Continual learning in cross-modal retrieval

Kai Wang[1], Luis Herranz[1], Joost van de Weijer[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, Spain
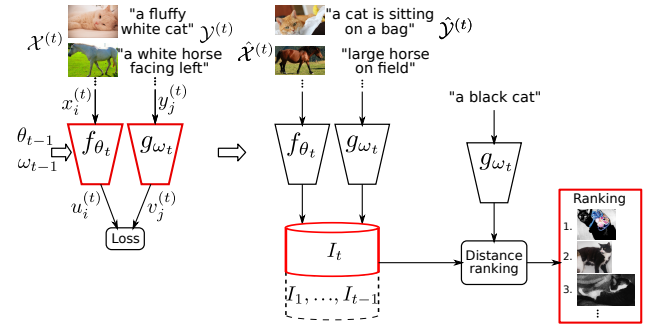
{kwang,lherranz,joost}@cvc.uab.es

Abstract

Multimodal representations and continual learning are two areas closely related to human intelligence. The former considers the learning of shared representation spaces where information from different modalities can be compared and integrated (we focus on cross-modal retrieval between language and visual representations). The latter studies how to prevent forgetting a previously learned task when learning a new one. While humans excel in these two aspects, deep neural networks are still quite limited. In this paper, we propose a combination of both problems into a continual cross-modal retrieval setting, where we study how the catastrophic interference caused by new tasks impacts the embedding spaces and their cross-modal alignment required for effective retrieval. We propose a general framework that decouples the training, indexing and querying stages. We also identify and study different factors that may lead to forgetting, and propose tools to alleviate it. We found that the indexing stage pays an important role and that simply avoiding reindexing the database with updated embedding networks can lead to significant gains. We evaluated our methods in two image-text retrieval datasets, obtaining significant gains with respect to the fine tuning baseline.

1. Introduction

Human intelligence requires integrating, processing and comparing information from multiple modalities. Ideally, mental representations should lie in an abstract common space that is decoupled from the specific modality of the perceived information. Language and vision already interact in simple tasks such as object classification, where images are mapped to concepts in a closed vocabulary of categories. However, multimodal representations [4] allow for richer interactions enabling cross-modal tasks such as cross-modal retrieval [11, 63, 9, 66, 13], image captioning [18, 12, 49], visual question answering [47, 23, 10, 65], and more recently text-to-image synthesis [32, 75]. Language models are also useful to extend visual classification

Figure 1. Stages in continual cross-modal retrieval (i.e. training feature extractors, indexing and query). The output of each stage is highlighted in red (i.e. feature extractors, index and ranking, respectively).

beyond the limited categories seen during training by projecting to language spaces, also known as zero-shot recognition [15, 71].

Another characteristic of humans is their ability for continual learning, which allows us to perform well tasks learned long back in time. In contrast, neural networks suffer from catastrophic interference [42, 39], which leads to almost complete forgetting of previous tasks when adapting to new ones, being a critical limitation to advance towards highly autonomous agents that can learn and adapt to changing environments. Continual learning (often referred to also as lifelong, sequential or incremental learning) in neural networks is an active research area, with recent methods addressing catastrophic forgetting with novel regularization [33, 35, 73], architectural [40, 5, 58] and (pseudo)-rehearsal [37, 59, 69, 62] mechanisms. Most continual learning methods focus on classification tasks.

Motivated by these two challenges, here we study continual learning in multimodal embedding spaces applied to cross-modal retrieval, and the specific problems that arise in this scenario. In continual learning, training (of new tasks) can happen at different points in time. In a retrieval scenario, we must consider also the indexing operation, where an embedded representation is extracted from the in-

put sample and stored in the database for future comparison. Since indexing in a continual setting could also happen at different points in time, we pay special attention to the role of this additional stage (see Fig. 1). An advantage of learning embedding networks instead of classification networks is that we operate in a single space shared by all tasks, so we can naturally retrieve data regardless whether we know or not the task related to that particular query sample (often referred to as task-aware and task-agnostic settings). Retrieval performance in continual learning is affected by how the embedded space may be distorted and cause representations to drift, as a result of the catastrophic interference. Additionally, these distortions and drifts may be unequal for each multimodality. Similarly, catastrophic forgetting affects differently to indexed data and query data.

In this paper we propose a continual cross-modal retrieval framework that can effectively perform retrieval in known and unknown domains. We identify and study the different factors that lead to forgetting in cross-modal embeddings and retrieval. Addressing those factors, we study modifications in the retrieval framework, network architecture and regularization that can help to alleviate them.

2. Related Work

2.1. Deep metric learning

Deep metric learning learns both feature extraction and a distance metric in an end-to-end fashion. It maps images to an embedding space in which a simple distance metric such as the Euclidean distance can be applied. For training, it requires positive pairs (PP), which should be close in the embedding space, and negative pairs (NP), which are mapped at least a margin apart. Initial work was based on Siamese networks [6] which consist of two identical neural networks with shared weights, each taking one of the two inputs and map them to an embedding space. They are widely used in patch matching [60], face verification [55], image retrieval [17], etc.

Regarding the training loss, two of the most widely used are contrastive loss [19] and triplet loss [21]. The former continually pushes similar instances closer, whereas negative pairs are only required to be at least a margin away. In the latter, similar samples are only required to be closer to each other than to any dissimilar ones. The training of Siamese and triplet networks is known to be difficult. Especially, since many of the negative pairs are already far apart in embedding space, they do not result in any training signal. Therefore, it was shown to be important to perform hard negative mining [60]. Later works observed that it was computationally advantageous to first pass the images through a single network, and only form the pairs in the loss layer [48, 36]. Other losses include center loss [67] and proxy-NCA [44].

