedia dataset, and we adopt the same parameters for all cross-modal domains. We adopt Adam optimizer during training with learning rate of 0.0001 and dropout rate of 0.5. The training is done with the batch size set to 32 and learning rate fixed as 1e−4. In the experiments, we adopt the same parameters for all cross-modal domains.

4.3 Evaluation Metric and Compared Methods

We conduct two types of **cross-modal retrieval** tasks on the multiple domains, namely, retrieving text by image (**Img→Txt**) and retrieving image by text (**Txt→Img**). Both types involve taking the queries of one modality and retrieving relevant instances of the other modality in the testing set. Concretely, after generating the common representations of image and text, we compute the cosine distances between the query and all instances of the other modality in the testing set, and then sort the distances in descending order to obtain the retrieval results.

The metric adopted for evaluation is mean average precision (MAP) score, which simultaneously reflects the precision and ranking of the returned retrieval results. The MAP score is the mean value of the average precision (AP) of all queries in the testing set and is calculated as follows:

$$AP = \frac{1}{R}\sum_{k=1}^{n}\frac{R_k}{k} \times rel_k, \tag{12}$$

where n is the number of retrieval set, R means the number of relevant items and R_k counts the number of relevant items in the top k results. When the k-th result is relevant, rel_k is set to 1, otherwise 0. Note that here, we consider all the returned retrieval when computing the MAP score rather than only top-50 results as in Corr-AE (Feng et al. 2014) and ACMR (Wang et al. 2017).

In addition, for MS-COCO and Im2p datasets, we report the Recall@K score as evaluation metric following (Peng et al. 2018a) for the **image-sentence matching task**, which includes matching text by image (**Img→Txt**) and matching image by text (**Txt→Img**). We calculate the recall rate for the top 1 result (R@1), top 5 results (R@5) and top 10 results (R@10). Considering that current works do not utilize the category labels in MS-COCO dataset, to ensure a fair comparison, we drop the semantic loss in our approach. But some methods including CMDN (Peng et al. 2016a), Deep-SM (Wei et al. 2017), JRL (Zhai et al. 2014) and LGCFL (Kang et al. 2015) cannot be compared on these datasets.

For the compared methods, we compare 12 state-of-the-art cross-modal retrieval methods in the experiments, including 5 traditional methods, namely CCA (Rasiwasia et al. 2010), CFA (Li et al. 2003), KCCA (Hardoon et al. 2004), JRL (Zhai et al. 2014) and LGCFL (Kang et al. 2015), and 7 deep learning based methods, namely Corr-AE (Feng et al. 2014), DCCA (Andrew et al. 2013), CMDN (Peng et al. 2016a), Deep-SM (Wei et al. 2017), ACMR (Wang et al. 2017), CCL (Peng et al. 2018a) and our previous CmLL (Qi et al. 2018). Brief introductions of these methods are as follows:

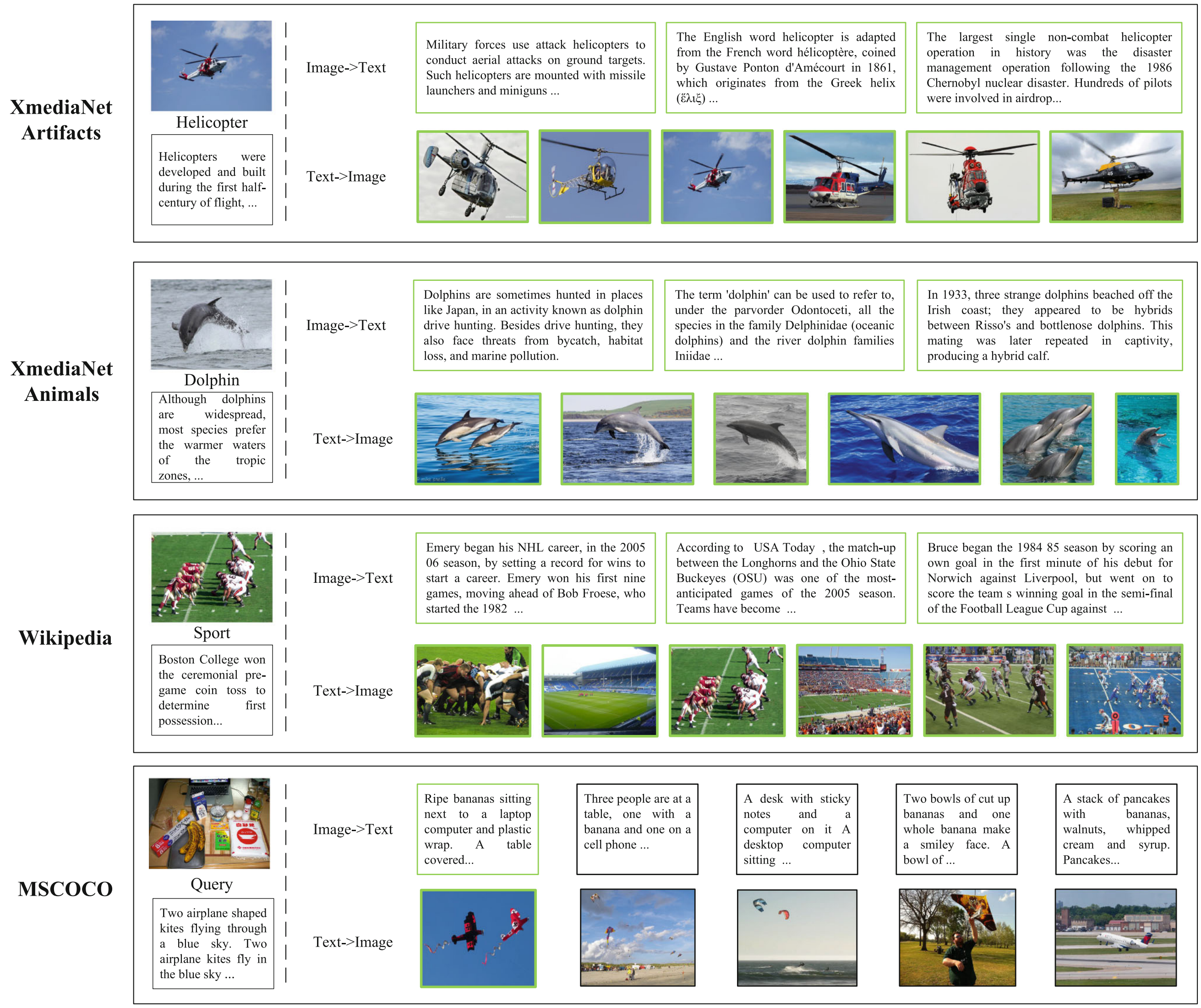
Table 1 The MAP scores of cross-modal retrieval on **XMediaNet-artifact**

