CL-CrossVQA: A Continual Learning Benchmark for Cross-Domain Visual Question Answering

Yao Zhang\textsuperscript{1}, Haokun Chen\textsuperscript{1,4}, Ahmed Frikha\textsuperscript{1,4}, Yezi Yang\textsuperscript{2}, Denis Krompass\textsuperscript{1,4}, Gengyuan Zhang\textsuperscript{1}, Jindong Gu\textsuperscript{3}, Volker Tresp\textsuperscript{1,4}
\textsuperscript{1} Institute of Informatics, LMU Munich \textsuperscript{2} Department of Informatics, Technical University of Munich \textsuperscript{3} Torr Vision Group, University of Oxford \textsuperscript{4} Corporate Technology, Siemens AG
yzhang@dbs.ifi.lmu.de, volker.tresp@siemens.com

Abstract

Visual Question Answering (VQA) is a multi-discipline research task. To produce the right answer, it requires an understanding of the visual content of images, the natural language questions, as well as commonsense reasoning over the information contained in the image and world knowledge. Recently, large-scale Vision-and-Language Pre-trained Models (VLPMs) have been the mainstream approach to VQA tasks due to their superior performance. The standard practice is to fine-tune large-scale VLPMs pre-trained on huge general-domain datasets using the domain-specific VQA datasets. However, in reality, the application domain can change over time, necessitating VLPMs to continually learn and adapt to new domains without forgetting previously acquired knowledge. Most existing continual learning (CL) research concentrates on unimodal tasks, whereas a more practical application scenario, i.e., CL on cross-domain VQA, has not been studied. Motivated by this, we introduce CL-CrossVQA, a rigorous Continual Learning benchmark for Cross-domain Visual Question Answering, through which we conduct extensive experiments on 4 VLPMs, 4 CL approaches, and 5 VQA datasets from different domains. In addition, by probing the forgetting phenomenon of the intermediate layers, we provide insights into how model architecture affects CL performance, why CL approaches can help mitigate forgetting in VLPMs to some extent, and how to design CL approaches suitable for VLPMs in this challenging continual learning environment. To facilitate future work on CL for cross-domain VQA, we will release our datasets and code.

1. Introduction

Visual Question Answering (VQA)\textsuperscript{3} is an exhilarating reasoning task at the crossroads of Natural Language Processing (NLP) and Computer Vision (CV), with the objec-

![CL-CrossVQA benchmark](image)

Figure 1. Illustration of CL-CrossVQA benchmark. Each VQA task is framed as a classification problem. The VQA system is trained on successive tasks in different domains and evaluated on all encountered tasks. The number of candidate answers, i.e., classes, is shown between the image and the question.
ationally intensive and requires a vast amount of training data, which might not be available in the application domain of interest. Moreover, the rapidly growing model size of VLPMs makes it impractical to store a new model for each incoming task. Thus, VLPMs are required to continually learn new concepts without forgetting previously acquired knowledge. Concretely, VLPMs need to maintain a balance between plasticity, i.e., the ability to acquire new knowledge, and stability, i.e., the ability to retain prior knowledge, which is known as the stability-plasticity dilemma [35].

In contrast to CV and NLP, CL is still underexplored in the field of Vision-and-Language (V-L) learning. While a few works [18, 28] address VQA in a continual learning environment, they focus on single-domain VQA tasks. However, the more practical application scenario, i.e., CL on cross-domain VQA, has not been studied yet.

To systematically investigate how large-scale VLPMs learn a stream of cross-domain VQA tasks, we construct CL-CrossVQA, a Continual Learning benchmark for Cross-domain Visual Question Answering, which covers 4 large-scale VLPMs, 4 CL approaches, and 5 cross-domain VQA datasets. Through comprehensive analysis of a variety of VLPMs and CL methods, we investigate how model design along multiple dimensions, e.g., text encoder, vision encoder, multimodal fusion module, affects CL performance, and which type of CL approach is more effective for VLPMs. Additionally, by dissecting the inner architecture of VLPMs using intermediate representation probing, we provide insights into why CL approaches can help mitigate forgetting in VLPMs to some extent, and how to design CL approaches suitable for VLPMs in this challenging, yet practical continual learning environment.

