Task Formulation Matters When Learning Continuously: A Case Study in Visual Question Answering

Anonymous ACL submission

Abstract

Continual learning is a promising alternative to the current pretrain-and-finetune paradigm: It aims to learn a model on a sequence of tasks without forgetting knowledge from preceding tasks. We investigate continual learning for Visual Question Answering and show that performance highly depends on task design, order, and similarity – where tasks may be formulated according to either modality. Our results suggest that incremental learning of language reasoning skills (such as questions about color, count etc.) is more difficult than incrementally learning visual categories. We show that this difficulty is related to task similarity, where heterogeneous tasks lead to more severe forgetting. We also demonstrate that naive fine-tuning of pretrained models is insufficient, and recent continual learning approaches can reduce forgetting by more than 20%. We propose a simple yet effective PSEUDO-REPLAY algorithm, which improves results while using less memory compared to standard replay. Finally, to measure gradual forgetting we introduce a new metric that takes into account the semantic similarity of predicted answers.

1 Introduction

The standard paradigm for Vision+Language (V+L) problems is to pretrain large-scale models, which are then finetuned and evaluated on independent and identically distributed (i.i.d.) data. In practice, the i.i.d. assumption often does not hold: New data becomes available sequentially, which often results in a change of data distribution. This is referred to as a new 'task' by the continual learning literature (Biesialska et al., 2020). Under this setting, continuously adapting an existing model via finetuning will lead to catastrophic forgetting, i.e. significant performance degradation on previous data (McCloskey and Cohen, 1989; Ratcliff, 1990). Continual learning provides a counterpart to i.i.d. learning by defining a class of algorithms aiming at incremental learning without forgetting. This line of work becomes increasingly relevant given the financial and environmental costs of (re-) training large models (Strubell et al., 2019; Bender et al., 2021), and the inability of static models to adequately generalize in a dynamic world (Lazaridou et al., 2021). While continual learning has been widely studied in the computer vision community, its use within V+L problems remains underexplored – with a few notable exceptions (Greco et al., 2019; Nguyen et al., 2019b; Hayes et al., 2020; Jin et al., 2020; Del Chiaro et al., 2020).

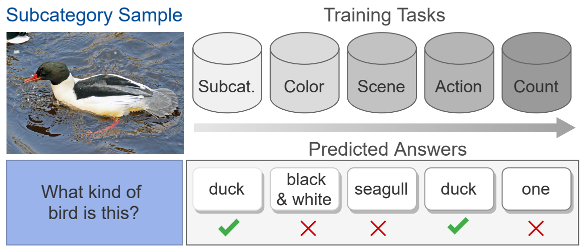


Figure 1: Predicted answers as the model continuously learns a sequence of tasks. Catastrophic forgetting causes incorrect predictions for preceding tasks.

V+L applications are a particularly challenging setting for continual learning since tasks can be formulated according to each modality. In particular, task definitions for Visual Question Answering (VQA) can either be based on the language reasoning skills (as defined by the question type, cf. Figure 1) or the objects in the image (Whitehead et al., 2021). While there is increasing evidence that continual learning performance is highly dependent on the task formulation, i.e. task design, order, and similarity (Van de Ven and Tolias, 2019; Yoon et al., 2020; Delange et al., 2021), tasks are often formulated in an ad-hoc fashion and vary widely for each application and dataset.

This paper addresses this challenge by conducting a systematic study on how different task formulations impact performance and forgetting in

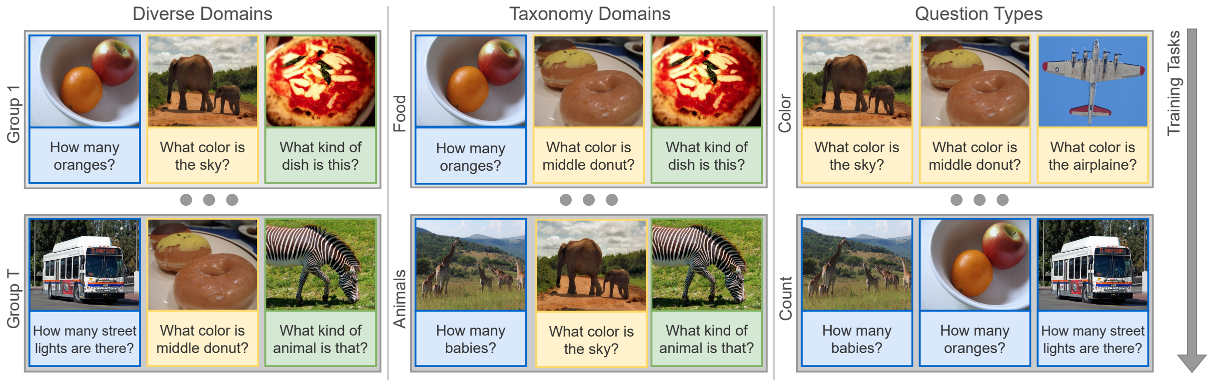


Figure 2: Tasks in continual VQA learning can be based on visual content, e.g. object categories split into into *Diverse* groups (left), or according to a *Taxonomy* such as 'food' items, 'animals' etc. (middle); Tasks can also be based on *Question Types* representing different reasoning skills, such as color recognition or counting (right).

