Learning with Adaptive Knowledge for Continual Image-Text Modeling

Yutian Luo
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
luoyutian2021@ruc.edu.cn

Yizhao Gao
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
gaoyizhao@ruc.edu.cn

Zhiwu Lu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
luzhiwu@ruc.edu.cn

ABSTRACT

In realistic application scenarios, existing methods for image-text modeling have limitations in dealing with data stream: training on all data needs too much computation/storage resources, and even the full access to previous data is invalid. In this work, we thus propose a new continual image-text modeling (CITM) setting that requires a model to be trained sequentially on a number of diverse image-text datasets. Although recent continual learning methods can be directly applied to the CITM setting, most of them only consider reusing part of previous data or aligning the output distributions of previous and new models, which is a partial or indirect way to acquire the old knowledge. In contrast, we propose a novel dynamic historical adaptation (DHA) method which can holistically and directly review the old knowledge from a historical model. Concretely, the historical model transfers its total parameters to the main/current model to utilize the holistic old knowledge. In turn, the main model dynamically transfers its parameters to the historical model at every five training steps to ensure that the knowledge gap between them is not too large. Extensive experiments show that our proposed DHA outperforms other representative/latest continual learning methods under the CITM setting.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**;

KEYWORDS

Image-text modeling, continual learning, contrastive learning, cross-modal retrieval

ACM Reference Format:
Yutian Luo, Yizhao Gao, and Zhiwu Lu. 2023. Learning with Adaptive Knowledge for Continual Image-Text Modeling . In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592297

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06…$15.00
https://doi.org/10.1145/3591106.3592297

Figure 1: (a) The results of catastrophic forgetting under the CITM setting. (b) The results of cross-dataset evaluation after independent training on four datasets.

1 INTRODUCTION

In the past few years, image-text modeling has drawn much attention from both academia and industry with a fundamental role in various cross-modal tasks, such as image-text retrieval [12, 24], image captioning [20, 44], and text-image generation [21, 33]. Although existing image-text modeling methods [5, 16, 18, 19, 25, 26, 30] have achieved great success in these tasks, most of them assume that a full (fixed) set of image-text pairs are provided for model training, which actually limits their deployment in realistic application scenarios. That is, the training data often comes in a stream way, and the current widely-used paradigm for image-text modeling faces two limitations: (1) training on all data (i.e., both previous and new data) severely increases the computational and storage overhead; (2) the full access to previous data may be invalid.

To overcome these limitations, we thus propose a continual image-text modeling (CITM) setting instead. Concretely, we recollect four diverse image-text datasets respectively from MSCOCO [28], CC3M [40], WIT [42] and GoodNews [6], each of which is split into the training, validation, and test sets. We adopt the SimCLR-based model [13] as the basic model which is also deployed in OpenAI CLIP [34]. Under the CITM setting, the model is sequentially trained on each of the four image-text datasets, and is finally evaluated on all datasets. To demonstrate the well-known catastrophic forgetting problem, we measure the image-to-text retrieval performance with the metric recall@1 (R@1) during sequential training on the four datasets. The results in Figure 1 (a) clearly show that every time the model is trained on a new dataset, its performance on previous datasets has a distinct degradation (i.e., catastrophic forgetting).

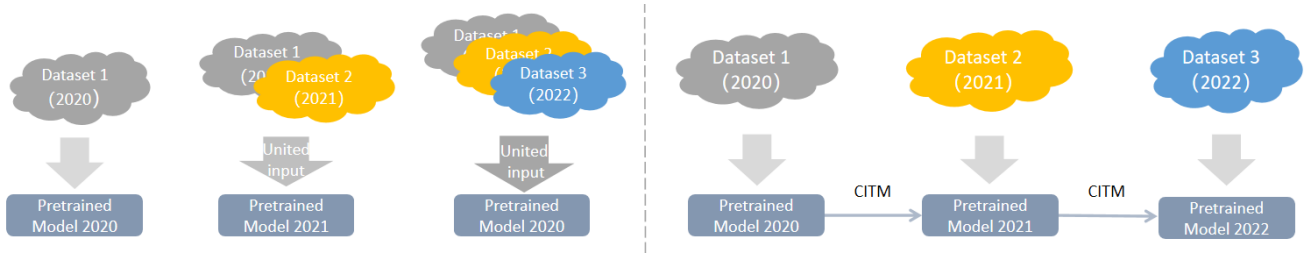

Figure 2: Schematic illustration of the realistic application of our proposed CITM setting in large-scale multi-modal pre-training (like OpenAI CLIP) with the pre-training data being updated every year. *Left:* The traditional setting for large-scale pre-training with annual data update. *Right:* Our CITM setting for large-scale pre-training with annual data update.

Among existing continual learning methods, rehearsal-based [7, 11] and regularization-based methods [8, 27, 36, 47] can be easily applied to the CITM setting, while architecture-related methods [31, 32, 38] generally need extra task-specific modules and are unsuitable for CITM with a unified architecture. In this paper, we thus devise baseline methods for CITM mainly by deploying rehearsal-based and regularization-based methods. Note that these two groups of continual learning methods have their own limitations. Specifically, rehearsal-based methods set up a memory buffer to replay previous data, and only preserve partial old knowledge due to the sample selection imposed on the memory buffer. Moreover, regularized-based methods can only convey the old knowledge by aligning the output distributions of the previous and new models, which indicates that the old knowledge from the previous model can only be indirectly transferred through data-driven guidance. Such an indirect approach is thus vulnerable to large domain shifts across the previous and new tasks.