2.2. Cross-modal retrieval

Cross-modal retrieval requires a coordinated representation [4] that allows computing a similarity measure between the query representation and that of the retrieved data, even when they belong to different modalities and extracted with different feature extractors. There are two main aproaches to this problem: canonical correlation analysis (CCA) [22] and metric learning [31].

CCA [22] learns linear projections to a space where the projections of two random vairables are maximally correlated, which makes it attractive to cross-modal retrieval. CCA has also been extended to deep networks [2, 64], and in particular to cross-modal retrieval [14, 16, 29]. A limitation of CCA approaches is the expensive computation of the covariance matrix that requires having all data in memory.

Metric learning has also been applied sucessfully to cross-modal retrieval. Early examples of joint text-image embeddings are WSABIE [68] and DeViSE [15] which map image and text embeddings into a single space using ranking losses. Kiros *et al.* [28] applied a similar approach to sentences using an LSTM model. Socher *et al.* [61] use an extended language model which includes dependency trees. Xu *et al.* [70] propose a joint representation for video and sentences. Two-branch networks [63] address image-text matching tasks with a bi-directional ranking loss. Multimodal representations have been also used for cross-modal retrieval of more structured visual-text documents, such as recipes [54, 43] and learning facts from images [13].

Efficient retrieval from large databases is also a concern, so cross-modal hashing [7] learns compact representations in binary spaces where indexing and retrieval can be performed efficiently. Cross-modal hashing has also been extended to deep models [8, 24].

2.3. Continual learning

A well known phenomenon in neural networks is catastrophic forgetting, where learning new tasks interferes with remembering previous ones [42, 39]. To enable networks to succeed in scenarios requiring continual learning, different techniques have been proposed.

A popular approach is to add regularization terms to the loss. Weight regularization methods [27, 74, 1, 35] add quadratic terms to penalize large differences to the solution for previous tasks, weighted by some importance measure so differences in more important parameters are penalized more. Elastic weight consolidation (EWC) [27] uses the diagonal approximation of the Fisher information matrix to estimate the importance. Rotated EWC [35] proposes a reparametrization that makes EWC more effective. Synaptic intelligence (SI) [74] estimates the importance measure during training by accumulating gradients. Memory aware synapses (MAS) [1] uses perturbation theory to estimate the importance in an unsupervised way. Forgetting can also

be prevented by regularizing the activations, as in learning without forgetting (LwF) [33], where a snapshot of the network right before starting to learn the new task (and therefore not suffering interference from it) is used as a teacher and a distillation loss [20] is used during the training of the new task. Encoder-based lifelong learning [50] uses distillation in task-specific projections, estimated by autoencoders.

Another way of avoiding forgetting is rehearsal [52, 51], where a fraction of data (i.e. exemplars) from previous tasks is kept and revisited during training, and pseudo-rehearsal [52, 3], where pseudo-exemplars are sampled from an auxiliary model trained to model previous tasks.

Recent pseudo-rehearsal methods include deep generative models models [59, 69]. Other approaches to continual learning include networks that expand their capacity to allocate new tasks [53, 72, 56] and task-attention mechanisms [58].

While most works focus on classification, continual learning has also been studied in other settings such as image generation [69, 57, 46], word embeddings [25, 41], Atari games [27] and continual adaptation of agents [45]. MAS[1] is evaluated in facts learning that involves image and structured text. However, to our knowledge, there is not any work specifically studying the implications of continual learning in a retrieval setting, and catastrophic forgetting from the perspective of cross-modal embeddings.

3. Continual cross-modal retrieval

3.1. Cross-modal deep metric learning

Our framework is based on a two-branch network [63], with image-specific and text-specific embedding branches that project images and text into a common space. The image embedding operation is $u = f_{\boldsymbol{\theta}}(x)$, where $u \in \mathbb{R}^E$ is the image embedding of an input image x, extracted by the image embedding network $f_{\boldsymbol{\theta}}$, parametrized by $\boldsymbol{\theta}$. Similarly, the text embedding $v \in \mathbb{R}^E$ of an input text y is obtained as $v = g_{\boldsymbol{\omega}}(y)$ by the text embedding network $g_{\boldsymbol{\omega}}$ parametrized by $\boldsymbol{\omega}$. Both u and v are normalized using l_2 norm. Images and text are compared in the embedding space using the Euclidean distance as $d(x, y) = \|u - v\| = \|f_{\boldsymbol{\theta}}(x) - g_{\boldsymbol{\omega}}(y)\|$.

The image set $\mathcal{X} = \{x_i\}_{i=1}^{N_{\text{I}}}$ is aligned with a text set $\mathcal{Y} = \{y_j\}_{j=1}^{N_{\text{T}}}$ via a pairwise similarity matrix S. This cross-modal pairwise similarity is indicated by a variable $s_{ij} \in S$ which takes value 1 when x_i and y_j are similar (i.e. *positive pair*) and 0 otherwise (i.e. *negative pair*). We want the distance between positive pairs to be significantly lower than the distance between negative pairs. In order to do that we use the bi-directional ranking loss of [63], which selects triplets and imposes constraints

$$d(x_i, y_j) + m \leq d(x_i, y_k)$$
$$\text{s.t. } s_{ij} = 1 \text{ and } s_{ik} = 0 \tag{1}$$

and (in the other direction)

$$d(y_{i'}, x_{j'}) + m \leq d(y_{i'}, x_{k'})$$
$$\text{s.t. } s_{i'j'} = 1 \text{ and } s_{i'k'} = 0 \tag{2}$$

where m is the predefined margin. The triplets are constructed based on a positive pair, and a negative pair creating by replacing either the image or the text by a dissimilar one. These triplet constraints are included using a margin-based loss function (where $[z]_+ = \max(0, z)$):

$$L_{\text{T}}(\mathcal{X}, \mathcal{Y}) = \lambda_1 \sum_{i,j,k} [d(x_i, y_j) + m - d(x_i, y_k)]_+$$
$$+ \lambda_2 \sum_{i',j',k'} [d(y_{i'}, x_{j'}) + m - d(y_{i'}, x_{k'})]_+ \tag{3}$$