VQA. We introduce three settings based on the VQA-v2 dataset (Goyal et al., 2017) as illustrated in Figure 2 – two defined by visual objects and one by reasoning skills as determined by the questions. We first characterize the difficulty of each setting by studying pairwise task relationships and relate the amount of forgetting, i.e. the accuracy decrease on the previous task, to task similarity. Our results show that dissimilar tasks exhibit more severe forgetting. We then evaluate several regularization and memory-based continual learning methods using randomly initialized and pretrained models across our three settings. Based on the observation that approaches which store samples from previous tasks in their 'memory' perform reliably well, we propose a simple yet effective PSEUDO-REPLAY algorithm that combines data augmentation and distillation for greater memory efficiency and better privacy. We also introduce a new metric, termed Semantic Backward Transfer, which penalizes semantically similar answer changes less than nonsensical ones. Finally, we demonstrate that task order leads to high performance variance per question type and analyze how representations from each modality change during continual learning.

2 Problem formulation

In continual learning, model parameters $\boldsymbol{\theta}$ are incrementally updated as new data become available. We assume that samples from tasks $t = 1 \ldots T$ arrive sequentially as $D_t = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_t}$, where N_t is the number of data for task t. Following previous work, VQA is formulated as a multi-label classification problem with soft targets $\boldsymbol{y}_i$ (Anderson et al., 2018). Starting from parameters $\boldsymbol{\theta}_{t-1}$ of the previous model, the updated parameters $\boldsymbol{\theta}_t$ are obtained by training on the new data D_t. Some approaches also use a memory M_t containing a subset of samples from previous tasks, e.g. $D_1, \ldots, D_{t-1}$. In our setup, all tasks share a common output head which is extended with new classes from each task. This allows inference to be task-agnostic but creates a more challenging setting than multi-head learning where separate heads are learned for each task (Hussain et al., 2021). At the end of the training sequence, the objective is to achieve strong performance across all tasks observed so far. This objective encloses two challenges: 1) minimizing catastrophic forgetting of tasks seen earlier in training, 2) facilitating positive transfer to improve performance on new tasks (Hadsell et al., 2020).

Setting	Task	Train	Val	Test	Classes
Diverse	Group 1	44254	11148	28315	2205
	Group 2	39867	10202	22713	1874
	Group 3	37477	9386	23095	1849
	Group 4	35264	8871	22157	2119
	Group 5	24454	6028	14490	1777
Taxonomy	Animals	37270	9237	22588	1331
	Food	26191	6612	15967	1365
	Interior	43576	11038	26594	2096
	Sports	32885	8468	19205	1471
	Transport	41394	10280	25416	1954
Question	Action	18730	4700	11008	233
	Color	34588	8578	21559	92
	Count	38857	9649	23261	42
	Scene	25850	6417	14847	170
	Subcategory	22324	8578	21559	659

Table 1: Statistics per task within each setting.

3 Settings for Continual VQA

We define three continual learning settings, which include different task splits, as summarized in Table 1 and illustrated in Figure 2.

3.1 Visual Settings

We design two settings based on visual object categories. We take advantage of the fact that images in the VQA-v2 dataset originate from the COCO dataset (Lin et al., 2014) which provides object-level image annotations. Following previous work in image captioning (Del Chiaro et al., 2020), we organize 50 object categories into five groups. Images with objects from multiple groups are discarded in order to create clean task splits D_t – resulting in a total of 181K train, 45K validation, and 110K test samples.

For the first setting, *Diverse Domains*, tasks are defined by grouping the object categories randomly. Each task is assigned a balanced count of 10 distinct objects resulting in five tasks. This type of setting corresponds to common practice of continual learning research within computer vision (Rebuffi et al., 2017; Lomonaco and Maltoni, 2017), and reflects a real-world scenario where sequential data do not necessarily follow a taxonomy.

The second setting, *Taxonomy Domains* groups objects based on their common super-category as in (Del Chiaro et al., 2020). This results in five tasks: Animals, Food, Interior, Sports, and Transport. Note that the number of object classes per task under this definition is unbalanced since splits depend on the size of the super-category. More details on each task can be found in Appendix A.

3.2 Language Setting

We create a third setting *Question Types*, where each task corresponds to learning to answer a different category of questions. We use a classification scheme developed by Whitehead et al. (2021) to form a sequence of five tasks: Count, Color, Scene-level, Subcategory, and Action recognition. The splits for Count, Color, and Subcategory questions are obtained from Whitehead et al. (2021). We create two additional tasks from the remaining questions. In particular, we cluster question embeddings from Sentence-BERT (Reimers and Gurevych, 2019) [1] so that each cluster has at least 15 questions and a minimum cosine similarity of 0.8 between all embeddings. We annotate clusters as 'scene', 'action' or 'irrelevant' question types. Based on a seed of 10K annotated questions, we retrieve all other questions with similarity above 0.8 and label them using the K-nearest neighbor algorithm ($K = 5$). Question Types have a total of 140K train, 35K validation and 84K test samples (cf. Table 1). Common question words and answers per task are presented in the Appendix (Figure 8).

[1]We use the 'all-MiniLM-L6-v2' model and Fast Clustering algorithm from the sentence-transformers package (`https://www.sbert.net/`).

4 Experimental Framework

In our experiments, we use the UNITER-base (Chen et al., 2020) model which has a single-stream transformer architecture and shows strong performance compared to state-of-the-art V+L model architectures (Bugliarello et al., 2021). In experiments where we finetune a pretrained model, we use the checkpoint from (Chen et al., 2020) which is pretrained among others on in-domain images, i.e. COCO captions (Lin et al., 2014).