To avoid the drawbacks of the above baseline methods for CITM, we thus propose a novel dynamic historical adaptation (DHA) method which can holistically and directly review the old knowledge from a historical model. The core idea of our DHA is to directly transfer knowledge between the old and new models through parameter interaction. In our DHA, we name the model trained on the current task as the main model, and the last (main) model on the previous task as the historical model. During parameter interaction, we directly transfer the parameters of the historical model to the main model and then train the main model with modified parameters on the current task. Meanwhile, we dynamically update the historical model with the guidance of the main model to ensure that the knowledge gap between them is not too large. Specifically, at every five steps, the parameters of the main model are passed to the historical model for parameter modification. Overall, these two parameter transfer strategies make up our DHA method. Compared with existing methods [7, 8, 11, 27], our proposed DHA method has two advantages: (1) DHA adopts direct parameter transfer instead of indirect model aligning (deployed by regularization-based methods), and thus it is more robust to large domain shifts across the previous and new tasks. (2) DHA holistically reviews the old knowledge from the historical model, which can overcome the drawback of rehearsal-based methods for partial data selection (i.e., only partial old knowledge is reused). To the best of our knowledge, we are the first to propose a direct parameter transfer method to cope with the forgetting problem in the continual learning field.

As we have mentioned, we construct a benchmark dataset for the CITM setting by recollecting four diverse image-text datasets respectively from MSCOCO [28], CC3M [40], WIT [42] and GoodNews [6]. Under a fair setting, we compare DHA with a number of baseline methods [7, 8, 11, 27] on this benchmark dataset. Extensive experiments prove that our DHA outperforms these baseline methods under the CITM setting.

Overall, the main contributions of this paper can be summarized as follows: (1) We propose a new continual image-text modeling (CITM) setting for image-text modeling on data stream, which has a realistic application in large-scale multi-modal pre-training (with annual data update) as shown in Figure 2. (2) We devise a novel dynamic historical adaptation (DHA) method under the CITM setting. For the first time, we identify the important role of direct parameter transfer (between the historical and main models) in continual learning. (3) We construct a benchmark dataset of four diverse sets of image-text pairs, which can facilitate the research on CITM. (4) Extensive experiments demonstrate the effectiveness of our DHA under the CITM setting.

2 RELATED WORK

Image-Text Modeling. Recent image-text modeling methods can be summarized into two groups: single-stream and two-stream methods. **(1)** Single-stream methods aim to learn the unified representation of the image-text pair with a fusion module. Most of existing single-stream methods [25, 26, 30, 43, 48] choose to concatenate the image and text embeddings as the input of the fusion module (e.g., cross-attention transformer). Although model training is easy for single-stream methods, it requires calculating the similarities of all the possible query-candidate pairs during inference. Therefore, they suffer from heavy computation burdens. **(2)** Two-stream methods [5, 16, 18, 19, 34] adopt independent image and text encoders to learn image and text embeddings that are aligned in a joint space. Compared to single-stream methods, two-stream methods allow different depths and designs of network architectures for the two modalities and enjoy much more efficient inference. In this work, we follow the two-stream architecture for image-text modeling: ResNet50 [17] is used as the image encoder, and BERT-base [15] is used as the text encoder. We adopt SimCLR [13] as the basic contrastive learning method for model training.

Continual Learning. By reviewing recent progress in conventional continual learning, we can divide main-stream approaches

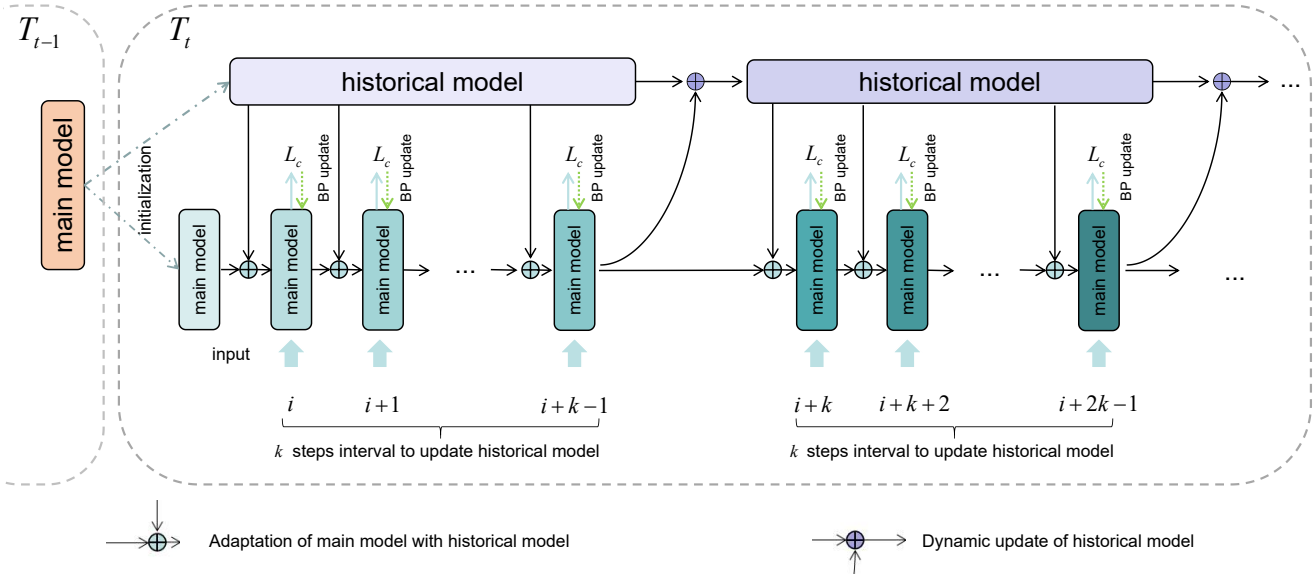

Figure 3: Overview of the proposed DHA method for the CITM setting. At the beginning of task T_t, the last main model in task T_{t-1} is used to initialize both the historical and main models. During each training iteration, the main model first receives the transferred parameters from the historical model and then learns on the dataset of task T_t. Moreover, for every k training iteration, the historical model is updated with the transferred parameters of the main model.