3.2. Training, indexing and query stages

In general, machine learning assumes two different stages, namely *training* and *evaluation* (or *test*), which take place in that exact order (although in continual learning it is not the case). We focus on retrieval with a learned feature extractor (i.e. embedding networks in our case). In this scenario we identify three stages (see Fig. 1):

Training (feature extractors(s)). Described in the previous section, the training stage learns the embedding networks from the image and text datasets $\mathcal{X}$ and $\mathcal{Y}$, and its result is the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$.

Indexing (database data). The database datasets $\hat{\mathcal{X}}$ and $\hat{\mathcal{Y}}$ to be indexed are processed using the embedding networks to obtain the text and image embeddings, which are subsequentially indexed in the database. Note that training data and database data are not required to be the same.

Querying (query data). This stage computes the similarity between a query sample and the indexed data. The result is a ranking with the most similar sample on top. In our cross-modal case there are two directions: querying with images, retrieving from indexed texts (*im2txt*) and querying with text, retrieving from indexed images (*txt2im*).

Note that these three stages are assumed to take place in that particular order, and a deployed system only performs the querying stage. For simplicity we consider that the database data is also used as training data, i.e. $\hat{\mathcal{X}} = \mathcal{X}$ and $\hat{\mathcal{Y}} = \mathcal{Y}$.

3.3. A framework for continual retrieval

Now we consider a continual learning setting, in which data is presented as a sequence of tasks $\{\mathcal{T}^{(t)}\}_{t=1}^T$. Each
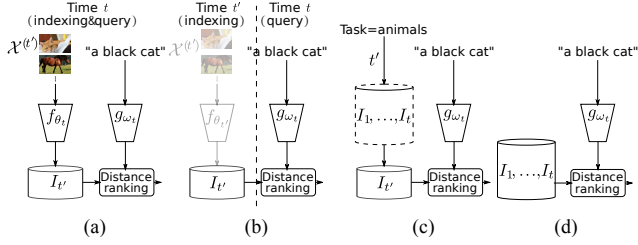
Figure 2. Variants of indexing data from a previous task t' when queried at time $t > t'$ (a-b) and retrieval (c-d): (a) reindexing, (b) not reindexing, (c) task known, (d) task unknown.
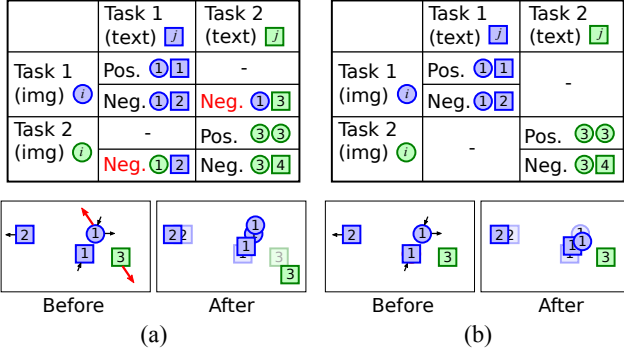

Figure 3. Types of pairs in continual cross-modal retrieval: (a) available in joint training, and (b) available in continual learning, i.e. without cross-task negative pairs (CTNP). CTNPs are crucial to avoid overlap between samples of different tasks (bottom). Best viewed in color.

task $\mathcal{T}^{(t)} = \left(\mathcal{X}^{(t)}, \mathcal{Y}^{(t)}, \mathcal{S}^{(t)} \right)$ involves data from a different domain (e.g. animals, vehicles). We assume that the embedding networks are updated (i.e. fine tuned) with data of a particular task (i.e. training stage) before indexing data of that task. The resulting parameters after training task t are $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_t$.

The retrieval system is evaluated in the querying stage with separate data from every task. We consider two settings for evaluation: *known task* and *unknown task*, depending on whether that information is available at query time (see Fig. 2a-b).

As described previously, the network is trained using cross-modal positive and negative pairs. When all data is presented jointly, all negative pairs are available for sampling. However, in the continual setting, pairs are formed within the same task, i.e. combining samples from $\mathcal{X}^{(t)}$ and $\mathcal{Y}^{(t)}$. Thus, we further classify a negative pair (x_i, y_j) as intra-task negative pair (ITNP), when $x_i \in \mathcal{X}^{(t)}$ and $y_j \in \mathcal{Y}^{(t)}$ ($s_{ij} = 0$, $s_{ij} \in \mathcal{S}^{(t)}$), or as cross-task negative pair (CTNP), when $x_i \in \mathcal{X}^{(t')}$ and $y_j \in \mathcal{Y}^{(t)}$, $t' \neq t$. Note that, for simplicity, we assume that all positive pairs are intra-task. In continual retrieval, CNTPs are not available during training (see Fig. 3).
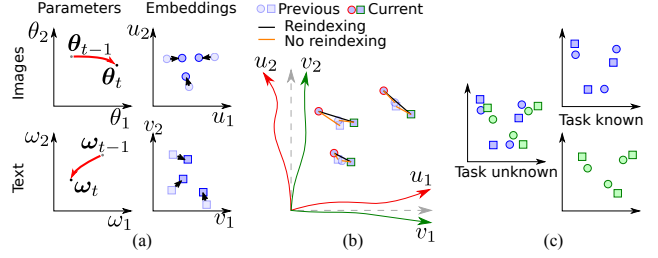

Figure 4. Causes of forgetting in cross-modal embeddings: (a) embedding networks become less discriminative due to drift in parameter space, and (b) unequal drift increases cross-modal misalignment, and (c) task overlap in embedded space (when task is unknown). Best viewed in color.