4.1 Defining Task Difficulty via Pairwise Task Relationships

We first characterize the difficulty of each setting by describing pairwise task relationships, following studies in transfer (Zamir et al., 2018) and multitask learning (Standley et al., 2020; Lu et al., 2020). In particular, we measure the extent to which each task is forgotten after training on a second task.

Diverse Domains

Task 1 \ Task 2	Group 1	Group 2	Group 3	Group 4	Group 5
Group 1	67.52	-6.58	-5.21	-4.84	-7.09
Group 2	-4.55	67.92	-5.61	-4.51	-4.99
Group 3	-4.64	-8.39	70.83	-7.37	-11.66
Group 4	-4.69	-7.10	-7.40	65.03	-9.63
Group 5	-4.29	-5.82	-6.09	-3.80	63.24

Taxonomy Domains

Task 1 \ Task 2	Animals	Food	Interior	Sports	Transport
Animals	73.29	-8.06	-3.63	-5.84	-4.35
Food	-16.38	63.00	-4.29	-17.08	-11.94
Interior	-5.75	-5.19	65.26	-7.63	-2.83
Sports	-11.63	-18.20	-9.60	73.36	-9.47
Transport	-4.19	-8.48	-2.62	-3.67	64.50

Question Types

Task 1 \ Task 2	Action	Color	Count	Scene	Subcat.
Action	78.01	-68.40	-90.45	-19.59	-12.58
Color	-88.89	81.01	-99.65	-27.75	-62.46
Count	-99.17	-99.68	61.68	-97.52	-87.00
Scene	-10.91	-34.40	-77.73	86.62	-15.22
Subcat.	-31.73	-85.45	-96.15	-30.55	58.43

Table 2: Task difficulty measured by forgetting in pairwise tasks. Diagonal elements show the accuracy after training on Task 1. Non-diagonal elements show relative BWT after finetuning on Task 2.

Experimental Setup. We finetune the pretrained UNITER model on Task T_1 and compute the ac-

Dissimilarity Factor	Diverse Domains	Taxonomy Domains	Questions Types
Answer distribution	0.567*	0.791*	0.795*
Image embedding	0.248	0.492*	-0.640*
Question embedding	0.184	0.531*	0.631*
Joint embedding	0.220	0.622*	-0.223

Table 3: Spearman correlation of pairwise performance drop and embedding dissimilarity (* where $p < 0.05$).

curacy A_{11} on its test set. Then, we finetune this model on another Task T_2 and compute the new accuracy A_{12} on the test set of T_1. Forgetting is measured as the relative accuracy drop: $(A_{12} - A_{11})/A_{11}$. Regardless of dataset size, we finetune on T_2 for a fixed number of 400 steps using a batch size of 512 and learning rate 5e-5.

Observations. Table 2 shows the relative accuracy drop for all task pairs. We observe that forgetting in Taxonomy Domains fluctuates more depending on the task pairing, compared to Diverse Domains. Question Types is evidently a more challenging setting, where several task combinations show more than 90% drop. In all settings, task relationships are asymmetric. We find that some relations reflect semantic similarity, e.g., low forgetting between Food and Interior, as the two tasks are expected to contain similar visual scenes and vocabulary. We also observe that the model is more robust against forgetting when Task T_2 has a wide range of possible answers (e.g., Interior); while T_2 with a narrow answer set (e.g., Food, Color, Count) lead to maximum forgetting.

Task similarity and forgetting. To gain further insight into which factors contribute to forgetting, we measure the correlation between accuracy drop and different proxies of task similarity. In particular, we consider the answer distributions P, Q of Tasks T_1, T_2 respectively, as well as average embeddings of the image, question and the joint pair. Since some answers of T_1 do not appear in T_2, we measure the skew divergence (Lee, 2001) between P and Q as the KL divergence between P and a mixture distribution $(1 - \alpha)P + \alpha Q$ with $\alpha = 0.99$ (Ruder and Plank, 2017). For the input embeddings, we measure the cosine distance between the average task representation. As image representations, we utilize Faster R-CNN features from (Anderson et al., 2018), while questions are embedded using Sentence-BERT. Joint embeddings for image-question pairs are obtained using the final layer representation of the [CLS] token of UNITER [2]. The detailed similarity measures are shown in the Appendix Table 9.

The correlation results in Table 3 indicate that the more similar two consecutive tasks are, the less forgetting occurs. The divergence of answer distributions consistently correlates with forgetting, but does not fully account for the performance drop. For example, the divergence of Interior from Animals and Sports answer distributions is the same, however Sports leads to 1.88% more forgetting. Regarding the embedding distances, image embeddings show the highest correlation in the visual Taxonomy Domain, meaning that the more visually similar two domains are, the less severe forgetting is. We observe the same relationship mirrored in Question Types for question embeddings. However, we find no factor to correlate significantly with Diverse Domains, where tasks are generally similar to each other (cf. Appendix 9). Looking across modalities, we find that question and joint similarities in Taxonomy Domains correlate with forgetting, showing that the shift of the visual domains results in changes of the referred objects and types of questions per task.[3]

4.2 Continual Learning Methods

We next benchmark common continual learning algorithms, including regularization- and replay-based approaches. We investigate two regularization-based approaches: *Learning without Forgetting* (LwF) (Li and Hoiem, 2018), which uses knowledge distillation (Hinton et al., 2015) in order to retain knowledge from previous tasks, and *Elastic Weight Consolidation* (EWC) (Kirkpatrick et al., 2017). The EWC regularization term discourages big changes of parameters that were important for previous tasks, where importance is approximated using the Fisher information matrix.