Method	XMediaNet-artifact		
	Img → Txt	Txt → Img	Average
VLKD(MS-COCO)	0.709	0.698	0.704
VLKD(Im2p)	0.707	0.704	0.706
VLKD-single	0.667	0.665	0.666
Our previous CmLL	0.604	0.608	0.606
CCL	0.504	0.488	0.496
ACMR	0.498	0.487	0.492
CMDN	0.448	0.484	0.466
Deep-SM	0.382	0.308	0.345
LGCFL	0.416	0.472	0.444
JRL	0.448	0.364	0.406
VLKD-noSe	0.467	0.504	0.486
DCCA	0.389	0.390	0.390
Corr-AE	0.431	0.468	0.449
KCCA	0.238	0.258	0.248
CFA	0.230	0.353	0.292
CCA	0.243	0.256	0.249

Table 2 The MAP scores for cross-modal retrieval on **XMediaNet-animal** dataset

Method	XMediaNet-animal		
	Img → Txt	Txt → Img	Average
VLKD(MS-COCO)	0.913	0.917	0.915
VLKD(Im2p)	0.915	0.916	0.916
VLKD-single	0.897	0.888	0.892
Our previous CmLL	0.768	0.781	0.775
CCL	0.684	0.684	0.684
ACMR	0.702	0.696	0.699
CMDN	0.660	0.652	0.656
Deep-SM	0.499	0.490	0.494
LGCFL	0.580	0.678	0.629
JRL	0.708	0.628	0.668
VLKD-noSe	0.775	0.763	0.769
DCCA	0.654	0.677	0.665
Corr-AE	0.625	0.661	0.643
KCCA	0.316	0.366	0.341
CFA	0.590	0.608	0.599
CCA	0.378	0.378	0.378

VLKD-single is trained individually on the current domain, and **VLKD-noSe** is trained without semantic loss



Helicopters were developed and built during the first half-century of flight, ...

Military forces use attack helicopters to conduct aerial attacks on ground targets. Such helicopters are mounted with missile launchers and miniguns ...

The English word helicopter is adapted from the French word hélicoptère, coined by Gustave Ponton d'Amécourt in 1861, which originates from the Greek helix (ἕλιξ) ...

The largest single non-combat helicopter operation in history was the disaster management operation following the 1986 Chernobyl nuclear disaster. Hundreds of pilots were involved in airdrop...

Although dolphins are widespread, most species prefer the warmer waters of the tropic zones, ...

Dolphins are sometimes hunted in places like Japan, in an activity known as dolphin drive hunting. Besides drive hunting, they also face threats from bycatch, habitat loss, and marine pollution.

The term 'dolphin' can be used to refer to, under the parvorder Odontoceti, all the species in the family Delphinidae (oceanic dolphins) and the river dolphin families Iniidae ...

In 1933, three strange dolphins beached off the Irish coast; they appeared to be hybrids between Risso's and bottlenose dolphins. This mating was later repeated in captivity, producing a hybrid calf.

Boston College won the ceremonial pre-game coin toss to determine first possession...

Emery began his NHL career, in the 2005 06 season, by setting a record for wins to start a career. Emery won his first nine games, moving ahead of Bob Froese, who started the 1982 ...

According to USA Today , the match-up between the Longhorns and the Ohio State Buckeyes (OSU) was one of the most-anticipated games of the 2005 season. Teams have become ...

Bruce began the 1984 85 season by scoring an own goal in the first minute of his debut for Norwich against Liverpool, but went on to score the team s winning goal in the semi-final of the Football League Cup against ...

Two airplane shaped kites flying through a blue sky. Two airplane kites fly in the blue sky ...

Ripe bananas sitting next to a laptop computer and plastic wrap. A table covered...

Three people are at a table, one with a banana and one on a cell phone ...

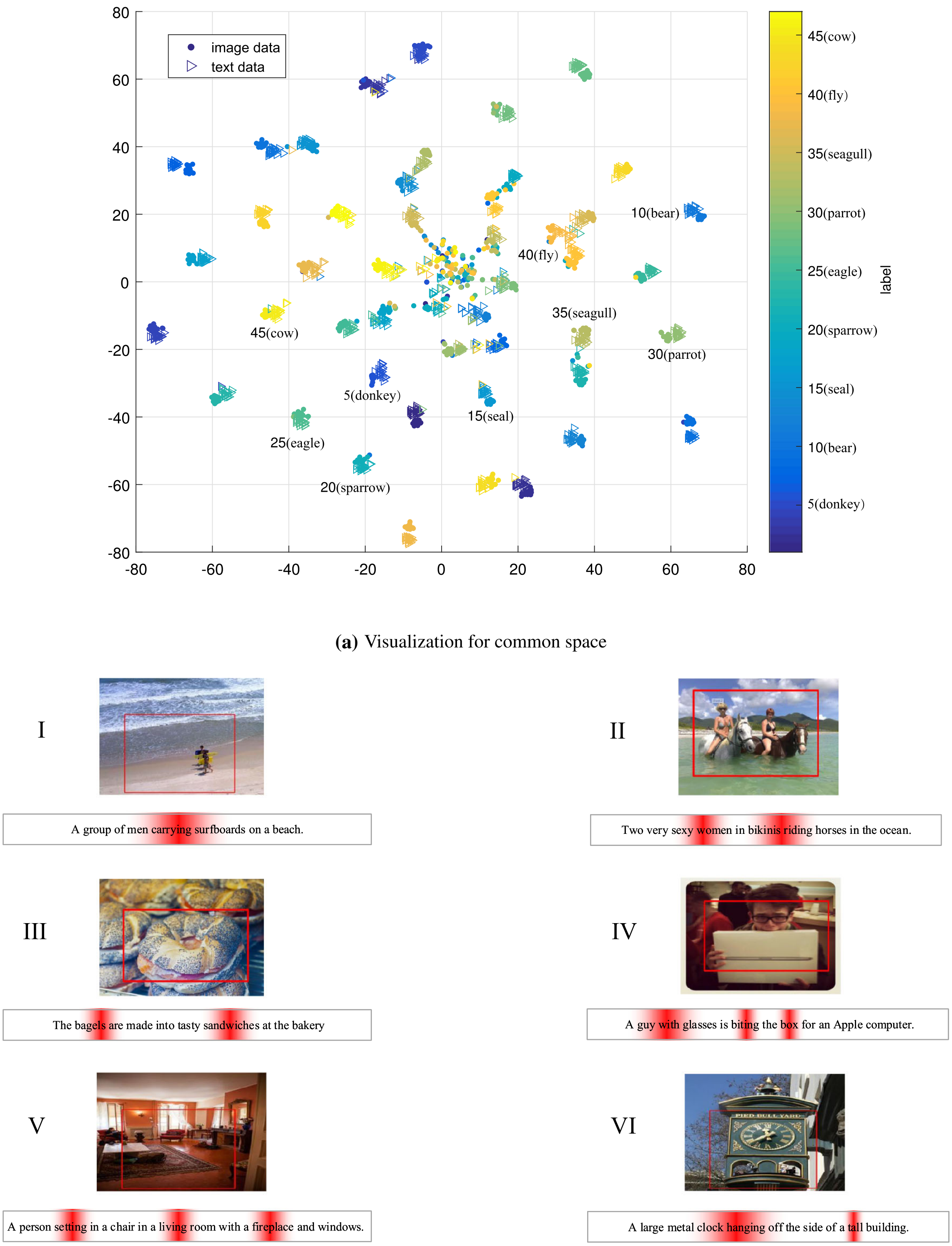
A desk with sticky notes and a computer on it A computer desktop sitting ...