Our key observations are as follows: (1) All VLPMs suffer from severe forgetting. Dual-stream encoder-decoder architecture exhibits alleviation of forgetting compared to single-stream encoder-only models; (2) For single-stream encoder-only models, the deeper layers are most affected by forgetting. For the dual-stream encoder-decoder model, the multimodal fusion module is less affected by forgetting; (3) Different VLPMs require different CL approaches to tackle the stability-plasticity dilemma. Replay-based approaches are the most effective for VLPMs and are less sensitive to task orders; (4) For replay-based approaches, the sampling strategy used for the replay buffer has a great impact on CL performance; (5) Adapter [20] is a strong competitor in this cross-domain multimodal continual learning scenario.

2. Related Work

Our work lies at the intersection of VLPMs, VQA, cross-domain learning, and CL. We provide here a compendious overview of the relevant background.

| Model       | Vision Encoder | Text Encoder | Modality Fusion Module | Decoder | Pre-training Dataset Size |
|-------------|----------------|--------------|------------------------|---------|---------------------------|
| ViLT [24]   | Patch Emb.     | BERT         | Single-stream          |         | 5M                        |
| VaulLT [12] | Patch Emb.     | BERT         | Single-stream          |         | 5M                        |
| FLAVA [42]  | ViT            | ViT          | Single-stream          |         | 70M                       |
| ALBEF [29]  | ViT            | BERT         | Dual-stream            | ✓       | 14M                       |

Table 1. Comparison of evaluated VLPMs.

2.1. Vision-and-Language Pre-trained Models

In recent years, pre-trained models have progressed at a dizzying speed in tandem with the evolution of transformers. Due to the significance of single-modal language/vision pre-training, some pioneering works [11, 30, 34, 44–46] have recently attempted to explore the joint representation of language and vision by pre-training large-scale models on vision and language modalities, which are referred to as Vision-and-Language Pre-trained Models (VLPMs). There are three primary components in VLPMs, namely vision encoder (VE), text encoder (TE), and modality fusion module (MF). VE and TE are pre-trained with images and texts, respectively. MF is pre-trained using image-text pairs and amalgamates the output embeddings of VE and TE. While almost all VLPMs extract text embeddings from a pre-trained language model, e.g., BERT [14], the pre-trained model used on the vision side is different. Unlike most existing VLPMs, which are built upon region-level features extracted from pre-trained object detectors, e.g., Faster RCNN [38]. All VLPMs evaluated in our work directly utilize patch features without bounding box annotations, which allows them to escape from capacity limitations due to the imperfections of the object detectors.

To thoroughly investigate the effects of model design in a CL setting, we evaluate 4 different types of VLPMs, which are summarized in Table 1. Specifically, ViLT [24] directly feeds image patch features and text token embeddings into a pre-trained ViT [15] model that learns the vision-language alignment with self-attention across both modalities. To address the impoverished language representations of ViLT, VaulLT [12], an extension of ViLT, propagates the output representations of a large language model, i.e., BERT, to the language input of ViLT. FLAVA [42], a foundational language and vision alignment model, learns strong representations through joint pre-training on both unimodal and multimodal data while encompassing cross-modal alignment objectives and multimodal fusion objectives. FLAVA adopts the ViT architecture for the visual encoder, text encoder, as well as multimodal encoder. Unlike the aforementioned VLPMs, FLAVA is pre-trained on Public Multimodal Datasets [42], which consist of 70M image-text pairs. As opposed to all aforementioned single-stream VLPMs, which learn one joint V+L representation, ALBEF [29] adopts dual-stream architecture to map vision and language embeddings into the same semantic space. Specifically, ALBEF employs two separate transformers for im-
ages, *i.e.*, ViT, and texts, *i.e.*, BERT, and introduces a contrastive loss to align the image and text representation before fusing them through cross-modal attention. Furthermore, all aforementioned VLPMs adopt the encoder-only architecture, where the cross-modal representations are directly fed into an output layer to generate the final outputs. ALBEF, on the other hand, utilizes an encoder-decoder architecture, where cross-modal representations are first fed into a decoder and then into an output layer.