into three groups: **(1) Rehearsal-Based Methods.** Early classic rehearsal-based method [37] proposes to store part of exemplars of previous classes in order to acquire better class means. [11] finds that retraining a subset of old data on new tasks can help address the forgetting problem and also provides several memory update strategies. [3, 7, 10] further explore the approaches to selecting representative samples from old tasks. In addition, pseudo-data rehearsal generating approaches [4, 23, 29, 35, 41] are proposed to avoid extra storage and generate more representative samples for training, whereas generating pseudo-data actually increases the training time. Note that the rehearsal-based methods suffer from the drawback that only partial historical knowledge is transferred by the memory buffer. **(2) Regularization-Based Methods.** This group of methods mainly aim to distill the knowledge of the previous models. [27, 36, 47] align the output features or logits between the previous and the current models with an extra regularization penalty. Since the domain shifts exist across different tasks, such regularization penalty brings additional training difficulty [14]. Other methods [1, 9, 22] constrain part of the parameters of the model. Since most of these methods are designed for classification tasks, they are hard to be directly applied to our CITM setting. **(3) Architecture-Related Methods.** This group of methods mitigate the difference in new tasks in two ways. [31, 32, 39] mask different parameters while training different tasks. [2, 38, 45] extend network architecture for new tasks. A potential drawback of these methods is that they generally need extra task-specific modules and are unsuitable for CITM with a unified architecture. Other than the above approaches with a single strategy, recent works [7, 8] start to design combined strategies for continual learning based on rehearsal-based and regularization-based methods. Finally, we notice that most of existing continual learning approaches have a common characteristic that the old knowledge is expressed with (partial) data, which means that the model update to mitigate forgetting may be affected by partial/indirect guidance. In contrast, our proposed DHA provides a new perspective of continual learning that the old knowledge could be holistically preserved by direct parameter transfer.

3 PROPOSED METHOD

3.1 Preliminary

We first define our proposed CITM setting formally. Given a sequence of n image-text datasets $\mathcal{D} = \{D_1, D_2, ..., D_n\}$ coming from n domain sources like a stream, a model for CITM is supposed to be sequentially trained on $\mathcal{D}$. Each dataset D_t $(1 \leq t \leq n)$ is defined as $D_t = \{(x_i^I, x_i^T)\}_{i=1}^{N_t}$, where x_i^I and x_i^T respectively denote the image and text samples in the i-th image-text pair, and N_t denotes the number of data pairs. The image-text retrieval task [12, 24] on each dataset D_t is denoted as T_t $(1 \leq t \leq n)$. For each task T_t, a model for image-text retrieval typically learns to align the image and text embeddings with contrastive loss [13]. Under the CITM setting, the model only concentrates on the current task during sequential training, leading to the catastrophic forgetting of previous knowledge. For performance evaluation, the obtained final model (trained across all tasks) is tested on each of the n tasks.

3.2 Network Architecture

Under the CITM setting, we propose a novel dynamic historical adaptation (DHA) method which can holistically and directly review the old knowledge from a historical model. The core idea of

our DHA is to directly transfer knowledge between the old and new models through parameter interaction. To this end, our DHA model is devised to have two key components: the historical model and the main model, which are illustrated in Figure 3, respectively. These two models share the same architecture while only the main model requires the backward update. We follow the two-stream network architecture like CLIP [34], which has achieved remarkable performance in image-text retrieval tasks. Concretely, the image encoder takes ResNet50 as the backbone and the text encoder takes BERT-Base as the backbone, which are both initialized with unimodal pre-trained models.

Image and Text Encoders. Formally, the backbone ResNet50 of the image encoder is denoted as f^I_{ResNet}. Meanwhile, the backbone BERT-Base of the text encoder is denoted as f^T_{Bert}. Given an input text x^T_i, we first tokenize it into a sequence as $[tk^1_i, tk^2_i, ..., tk^{l_i}_i]$, where l_i denotes the length of x^T_i. To ensure that the text and image embeddings have the same dimension, we append linear projection layers f^I_P and f^T_P to the image encoder ResNet50 and the text encoder BERT-Base, respectively. Given an image-text pair (x^I_i, x^T_i), the final image and text embeddings are given by:

$$e^I_i = f^I_P(f^I_{ResNet}(x^I_i)), \tag{1}$$

$$e^T_i = f^T_P(f^T_{Bert}(tk^1_i, tk^2_i, ..., tk^{l_i}_i)). \tag{2}$$

Contrastive Loss Function. Since our DHA has the two-stream architecture, it can be effectively trained by the well-known contrastive learning method SimCLR [13]. Concretely, given a batch of B image-text pairs $\{x^I_i, x^T_i\}^B_{i=1}$ during training, the loss function is constructed as follows. For each input image x^I_i, we define the contrastive loss between its image embedding e^I_i and the embeddings of all positive/negative texts in the batch as an InfoNCE loss:

$$L^{i2t}_c = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(e^I_i \cdot e^T_i/\tau)}{exp(e^I_i \cdot e^T_i/\tau) + \sum_{j\neq i} exp(e^I_i \cdot e^T_j/\tau)}, \tag{3}$$

where τ denotes the temperature hyperparameter, and the vector similarity is measured by dot product $(\cdot)$. Similarly, for each input text x^T_i, the InfoNCE loss is given by:

$$L^{t2i}_c = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{exp(e^I_i \cdot e^T_i/\tau)}{exp(e^I_i \cdot e^T_i/\tau) + \sum_{j\neq i} exp(e^I_j \cdot e^T_i/\tau)}. \tag{4}$$

The total contrastive loss for training our DHA is thus defined as:

$$L_c = L^{i2t}_c + L^{t2i}_c. \tag{5}$$

In this work, for fair comparison, all the competitors for CITM adopt the same network architecture and the same basic contrastive loss function as our DHA. More details can be found in Sec. 4.