Two bowls of cut up bananas and one whole banana make a smiley face. A bowl of ...

A stack of pancakes with bananas, walnuts, whipped cream and syrup. Pancakes...

Fig. 6 Retrieval results of our proposed VLKD approach. In these examples, all the correct retrieval results are marked with green boxes. We can see that our proposed approach achieves effective life-long cross-modal retrieval among 4 domains

- **CCA** (Rasiwasia et al. 2010) projects image and text into one common space by project matrices, which maximize the correlation between their projected features.
- **CFA** (Li et al. 2003) minimizes the Frobenius norm and projects the cross-modal data into one common space.
- **KCCA** (Hardoon et al. 2004) extends CCA by adopting kernel functions to project features of different modalities into a higher-dimensional space before projecting them into one common space. In the experiments, we use Gaussian kernel as the kernel function.
- **JRL** (Zhai et al. 2014) utilizes semi-supervised regularization as well as sparse regularization to learn the common space with semantic information.
- **LGCFL** (Kang et al. 2015) takes advantage of popular block based features to learn correlation between the data of different modalities, and jointly learns basis matrices of different modalities.
- **Corr-AE** (Feng et al. 2014) introduces the reconstruction error and correlation loss, constructing two subnetworks linked at the code layer. Note that Corr-AE has two extensions, and we compare with their best results in the experiments.
- **DCCA** (Yan and Mikolajczyk 2015) combines CCA with deep neural network, which aims to maximize the correlation between two separate subnetworks.
- **CMDN** (Peng et al. 2016a) constructs multiple deep networks to learn common representation with two-stage training strategy.
- **Deep-SM** (Wei et al. 2017) conducts deep semantic matching by utilizing the representation learning ability of convolutional neural network for image.



(a) Visualization for common space

I A group of men carrying surfboards on a beach.

II Two very sexy women in bikinis riding horses in the ocean.

III The bagels are made into tasty sandwiches at the bakery

IV A guy with glasses is biting the box for an Apple computer.

V A person setting in a chair in a living room with a fireplace and windows.

VI A large metal clock hanging off the side of a tall building.

(b) Visualization for attention mechanism

Fig. 7 Visualization results of our proposed VLKD approach for both common space learning and attention modeling. a We visualize the distribution of learned common representations on XMediaNet-animal by using t-SNE tool. b We visualize the learned attention information in pairwise image and text data, which highlights the discriminative parts with larger attention weights

Table 3 The MAP scores for cross-modal retrieval on **Wikipedia**

Method	Wikipedia		
	Img → Txt	Txt → Img	Average
VLKD(MS-COCO)	0.540	0.519	0.529
VLKD(Im2p)	0.535	0.511	0.523
VLKD-single	0.523	0.486	0.505
Our previous CmLL	0.518	0.462	0.490
CCL	0.505	0.457	0.481
ACMR	0.468	0.412	0.440
CMDN	0.487	0.427	0.457
Deep-SM	0.478	0.422	0.450
LGCFL	0.466	0.431	0.449
JRL	0.479	0.428	0.454
VLKD-noSe	0.451	0.432	0.442
DCCA	0.445	0.399	0.422
Corr-AE	0.442	0.429	0.436
KCCA	0.438	0.389	0.414
CFA	0.319	0.316	0.318
CCA	0.298	0.273	0.286

Note that **VLKD-single** is trained individually on the current domain, and **VLKD-noSe** is trained without semantic loss

Table 4 The recall scores for cross-modal retrieval on **MS-COCO**

Method	Img → Txt			Txt → Img		
	R@1	R@5	R@10	R@1	R@5	R@10
VLKD	0.230	0.523	0.642	0.228	0.549	0.673
VLKD-single	0.230	0.520	0.636	0.225	0.545	0.670
CCL	0.186	0.474	0.625	0.196	0.469	0.623
ACMR	0.209	0.478	0.603	0.185	0.437	0.580
DCCA	0.069	0.211	0.318	0.066	0.209	0.322
Corr-AE	0.154	0.397	0.532	0.138	0.353	0.478
KCCA	0.072	0.202	0.305	0.020	0.074	0.122
CFA	0.086	0.258	0.371	0.150	0.381	0.514
CCA	0.041	0.142	0.226	0.041	0.155	0.251
SCO*	0.428	0.723	0.830	0.331	0.629	0.755
PVSE*	0.452	0.743	0.845	0.324	0.630	0.750
SCAN*	0.464	0.774	0.872	0.344	0.637	0.757

Note that **VLKD-single** is trained individually on the current domain. The methods with * use a different evaluation protocol that focuses only on the image-sentence matching task

- **ACMR** (Wang et al. 2017) performs adversarial training to learn discriminative and modality-invariant common representations with a modality classifier and a feature projector.
- **CCL** (Peng et al. 2018a) adopts multi-grained fusion and multi-task learning with both intra-modality semantic category constraint and inter-modality pairwise similarity constraint, which fully explores both intra-modality and inter-modality correlation simultaneously.