### 2.2. Cross-Domain Learning in VQA

The cross-domain learning remains underinvestigated in VQA, only a few works research in this direction: [8] found that domain shift across VQA datasets mostly lies in their questions and answers and proposed an algorithm to align the source and target feature embeddings. [47] trained an additional domain discriminator to adversarially penalize the mismatch between the multi-modal embedding from the source domain and the target domain. [50] focused on domain shift in the visual space and proposed a domain-adaptive visual feature extractor to better capture image semantics from different domains. [7] generated synthetic images and incorporated them in the training by object-level feature swapping across domains.

### 2.3. Continual Learning

The main goal of continual learning research is to mitigate the problem of catastrophic forgetting (CF), *i.e.*, the tendency of neural networks to forget existing knowledge when learning new tasks with novel input patterns. To tackle this problem, numerous studies have been conducted in different directions.

#### Rehearsal-based Methods

To mitigate forgetting, Replay methods [10, 21, 36, 39] store and replay samples from previous tasks when learning new tasks. Dark Experience Replay (DER and DERPP) [6] expands ER [39] by distilling the dark knowledge from models on previous tasks using replayed samples. Gradient Episodic Memory (GEM) [33] and its lightweight variant A-GEM [9] constrain the model to update on new tasks without interfering with the previous tasks. Instead of storing samples from previous tasks, pseudo replay methods [4, 26, 41] optimize a generative network to produce synthetic samples depicting the knowledge of previous tasks.

#### Regularization-based Methods

Regularization-based methods focus on estimating a distribution over the model parameters, and use prior when learning new tasks. Elastic Weight Consolidation (EWC) [25] is the first to estimate the importance of network parameters and penalize parameter drifting accordingly when learning new tasks. Instead of measuring the parameter importance after each task, Syntactic Intelligence (SI) [48] extends EWC by maintaining an online importance estimation. There are also methods that obtain priors by conducting unsupervised estimation [2], knowledge distillation [23, 31, 49] or incrementally merging Gaussian posteriors for the task parameters [27].

### 3. CL-CrossVQA Benchmark

We first illustrate the CL-CrossVQA benchmark, then we introduce the metrics to measure the knowledge preservation and transfer abilities of different CL models.

#### 3.1. Task Formulation

Given an image-question pair, the VQA system is required to continually adapt to new domains. In this paper, we consider a more practical CL setting, in which a kid obtains question types is beneficial to computational models. [28] reorganizes the existing VQA dataset to construct CLOVE, which contains two CL setting *i.e.*, scene-incremental and function-incremental, to test the ability of the model to adapt to new scenes (*e.g.* ShopAndDining, Workplace *etc.*) and acquire new features (*e.g.* object recognition, attribute recognition *etc.*). In addition to these VQA benchmarks, a multimodal multi-task CL benchmark CLI MB [43] has recently been introduced. In this benchmark, VLPMs are trained on a sequence of different V+L tasks. In this paper, we consider a more practical CL setting, in which a VQA system is required to continually adapt to new domains.
could be different for different tasks. The goal is to optimize a set of parameters $\Theta$ of the learner towards an objective on each task $T_k$.

3.2. Datasets

CL-CrossVQA consists of 5 datasets across different domains, namely the abstract domain, i.e., VQA Abstract [1], the general domain, i.e., Toronto COCO QA [37], the artistic domain, i.e., AQUA [17], the pathological domain, i.e., PathVQA [19] and the radiological domain, i.e., VQA-Med-2019 [5]. Examples from these datasets are provided in Fig. 1.

VQA Abstract consists of images of abstract scenes. Each image is accompanied by three questions, with 10 human-provided ground truth answers, ensuring a diverse and interesting set of questions and answers. By its abstraction, it requires high-level reasoning skills for VQA.

COCO QA contains questions automatically generated from captions in Microsoft COCO dataset [32] and therefore may contain syntactic and semantic errors. There are in total 78,736 training questions and 38,948 test questions, of which the answers are all one word.

AQUA Artwork has been an important part of human history. Different styles of artwork such as naturalism, realism, and cubism represent different levels of abstraction, posing an extra challenge for VQA system. Aiming to investigate this, AQUA is built on the SemArt dataset [16] and its QA pairs are generated from the associated comments of the paintings. We use the Visual QA subset from AQUA, including 29,568, 1,507, and 1,270 question-answer (QA) pairs for training, validation, and testing, respectively.