3.4. Do or do not reindex?

The conventional retrieval scenario assumes that training and indexing are performed once. In this case, there is only a static set of embeddings, extracted with the same network at the same time. The same network is used to extract embeddings from queries. In continual retrieval this may not be the case, since training and indexing are performed every time a new task is presented.

Reindexing. We first consider the straightforward extension of cross-modal retrieval that assumes that current and previous tasks all are reindexed with the version of the embedding networks with updated parameters $f_{\boldsymbol{\theta}_t}$ and $g_{\boldsymbol{\omega}_t}$ after a new task t is learned (see Fig. 2a). We refer to this case as *reindexing*. However, it has the drawbacks of being time and resource consuming, since it requires indexing the same data multiple times, and always requiring access to the image and text samples of previous tasks. It has the advantage that database and query samples are processed with the same networks.

No reindexing. We also propose the variant *no reindexing* that only indexes the data of current task t after training task t (see Fig. 2). This variant is more efficient, since database samples are processed only once, and flexible since it does not require access to previous images and text (only to their indexed embeddings for retrieval). On the other hand, no reindexing introduces asymmetry, since query embeddings are extracted with $f_{\boldsymbol{\theta}_t}$ (or $g_{\boldsymbol{\omega}_t}$), while database embeddings with $g_{\boldsymbol{\omega}_{t'}}$ (or $f_{\boldsymbol{\theta}_{t'}}$, with $t' \leq t$).

4. Catastrophic forgetting in cross-modal embeddings

Learning a new task implies that the values of the network parameters will shift away from the previous ones. This is particularly important when the new task is very different from previous, causing interference between new and previous tasks that leads to lower performance in the latter. For simplicity, we will refer to this drop in performance as *catastrophic forgetting*. In the following, we identify

several phenomena that may lead to forgetting in continual cross-modal retrieval.

Embedding networks. We first consider forgetting in each embedding network separately, without considering pairwise interactions. As their parameters move away from the optimal values for $t-1$ (see Fig. 4a), the embeddings u and v will also drift from their previous values. In general, the new values $f_{\boldsymbol{\theta}_t}$ and $g_{\boldsymbol{\omega}_t}$ are less discriminative than previous previous $f_{\boldsymbol{\theta}_{t-1}}$ and $g_{\boldsymbol{\omega}_{t-1}}$, causing lower performance, because the embedding spaces of $u^{(t)}$ and $v^{(t)}$ are also less discriminative.

Embedding misalignment. In the particular case of cross-modal networks, embeddings of different modalities may drift differently (see Fig. 4a). This unequal drift in u and v spaces causes additional misalignment that also leads to higher distances than in the optimal case. Note that uni-modal retrieval with Siamese networks or Triplet networks does not suffer from this problem, since parameters are shared across de various branches.

Task overlap. Negative pairs pull dissimilar samples away in the embedded space. However, in continual retrieval CT-NPs cannot be sampled (unless we include some samples from previous tasks). CNTPs are the only repulsive force between samples of different tasks. Without them, it is likely that samples from different tasks will overlap in the embedded space (see Fig. 4a). Knowing the task at query time makes this problem less important, since data from other tasks are not considered at query time.

5. Preventing forgetting

In the following we propose several tools to alleviate forgetting by addressing the previous causes.

5.1. Preventing embedding drift

A common approach to prevent forgetting is regularizing the weights with a quadratic term in the loss that penalizes the weighted Euclidean distance (in the parameter spaces) to the solution for previous tasks [27, 74, 1, 35]. This can help to avoid significant drift in the embeddings and to keep them discriminative for previous tasks. We can write the particular regularization term for our case as

$$
L_{\mathrm{R}} = \sum_{k} \Theta_k^{(t-1)} \left(\theta_k^{(t-1)} - \theta_k \right)^2 + \\
+ \sum_{k'} \Omega_{k'}^{(t-1)} \left(\omega_{k'}^{(t-1)} - \omega_{k'} \right)^2
\tag{4}
$$

where Θ_k and $\Omega_{k'}$ control the regularization strength depending on the importance of $\boldsymbol{\theta_k}$ and $\boldsymbol{\omega_k}$, respectively, for previous tasks. During the training of the first task there is no regularization, i.e. $\Theta_k^{(0)} = 0$ and $\Omega_{k'}^{(0)} = 0$. The way

to compute the importance differs in different methods. We consider two variants:

Global. Here we estimate the importance with respect to the loss, adapting elastic weight consolidation (EWC) to our particular triplet loss as (L_{TR} represents the triplet loss):

$$
\Theta_k^{(t)} = \mathbb{E}_{x,y} \left[\left(\frac{\partial}{\partial \theta_k} L_{\mathrm{TR}} \left(\mathcal{X}^{(t)}, \mathcal{Y}^{(t)} | \boldsymbol{\theta}_t, \boldsymbol{\omega}_t \right) \right)^2 \right]
\tag{5}
$$

which is computed by sampling triplets as in 1 and 2, and analogously for $\Omega_{k'}$. This loss already takes into account triplets and their interactions.