Table 5 The recall scores of cross-modal retrieval on **Im2p**

Method	Img → Txt			Txt → Img		
	R@1	R@5	R@10	R@1	R@5	R@10
VLKD	0.151	0.398	0.538	0.136	0.374	0.529
VLKD-single	0.150	0.394	0.531	0.132	0.367	0.520
CCL	0.135	0.386	0.519	0.112	0.349	0.505
ACMR	0.119	0.364	0.498	0.101	0.337	0.482
DCCA	0.051	0.171	0.246	0.052	0.135	0.272
Corr-AE	0.106	0.333	0.424	0.093	0.254	0.378
KCCA	0.043	0.123	0.254	0.028	0.098	0.153
CFA	0.065	0.218	0.322	0.095	0.258	0.385
CCA	0.019	0.064	0.102	0.024	0.071	0.114

Note that **VLKD-single** is trained individually on the current domain

- **CmLL** (Qi et al. 2018) is our previous conference work, which performs cross-modal life-long learning with both intra-domain distribution alignment and inter-domain knowledge distillation.

To ensure a fair comparison, we replace the original input settings of the compared methods with the exact global representations used in our proposed approach, i.e., 4096-dimensional image features extracted from the fc7 layer in 19-layer VGGNet (Simonyan and Zisserman 2014) and 300-dimensional text features extracted from Word-CNN with the same configuration of Kim (2014). We directly rerun the source code of the compared methods released by their authors. Each compared method is evaluated on 4 domains respectively with the steps described above.

4.4 Comparisons with State-of-the-Art Methods

In this subsection, we introduce the visual-textual cross-modal retrieval performance of our proposed VLKD approach compared with 12 state-of-the-art methods. "VLKD (MSCOCO)" and "VLKD(Im2p)" represent different training paths as mentioned in Section 4.1. And we denote them as "VLKD" for brevity in the following part. To ensure a fair and comprehensive comparison with other methods, we report not only the results of our "VLKD" approach under life-long scenario, but also the results trained individually on the corresponding domain, denoted as "VLKD-single". The main difference between "VLKD-single" and "VLKD" lies in the training strategy, where VLKD-single is trained only with the common space learning strategy as proposed in Sect. 3.2.3, which addresses only intra-domain characteristics and cannot support life-long scenarios for continual learning across multiple domains as the full VLKD approach. We train multiple independent models for different domains separately to present the results for "VLKD-single". Tables 1, 2, 3, 4, and

Table 6 Comparisons with fine-tuning

Method	XMediaNet-artifact Average MAP score	XMediaNet-animal Average MAP score	Wikipedia Average MAP score	MS-COCO Average R@10 score
VLKD	0.704	0.915	0.529	0.658
Ft(XMN-animal)	0.447	0.907	–	–
Ft(Wiki)	0.255	0.722	0.516	–
Ft(MS-COCO)	0.141	0.706	0.505	0.642
VLKD-reverse	0.698	0.902	0.526	0.563

We assume that the 4 domains arrive in the following sequence: "**XMediaNet-artifact → XMediaNet-animal → Wikipedia → MS-COCO**". "Ft(XMN-animal)" denotes fine-tune for the XMediaNet-animal dataset without inter-domain knowledge distillation and attention-level knowledge transfer on the model trained by previous domain. Thus, the results on Wikipedia are not available in this situation. "Ft(Wiki)" denotes fine-tune for Wikipedia on the model trained by XMediaNet-artifact&animal. The fine-tuning process for "Ft(MS-COCO)" is similar. Furthermore, VLKD-reverse denotes that the 4 domains arrive in the reverse sequence: "**MS-COCO → XMediaNet-artifact → XMediaNet-animal → Wikipedia**'

5 show the results on 5 cross-modal domains, which present the retrieval accuracies of the two retrieval tasks. Detailed experimental analyses and discussions are given in the following paragraphs.

From the experimental results, we can see that both "VLKD" and "VLKD-single" show clear advantages over the other compared state-of-the-art methods, including our previous CmLL, on all 5 cross-modal domains. In particular, our proposed VLKD approach improves the average MAP score from 0.775 to 0.910, on the cross-modal domain of XMediaNet animal. This result is primarily due to the effective cross-modal context correlation modeling by the hierarchical recurrent network, and the knowledge distillation from both semantic and attention levels. Among all the compared methods, we can draw the following 2 observations: (1) Most of the deep learning based methods achieves better performance compared with traditional methods; our previous CmLL and CCL achieve two of the best retrieval accuracies among all the compared methods. This indicates that deep neural network have a greater ability to learn correlation modeling compared with traditional frameworks. (2) The traditional methods achieve better performance with CNN features, compared to their original results with handcrafted features. In fact, some traditional methods, such as LGCFL, obtain comparable results with deep learning based methods. It indicates the powerful representation ability of CNN features. We observe the similar trends in the cross-modal retrieval on the other domains, where our proposed VLKD approach retains its advantages. Furthermore, for intuitive comparison, we show some retrieval results on multiple cross-modal domains in Fig. 6. These results show that our proposed approach effectively achieves cross-modal retrieval from multiple domains under life-long scenarios.

By comparing the results of the "VLKD" and "VLKD-single", the proposed life-long learning mechanism boosts model performance effectively on XMediaNet-artifact, XMediaNet-animal and Wikipedia, but it only slightly improves the performance on MS-COCO and Im2p. The performances of VLKD and VLKD-single on MS-COCO and Im2p are almost the same, for the following reasons: (1) The image-sentence matching task in MS-COCO and Im2p is quite different from the other domains, because the goal is to find exactly matching between image and sentence. Thus, the high-level semantic knowledge from other domains provide little contribution to the image-sentence matching task. (2) The inter-domain information from the other domains is relatively less than intra-domain information in MS-COCO and Im2p due to the training data scale, which leads to a relatively limited improvement of inter-domain knowledge transfer.

We also report some recent methods that address the image-sentence matching task of the COCO dataset in Table 4, such as SCO (Huang et al. 2018b), PVSE (Song and Soleymani 2019) and SCAN (Lee et al. 2018). These methods adopt complicated mechanism to obtain detailed information for the image-sentence matching task and achieve better performance. In contrast, our approach focuses primarily on cross-modal life-long learning and transferring high-level semantic category knowledge. The proposed architecture is designed for semantic category based cross-modal retrieval, which aims to return the same category that have content similar to the query instance, not to find an exact content match between the query and the results.