PathVQA is created with the goal to develop an AI pathologist and foster research in pathological VQA. It consists of 32,799 question-answer pairs automatically generated from captions of 4998 pathology images in pathology textbooks and online digital libraries. We note that the answers in PathVQA can be long, clinical expressions, adding an extra challenge for VQA system. To analyze this, PathVQA is built on the SemArt dataset [16] and its QA pairs are generated from the associated comments of the paintings. We use the Visual QA subset from AQUA, including 29,568, 1,507, and 1,270 question-answer (QA) pairs for training, validation, and testing, respectively.

PathVQA-Med-2019 aims to support clinical education, clinical decision, and patient education. It focuses on radiology images selected from the MedPix database and the QA pairs are generated based on patterns from naturally asked and validated questions by medical students. There are 3,200 images and 12,792 QA pairs for the training set and 500 images with 2,000 QA pairs for the validation set. The test set includes 500 images with 500 questions.

To avoid bias caused by different dataset sizes, we downsample the training sets of all datasets to be consistent with VQA-Med-2019 while maintaining their original class distributions. The same train-test split ratio is adopted for all datasets. Detailed dataset pre-processing is provided in the Appendix. To better reflect the challenging real-world scenarios, the selected cross-domain datasets depict minor overlap in the answer space, i.e., only a few answers are shared across datasets.

3.3. Evaluation Metrics

To evaluate the knowledge transfer ability of the models and their possible catastrophic forgetting of the previously learned tasks, we assume access to the test set of each task and report the following three evaluation metrics [33]:

Forward Transfer

$$FWT = \frac{1}{T-1} \sum_{i=1}^{T-1} S_{i,i+1} - \tilde{b}_{i+1},$$

Backward Transfer

$$BWT = \frac{1}{T-1} \sum_{i=1}^{T-1} S_{T,i} - S_{i,i},$$

Average Accuracy

$$Acc = \frac{1}{T} \sum_{i=1}^{T} S_{T,i},$$

where $S_{i,i}$ is the evaluation score of the model on the test set of $j$-th task after training on $i$-th task, and $\tilde{b}_i$ is the evaluation score of the pre-trained model, i.e., model without any specific fine-tuning for VQA tasks, on the test set of $i$-th task. Specifically, we feed the task-specific answer list into the decoder of pre-trained ALBEF to compute $\tilde{b}_i$, while for the encoder-only backbones, we attach a randomly initialized classification head on the top for each task.

4. Experiments

We provide a comprehensive analysis of VLPMs with our CL-CrossVQA benchmark. Concretely, we investigate a CL baseline and 4 CL approaches from different categories: (1) Sequential uses the model learned on previous tasks as initialization and then optimizes the parameters for the current task. (2) ER [39] is a rehearsal-based method, which stores a portion of samples from the previous task and replays them regularly with the training batch of the current task. (3) DER [6] is an extension of ER, which additionally stores the model prediction of the replayed samples at the end of the previous task as dark knowledge regularization. (4) DERPP [6] combines ER and DER and is a strong hybrid method. (5) EWC [25] is a regularization-based method, which penalizes the drifting of important network parameters when training on new tasks.

To analyze the effects of VLPM architecture in the CL setting, we train 4 representative VLPMs with the aforementioned CL methods using our benchmark. To handle the distribution shift in the ground-truth answers of different tasks, we incorporate task-specific classification heads with the encoder-only backbones, i.e., ViLT, VAlLT, and FLAVA. For the encoder-decoder leaner, i.e., ALBEF, we implement a static model architecture for all tasks and utilize different answer lists for each task. The hyperparameters used in the training are detailed in the Appendix.
## 4.1 Results and Discussion

In Table 2, we provide the evaluation results of the selected VLPMs in our CL-CrossVQA benchmark. We report three metrics, i.e., average accuracy (Acc), backward transfer (BWT), and forward transfer (FWT). All experiments are conducted with 6 randomly sampled task orders. In the following, we discuss our findings.