Branch. Instead of estimating importance values that depend on a joint loss, we consider regularizing each branch independently. In this case we estimate the importance using the approach memory aware synapses (MAS), which can be computed unsupervisedly for each branch with images or text. The importance for the image branch is estimated as:

$$
\Theta_k^{(t)} = \Theta_k^{(t-1)} + \mathbb{E}_{x_i \sim \mathcal{X}^{(t)}} \left[\frac{\partial}{\partial \boldsymbol{\theta}_k} l_2^2 (f_{\theta_t}(x_i)) \right]
\tag{6}
$$

which is accumulated over previously computed one. For the text branch the estimation of $\Omega_{k'}$ is analogous. In this equation, l_2^2 is the squared l^2 norm of the function outputs, which is used to estimate the importance of parameters in MAS method.

The final loss combines (3) and (4) as $L = L_{\mathrm{T}} + \lambda_3 L_{\mathrm{R}}$.

5.2. Preventing unequal drift

In order to prevent unequal drift we propose tying the networks by sharing layers at the top (bottom layers must remain modality-specific). In this way, the unequal drift can be alleviated since the gradients are tied and only differ in the lower layers.

In some cases when the drifts in text and image embedding are in opposite directions, refraining from reindexing the database can be an effective tool to alleviate drift, since only one of the embeddings is affected while the other remains fixed. Fig. 4b illustrates how in that case no reindexing keeps matching pairs at lower distances.

5.3. Decoupling retrieval directions

So far we assumed only a single model is trained to perform both text to image and image to text retrieval. This is reasonable when embeddings are reindexed since the architecture and the loss are symmetric. However, when database data is not reindexed and query is, the forgetting is asymmetric. In that case we can decouple both directions and train one model for each direction, only regularizing the

weights in the query branch. This can also be beneficial in some cases when the image and text embeddings drift in different directions, keeping one fixed in the previous position can keep the distance lower (see example in Fig. 4b).

5.4. Preventing cross-task overlap

The lack of CTNPs can lead to cross-task overlap, since there is no force separating them. However, reducing the drift, keeping the embeddings discriminative via weight regularization and sharing layers may indirectly help to keep tasks separated (we observed that in our experiments).

Nevertheless, we made some preliminary experiments creating pseudo-CTNPs (u_i, x_j) in models with decoupled retrieval directions using the already indexed embeddings (analogously for text to image retrieval for the other direction), but we found they did not help in our experiments, probably because the asymmetric force that only pushes the embeddings of one branch. In this case the gradients are only backpropagated through one branch. We leave their study more in depth for future work.

6. Experiments

Baselines and variants We evaluate the different variants of our continual cross-modal framework in two tasks involving images and text, one focusing on regions and the other on scenes. We follow the implementation of the two-branch networks in [63] where 4096-dim image features are extracted from a VGG-19 model trained on ImageNet, and text features are 6000-dim from HGLMM features (reduced with PCA from initial 18000-dim) [29]. The image branch includes two additional fully connected layers with sizes 2048 and 64 (for SeViGe, 2048 and 512 for SeCOCO) on top and l_2 normalization, and the same for the text branch. We focus our study on the two fully connected layers on top, while the initial feature extractors remain fixed. As in [63] we set $\lambda_1 = 1.0$, $\lambda_2 = 1.5$ and the margin $m = 0.05$. The resulting model is trained with Adam[26] and a learning rate of 0.0001, and using dropout after ReLu with probability 0.5. We evaluate different variations of this architecture:

- **Joint vs continual.** We compare the variants of the proposed framework (*continual*) with two baselines that learn all tasks jointly (*joint*), differing on whether CTNPs are sampled or not during training.
- **Retrieval direction.** We evaluate both text to image retrieval (*txt2im*) and image to text retrieval (*im2txt*).
- **Task knowledge.** We evaluate both the cases where the task is *known* and *unknown*.
- **Reindexing.** We consider the embeddings for database samples are extracted when the corresponding task was learned (*no reindex*), or are at the same time as the query embeddings (*reindex*).

- **Weight regularization.** We consider fine tuning with no regularization (*ft*), with joint regularization on the loss (*EWC*) and with regularization on each u and v embedding independently (*MAS*). We set $\lambda_3 = 10^6$.
- **Decoupled directions.** For *no reindex* we also consider variants where EWC or MAS are only computed in the branch extracting query embeddings (e.g. *MAS-txt* when MAS is computed only on the text branch). In this case we run two different experiments, each specialized for one particular retrieval direction.
- **Layer sharing.** We consider keeping both embedding networks independent (*no sharing*) or sharing the top fully connected layer (*sharing*).

We consider experiments where each task consist in updating the embedding networks by learning a new domain. After training the model, the same training data is indexed (i.e. we extract image and text embeddings) and then the retrieval performance can be evaluated. We report the final results after all tasks are learned. We use Recall@K as evaluation metric (with $K = 10$, results for other K in the supplementary material), with respect to the indexed data of the same domain (*known*) or to the whole indexed data with all domains (*unknown*). We repeat each experiment five times and report the average.

6.1. Sequential Visual Genome

Sequential Visual Genome (SeViGe) dataset. We created a dataset based on the regions with object-description pairs in the Visual Genome dataset [30]. Based on the object categories of those regions, we selected pairs related with the domains *animals* (9 categories), *vehicles* (6 categories) and *clothes* (6 categories), which are learned in sequence as tasks in our experiments. Each task has a total of 10481, 7531 and 10200 training images, respectively, and additional 900/900, 600/600 and 600/600 for validation/test, respectively (100/100 per category).