In addition, we visualize some mapping results to intuitively demonstrate the effectiveness of our proposed approach. We visualize the results from 2 aspects, as shown in Fig. 7: (1) *Visualization of common space*: We first visualize the distribution of the learned common representations on XMediaNet-animal using t-SNE tool. The figure shows that visual and textual data from the same category are intermixed in the common semantic space for cross-modal retrieval. It can verify the effectiveness of common space learning to maintain the cross-modal semantic consistency. Some points (in different colors) are distributed uniformly in the area around 0. It indicates some difficult samples for representation learning from different categories, which cannot be categorized into any single cluster by t-SNE. (2) *Visualization for attention mechanism*: We also visualize the learned attention information in pairwise image and text data. The highlighted image region and textual phrase indicate the discriminative parts that have larger attention weights. Specifically, we selected the region with the highest attention value and highlight the words with relatively higher value in different locations. The attention mechanism in our proposed approach can focus on the main object in most cases, as shown in Fig. 7b(I–IV). For example, in Fig. 7b(III), it focuses on the word "sandwich" and the image region with corresponding object in red box. In Fig. 7b(IV), the textual attention focuses on "guy with glasses", "biting" and "box", which can accurately describe the image region in red box. These samples can verify the effectiveness of cross-modal attention modeling to learn discriminative fine-grained information. However, a few failure cases also exist, as shown in Fig. 7b(V–VI). In Fig. 7b(V), the image attention fails to focus on the specific object, due to the object is much smaller than the receptive field of Pool5 features. In Fig. 7b(VI), the textual attention ignores the "building", because the "clock" is much more significant in both the image and text.

Then, we present some in-depth experimental analyses from the following 2 aspects. First, among the traditional methods, CCA has the worst accuracy because it models only statistical values between the different modalities. KCCA and CFA outperform CCA due to their use of the kernel function and Frobenius norm to construct common space for cross-modal data. JRL constructs separate graphs for different modalities to perform correlation learning, while LGCFL has similar accuracy to JRL but adopts local group based priori to learn the basis matrices, which makes these models to outperform the other compared traditional methods. Although the performance of the traditional methods is promoted by the deep features, our proposed VLKD approach achieves better accuracies compared with them, mainly due to its hierarchical network architecture that fully exploits the complex cross-modal correlations. The performance of traditional framework is limited by with projection matrix learning.

Second, among the deep learning based methods, most of the compared methods have the network architecture with two pathways, such as DCCA, Corr-AE and Deep-SM. DCCA and Corr-AE mainly maximize the correlation between the two separate networks, which makes them have similar accuracies. Deep-SM utilizes the strong representation learning ability of the convolutional network and integrates semantic category information to achieve better performances than the above two methods. CMDN, ACMR and CCL consider both inter-modality and intra-modality modeling to obtain higher retrieval accuracies. CMDN constructs a hierarchical network architecture, and ACMR introduces adversarial learning into cross-modal correlation modeling. CCL outperforms both ACMR and CMDN due to its integration of multigrained and multitask modeling. Finally, CmLL achieves the best performance because it performs knowledge transfer across the different cross-modal domains under life-long scenarios.

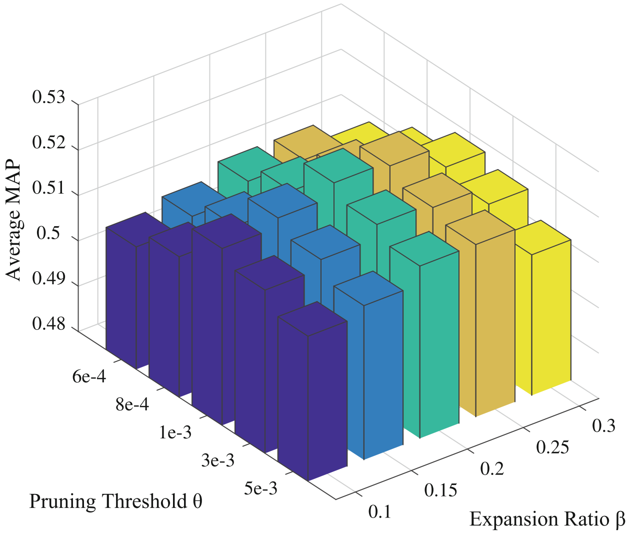


Fig. 8 The experimental results for different parameters in adaptive network expansion on the domain of Wikipedia, including expansion ratio β and pruning threshold θ

Compared with the state-of-the-art methods and our previous CmLL, our proposed VLKD approach achieves promising improvement for the following 3 reasons: (1) Hierarchical recurrent modeling effectively models the fine-grained context information, shares knowledge across the different modalities, and expands the network capacity adaptively to absorb knowledge from new domains. (2) Semantic-level knowledge distillation not only aligns the semantic distribution between different modalities, but also transfers semantic knowledge across different domains to boost correlation performance. (3) Attention-level knowledge transfer utilizes both intra-modality and inter-modality attention information to enhance discriminative fine-grained correlation learning.

4.5 Baseline Experimental Comparisons

In this section, we conduct comprehensive baseline experiments as follows: (1) Effectiveness experiment with the object category labels in MS-COCO dataset. (2) Comparisons with different training paths. (3) Comparison with conventional the fine-tuning strategy. (4) Parameter experiments to investigate the effects of the important parameters in the loss functions. (5) Ablation studies to verify the effectiveness of each component in our proposed VLKD approach.

4.5.1 The Effectiveness of the Object Category Labels

We conduct "VLKD-label" using the object category labels in MS-COCO dataset with semantic loss. The results are shown in Table 10. Although 80 object category labels are provided in MS-COCO dataset for images, they describe only high-level semantics, which cannot distinguish the status of image. Thus, these object category labels have little contribution to the matching task in MS-COCO dataset, which aims to find an exact match between an image and a sentence.

4.5.2 Comparison of Different Training Paths

We also conduct life-long learning experiments following the training path of "**MS-COCO** → **XMediaNet-artifact** → **XMediaNet-animal** → **Wikipedia**" (VLKD-reverse), which is trained using the object category labels of MS-COCO for semantic knowledge distillation during life-long learning. The results are reported in Table 6. When starting from MS-COCO dataset, the approach achieves better performance than that of a single baseline in PKU XMediaNet and Wikipedia domains. However. MS-COCO suffers performance losses after training on the other three domains. The result occurs because the task in MS-COCO is considerably different from those of the other domains. In MS-COCO, the task is to perform image-sentence matching, by finding an exact match between an image and a sentence. Thus, when training on PKU XMediaNet and Wikipedia, the model focuses on the high-level semantic knowledge instead of on the exact match required by MS-COCO, which affects the performance on MS-COCO. However, our approach also benefits from the knowledge gleaned from MS-COCO, which improves the performance on PKU XMediaNet and Wikipedia domains.