We apply three types of replay approaches that allow access to a memory of past samples. *Experience Replay* (ER) (Chaudhry et al., 2019b) is the most straightforward approach, as it samples training data from both the current task and memory at each training step. *Average Gradient Episodic Memory* (A-GEM) (Lopez-Paz and Ranzato, 2017;

[2]The [CLS] token aggregates multimodal information. It is the first token of the input sequence and the final transformer layer passes only its representation to the classifier.

[3]We notice that the more similar images of two Question Types tasks are, the more forgetting occurs. A possible explanation is that new questions for similar images 'overwrite' previous knowledge. However, all cosine distances of image embeddings are too low (<0.05) to lead to any conclusions.

Chaudhry et al., 2019a) utilizes the memory of past data to ensure that gradient updates on past and new data are aligned.

We also experiment with a PSEUDO-REPLAY method for the Question Types setting. Instead of storing raw data from previous tasks, we use a data augmentation method, inspired by (Kafle et al., 2017; Kil et al., 2021). When training on task t, we augment the data D_t by retrieving past questions based on their shared detected objects classes. For example, if an elephant is detected on the current picture, we retrieve a past question about an elephant. We then use the previous model $f_{\theta_{t-1}}$ to generate a distribution $\tilde{\boldsymbol{y}} = f_{\theta_{t-1}}(\tilde{\boldsymbol{x}})$ which serves as soft targets for the new sample $\tilde{\boldsymbol{x}}$. By not storing the original answers, we address privacy and efficiency concerns of replay approaches (Van de Ven and Tolias, 2018; Delange et al., 2021).

4.3 Evaluation Metrics

After training on task t, we compute the VQA accuracy $A_{t,i}$ on data from the previous task i. We report the macro-average accuracy at the end of the training sequence: $\mathrm{A} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}$. Following Riemer et al. (2019), we report the learned accuracy $\mathrm{LA} = \frac{1}{T}\sum_{i=1}^{T} A_{i,i}$, which measures the ability to learn the new task i. We also compute backward transfer $\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} A_{T,i} - A_{i,i}$ (Lopez-Paz and Ranzato, 2017), that captures the impact of catastrophic forgetting.

In addition, we introduce a new metric, we term *semantic backward transfer* (SBWT), that weights backward transfer with the semantic distance of the predicted answers. The motivation for this metric is that some forgetting is worse than others. Consider the example in Figure 1, where the ground truth is 'duck'. After training on subsequent tasks, the sample gets misclassified as 'seagull' which might have a milder impact on the downstream application than completely unsuited answers such as 'black and white' or 'one'. For each sample $j = 1 \dots, N$ of task i, we measure the accuracy difference Δ_j^{Ti} of the answers predicted by the T-th and i-th models and weigh it by cosine distance of the two answer embeddings $\boldsymbol{e_{Tj}}$ and $\boldsymbol{e_{ij}}$. The final SBWT is computed as :

$$\mathrm{SBWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} S_{T,i} \tag{1}$$

where $S_{T,i}$ is the average weighted accuracy difference for task i:

$$S_{T,i} = \frac{1}{N}\sum_{j=1}^{N}(1 - \cos(\boldsymbol{e_{Tj}}, \boldsymbol{e_{ij}})) \cdot \Delta_j^{Ti} \tag{2}$$

In our implementation, we use averaged 300-dimensional GloVE embeddings (Pennington et al., 2014), since most answers are single words.

4.4 Experimental Setup

We investigate our three task settings on the VQA-v2 dataset (Goyal et al., 2017). Since ground truths are publicly available for the train and validation sets, we use validation samples as our test set, and create a new validation set by randomly sampling 20% of the training images. We follow a single head setting to allow for task-agnostic inference but assume knowledge of task boundaries during training. Memory-based approaches store 500 randomly selected samples per past task. For further implementation details, please refer to Appendix B.

We consider two baselines: The *Fix Model* baseline represents the generalization ability of the model across all tasks after being trained on only the first task D_1. The vanilla *Finetuning* baseline represents the performance degradation if no measures are taken to prevent forgetting. We also report the performance of joint training on all the data simultaneously (*Joint*) as an upper bound.

5 Results

5.1 Continual Learning Results

Table 4 summarizes the results averaged over five task orders. The results show an increasing difficulty for the three incremental learning task definitions, i.e. Diversity Domains < Taxonomy Domains < Question Types, which is in line with the results from our pairwise task characterization in Section 4.1. Although Question Types has the highest Joint accuracy, naive finetuning shows poor performance: it has the lowest final accuracy and large negative BWT. The low Fixed Model accuracy corroborates that tasks are highly dissimilar as a model trained on a single task fails to generalize.