3.3 Dynamic Historical Adaptation

As we have mentioned, our motivation of method design is to transfer the holistic knowledge contained in the historical model to the new model, without suffering from the drawbacks of existing continual learning methods. Concretely, data rehearsal approaches [7, 11] attempt to preserve the previous data distribution, but retaining a memory buffer of limited size may cause the overfitting to the partial samples of the previous task. Moreover, regularization-based approaches [8, 27, 36, 47] attempt to store the historical knowledge by aligning the historical and main models with regularization-based penalty terms, but such an indirect way to addressing the forgetting problem is thus vulnerable to large domain shifts across the previous and new tasks. In this paper, we thus propose a novel dynamic historical adaptation (DHA) method which can holistically and directly review the old knowledge from a historical model. Below we introduce the details of the two update strategies applied in our DHA throughout training.

Adaptation of Main Model with Historical Model: Since all the learned knowledge has been held and expressed by model parameters, we believe that directly transferring the parameters of the historical model to the main model is a direct and effective approach to preserving the historical knowledge. The direct parameter transfer process is shown in Figure 3. Formally, let θ_H, θ_M, θ^*_M denote the parameters of the historical model, the main model, and the last main model in the previous task, respectively. Moreover, let θ^i_H and θ^i_M denote the parameters of the historical model and the main model at the end of the i-th training iteration in the current task, respectively. At the beginning of the current task, we initialize the main model and the historical model with the parameters of the last main model in the previous task (i.e., $\theta^0_M = \theta^*_M$ and $\theta^0_H = \theta^*_M$). For each training iteration ($i \geq 1$) before data load, we choose to update θ^{i-1}_M with part of θ^{i-1}_H and obtain $\theta^{i_{pre}}_M$ as the new intermediate parameters of the main model. After such parameter update, the main model is trained on the input data and backward updated normally to obtain θ^i_M as the final parameters of the i-th training iteration. We define the gradient function w.r.t. θ_M as:

$$G_{L_c}(\hat{\theta}_M) = \left.\frac{\partial L_c}{\partial \theta_M}\right|_{\theta_M=\hat{\theta}_M}. \tag{6}$$

The above adaptation strategy for the main model with the historical model can be formulated as:

$$\theta^{i_{pre}}_M = \lambda_1\theta^{i-1}_M + (1-\lambda_1)\theta^{i-1}_H, \tag{7}$$

$$\theta^i_M = \theta^{i_{pre}}_M - \eta G_{L_c}(\theta^{i_{pre}}_M), \tag{8}$$

where η denotes the learning rate, and λ_1 denotes the weighting coefficient. By combining Eq. (7) and Eq. (8), we have the adaptation process from θ^{i-1}_M to θ^i_M as follows:

$$\theta^i_M = \lambda_1\theta^{i-1}_M + (1-\lambda_1)\theta^{i-1}_H - \eta G_{L_c}(\lambda_1\theta^{i-1}_M + (1-\lambda_1)\theta^{i-1}_H). \tag{9}$$

Dynamic Update of Historical Model: Currently, the main model has received the guidance from the historical model. However, since the parameters of the historical model remain static in the current task, this may cause two concerns: (1) Since the main model always learns better on the current task as the training process goes on, the knowledge gap between the historical and main models is gradually enlarged. Therefore, the parameters transferred from the unchanged historical model tend to cause degradation to the retrieval performance of the main model on the current task. (2) Such performance degradation to the main model on the current task would finally affect the performance of the final model (i.e. the last main model across all tasks) when it is evaluated on this task. To address these concerns, we choose to make the parameters of the historical model gradually change by updating it with the

Algorithm 1: Sequential Training with DHA

Input: the dataset for sequential tasks $\{D_i\}_{i=1}^{T}$
the main model with parameters θ_{main}
the historical model with parameters θ_{hist}
the last model of the last task with parameters θ_{last}
max iterations i_{max} in each task
hyperparameters k, λ_1, λ_2
Output: the learned θ^*_{main}
initialize θ_{main} by training the main model on D_1;
initialize $\theta_{last} \leftarrow \theta_{main}$;
for $D_t \in \{D_2, ..., D_T\}$ **do**
  initialize $\theta_{hist} \leftarrow \theta_{last}$, $\theta_{main} \leftarrow \theta_{last}$;
  for $i \leftarrow 1$ *to* i_{max} **do**
    if $i\%k = 0$ **then**
      $\theta^i_{hist} \leftarrow \lambda_2\theta^{i-1}_{hist} + (1-\lambda_2)\theta^{i-1}_{main}$
    else
      $\theta^i_{hist} \leftarrow \theta^{i-1}_{hist}$
    end
    update θ^i_{main} according to Eq. (9);
  end
  obtain θ_{main} $\theta_{last} \leftarrow \theta_{main}$;
end
return the θ^*_{main};

parameters of the main model (but not so frequently). This dynamic update of the historical model at the i-th iteration is given by:

$$\theta^i_H = \begin{cases} \lambda_2\theta^{i-1}_H + (1-\lambda_2)\theta^{i-1}_M, & \text{if } i = mk,\ m \in \mathbb{N} \\ \theta^{i-1}_H, & \text{otherwise} \end{cases}, \tag{10}$$

where k denotes the step interval for model update, and λ_2 denotes the weighting coefficient.