Cross-modal retrieval. We evaluate the different methods in cross-modal region image-text retrieval. The results for both directions are shown in Table 1. We focus on *average* for evaluation when the task is known, and *A+V+C* to evaluate when the task is unknown (i.e. the aggregate of all domains). We first observe that training all tasks jointly delivers higher performance than the continual setting, as expected. Significant part of that superiority is due to CTNP, since the performance of joint training drops significantly when not sampled. This provides a more realistic and tighter upper bound, since in the continual scenario CT-NPs are not available. This drop is more moderate when the task is known (∼1-2% known, ∼3-4% unknown), since task overlapping is not a problem. This drop still suggests that CTNPs still contribute to shape the embedding spaces to be discriminative beyond simply avoiding task overlap. When

the top layer is shared, the drop is also smaller, although the overall performance is also lower than not sharing the layer.

Focusing on continual learning we observe that the single modification that most reduces forgetting is not reindexing the database, which provides 3% and 5.1% boosts in *im2txt* and *txt2im* directions, respectively (1.6% and 3.4% if task is unknown). This surprising result, showing that reindexing can be harmful, suggests that the misalignment caused by the unequal drift of image and text embeddings is more critical than the misalignment caused by not extracting embeddings at the same time, and that keeping good and discriminative representations in the database is also important (recall that in *no reindexing* only the query embedding has endured catastrophic interference). Note that this is specific to cross-modal retrieval because branches do not share parameters. This may not be the case in image-to-image or text-to-text retrieval with Siamese or Triplet networks because the embeddings of the two branches drift equally.

Sharing layers by itself gives an improvement of 1.4% in *im2txt*, while not having impact in *txt2im*. Not reindexing also gives a similar boost as in the previous case. Interestingly, sharing layers harms performance in joint training, while for the continual setting it improves the performance, probably because it reduces the unequal drift by tying the drift of both modalities at least in the shared layers.

Weight regularization has moderate impact and could harm the performance sometimes. Decoupling both modalities and applying regularization only in the query network extractor seems to help in some cases (e.g. +1.3%/+1.6% gain with *MAS-txt* vs *MAS* in *txt2im*, and +0.7%/+0.9% in *im2txt* in the *sharing* architecture). In this dataset regularizing embeddings independently with MAS instead of the whole network with EWC seems to work better, although the differences are very marginal. Here we can see that the forgetting in embedding network is not a significant problem in cross-modal retrieval setting.

Overall, the best combination provides improvements of 6%/6.3% in known/unknown *txt2im* retrieval, and more moderate improvements of 2.9%/2% in *im2txt* retrieval.

Insights about the embedding space. We use t-SNE [38] to visualize the embedding space of variants with shared layers. Although distances in t-SNE do not reflect real distances, it is useful to identify structure. We combine text and image embeddings and run t-SNE, color coding data with modality and task labels. Joint training (see Fig. 5a) generates embeddings where data is structured clearly in separated tasks, and within tasks, in separated clusters (probably the categories within each task in SeViGe). This happens in both modalities, which also overlap, aligned according to the related clusters. Not sampling CTNPs still results in intra-task structure (see Fig. 5b, e.g. category clusters are clear), but the modalities are significantly more misaligned and with larger overlap, showing the important
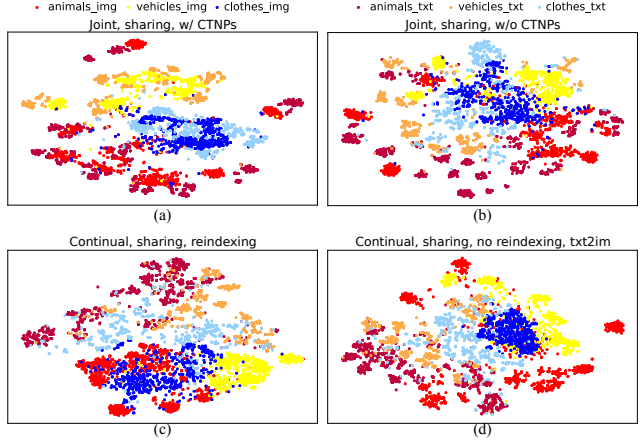

Figure 5. t-SNE visualization of the cross-modal embedding space of SeViGe, with the *sharing* architecture: (a) joint training (with CTNPs), (b) joint training (without CTNPs), (c) continual (reindexing), and (d) continual (no reindexing). Best viewed in color.

role of CTNPs in aligning modalities and separating tasks. When learned in a continual fashion (see Fig. 5c), the misalignment is more extreme, even resulting in text and image samples distributed in different halves of the space. No reindexing (see Fig. 5d for *txt2im* direction) seems to keep the image embeddings (database) more discriminative, which may explain the improved results compared to reindexing. For example, the image embeddings of animals and vehicles seem much better separated in Fig. 5d than in Fig. 5c.

6.2. Sequential MS-COCO

Sequential MS-COCO (SeCOCO) dataset. We created a second dataset with image-description pairs of MS-COCO[34]. Each image in MS-COCO is annotated with five image-level descriptions of the scene and a variable number of object annotations localized to specific regions and labeled with one of 80 disjoint object categories. Object categories are further organized in 12 disjoint super-categories. Organizing the data into tasks is challenging in this case since we want to avoid overlap between tasks, but there are many object annotations in each image. We organized the data into groups of super-categories and removed the images with object annotations in more than one group. After removing overlapping images we use those groups as tasks. We finally selected *animal*, *accessory*, *kitchen*, *food* and *furniture* for task 1, *vehicle*, *outdoor*, *electronic*, *appliance* and *indoor* for task 2 and *person* and *sports* for task 3, with 22475, 13903 and 13919 training images respectively, in addition to 1000/1000 images for validation/test for each task. Note that many other objects and concepts remain unannotated, so there is still semantic overlap across tasks that we cannot control.