4.5.3 Comparison of Fine-Tuning Strategies

We first compare our proposed cross-modal life-long learning strategy with directly fine-tuning the model with data from the new domain without inter-domain knowledge distillation in Eq. (8) and attention-level knowledge transfer, while remaining the same architecture. We fine-tune the model using the training path as "XMediaNet-artifact → XMediaNet-animal → Wikipedia → MS-COCO" under the assumption that the 4 cross-modal domains arrive in sequence. First, the data from XMediaNet-animal is used to fine-tune the model trained on the old XMediaNet-artifact domain, denoted as "Ft(XMN-animal)". Under this situation, the results on Wikipedia are not yet available. Then we further fine-tuned with the Wikipedia dataset using the model trained on both XMediaNet-artifact and XMediaNet-animal, denoted as "Ft(Wiki)". Finally, we fine-tuned the model on MS-COCO, denoted as "Ft(MS-COCO)".

As shown in Table 6, the model directly fine-tuned without conducting regularizations to preserve the knowledge of the old domains suffers from severe forgetting. Thus, our proposed approach is effective at learning the cross-modal correlations in new domains while preserving the knowledge of the old domains. Additionally, the performance of our approach on the new domain outperforms the fine-tuning strategy, which indicates that transferring the knowledge in aspect of semantic and attention among domains can mutually boost the correlation ability.

Table 7 Baseline comparisons for each component in our proposed VLKD approach on **XMediaNet-artifact**

Method	XMediaNet-artifact		
	Img → Txt	Txt → Img	Average
VLKD(MS-COCO)	0.709	0.698	0.704
VLKD-global	0.682	0.680	0.681
VLKD-local	0.693	0.686	0.690
VLKD-noShare	0.699	0.693	0.696
VLKD-noKd	0.678	0.657	0.668
VLKD-noAtt	0.696	0.680	0.688
VLKD-noEx	0.694	0.692	0.693
VLKD-noAttEx	0.623	0.617	0.620
VLKD-noSe	0.467	0.504	0.486
VLKD-noJe	0.698	0.681	0.690
VLKD-noMMD	0.674	0.663	0.669
VLKD-noLSTM	0.687	0.684	0.686

4.5.4 Parameter Experiments

We conduct parameter sensitivity experiments to evaluate the impacts of the important parameters in the loss functions, including the expansion ratio β and the pruning threshold θ for adaptive network expansion.

The experimental results for the parameters in the adaptive network expansion are shown in Fig. 8. These parameter values are evaluated under the life-long scenario on the last-arriving cross-modal domain, namely, Wikipedia. The results show that on the one hand, the retrieval accuracies diminishes when expansion ratio β and the pruning threshold θ have small values, because small values add little additional capacity to the network. This result indicates that without sufficient expansion, the life-long model cannot fully absorb the knowledge from the new cross-modal domain and achieve effective correlation learning. On the other hand, the retrieval accuracies also decrease when both the expansion ratio β and the pruning threshold θ become larger, because large values add too much capacity and cause more redundancy in the network, which has an adverse effect on the performance. Therefore, it is necessary to balance these two parameters to achieve effective network expansion.

Table 8 Baseline comparisons for each component in our proposed VLKD approach on **XMediaNet-animal**

Method	XMediaNet-animal		
	Img→Txt	Txt→Img	Average
VLKD(MS-COCO)	0.913	0.917	0.915
VLKD-global	0.853	0.845	0.849
VLKD-local	0.864	0.864	0.864
VLKD-noShare	0.870	0.862	0.866
VLKD-noKd	0.841	0.830	0.835
VLKD-noAtt	0.864	0.861	0.862
VLKD-noEx	0.867	0.863	0.865
VLKD-noAttEx	0.824	0.813	0.819
VLKD-noSe	0.775	0.763	0.769
VLKD-noMMD	0.871	0.865	0.868
VLKD-noJe	0.861	0.848	0.855
VLKD-noLSTM	0.861	0.859	0.860

Table 9 Baseline comparisons for each component in our proposed VLKD approach on **Wikipedia**

Method	Wikipedia		
	Img→Txt	Txt→Img	Average
VLKD(MS-COCO)	0.540	0.519	0.529
VLKD-global	0.517	0.502	0.509
VLKD-local	0.524	0.502	0.513
VLKD-noShare	0.533	0.512	0.522
VLKD-noKd	0.519	0.499	0.509
VLKD-noAtt	0.521	0.501	0.511
VLKD-noEx	0.531	0.506	0.518
VLKD-noAttEx	0.504	0.501	0.503
VLKD-noSe	0.451	0.432	0.442
VLKD-noMMD	0.523	0.512	0.518
VLKD-noJe	0.510	0.504	0.507
VLKD-noLSTM	0.514	0.501	0.508

4.5.5 Ablation Studies

We further conduct comprehensive ablation studies to verify the effectiveness of each component in our proposed VLKD approach based on VLKD(MS-COCO). The experimental results are shown in Tables 7, 8, 9, and 10, which can be divided into the following 6 aspects:

(1) *Separate performances for different pathways in hierarchical recurrent network* We evaluate the performance when using only the global representations of image and text from the convolutional network, denoted as "VLKD-global", and when using only the fine-grained local representations from two levels of the recurrent networks, denoted as "VLKD-local". The experimental results show that "VLKD-local" slightly outperforms "VLKD-global", which indicates that local level modeling exploits more fine-grained information within the image and text to promote cross-modal correlation learning. "VLKD-global" and "VLKD-local" can be mutually boosted to further improve the accuracies of cross-modal retrieval. In addition, we also evaluate the effect of the weight-sharing strategy in the second level of the recurrent network, denoted as "VLKD-noShare". The experimental results indicate that the weight-sharing strategy effectively boosts the retrieval accuracy by sharing contextual knowledge between the different modalities to exploit the cross-modal context correlation.

(2) *Effectiveness of each loss term for common space learning* We conduct ablation studies, including "VLKD-noSe" (without the semantic loss in Eq. (6)), "VLKD-noMMD" (without the MMD criterion in Eq. (4)) and "VLKD-noJe" (without the ranking loss in Eq. (5)) to evaluate the effectiveness of each loss term for common space learning. The results are shown in Tables 7, 8, 9, and 10, where we can observe that each of them contributes to the final results, and they mutually boost the performance of cross-modal retrieval. In addition, ranking loss is relatively more important than MMD loss, due to the ranking information is more important for retrieval.