Overall, our proposed DHA is composed of the above two update strategies, which have been shown to be effective in Sec. 4.3 and Sec. 4.4. We believe that direct parameter transfer is another promising way to handling the continual learning problem in image-text modeling. The pseudocode of the full algorithm for our proposed DHA is given in Algorithm 1.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets. To mimic the realistic application of the CITM setting in multi-modal pre-training (like OpenAI CLIP), we recollect four image-text datasets for benchmark construction from the following large diverse datasets of image-text pairs: **(1) MSCOCO [28]** is an image-text dataset that consists of 123, 287 images with their captions. Each image is annotated with 5 captions. Most images are related to the nature and common objects in daily life. **(2) CC3M [40]** is a well-known image-captioning dataset for image-text pre-training. It is composed of about $3M$ image-text pairs, which are collected from the Internet with weak relation between images and their textual descriptions. **(3) WIT [42]** is a large multimodal multilingual dataset collected from the Wikipedia website. This dataset has a total of $11.5M$ images. Each image is annotated with the corresponding textual description or contextual information. **(4) GoodNews [6]** is a large news image-captioning dataset. It is collected from the New York Times. Unlike the other datasets, the captions in GoodNews are written by professional journalists and are thus claimed to have implications for the style and richness of the news. In this paper, based on the aforementioned four image-text datasets, our benchmark dataset of four sequential tasks is constructed as follows: (1) For task T_1, we randomly select 100, 000 images with corresponding captions from MSCOCO as the training set, 13, 287 as the validation set, and 5, 000 as the test set. (2) For the other tasks T_2 – T_4, we construct the three task-specific datasets from CC3M, WIT, and GoodNews, respectively. Concretely, the training/validation/test set is formed to have 130, 000/13, 000/5, 000 image-text pairs uniformly for each task of T_2 – T_4.

To directly demonstrate the domain shift across the four datasets used in our CITM setting, we conduct cross-dataset evaluation experiments. Concretely, we train the model (with the same architecture described in Sec. 3.2) independently on the train set of each dataset, and then evaluate it on the test sets of all four datasets to show its performance on the seen dataset and the other unseen datasets. As shown in Figure 1 (b), the model achieves the highest performance on its seen dataset and much lower performance on unseen datasets. Therefore, we validate that there do exist domain shifts, in other words, the domain gap across the four datasets. Overall, our CITM setting reasonably mimics the realistic application scenarios of image-text modeling.

Evaluation Metrics. We adopt **Recall@mean** (R@mean) and **Forgetting Rate** (FR) as our evaluation metrics. **R@mean** indicates the mean value of Recall@1, Recall@5, and Recall@10, where Recall@K (K=1,5,10) denotes the percentage of correct matching in the top-K retrieved results. The R@mean on each task indicates the retrieval performance of the final model on this task. Moreover, for the final model tested on task T_t, **FR** is defined as $\text{FR}^n_t = \frac{R^t_t - R^n_t}{R^t_t}$, where R^t_t denotes the R@mean of the last main model in task T_t on the test set of task T_t, and R^n_t denotes the performance of the final model on the test set of task T_t. The average FR is $\text{FR} = \frac{1}{n-1}\sum_{t=1}^{n-1}\frac{R^t_t - R^n_t}{R^t_t}$.

4.2 Implementation Details

Under the CITM setting, we train our DHA model on a sequence of four datasets: MSCOCO (Task T_1), CC3M (Task T_2), WIT (Task T_3), and GoodNews (Task T_4). After the main model has completed its training on task T_{t-1}, we find the last main model on the validation set of T_{t-1}. At the beginning of task T_t, this main model is used to initialize both the main and history models on this new task. For a fair comparison, we set the memory buffer to have 5% samples of the train set of each task for our DHA (if a buffer is used) and all competitors. The core idea of our memory buffer update strategy is similar to ER-ring [11] and we keep the buffer evenly containing image-text pairs from all the previous datasets instead. To make a comprehensive study, we implement our DHA with and without memory buffer to validate its effectiveness under the CITM setting.

We adopt BERT-Base [15]/ResNet50 [17] as the backbone of text/image encoder. They both use corresponding unimodal pre-trained models for initialization. The images are resized to 224x224 pixels, and the max length of the text descriptions is set to 256

Table 1: Comparative results between our proposed DHA and other representative/latest methods under the CITM setting. 'T2I' denotes text-to-image retrieval and 'I2T' denotes image-to-text retrieval. All methods adopt the same network architecture. 'Mem' denotes the data rehearsal with 5% buffer.

	Method	Mem	Task T_1		Task T_2		Task T_3		Task T_4	Average	
			R@mean	FR	R@mean	FR	R@mean	FR	R@mean	R@mean	FR
T2I	Joint training	-	31.61	-	32.39	-	18.86	-	22.98	26.46	-
	Baseline	N	13.64	68.30	12.78	64.82	13.63	40.25	22.62	15.67	57.79
	LwF [27]	N	16.81	60.94	15.69	56.31	15.33	31.41	22.68	17.63	49.55
	ER [11]	Y	16.30	62.12	16.15	55.08	14.37	36.42	21.57	17.10	51.21
	DER [7]	Y	20.52	52.31	20.92	41.74	**16.31**	**24.07**	21.04	19.70	39.37
	CO2L [8]	Y	19.64	54.35	18.95	46.14	16.13	24.94	**22.95**	19.42	41.81
	DHA† (ours)	N	21.31	50.48	21.82	37.15	15.64	27.09	21.37	20.04	38.24
	DHA (ours)	Y	**24.58**	**42.88**	**22.95**	**33.82**	16.15	24.50	21.22	**21.29**	**33.73**
I2T	Joint training	-	41.89	-	32.27	-	20.82	-	23.83	29.70	-
	Baseline	N	17.26	66.48	11.55	68.60	13.48	42.02	**23.72**	16.50	59.03
	LwF [27]	N	21.59	58.07	15.36	57.58	15.42	34.69	23.29	18.92	50.11
	ER [11]	Y	21.23	58.57	15.17	57.79	14.79	37.83	22.03	18.31	51.40
	DER [7]	Y	27.55	46.49	19.08	47.31	16.34	26.43	21.99	21.24	40.08
	CO2L [8]	Y	26.23	49.06	17.09	52.18	16.33	27.67	23.59	20.81	42.97
	DHA† (ours)	N	27.91	45.80	17.69	50.28	15.60	30.11	22.02	20.81	42.06
	DHA (ours)	Y	**32.72**	**36.45**	**21.01**	**39.50**	**16.45**	**24.71**	22.25	**23.11**	**33.55**