### All VLPMs evaluated suffer from severe catastrophic forgetting.

Comparing the backward transfer of all CL methods, we observe that sequential fine-tuning of all VLPMs results in the largest forgetting, indicating that the selected VLPMs cannot preserve the knowledge learned on the previous tasks. Therefore, continually fine-tuning VLPMs without applying any CL approaches would result in a considerable performance loss.

### Replay-based approaches are the most effective for VLPMs.

Comparing the different CL methods, we can observe that replay-based CL approaches, i.e., ER, DER, and DERPP, achieve higher accuracy and higher backward transfer than the regularization-based method, i.e., EWC, for all VLPMs. ER, and DERPP, which adopts both regularization and experience replay, achieve comparable results on our benchmark. Note that, we only use 1% of the training data, which implies that a small number of replayed data is already able to alleviate forgetting to a large extent. Despite the careful tuning of the regularization coefficient, EWC performs as poorly as the sequential fine-tuning baseline, demonstrating that EWC is not suitable for VLPMs in the cross-domain continual learning setting.

### Dual-stream encoder-decoder architecture is more resistant to forgetting.

ALBEF shows the highest CL performance among all VLPMs, suggesting that the dual-stream encoder-decoder model is significantly more resistant to forgetting than the encoder-only model. Note that, ALBEF allows one to perform continual learning without adding new task-specific parameters. However, for single-stream models, new classification heads have to be added each time we train on a new task since each task in our benchmark has a different number of classes. Such differences may partially explain why ALBEF is more suitable for CL. Also, ALBEF produces the highest FWT score among all backbones, which further indicates the benefit of applying the same parameters for all tasks. In order to examine whether adding task-specific parameters is responsible for the low performance of single-stream models, we take steps toward understanding why ALBEF seems to be resistant to forgetting by probing the inner layer of ALBEF and comparing it to the single-stream representative VLPM, i.e., ViLT in Section 5.

### Impact of task orders.

To investigate how task orders affect CL performance, we evaluate 6 task orders selected at random from all possible combinations. We find that replay-based approaches are less sensitive to task orders in comparison to regularization-based approaches (see standard errors in Table 2). Among all VLPMs, ALBEF is more robust to task orders using the best-performing CL approaches, i.e., ER, and DERPP (see Fig. 2).

### Table 2: Evaluation of selected VLPMs combined with different CL approaches.

| Model      | ViLT | VAutoT | FLAVA | ALBEF |
|------------|------|--------|-------|-------|
|            | Acc  | BWT    | FWT   | Acc   | BWT    | FWT   | Acc   | BWT    | FWT   |
| Sequential | 26.82±0.29 | -42.49±3.41 | -0.05±0.06 | 26.21±0.79 | -42.58±3.64 | -0.14±0.24 | 26.61±0.27 | -34.31±2.36 | -0.02±0.40 |
| ER         | 54.15±1.36 | -12.38±1.82 | 0.03±0.06 | 51.51±0.91 | -12.67±1.07 | 0.08±0.12 | 44.52±0.80 | -10.53±1.14 | -0.11±0.18 |
| DER        | 51.42±1.71 | -12.56±1.58 | -0.15±0.38 | 49.35±1.29 | -14.86±1.63 | -0.15±0.41 | 44.52±1.09 | -10.91±2.75 | -0.07±0.16 |
| DERPP      | 54.21±1.31 | -12.34±1.70 | 0.00±0.28 | 51.30±1.03 | -13.08±1.12 | 0.23±0.31 | 44.52±1.57 | -11.30±2.64 | -0.06±0.33 |
| EWC        | 26.94±2.75 | -42.86±3.96 | 0.03±0.06 | 25.43±3.79 | -43.64±4.02 | -0.03±0.00 | 25.83±4.89 | -34.09±5.87 | -0.01±0.18 |