(3) *Effectiveness of semantic-level knowledge distillation* Note that inter-domain knowledge distillation is essential for life-long learning, which cannot be dropped. Thus, we remove only the other term of intra-domain distribution alignment for experimental evaluation, denoted as "VLKD-noKd". The experimental results verify that the semantic distribution alignment and relative semantic similarity ranking information make effective contributions to the final retrieval performance.

(4) *Effectiveness of attention-level knowledge transfer* We also drop the two levels of attention layers, including both intra-modality and inter-modality attention models, as well as the attention transfer learning, denoted as "VLKD-noAtt". By comparing the full model with the results of "VLKD-noAtt", we can find that the improvement from the attention mechanism is relatively small. This result is due to the complex textual content from Wikipedia, which introduces huge challenges for joint attention modeling and transference between different modalities and different domains. We also note that the attention mechanism still consistently provides improvements of approximately 1.6–1.9%, because of the effective modeling of attention information from both intra-modality and inter-modality perspectives. Attention knowledge transfer also promote the discriminative fine-grained information learning in the new cross-modal domain, which results in higher retrieval accuracies.

(5) *Effectiveness of adaptive network expansion* To verify the effectiveness of network expansion, we fixed the capacity of the hierarchical recurrent network, denoted as "VLKD-noEx". From the comparison results in Fig. 8, we can see

Table 10 Baseline comparisons for each component in our proposed VLKD approach on **MS-COCO**

Method	Img → Txt			Txt→Img		
	R@1	R@5	R@10	R@1	R@5	R@10
VLKD	0.230	0.523	0.642	0.228	0.549	0.673
VLKD-label	0.236	0.527	0.645	0.220	0.499	0.641
VLKD-global	0.210	0.492	0.617	0.207	0.512	0.638
VLKD-local	0.215	0.509	0.623	0.216	0.526	0.654
VLKD-noShare	0.226	0.520	0.638	0.222	0.541	0.666
VLKD-noKd	0.218	0.511	0.624	0.225	0.536	0.659
VLKD-noAtt	0.222	0.519	0.630	0.223	0.532	0.656
VLKD-noEx	0.225	0.521	0.636	0.222	0.533	0.661
VLKD-noAttEx	0.220	0.510	0.621	0.215	0.512	0.647
VLKD-noMMD	0.226	0.519	0.632	0.217	0.537	0.665
VLKD-noJe	0.212	0.502	0.616	0.203	0.514	0.648
VLKD-noLSTM	0.215	0.507	0.623	0.207	0.526	0.658

Table 11 The MAP scores of XMediaNet-artifact domain on different life-long learning stages

Task	Life-long learning stage		
	XMediaNet-artifact	XMediaNet-animal	Wikipedia
Img → Txt	0.667	0.740	0.741
Txt → Img	0.665	0.737	0.738

that the performance can be boosted by expanding the network capacity appropriately to improve the ability to absorb the knowledge from new cross-modal domain, as well as to eliminate redundancy for effective training.

In addition, we further conduct the ablation study "VLKD-noAttEx", which combines the settings of "VLKD-noAtt" and "VLKD-noEx". The results are shown in Tables 7, 8, 9, and 10. We can observe that the method "VLKD-noAttEx" achieves slightly higher accuracies than the previous CmLL due to the further optimization of network training. However, it is lower than both "VLKD-noAtt" and "VLKD-noEx", which indicates the effectiveness of the attention and adaptive network expansion strategies.

(6) *Performance at different learning stages* We conduct the online training experiments to record the cross-modal retrieval performances in a life-long scenario to verify the capability of life-long learning.

We adopt the first domain (XMediaNet-artifact) in the training path of cross-modal retrieval, and evaluate the performance on this domain at different life-long learning stages. As shown in Table 11, the retrieval performance remains constant or improves throughout the life-long learning process. The adoption of knowledge distillation and attention transfer in different domains mutually boosts the semantic distribution of the old domains for better cross-modal correlation learning.

Table 12 The MAP scores of single category queries at different life-long learning stages

Category	Life-long learning stage	
	XMediaNet-animal	Wikipedia
Bee	0.959	0.965
Mouse	0.959	0.961
Seagull	0.945	0.948
Goat	0.941	0.939
Duck	0.939	0.943
Turkey	0.937	0.941
Chipmunk	0.936	0.943
Woodpecker	0.931	0.944
Eagle	0.931	0.934
Magpie	0.916	0.944

We report Top-10 MAP scores (Img→Txt) of single category queries on XMediaNet-animal domain at different life-long learning stages

To verify the influence of learning relevant categories, we choose two domains, namely XMediaNet-animal and Wikipedia, where Wikipedia contains the category "biology", which is similar to the animal categories in XMediaNet-animal. We compare the performance before and after learning similar concepts. As shown in Table 12, when we learn the new concept "biology" in Wikipedia domain, most of the categories in XMediaNet-animal show increased performance during the life-long learning process, which demonstrates the effectiveness of our approach.

From the above baseline experimental comparisons, we can conclude that our proposed VLKD approach fully exploits the discriminative fine-grained context information using the hierarchical recurrent network and effectively performs cross-modal life-long correlation learning by leveraging the knowledge from both semantic and attention levels. In addition, it expands the network capacity adaptively to accommodate the life-long training process. Among all model components, the most critical is the semantic-level knowledge distillation. Semantic-level knowledge distillation attempts to maintain the semantic distribution of old domains when learning new domains, which is a key aspect of life-long learning because it preserves the original semantic knowledge in the old domains and further transfers knowledge across domains.

5 Conclusion

In this paper, we have proposed a visual-textual life-long knowledge distillation (VLKD) approach that leverages knowledge from both semantic and attention levels to preserve the original correlation ability in existing cross-modal domains, and achieve better performances in the new domain. Our conclusions are as follows: First, visual-textual hierarchical recurrent network can share knowledge at a high level to boost cross-modal context correlation modeling. Second, cross-domain semantic-level knowledge distillation can effectively match semantic distributions between different modalities to boost correlation learning. Third, cross-modal attention-level knowledge transfer enhances the discriminative fine-grained correlation learning from both intra-modality and inter-modality perspectives. Forth, life-long adaptive network expansion expands the network capacity and helps eliminate redundancy to absorb correlation knowledge from new domains.

In the future work, we plan to import an external knowledge base for knowledge transfer during life-long correlation learning, and further move toward unsupervised learning situation to enable practical cross-modal retrieval applications.