Pretraining. Our results also confirm that pretraining leads to models that are more robust to forgetting (Mehta et al., 2021): all metrics consistently improve starting from a pretrained model. Pretraining combined with naive finetuning achieves on average 58% relative accuracy improvement over finetuning a model from scratch. Interestingly, the

		w/o Pretraining				w/ Pretraining			
Split	Method	Accuracy	LA	BWT	SBWT	Accuracy	LA	BWT	SBWT
Diverse	Fixed Model	41.60 ± 0.84	-	-	-	57.38 ± 0.83	-	-	-
	Finetuning	49.64 ± 0.78	56.69 ± 0.28	-8.80 ± 0.89	–5.35 ± 0.61	64.59 ± 0.56	**67.77** ± 0.22	-3.97 ± 0.59	-1.93 ± 0.39
	LwF	50.70 ± 0.56	54.67 ± 0.42	-4.96 ± 0.29	-2.89 ± 0.17	65.23 ± 0.42	67.62 ± 0.25	-3.02 ± 0.44	-1.50 ± 0.28
	AGEM	51.56 ± 0.78	**56.72** ± 0.30	-6.45 ± 0.87	-3.84 ± 0.60	65.65 ± 0.85	67.72 ± 0.30	-2.60 ± 0.71	-1.22 ± 0.38
	EWC	52.05 ± 0.30	56.49 ± 0.22	-5.55 ± 0.60	-3.12 ± 0.40	66.26 ± 0.55	67.58 ± 0.27	-1.65 ± 0.45	-0.67 ± 0.29
	ER	**54.36** ± 0.33	56.31 ± 0.51	**-2.45** ± 0.49	**-1.42** ± 0.26	**66.66** ± 0.50	67.55 ± 0.23	**-1.11** ± 0.41	**-0.51** ± 0.27
	Joint	60.41 ± 0.03	-	-	-	69.76 ± 0.18	-	-	-
Taxonomy	Fixed Model	39.96 ± 1.05	-	-	-	55.00 ± 0.95	-	-	-
	Finetuning	47.72 ± 0.72	57.75 ± 0.24	-12.53 ± 0.65	-8.45 ± 0.38	63.65 ± 0.63	68.77 ± 0.12	-6.40 ± 0.67	-3.89 ± 0.53
	LwF	48.05 ± 0.24	55.25 ± 0.27	-9.00 ± 0.38	-6.13 ± 0.44	64.83 ± 0.50	68.73 ± 0.17	-4.88 ± 0.69	-2.88 ± 0.43
	AGEM	50.51 ± 0.66	**57.80** ± 0.25	-9.10 ± 0.79	-5.77 ± 0.55	66.52 ± 0.34	**68.86** ± 0.12	-2.92 ± 0.50	-1.63 ± 0.33
	EWC	52.17 ± 0.54	57.49 ± 0.19	-6.65 ± 0.44	-4.33 ± 0.28	**67.70** ± 0.29	68.57 ± 0.16	**-1.09** ± 0.33	**-0.62** ± 0.19
	ER	**54.60** ± 0.14	57.67 ± 0.28	**-3.84** ± 0.42	**-2.38** ± 0.27	66.76 ± 0.16	68.61 ± 0.13	-2.32 ± 0.16	-1.22 ± 0.10
	Joint	60.82 ± 0.02	-	-	-	70.08 ± 0.18	-	-	-
Questions	Fixed Model	18.81 ± 5.90	-	-	-	25.54 ± 8.75	-	-	-
	Finetuning	23.30 ± 8.83	65.24 ± 0.42	-52.42 ± 10.88	-39.86 ± 12.08	48.81 ± 5.56	72.94 ± 0.20	-30.17 ± 7.07	-22.43 ± 7.02
	LwF	26.23 ± 8.56	60.69 ± 1.43	-43.08 ± 11.22	-34.32 ± 9.94	46.61 ± 3.95	72.06 ± 0.44	-31.82 ± 5.42	-25.13 ± 5.35
	AGEM	50.73 ± 1.92	**65.38** ± 0.56	-18.31 ± 3.04	-10.02 ± 1.39	68.30 ± 0.74	72.96 ± 0.24	-5.83 ± 1.08	-2.95 ± 0.63
	EWC	36.77 ± 5.01	49.05 ± 3.82	-15.35 ± 5.85	-11.76 ± 5.41	66.77 ± 3.54	70.03 ± 1.03	**-4.08** ± 3.58	-2.62 ± 2.28
	Pseudo-Replay	55.22 ± 1.75	65.12 ± 0.46	-12.37 ± 2.57	-7.29 ± 1.64	67.66 ± 1.15	**72.97** ± 0.26	-6.63 ± 1.74	-3.27 ± 0.98
	ER	**59.54** ± 0.32	65.09 ± 0.52	**-6.93** ± 0.71	**-3.50** ± 0.35	**69.18** ± 0.38	72.82 ± 0.22	-4.56 ± 0.56	**-1.82** ± 0.34
	Joint	66.35 ± 0.24	-	-	-	72.54 ± 0.15	-	-	-

Table 4: Results from VQA Incremental Learning. We report the average and standard deviation over five random task orders. LA: Learned Accuracy, BWT: Backward Transfer, SBWT: Semantic Backward Transfer.

pretrained Fixed Model is able to generalize reasonably well to other domains for both image-based settings, and the final Pretraining+Finetuning accuracy exceeds the Joint accuracy without pretraining. These results indicate that learning generic V+L representations via pretraining has persistent benefits. However, pretraining is insufficient for ensuring continual learning and additional strategies improve the final accuracy by 8.83% on average.

Continual Learning Methods. Among continual learning methods, LwF offers the smallest gains in terms of final accuracy and forgetting. [4] This shortcoming is reasonable considering that LwF generates pseudo-labels using the current data, which may be too noisy if the answers for the current and previous tasks differ substantially. In contrast, our Pseudo-Replay method, which combines distillation and replay, does not suffer from the same limitation and achieves almost 20% improvement of the accuracy in Question Types.