Table 2. Evaluation of selected VLPMs combined with different CL approaches. Average accuracy (Acc), backward transfer (BWT), and forward transfer (FWT) is reported. 6 task orders are adopted to produce the mean ± standard deviation. All models are fine-tuned for a maximum of 15 epochs. We apply a random sampling strategy for ER, DER, and DERPP. For all replay-based approaches, the memory size is set to 1% of training data. We tune the weight for logits distillation and the weight for experience replay of DERPP by grid search in {0.1, 0.5, 1, 10, 100}. We tune the regularization coefficient of EWC by grid search in {0.1, 1, 10, 100, 1000, 10,000, 100,000, 1,000,000}. 
Impact of memory buffer. To investigate the effect of the answer distribution in the replay buffer on the CL performance, we compare two different methods for sampling replay instances: random sampling, i.e., sampling all instances from the previous tasks with equal probability, and balanced sampling, i.e., memory buffer exhibits a class-balanced distribution regardless of the ground-truth class distribution in previous answer lists. The evaluation results are shown in Fig. 3. We observe a clear performance gap between the two sampling strategies for all 3 CL methods, indicating that replaying samples based on the class distribution of previous tasks is beneficial for continual learning. It is also important to note that this effect of the replay buffer on CL performance does not seem to be architecture-dependent. This shows that the selection of sampling strategy has a great impact on CL performance, and improving the sampling strategy could be one direction to obtain better CL performance.

Comparison with Adapter In addition to conventional CL approaches, we also conduct experiments with Adapter [20], which adds new trainable modules between the intermediate layers of VLPMs. We implement the same adapter architecture for two representative VLPMs, i.e., ALBEF and ViLT. We freeze the parameters of all layers and continually optimize the task-specific adapters and the classification head for ViLT. As shown in Table 3 and Table 4, Adapter enables both models to learn new tasks without any performance degradation on the previously learned tasks, i.e., no catastrophic forgetting. Adapter is beneficial for the single-stream model as it boosts performance by ca. 11% accuracy while being ca. 3 times more computationally efficient. In contrast, for the dual-stream model, Adapter leads to a drop of 8% in retained accuracy compared to the best-performing CL approach. Therefore, Adapter is a strong competitor for the more computationally expensive CL methods.

5. Model Analysis

In this section, we provide a comprehensive analysis of the impact of each individual module of VLPMs when trained in a CL environment. Specifically, we investigate the influence of vision and text encoders in Section 5.1, and examine the impact of multimodal fusion designs in Section 5.2. Finally, we provide some Grad-CAM visualizations for VQA after training on each task in Section 5.3. For ablation and analysis, we mainly focus on ViLT and ALBEF. The task order we use is illustrated in Fig. 1.

5.1. Impact of Vision and Text Encoder

In this section, we focus exclusively on vision and text encoders adopted in ALBEF. In order to further investigate the roles of the individual modules of ALBEF, we design the following experiments: 1) Freeze VE/TE: we freeze the pre-trained vision encoder (or text encoder) and fine-tune the rest; 2) Freeze VE+TE: we freeze both vision and text encoders, and fine-tune the rest; 3) Freeze B9: besides vision and text encoder, we also freeze the bottom 3 layers of the multimodal encoder; 4) Freeze B12: we freeze the full encoder, and only fine-tune the decoder.

As shown in Table 4, the more modules are frozen, the less accurate the model is. Freeze B12 has the lowest ac-

---

**Table 3. Comparison of Adapter and the best-performing CL approach for ViLT.**

| Method  | Trainable Parameters | Acc | BWT | FWT |
|---------|---------------------|-----|-----|-----|
| DERPP   | 125.97M (100.00%)   | 54.74 | -12.22| 0.11 |
| Adapter | 42.72M (33.91%)     | 65.10 | /   | /   |

**Table 4. Model analysis of ALBEF with partial freeze.**

| Method   | Trainable Parameters | Acc | BWT | FWT |
|----------|---------------------|-----|-----|-----|
| Seq.     | 290.34M (100.00%)   | 43.32 | -30.16| 10.16 |
| ER       | 290.34M (100.00%)   | 60.19 | -10.73| 12.71 |
| Freeze VE| 204.25M (70.35%)    | 38.50 | -36.12| 11.53 |
| Freeze TE| 223.97M (77.14%)    | 39.15 | -36.73| 11.60 |
| Freeze VE+TE| 137.88M (47.49%) | 40.57 | -34.20| 11.42 |
| Freeze B9| 109.53M (37.72%)    | 38.72 | -34.10| 11.35 |
| Freeze B12| 81.17M (27.96%)    | 34.69 | -33.23| 8.46  |
| Adapter  | 85.67M (29.51%)     | 52.44 | /   | /    |
accuracy, reflecting that due to its low plasticity, *Freeze B12* prevents its feature extractor from adapting to new domains by learning new representations. On the other hand, thanks to its high stability, it suffers less from forgetting. Besides, its backward transfer performance indicates that the decoder also suffers from severe forgetting. Freezing the vision encoder or text encoder results in a similar performance to freezing the multimodal encoder.