References

Akaho, S. (2006). A kernel method for canonical correlation analysis. arXiv preprint arXiv:cs/0609071

Aljundi, R., Chakravarty, P., & Tuytelaars, T. (2017). Expert gate: Life-long learning with a network of experts. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 7120–7129).

Andrew, G., Arora, R., Bilmes, J. A., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning (ICML)* (pp. 1247–1255).

Eisenschtat, A., & Wolf, L. (2017). Linking image and text with 2-way nets. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 1855–1865).

Feng, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *ACM conference on multimedia (ACM-MM)* (pp. 7–16).

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Conference on empirical methods in natural language processing (EMNLP)* (pp. 457–468).

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *Computer Science*, *84*(12), 1387–91.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. J. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*, 723–773.

Hardoon, D. R., Szedmák, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 770–778).

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Huang, Y., Wu, Q., Song, C., & Wang, L. (2018a). Learning semantic concepts and order for image and sentence matching. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 6163–6171).

Huang, Y., Wu, Q., & Wang, L. (2018b). Learning semantic concepts and order for image and sentence matching. In *Computer vision and pattern recognition (CVPR)*.

Kang, C., Xiang, S., Liao, S., Xu, C., & Pan, C. (2015). Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia (TMM)*, *17*(3), 370–381.

Karpathy, A., & Li, F. (2015). Deep visual-semantic alignments for generating image descriptions. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 3128–3137).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Conference on empirical methods in natural language processing (EMNLP)* (pp. 1746–1751).

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2016). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, *114*(13), 3521–3526.

Krause, J., Johnson, J., Krishna, R., & Fei-Fei, L. (2017). A hierarchical approach for generating descriptive image paragraphs. In *Computer vision and pattern recognition (CVPR)*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)* (pp. 1106–1114).

Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In *European conference on computer vision (ECCV)* (pp. 212–228).

Li, D., Dimitrova, N., Li, M., & Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In *ACM conference on multimedia (ACM-MM)* (pp. 604–611).

Li, Z., & Hoiem, D. (2018). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *40*(12), 2935–2947.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)* (pp. 740–755).

Lin, Y., Pang, Z., Wang, D., & Zhuang, Y. (2017). Task-driven visual saliency and attention-based visual question answering. arXiv preprint arXiv:1702.06700.

Mallya, A., & Lazebnik, S. (2018). Packnet: Adding multiple tasks to a single network by iterative pruning. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 7765–7773).

Mitchell, T. M., Cohen, W. W, Jr., E. R. H., Talukdar, P. P., Yang, B., Betteridge, J., et al. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103–115.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *International conference on machine learning (ICML)* (pp. 689–696).

Peng, Y., Huang, X., & Qi, J. (2016a). Cross-media shared representation by hierarchical learning with multiple deep networks. In *International joint conference on artificial intelligence (IJCAI)* (pp. 3846–3853).

Peng, Y., Zhai, X., Zhao, Y., & Huang, X. (2016b). Semi-supervised cross-media feature learning with unified patch graph regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, *26*(3), 583–596.

Peng, Y., Huang, X., & Zhao, Y. (2017). An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, *28*(9), 2372–2385.

Peng, Y., Qi, J., Huang, X., & Yuan, Y. (2018a). CCL: Cross-modal correlation learning with multi-grained fusion by hierarchical network. *IEEE Transactions on Multimedia (TMM)*, *20*(2), 405–420.

Peng, Y., Qi, J., & Yuan, Y. (2018b). Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing (TIP)*, *27*(11), 5585–5599.

Qi, J., Peng, Y., Zhuo, Y. (2018). Life-long cross-media correlation learning. In *ACM conference on multimedia (ACM-MM)*. ACM (pp. 528–536).

Ranjan, V., Rasiwasia, N., & Jawahar, C. V. (2015). Multi-label cross-modal retrieval. In *IEEE international conference on computer vision (ICCV)* (pp. 4094–4102).

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet. G. R., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *ACM conference on multimedia (ACM-MM)* (pp. 251–260).

Reed, S.E., Akata, Z., Lee, H., & Schiele, B. (2016). Learning deep representations of fine-grained visual descriptions. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 49–58).

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. arXiv preprint arXiv:1606.04671.

Shin, H., Lee, J. K., Kim, J., & Kim, J. (2017). Continual learning with deep generative replay. In *Advances in neural information processing systems (NeurIPS)* (pp. 2994–3003).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations (ICLR)*.

Song, Y., & Soleymani, M. (2019). Polysemous visual-semantic embedding for cross-modal retrieval. In *Computer vision and pattern recognition (CVPR)*.

Triki, A. R., Aljundi, R., Blaschko, M. B., & Tuytelaars, T. (2017). Encoder based lifelong learning. In *IEEE international conference on computer vision (ICCV)* (pp. 1329–1337).

Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Hengtao, S. (2017). Adversarial cross-modal retrieval. In *ACM conference on multimedia (ACM-MM)* (pp. 154–162).

Wang, K., He, R., Wang, L., Wang, W., & Tan, T. (2016a). Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *38*(10), 2010–2023.

Wang, L., Li, Y., & Lazebnik, S. (2016b). Learning deep structure-preserving image-text embeddings. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 5005–5013).

Wang, S., Chen, Y., Zhuo, J., Huang, Q., & Tian, Q. (2018). Joint global and co-attentive representation learning for image-sentence retrieval. In *ACM conference on multimedia (ACM-MM)* (pp. 1398–1406).

Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., et al. (2017). Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics (TCYB)*, *47*(2), 449–460.

Xu, J., & Zhu, Z. (2018). Reinforced continual learning. In *Advances in neural information processing systems (NeurIPS)* (pp. 907–916).

Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *Conference on computer vision and pattern recognition (CVPR)* (pp. 3441–3450).

Yoon, J., Yang, E., Lee, J., & Ju Hwang, S. (2017). Lifelong learning with dynamically expandable networks. arXiv preprint arXiv:1708.01547.

Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

Zenke, F., Poole, B., & Ganguli, S. (2017). Continual learning through synaptic intelligence. In *International conference on machine learning (ICML)* (pp. 3987–3995).

Zhai, X., Peng, Y., & Xiao, J. (2013). Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In *AAAI conference on artificial intelligence (AAAI)* (pp. 1198–1204).

Zhai, X., Peng, Y., & Xiao, J. (2014). Learning cross-media joint representation with sparse and semi-supervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, *24*, 965–978.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems (NeurIPS)* (pp. 649–657).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.