(tokens). We set the learning rate at the beginning of each task to 5e-5 and multiply it by 0.1 as the validation loss does not decrease. We adopt the optimizer Adam for gradient propagation, with the weight decay $1e^{-5}$. The batch size is set to 320 for each training iteration. λ_1, λ_2, and k are empirically selected as 0.995, 0.985, and 5, respectively. The main model is trained for 15 epochs on the training set of each task. The total training time on four datasets is around 12 hours with 8 Tesla V100 GPUs. The dataset and code will be released soon.

4.3 Main Results

We compare our DHA with other representative/latest methods, including the classic regularized-based method LwF [27], the classic rehearsal-based method ER [11], and two fusion methods DER [7] and CO2L [8] which combine the regularized-based and rehearsal-based strategies (the implementation details of these competitors are included in the suppl. material). For a clear comparison, we first provide the results of joint training on four datasets as the approximate upper bound. Besides, the basic method (denoted as 'Baseline') denotes training the same network sequentially on four tasks but without any continual learning strategy. The comparative results in Table 1 (see more results in the suppl. material) show that: (1) Our DHA beats all the competitors according to average R@mean and average FR over all tasks. The margins between our DHA and all the competitors are especially significant on average FR. This suggests that our direct parameter transfer strategy used for designing DHA is indeed effective for the CITM setting. (2) Our DHA outperforms the second best method DER by 1.59% – 1.87% on average R@mean and 5.64% – 6.53% on average FR. This further validates the effectiveness of our direct parameter transfer strategy used for designing DHA. (3) Our DHA† (without memory buffer) achieves better results than most of the other approaches. When the rehearsal-based strategy is fused, our DHA achieves the state-of-the-art results. That is, our DHA provides a new promising approach to continual image-text modeling. (4) On the most previous tasks (e.g., T_1 and T_2), our DHA performs significantly better than all the competitors in preserving much earlier knowledge. This superior ability would make a greater difference in realistic applications when there are more tasks in the data stream. (5) On the newest task T_4, nearly all the methods cause a drop on R@mean as compared to 'Baseline'. Such performance drop is mainly due to the trade-off between preserving previous knowledge and learning the current task, which is a common practice in continual learning scenarios. (6) We can find that there exist 5-6% gaps between our DHA and the upper bound in terms of the average performance (over all tasks), which implies further explorations could be made in future work.

To show more detailed performance of all methods in alleviating forgetting, we provide the results of the main model (of all methods) on task T_1 during sequential training on the four tasks in Figure 4 (more results on task T_2 and task T_3 are shown in the suppl. material). It actually shows the change tendency of R@mean on task T_1 of the main model when it is being trained on the later tasks sequentially. Specifically, the left sub-figure shows the text-to-image retrieval performance on task T_1 when the main model is trained from task T_1 to task T_4, while the right sub-figure shows the corresponding image-to-text performance. It can be clearly seen that: (1) Our DHA helps the main model forget with the slowest speed during sequential training among all the methods under the CITM setting. (2) Even without memory data, the forgetting speed of our DHA† is still slower than that of most of the other competitors. Overall, these observations provide further evidence that our direct parameter transfer strategy (used in DHA) is indeed effective in alleviating forgetting, and our DHA can be deployed as a new promising approach to continual image-text modeling.

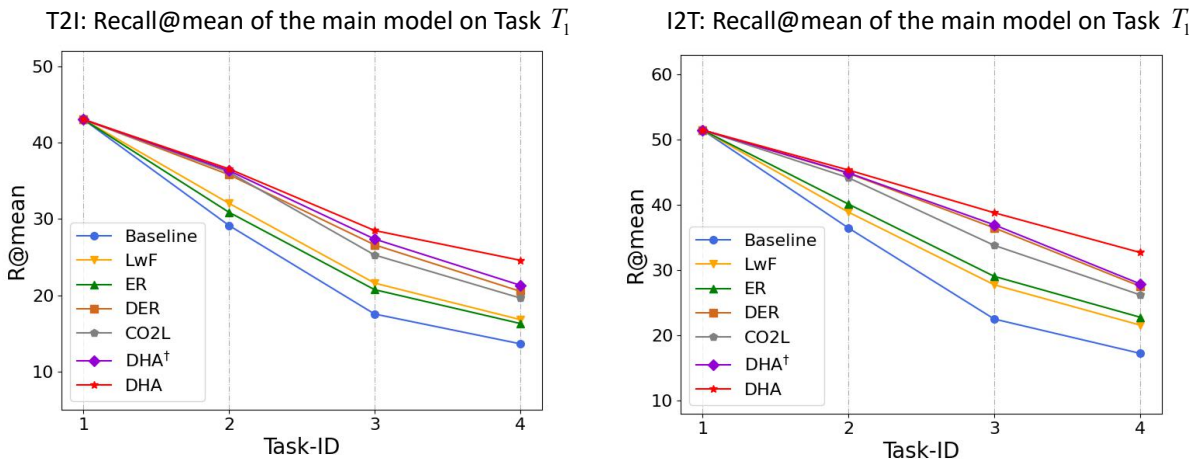

Figure 4: Illustration of the results of the main model on Task T_1 during sequential training on the four tasks. 'T2I' denotes text-to-image retrieval and 'I2T' denotes image-to-text retrieval. It can be seen that our DHA forgets with the lowest speed.