5.2. Analysis of Multimodal Fusion Module

In this section, we investigate the impact of *multimodal fusion module* on CL performance.

5.2.1 Forgetting Analysis at Representation Level

Following [13], we first compute the average embedding for each class using the ground-truth label of all embeddings. Afterward, we re-classify all the embeddings by assigning the label of the closest class-wise embedding prototype and report the classification accuracy. We use the test set of each task and assume the ground-truth labels are available. Here, we provide the probing results for all 12 layers of ViLT as well as 6 layers in the multimodal encoder of ALBEF. We report the *representation forgetting*, i.e., the average accuracy drop from the best performance on previous tasks, in Fig. 4. Here, numbers smaller than zero mean there is positive knowledge transferred and benefits the performance of previous tasks. In Fig. 5, we provide the *representation accuracy* of different methods and discuss our findings in the following section.

5.2.2 Results and Discussions

The deeper layers of the single-stream model are most affected by forgetting. For single-stream architecture ViLT, the amount of representation forgetting increases with the increase of layers (see Fig. 4a). For the lower layers, the representation forgetting stays relatively flat or decreases
slightly, showing that the bottom layers barely suffer from forgetting and a positive backward transfer is occurring. However, the representation forgetting increases significantly at layer 7, and then the deeper the layer, the more representation forgetting we observe. This phenomenon may be due to the fact that in the lower layers, the model only learns general knowledge. However, the model tends to learn more task-specific features in the upper layers, which are easily forgotten and corrupted when constantly learning new tasks. Moreover, we find that the representation accuracy of the last layer is lower than the second last layer (see Fig. 5a and Fig. 5c), suggesting that applying another classification method, e.g., metric learning, on the representations extracted from the second to last layer might be a more viable approach to adapt the single-stream models to the CL environment. We assume that this is caused by the introduction of a task-specific classifier, which harms the representation ability of the final layer embedding when switching between tasks.

**ER helps align the single-stream encoder with the task-specific classifier.** Comparing Fig. 4a with Fig. 4b, we notice that the representation forgetting of ER is similar to that of the sequential fine-tuning baseline, however, ER achieves much higher backward transfer than the baseline (see Table 2). We believe that ER and other replay-based approaches may assist in aligning the encoder with the task-specific classifier. In particular, it is worth noting that, although the forgetting in terms of the traditional measure is high for EWC (see Table 2), the representation forgetting shown in Fig. 4c indicates that EWC is more effective at mitigating forgetting of the middle layers. However, it lacks the ability to constrain the classifier when a task-specific classifier needs to be utilized.

**Fusion module of a dual-stream model forgets less than that of a single-stream model.** Comparing Fig. 4b Fig. 4c with Fig. 4e and Fig. 4f, we observe that ALBEF shows much milder representation forgetting than ViLT, which is also evidenced by the best CL performance achieved by ALBEF (see Table 2). Besides, we notice that across all CL methods, the representativeness of the layers within the multimodal encoder of ALBEF does not vary as dramatically as ViLT, possibly because the forgetting of ALBEF occurs mainly in the vision encoder, text encoder, and the decoder, which is also evidenced by Table 4. Also, as shown in Fig. 5d, the multimodal encoder of ALBEF extracts representations with higher separability.

**5.3. Visualization**

Fig. 6 shows the Grad-CAM [40] visualization changes of the continually adapted VLPM on a selected image from the first task. Following [29], we compute the Grad-CAM on the cross-attention maps in the 3rd layer of the multimodal encoder in ALBEF for the VQA task. From the first row in Fig. 6, we see that the model attention shifts from the target object to almost every possible object existing in the image after being trained on 2 medical datasets, which we assume are due to the large domain shift in the visual aspect. Fig. 6d shows that the model attention again focuses on a certain object but fails to target the correct one, which is a predictable behavior since, in the AQUA dataset, objects are also abstractions of real-life objects. The second row of Fig. 6 shows the same Grad-CAM visualization of ALBEF optimized with ER. We notice that by replaying instances from the VQA Abstract dataset, the model maintains the ability to distinguish abstract objects and connect them with the questions while learning new tasks.