Cross-modal retrieval. Table 2 shows the recall@10 for

Domain	im2txt										txt2im									
	Joint CTNP		Continual								Joint CTNP		Continual							
			reindexing			no reindexing							reindexing			no reindexing				
	Yes	No	ft	EWC	MAS	ft	EWC	EWC-im	MAS	MAS-im	Yes	No	ft	EWC	MAS	ft	EWC	EWC-txt	MAS	MAS-txt
Architecture: *no sharing*																				
animals	29.1	26.0	16.1	16.8	16.9	24.5	24.6	24.2	24.7	24.3	27.8	25.9	15.4	15.2	15.4	20.8	20.8	20.9	19.8	20.7
vehicles	30.9	27.7	20.8	23.3	22.7	24.0	25.1	24.8	26.0	24.8	30.9	27.0	17.5	18.6	19.5	27.2	29.4	28.0	28.8	28.7
clothes	27.9	27.5	27.4	27.0	27.5	27.4	27.0	27.3	27.5	26.3	29.3	27.7	28.1	27.5	28.0	28.1	27.5	27.4	28.0	28.5
average	**29.3**	27.0	21.5	22.3	**22.4**	24.5	24.6	24.2	**24.7**	24.3	**29.3**	26.8	20.3	20.5	**21.0**	25.4	25.9	25.4	25.6	**26.0**
A+V+C	**28.5**	24.4	17.0	**18.4**	17.8	18.6	17.9	17.5	**19.0**	18.3	**28.0**	23.8	16.3	16.3	**16.9**	20.7	21.3	20.9	20.9	**21.4**
Architecture: *sharing*																				
animals	28.3	25.3	18.4	17.1	16.4	23.1	21.2	21.4	21.1	21.4	26.8	24.4	16.6	14.8	14.3	22.1	20.7	21.1	20.6	22.2
vehicles	30.2	28.6	22.6	24.7	23.5	23.0	24.9	25.0	23.8	26.0	31.2	27.9	16.9	17.8	16.3	27.3	29.4	29.5	28.4	28.7
clothes	26.7	27.4	27.7	26.9	27.1	27.7	26.9	27.3	27.1	26.7	27.5	26.8	27.2	27.0	26.0	27.2	27.0	27.5	26.0	28.0
average	28.4	27.1	22.9	**22.9**	22.3	24.6	24.3	24.6	24.0	**24.7**	28.5	26.4	**20.3**	19.9	18.9	25.6	25.7	26.0	25.0	**26.3**
A+V+C	27.8	24.5	18.2	**18.2**	17.6	**19.0**	17.9	18.2	17.9	18.8	27.2	23.7	**15.9**	15.5	14.9	21.8	21.5	22.2	21.0	**22.6**

Table 1. Results in SeViGe after learning all tasks (Recall@10 in %). *average* measures performance with *known* task, while *A+V+C* with *unknown* task. Best joint learning result in **green**, best continual learning result in **red**.

Domain	im2txt										txt2im									
	Joint CTNP		Continual								Joint CTNP		Continual							
			reindexing			no reindexing							reindexing			no reindexing				
	Yes	No	ft	EWC	MAS	ft	EWC	EWC-im	MAS	MAS-im	Yes	No	ft	EWC	MAS	ft	EWC	EWC-txt	MAS	MAS-txt
Architecture: *no sharing*																				
task1	65.7	63.8	33.6	32.0	33.0	49.8	48.1	47.2	50.5	47.1	69.7	68.2	40.1	38.0	38.2	59.8	59.2	58.3	60.0	59.7
task2	56.5	54.9	39.8	38.5	40.0	47.0	46.6	46.4	47.0	46.9	65.2	62.6	46.8	44.7	46.9	54.6	55.5	55.1	55.5	55.9
task3	38.2	39.9	39.7	40.1	40.2	39.7	40.1	39.9	40.5	39.7	44.6	45.7	46.7	46.7	46.0	46.7	46.7	46.7	46.0	46.2
average	**53.5**	52.9	**37.7**	36.9	**37.7**	45.5	44.9	44.5	**46.0**	44.6	59.8	58.9	**44.5**	43.1	43.7	53.7	53.8	53.4	53.8	**54.0**
total	**52.4**	49.8	**33.0**	32.1	**33.0**	37.1	36.2	35.6	**37.4**	36.0	58.5	56.3	**40.4**	38.7	39.7	48.3	48.0	47.3	48.2	**48.4**
Architecture: *sharing*																				
task1	65.3	63.9	32.9	31.9	34.1	48.4	47.7	47.7	47.8	45.1	70.2	67.7	38.2	37.4	39.8	58.6	56.3	58.4	57.1	57.5
task2	55.7	55.3	40.6	39.9	40.4	46.3	46.0	45.2	44.0	44.4	64.7	63.1	46.0	45.7	46.3	54.6	54.2	55.6	54.6	54.9
task3	37.6	40.1	39.6	39.7	39.3	39.6	39.7	39.9	40.0	39.7	44.8	46.5	46.2	45.8	45.7	46.2	45.8	45.7	46.7	46.1
average	52.9	53.1	37.7	37.2	**37.9**	**44.8**	44.5	44.3	43.9	43.1	**59.9**	59.1	43.5	43.0	**43.9**	53.1	52.1	**53.2**	52.8	52.8
total	51.8	50.1	33.2	32.5	**33.5**	**36.1**	35.9	35.4	35.5	35.3	**58.7**	56.4	39.3	38.9	**39.9**	47.7	46.8	**48.1**	47.1	47.5

Table 2. Results in SeCOCO after learning all tasks (Recall@10 in %). *average* measures performance with *known* task, while *total* with *unknown* task. Best joint learning result in **green**, best continual learning result in **red**.

different methods on SeCOCO. In this case joint training also performs better than continual learning methods. The drop due to not training with CTNPs is relatively lower than in SeViSe. This can be explained by a higher semantic overlap between tasks that makes CTNPs less critical. The relative importance of sharing layers is also less important in this case, with very little difference in the results.