Table 2: Direct retrieval results on the test set of Flickr30K obtained by our DHA and other representative methods. Note that all methods are trained sequentially on four image-text datasets (i.e., {MSCOCO, CC3M, WIT, GoodNews}) under the CITM setting.

Method	T2I				I2T			
	R@1	R@5	R@10	R@mean	R@1	R@5	Recall@10	R@mean
Baseline	11.04	29.70	40.42	27.05	16.30	36.60	47.00	33.30
ER [11]	12.46	32.74	44.30	29.83	17.90	40.80	51.30	36.67
DER [7]	16.52	39.28	51.46	35.93	22.70	47.60	59.40	43.23
DHA† (ours)	16.02	36.58	47.52	33.52	21.50	47.10	58.80	42.47
DHA (ours)	**17.76**	**41.30**	**54.22**	**37.76**	**24.00**	**48.90**	**63.10**	**45.33**

Additionally, we conduct direct retrieval experiments on the test set of Flickr30K [46], which has no overlap with the sequence of four image-text datasets (i.e., {MSCOCO, CC3M, WIT, GoodNews}) under the CITM setting. Such a downstream task is commonly adopted as an evaluation task for current pre-training works. We compare our DHA with Baseline, ER, and the best competitor DER, which are all sequentially trained on the four datasets. The comparative results in Table 2 show that our DHA achieves the best performance, i.e., our DHA has the strongest generalization ability due to the direct parameter transfer strategy used for alleviating forgetting.

4.4 Ablation Study

Our proposed DHA is composed of two main strategies: (1) adaptation of the main model with the historical model (shortened as 'Adapt with Hist'), i.e., the main model keeps reviewing the historical knowledge by receiving the parameters of the historical model; (2) dynamic update of the historical model (shortened as 'Dynamic Hist'), i.e., the historical model is renewed by updating its parameters with the parameters of the main model. To clearly show the influence of each strategy on the model performance and also study the effect of the memory buffer, we provide the ablation study results for our full DHA on image-to-text retrieval in Table 3. We only show the results (R@mean) of the final model (trained across all four tasks) on each task under the CITM setting. We can observe that: (1) The most basic method with no DHA strategies and no memory buffer has the lowest performance on average R@mean. (2) Only adopting the strategy of 'Adapt with Hist' yields a 3.85% improvement on average R@mean, showing that it can well retain the knowledge of the previous tasks. (3) Adopting both 'Adapt with Hist' and 'Dynamic Hist' strategies brings further improvements on average R@mean. Particularly, such fusion yields performance gains on tasks T_2, T_3, and T_4 out of all the four tasks. This actually validates the effectiveness of 'Dynamic Hist': by controlling the gap between the historical and main models, the last main model of task T_t suffers from much less degradation (caused by 'Adapt with Hist') on task T_t. In other words, adopting both of the two strategies can ensure a good trade-off between the previous tasks and the current task during sequential training. (4) The extra memory buffer yields improvements in most cases. Moreover, even if the memory buffer is used, the two strategies of our DHA are still effective. This means that our DHA is complementary to the rehearsal-based methods.

We further conduct experiments to explore the effect of step k on the performance of our DHA. Intuitively, if k is too small, the historical model would be updated too frequently with the parameters of the main model. Therefore, although the knowledge gap is too small to affect the performance of the main model on the current task, the historical model is hard to preserve the historical knowledge. On the contrary, if k is too large, the historical model

Table 3: Ablation study results (R@mean) for our full DHA under the CITM setting. 'Adapt with Hist' denotes adaptation of the main model with the historical model. 'Dynamic Hist' denotes dynamic update of the historical model. 'Mem' denotes the data rehearsal with 5% buffer. The second best results are highlighted by underline.

Adapt with Hist	Dynamic Hist	Mem	Task T_1	Task T_2	Task T_3	Task T_4	Average
			17.59	12.33	12.12	**22.80**	16.21
✓			32.30	17.09	12.29	18.57	20.06
✓	✓		27.91	17.69	<u>15.60</u>	22.02	20.81
		✓	21.23	15.17	14.79	22.03	18.31
✓		✓	**36.29**	<u>20.41</u>	13.69	18.30	<u>22.17</u>
✓	✓	✓	<u>32.72</u>	**21.01**	**16.45**	<u>22.25</u>	**23.11**

Table 4: Effect of step k (for updating the historical model) on the performance of our proposed DHA. We only show the I2T results (R@mean) of the final model (trained across all four tasks) on each task under the CITM setting.

k	T_1	T_2	T_3	T_4	Avg.
1	22.62	15.21	14.20	**22.39**	18.60
3	28.01	17.69	14.45	22.22	20.59
5	32.72	21.01	**16.45**	22.25	**23.11**
7	**33.01**	**21.28**	16.23	21.69	23.01

Table 5: Effect of coefficients λ_1 and λ_2 on the performance of our proposed DHA. We only show the I2T results (R@mean) of the final model (trained across all four tasks).

λ_1	λ_2			
	0.98	0.985	0.99	0.999
0.98	20.68	21.01	20.93	20.87
0.99	21.24	22.23	21.64	21.33
0.995	22.15	**23.11**	22.28	22.12
0.999	21.43	22.84	21.97	22.08

would only have few updates with the parameters of the main model. As a result, a huge knowledge gap between the historical and main models would harm the performance of the main model on the current task. Overall, a good trade-off can be ensured by selecting the best k. Indeed, this analysis is validated by the I2T results (Average R@mean) in Table 4. Specifically, the performance of our DHA on task T_4 is the best when $k = 1$ and gradually decreases when k increases from 1 to 7, while the performance on the previous tasks T_1-T_3 grows higher at the same time. We thus select $k = 5$ with the highest average R@mean in this paper.