**6. Conclusion**

In this paper, we present CL-CrossVQA, the first VQA benchmark for cross-domain continual learning. We systematically investigate which type of CL method is the most effective for VLPMs, and how model design impacts CL performance. Our comprehensive experimental results reveal that: 1) Dual-stream encoder-decoder model exhibits
alleviation of forgetting; (2) The deeper layers of single-stream encoder-only models are most affected by forgetting; 3) Replay-based approaches are the most effective and are less sensitive to task orders; 4) Adapter is a promising approach in this cross-domain continual learning scenario. We believe that the insights gained by dissecting the inner architecture of VLPMs will inform the development of CL approaches suitable for multimodal continual learning.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2015. 3, 4
[2] 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. 3
[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015. 1
[4] Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony Robins. Pseudo-recursral: Solving the catastrophic forgetting problem in deep neural networks. arXiv preprint arXiv:1802.03875, 2018. 3
[5] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vquad: Overview of the medical visual question answering task at imageclef 2019. In CLEF 2019 Working Notes, CEUR Workshop Proceedings, Lugano, Switzerland, September 9–12 2019. CEUR-WS.org <http://ceur-ws.org/>. 4
[6] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. Advances in neural information processing systems, 33:15920–15930, 2020. 3, 4
[7] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio S Feris, and Vicente Ordonez. Simvqa: Exploring simulated environments for visual question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5056–5066, 2022. 2
[8] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5716–5725, 2018. 3
[9] Arslan Chaudhry, Marc’Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with agem. arXiv preprint arXiv:1812.00420, 2018. 3
[10] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. 2019. 3
[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. 2
[12] Georgios Chochlakis, Tejas Srinivasan, Jesse Thomason, and Shrikanth Narayanan. Vault: Augmenting the vision-and-language transformer with the propagation of deep language representations. arXiv preprint arXiv:2208.09021, 2022. 2
[13] Mohammad Reza Davari, Nader Asadi, Sadhir Mudur, Rahaf Aljundi, and Eugene Beilislovsky. Probing representation forgetting in supervised and unsupervised continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16712–16721, 2022. 7
[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 2
[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 2
[16] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. 2018. 4
[17] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286, 2020. 3
[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, pages 2790–2799. PMLR, 2019. 2, 6
[19] David Isele and Akansel Cosgun. Selective experience replay for lifelong learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018. 3
[20] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2901–2910, 2017. 3
[21] Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. arXiv preprint arXiv:1607.00122, 2016. 3
[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In International Conference on Machine Learning, pages 5583–5594. PMLR, 2021. 2
[25] 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. 3, 4

[26] Frantizska Lavda, Jason Ramapuram, Magda Gregorova, and Alexandros Kalousis. Continual classification learning using generative models. *arXiv preprint arXiv:1810.10612*, 2018. 3

[27] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017. 3

[28] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. Symbolic replay: Scene graph as prompt for continual learning on vqa task. *arXiv preprint arXiv:2208.12037*, 2022. 2, 3

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 8

[30] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[31] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 3

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 4

[33] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017. 3, 4

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[35] Martial Merimlodd, Aurelia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects, 2013. 2

[36] 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

[37] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[38] Shaqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[39] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 4

[40] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. 8

[41] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 3

[42] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[43] Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *arXiv preprint arXiv:2206.09059*, 2022. 3

[44] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[45] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[46] Zirui Wang, Jiuhui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[47] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Michael Cogswell. Dynamic-glm: A generative model for the visual world. *Advances in neural information processing systems*, 34:9706–9717, 2021. 2

[48] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3995. PMLR, 2017. 3

[49] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafetin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020. 3

[50] Mingda Zhang, Tristan Maimdent, Ahmad Diab, Adriana Kovashka, and Rebeca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7046–7056, 2021. 3