Pretraining+EWC achieves the highest accuracy in the Taxonomy Domains. However, when dealing with heterogeneous tasks (i.e. within Question Types) the high regularization weights, which are required to prevent forgetting, end up limiting the model's ability to adapt to new and dissimilar tasks. This over-stability is also reflected in the low LA of EWC, which indicates that the model struggles to learn new tasks. On the other hand, memory-based approaches have consistently high LA. In addition, ER shows the best performance with models trained from scratch as well as for the challenging setting of Question Types.

Measuring Forgetting. Next, we compare our newly introduced metric SBWT, which takes semantic similarities into account, to the standard BWT, which measures absolute forgetting. We observe some notable differences, which indicate that SBWT favors strong models that forget gradually. For instance, EWC w/o pretraining shows lower performance and LA under the Question Types setting compared to, e.g. AGEM w/o pretraining. However, it receives a better BWT score. We make similar observations for LwF vs. AGEM in Taxonomy Domains w/o pretraining, and EWC vs. ER in Taxonomy Domains with pretraining. Table 9 in the Appendix provides an example-based analysis, showing that semantically more similar answers have higher SBWT scores.

5.2 Effect of Memory Size

Here, we compare the memory size for ER and our new Pseudo-Replay method. Pseudo-Replay only stores questions and uses the previous checkpoint to generate soft pseudo-labels. We choose the Question Types setting, as it is most prone to forgetting. In general, more memory means less forgetting but at a higher computation and storage cost. Figure 3 shows the average accuracy for three memory sizes across training. At each step, we compute the average accuracy of the experi-

[4] Despite searching a wide range of values, we were unable to find a distillation weight that improves the final accuracy of the pretrained model in Question Types.

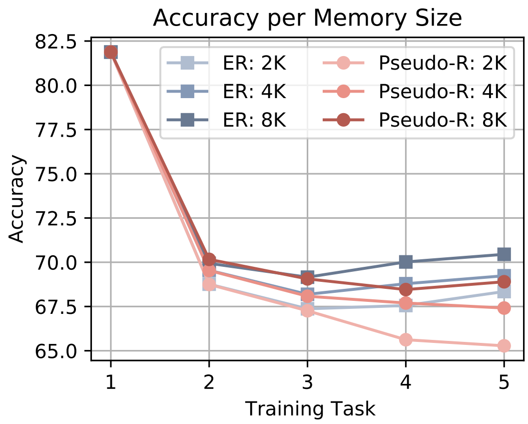


Figure 3: Average accuracy of seen tasks per memory size. PSEUDO-REPLAY performs competitively up the the third task despite only storing questions.

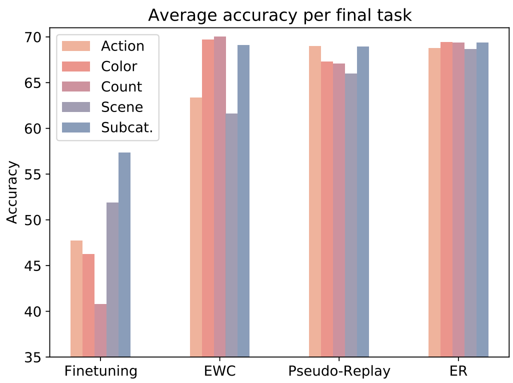


Figure 4: Sensitivity to task order as illustrated for Question Types. Each bar shows the accuracy of a task sequence ending with a different task.

enced tasks up to that point. As expected, both methods benefit from access to a larger memory. PSEUDO-REPLAY shows comparable performance for up to three tasks, while raw ER replay becomes more advantageous as more tasks are added. We attribute this convergence in performance to errors by PSEUDO-REPLAY's pseudo-labeling causing confirmation bias (Tarvainen and Valpola, 2017). Despite this limitation, PSEUDO-REPLAY exceeds the performance of naive finetuning by over 18% when storing only 500 samples per task.

5.3 Sensitivity to Task Order

Next, we investigate the impact of task order. Results in Table 4 were averaged over five random task orders. In real scenarios, however, tasks would appear in a specific order. The high variance of the results in Question Types already indicates that task order can influence performance. To verify this, we plot in Figure 4 the final accuracy of a pretrained

w/o Pretraining			
Method	**What animal**	**What room**	**What sport**
Finetuning	33.09 $\pm$ 13.38	54.38 $\pm$ 32.42	25.14 $\pm$ 32.11
EWC	48.18 $\pm$ 15.67	83.48 $\pm$ 7.61	62.81 $\pm$ 13.67
ER	73.11 $\pm$ 0.70	89.04 $\pm$ 2.80	87.20 $\pm$ 1.84
w/ Pretraining			
Method	**What animal**	**What room**	**What sport**
Finetuning	75.07 $\pm$ 3.54	83.26 $\pm$ 12.47	69.92 $\pm$ 14.14
EWC	81.75 $\pm$ 1.42	94.32 $\pm$ 0.88	90.82 $\pm$ 1.36
ER	80.73 $\pm$ 0.37	94.10 $\pm$ 1.39	90.92 $\pm$ 0.71

Table 5: Accuracy and standard deviation of the best performing models on different sub-questions in Taxonomy Domains.

model for five training sequences, each ending with a different task. Our results show that task order can lead to Finetuning accuracy that varies more than 15%. Although EWC improves the average accuracy, there is still a 10% fluctuation depending on the order. However, replay-based methods are able to improve performance and mitigate the sensitivity to task order.