Regarding continual learning methods, no reindexing is again the most helpful tool to prevent forgetting. Comparing with the *ft* baselines it gives important boosts of roughly 7-9%/3-8% in known/unknown tasks, for both sharing and not sharing layers. As in joint training, sharing layers does not have significant impact in this dataset. Similarly, weight regularization only brings marginal gains. In total, the best result for *txt2im* retrieval improves 9.5%/8% over the baseline for known/unknown tasks. For *im2txt* retrieval the improvement is 8.3%/4.4%. The results are still far from joint training, so there is space for improvement in future works.

7. Conclusion

In this paper we propose, to our knowledge, the first study on how forgetting affects multimodal embedding spaces, focusing on cross-modal retrieval. We propose a continual cross-modal retrieval model that emphasizes the important role of the indexing stage. Cross-modal drifts are also key factors in forgetting in cross-modal tasks. We evaluated several specific tools to alleviate forgetting.

Acknowledgements

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 2, 3, 5

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013. 2

[3] Bernard Ans and Stéphane Rousset. Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 320(12):989–997, 1997. 3

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 1, 2

[5] David Berga, Marc Masana, and Joost Van de Weijer. Disentanglement of color and shape representations for continual learning. *arXiv preprint arXiv:2007.06356*, 2020. 1

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994. 2

[7] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3594–3601. IEEE, 2010. 2

[8] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S Yu. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1445–1454. ACM, 2016. 2

[9] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[10] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[11] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. *arXiv preprint arXiv:2101.05068*, 2021. 1

[12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[13] Mohamed Elhoseiny, Scott Cohen, Walter Chang, Brian Price, and Ahmed Elgammal. Sherlock: Scalable fact learning in images. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1, 2

[14] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014. 2

[15] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 1, 2

[16] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*, pages 529–545. Springer, 2014. 2

[17] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 2

[18] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[19] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[21] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 2

[22] HAROLD HOTELLING. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 12 1936. 2

[23] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[24] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3232–3240, 2017. 2

[25] Nobuhiro Kaji and Hayato Kobayashi. Incremental skip-gram model with negative sampling. *arXiv preprint arXiv:1704.03956*, 2017. 3

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2, 3, 5

[28] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[29] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014. 2, 6

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 6

[31] Brian Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013. 2

[32] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019. 1

[33] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 1, 3

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[35] Xialei Liu, Marc Masana, Luis Herranz, Joost Van de Weijer, Antonio M Lopez, and Andrew D Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2262–2268. IEEE, 2018. 1, 2, 5

[36] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1040–1049, 2017. 2

[37] Xialei Liu, Chenshen Wu, Mikel Menta, Luis Herranz, Bogdan Raducanu, Andrew D. Bagdanov, Shangling Jui, and Joost van de Weijer. Generative feature replay for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1

[38] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[39] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation. *arXiv preprint arXiv:2010.15277*, 2020. 1, 2

[40] Marc Masana, Tinne Tuytelaars, and Joost van de Weijer. Ternary feature masks: continual learning without any forgetting. *arXiv preprint arXiv:2001.08714*, 2020. 1

[41] Chandler May, Kevin Duh, Benjamin Van Durme, and Ashwin Lall. Streaming word embeddings with the space-saving algorithm. *arXiv preprint arXiv:1704.07463*, 2017. 3

[42] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 1, 2

[43] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration. *IEEE Transactions on Multimedia*, 19(5):1100–1113, 2017. 2

[44] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 360–368, 2017. 2

[45] Anusha Nagabandi, Chelsea Finn, and Sergey Levine. Deep online learning via meta-learning: Continual adaptation for model-based rl. *arXiv preprint arXiv:1812.07671*, 2018. 3

[46] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017. 3

[47] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020. 1

[48] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 2

[49] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[50] Amal Rannen, Rahaf Aljundi, Matthew B Blaschko, and Tinne Tuytelaars. Encoder based lifelong learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1320–1328, 2017. 3

[51] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 3

[52] Anthony Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146, 1995. 3

[53] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 3

[54] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028, 2017. 2

[55] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2

[56] Jonathan Schwarz, Jelena Luketina, Wojciech M Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. *arXiv preprint arXiv:1805.06370*, 2018. 3

[57] Ari Seff, Alex Beatson, Daniel Suo, and Han Liu. Continual learning in generative adversarial nets. *arXiv preprint arXiv:1705.08395*, 2017. 3

[58] Joan Serra, Dídac Surís, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423*, 2018. 1, 3

[59] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2017. 1, 3

[60] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015. 2

[61] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 2

[62] Kai Wang, Luis Herranz, Anjan Dutta, and Joost van de Weijer. Bookworm continual learning: beyond zero-shot learning and continual learning. *arXiv preprint arXiv:2006.15176*, 2020. 1

[63] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2019. 1, 2, 3, 6

[64] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015. 2

[65] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[66] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[67] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 2

[68] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 2

[69] Chenshen Wu, Luis Herranz, Xialei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31:5962–5972, 2018. 1, 3

[70] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 2

[71] Shiqi Yang, Kai Wang, Luis Herranz, and Joost van de Weijer. Simple and effective localized attribute representations for zero-shot learning. *arXiv preprint arXiv:2006.05938*, 2020. 1

[72] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 3

[73] Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[74] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3987–3995. JMLR. org, 2017. 2, 5

[75] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 1

11