Moreover, we show the ablative I2T results (Average R@mean) of our DHA model with different values of λ_1 and λ_2 in Table 5. According to their definitions in Sec. 3.3, λ_1 controls the update speed of the main model with the historical model, and λ_2 controls the update speed of the historical model with the main model. Thus, the balance should be made between the two coefficients. We can see that the best performance could be obtained when $\lambda_1 = 0.995$ and $\lambda_2 = 0.985$. Moreover, most of the combination groups in Table 5 lead to better results than DER [7] (21.24), which further indicates that our DHA is indeed a promising approach to CITM.

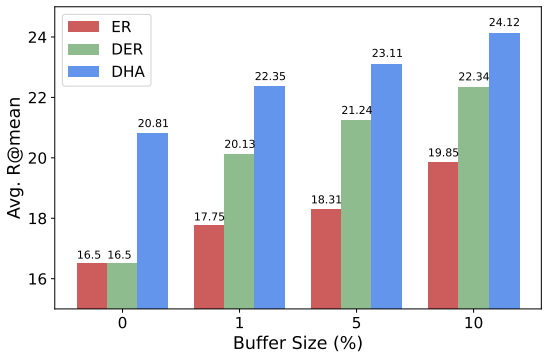

Figure 5: Results of DHA with extra memory buffer.

Finally, to investigate the effect of the buffer size, we make a comparison among ER [11], DER [7], and our DHA with different buffer sizes (0%, 1%, 5%, and 10% of the training data). We show the comparative results (average R@mean) in Figure 5. It can be seen that DHA beats ER and DER in all cases. Furthermore, our DHA with 1% buffer and 0% buffer even perform better than DER and ER with up to 10% buffer, respectively. This validates the effectiveness of direct parameter transfer (used in DHA) in continual learning.

5 CONCLUSION

In this paper, we propose a continual image-text modeling (CITM) setting, under which the model is required to be trained sequentially on four diverse image-text datasets and finally evaluated on all previous datasets. This new continual setting has a realistic application in large-scale image-text pre-training. We devise an effective dynamic historical adaptation (DHA) approach to coping with the forgetting problem in CITM. Different from existing continual learning methods, our DHA proposes to preserve the historical knowledge with direct parameter interaction between the historical and main models. Extensive experiments demonstrate the effectiveness of our DHA under the CITM setting.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (61976220). Zhiwu Lu is the corresponding author.

REFERENCES

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*. 139–154.

[2] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. 2017. Expert gate: Lifelong learning with a network of experts. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3366–3375.

[3] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for online continual learning. *Advances in Neural Information Processing Systems* 32 (2019), 11817–11826.

[4] Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony Robins. 2018. Pseudo-recursal: Solving the catastrophic forgetting problem in deep neural networks. *arXiv preprint arXiv:1802.03875* (2018).

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE/CVF International Conference on Computer Vision*. 1728–1738.

[6] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475.

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in Neural Information Processing Systems* 33 (2020), 15920–15930.

[8] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. 2021. Co2l: Contrastive continual learning. In *IEEE International Conference on Computer Vision*. 9516–9525.

[9] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision*. 532–547.

[10] Arslan Chaudhry, Albert Gordo, Puneet Kumar Dokania, Philip H. S. Torr, and David Lopez-Paz. 2021. Using Hindsight to Anchor Past Knowledge in Continual Learning. In *AAAI Conference on Artificial Intelligence*. 6993–7001.

[11] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486* (2019).

[12] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12652–12660.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. 1597–1607.

[14] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *TPAMI* 44, 7 (2021), 3366–3385.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. 2020. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in Neural Information Processing Systems* 33 (2020), 22605–22618.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[18] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. 2021. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561* (2021).

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. 4904–4916.

[20] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the Long-Short Term Memory Model for Image Caption Generation. In *IEEE International Conference on Computer Vision*. 2407–2415.

[21] Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image Generation from Scene Graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1219–1228.

[22] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences* 114, 13 (2017), 3521–3526.

[23] Frantzeska Lavda, Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. 2018. Continual classification learning using generative models. *arXiv preprint arXiv:1810.10612* (2018).

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. *ArXiv* abs/1803.08024 (2018).

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7331–7341.

[26] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence*. 11336–11344.

[27] Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (2017), 2935–2947.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 740–755.

[29] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. 2020. Mnemonics training: Multi-class incremental learning without forgetting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12245–12254.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019), 13–23.

[31] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *European Conference on Computer Vision*. 67–82.

[32] Arun Mallya and Svetlana Lazebnik. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7765–7773.

[33] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. MirrorGAN: Learning Text-To-Image Generation by Redescription. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1505–1514.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. 8748–8763.

[35] Jason Ramapuram, Magda Gregorová, and Alexandros Kalousis. 2020. Lifelong Generative Modeling. *ArXiv* abs/1705.09847 (2020).

[36] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. 2017. Encoder based lifelong learning. In *IEEE International Conference on Computer Vision*. 1320–1328.

[37] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2001–2010.

[38] Amir Rosenfeld and John K Tsotsos. 2018. Incremental learning through deep adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 3 (2018), 651–663.

[39] Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*. 4548–4557.

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*. 2556–2565.

[41] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. *Advances in Neural Information Processing Systems* 30 (2017), 2994–3003.

[42] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *ACM SIGIR Conference on Research and Development in Information Retrieval*. 2443–2449.

[43] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[44] Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[45] Ju Xu and Zhanxing Zhu. 2018. Reinforced continual learning. *Advances in Neural Information Processing Systems* 31 (2018), 907–916.

[46] Peter Young, Alice Lai, Micah Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[47] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. 2020. Class-incremental learning via deep model consolidation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*. 1131–1140.

[48] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8746–8755.