While Table 4 shows low variance in Taxonomy Domains, we find high variance when examining the performance on specific questions. In particular, we find that certain question types, such as Animals, Interior, and Sports, have high variance. Table 5 reveals a standard deviation which is up to 30 times higher compared to the average results in Table 4. High standard deviation across randomized task orders is problematic since models can have different behavior in practice despite similar (aggregated) performance. In other words, the current task performance will highly depend on the previous task order, even though the overall accuracy from the randomized trials appears similar.

5.4 Representation Analysis

Finally, we ask how representations from each modality evolve throughout the training sequence and compare this evolution across our continual learning settings. We use centered kernel alignment (CKA) (Kornblith et al., 2019) to track the representation similarity of sequentially finetuned models. We extract representations X_t^1 of the validation data of the first task after training for each task $t = 1 \cdots T$, and measure the CKA similarity of $X_{t>1}^1$ to the original representations X_1^1. Figure 5 shows the evolution of the representation of the [CLS] token from the final transformer layer as well as the average representation of visual and textual tokens from the embedding and final layers.

Across all settings, the representations of ques-

Figure 5: Representation similarity for the first task under the three settings.

tion tokens retain higher similarity than the image tokens. This suggests that the features extracted from the visual inputs in order to predict an answer are more dependent on the current task than the features extracted from the more reusable question tokens. We also corroborate previous findings (Ramasesh et al., 2021) showing that representations from deeper layers change more during continual learning. These results highlight the importance of stabilizing visual representations in deeper layers.

6 Related Work

To the best of our knowledge, this is the first work studying the impact of task formulation for continual learning in V+L models. Past studies examined the relationship between catastrophic forgetting and different aspects of a continual learning algorithm, such as the activation function, dropout, and learning rate schedule (Goodfellow et al., 2013; Mirzadeh et al., 2020). Other work has investigated which layers of deep neural networks forget more (Nguyen et al., 2021), the role of task similarity (Ramasesh et al., 2021; Lee et al., 2021) and which properties of task sequences amplify forgetting (Nguyen et al., 2019a). However, all of these studies have focused on image classification tasks.

Previous work on V+L continual learning has studied a range of different tasks. Del Chiaro et al. (2020) and Nguyen et al. (2019b) study continual learning for domain- and class-incremental image captioning, while Jin et al. (2020) provide a benchmark for task-agnostic phrase prediction to test compositionality and soft task boundaries. Kemker et al. (2018) propose a multimodal continual learning setting, where audio and image classification tasks are learned sequentially.

More closely related to our work, Greco et al. (2019) explore the effect of forgetting in VQA with two question types (‘Wh-’ and binary questions). Consistent with our findings, they show that task order influences forgetting and that continual learning methods can alleviate forgetting. However, their study is limited to only two tasks and does not test the impact of pretrained models, which, as we show, can mitigate forgetting. Hayes et al. (2020) also study continual learning of question-based tasks focusing on a challenging low-resource online setting, where new samples are available for a single update. Our study focuses on a less strict yet practical scenario where models are updated periodically with all data for the new task until convergence.

7 Conclusion

We empirically investigate the impact of task formulation, i.e. task design, order and similarity, on continual learning in VQA. We evaluate a transformer-based model and benchmark several methods, including a new PSEUDO-REPLAY approach which combines data augmentation and distillation. Our results show that both task order and similarity influence results. These results are important for designing continual learning experiments for real-world settings, where task formulation depends on the application scenario. For example, the Taxonomy Domains resembles applications where data is continuously collected in different visual surroundings, whereas Question Types corresponds to ‘teaching’ the system new reasoning capabilities. Our results suggest that the latter is the most challenging. The easiest and thus ‘best-case’ scenario is a Diverse data collection setup, where the system incrementally learns to recognize new objects which are randomly sampled from different domains. Moreover, the strong performance of the relatively simple PSEUDO-REPLAY method suggests that more advanced strategies for selecting or generating samples representative of past tasks can yield further improvements.

8 Ethical Impact

The proposed continual learning approach to V+L problems offers a promising alternative to the current pretraining-and-finetuning paradigm, which has the potential to mitigate the financial and environmental costs of (re-)training large models (Strubell et al., 2019; Bender et al., 2021). In addition to demonstrating performance gains of continual learning over vanilla finetuning, our paper also proposes a novel PSEUDO-REPLAY algorithm. PSEUDO-REPLAY not only uses less memory than standard memory-based approaches, but also is better at preserving privacy. Preserving privacy is especially important for federated data settings (Jiang et al., 2021) or for sensitive applications such as medical imaging (Ravishankar et al., 2019).

The paper also highlights potential negative impacts related to the high variability in performance, where performance can vary up to 15% depending on the task order. Robust performance is especially important in the context of applying this technology with real-users, such as supporting users with visual impairments (Gurari et al., 2018). We thus see the robustness of continual learning approaches as a main challenge for future research.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019a. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. 2019b. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Riccardo Del Chiaro, Bartł omiej Twardowski, Andrew Bagdanov, and Joost van de Weijer. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. In *Advances in Neural Information Processing Systems*, volume 33, pages 16736–16748. Curran Associates, Inc.

Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Claudio Greco, Barbara Plank, Raquel Fernández, and Raffaella Bernardi. 2019. Psycholinguistics meets continual learning: Measuring catastrophic forgetting in visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3601–3605, Florence, Italy. Association for Computational Linguistics.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Raia Hadsell, Dushyant Rao, Andrei A. Rusu, and Razvan Pascanu. 2020. Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences*, 24(12):1028–1040.

Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. 2020. Remind your neural network to prevent catastrophic forgetting. In *Computer Vision – ECCV 2020*, pages 466–483, Cham. Springer International Publishing.

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

Aman Hussain, Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Ziyue Jiang, Yi Ren, Ming Lei, and Zhou Zhao. 2021. Fedspeech: Federated text-to-speech with continual learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3829–3835. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. 2020. Visually grounded continual learning of compositional phrases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2018–2029, Online. Association for Computational Linguistics.

Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. 2017. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, Santiago de Compostela, Spain. Association for Computational Linguistics.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jihyung Kil, Cheng Zhang, Dong Xuan, and Wei-Lun Chao. 2021. Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. *arXiv preprint arXiv:2109.06122*.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomáš Kočiský, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, volume R3 of *Proceedings of Machine Learning Research*, pages 176–183. PMLR. Reissued by PMLR on 31 March 2021.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. 2021. Continual learning in the teacher-student setup: Impact of task similarity. In *2021 International Conference on Machine Learning*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pages 17–26. PMLR.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6470–6479.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. 2021. An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. 2020. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 7308–7320. Curran Associates, Inc.

Cuong V. Nguyen, Alessandro Achille, Michael Lam, Tal Hassner, Vijay Mahadevan, and Stefano Soatto. 2019a. Toward understanding catastrophic forgetting in continual learning. *CoRR*, abs/1908.01091.

Giang Nguyen, Shuan Chen, Tae Joon Jun, and Daeyoung Kim. 2021. Explaining how deep neural networks forget by deep visualization. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 162–173, Cham. Springer International Publishing.

Giang Nguyen, Tae Joon Jun, Trung Tran, Tolcha Yalew, and Daeyoung Kim. 2019b. Contcap: A scalable framework for continual image captioning. *arXiv preprint arXiv:1909.08745*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. 2021. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Hariharan Ravishankar, Rahul Venkataramani, Saiharessh Anamandra, Prasad Sudhakar, and Pavan Annangi. 2019. Feature transformers: Privacy preserving lifelong learners for medical imaging. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 347–355, Cham. Springer International Publishing.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, , and Gerald Tesauro. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*.

Sebastian Ruder and Barbara Plank. 2017. Learning to select data for transfer learning with Bayesian optimization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 372–382, Copenhagen, Denmark. Association for Computational Linguistics.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NEURIPS'17, page 1195–1204, Red Hook, NY, USA. Curran Associates Inc.

Gido M Van de Ven and Andreas S Tolias. 2018. Generative replay with feedback connections as a general strategy for continual learning. *arXiv preprint arXiv:1809.10635*.

Gido M Van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. 2021. Separating skills and concepts for novel visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5632–5641.

Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. 2020. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations*.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Data Details

We investigate three continual learning settings based on the VQA-v2 dataset (Goyal et al., 2017), a collection of visual question annotations in English. Tasks in the Diverse Domains setting are created by grouping 10 objects from COCO annotations (Lin et al., 2014) as follows:

- Group 1: bird, car, keyboard, motorcycle, orange, pizza, sink, sports ball, toilet, zebra
- Group 2: airplane, baseball glove, bed, bus, cow, donut, giraffe, horse, mouse, sheep
- Group 3: boat, broccoli, hot dog, kite, oven, sandwich, snowboard, surfboard, tennis racket, TV
- Group 4: apple, baseball bat, bear, bicycle, cake, laptop, microwave, potted plant, remote, train
- Group 5: banana, carrot, cell phone, chair, couch, elephant, refrigerator, skateboard, toaster, truck

We also provide a few example questions for each task in Question Types:

- Action: What is the cat doing?, Is the man catching the ball?, What is this sport?
- Color: What color is the ground?, What color is the right top umbrella?
- Count: How many skaters are there?, How many elephants?, How many rooms do you see?
- Scene: Is the picture taken inside?, Is this photo black and white?, What is the weather like?
- Subcategory: What type of vehicle is this?, What utensil is on the plate?, What kind of car is it?

Figures 6-8 show the distribution of the 20 most common question words and answers for each task. The counts are computed on the combined train and validation data, excluding stopwords from the question vocabulary. These plots support our general findings about the characteristics of each task and the relationships between them. For example, answers in Diverse Domains are highly similar across tasks, while the most considerable difference of common answers is observed in Question Types. In addition, frequent nouns in Diverse and Taxonomy Domains reflect the typical objects from the image annotations of each task. Common words in Question Types also follow the definition of each

Dissimilarity	Diverse	Taxonomy	Questions
Answers	0.567 (0.009)	0.791 (0.000)	0.795 (0.000)
Image embed.	0.248 (0.293)	0.492 (0.028)	-0.640 (0.002))
Question embed.	0.184 (0.437)	0.531 (0.016)	0.631 (0.003)
Joint embed.	0.220 (0.350)	0.622 (0.003)	-0.223 (0.344)

Table 6: Spearman correlation of pairwise performance drop and and different dissimilarity heuristics. In addition to the results in table 3, we show in parentheses the corresponding p-values. We underline statistically significant results ($p < 0.05$).

Setting	Batch Size	Learning Rate	LwF λ	EWC λ
Diverse	512	8e-5	1	400
Diverse+PT	1024	8e-5	0.7	500
Taxonomy	512	8e-5	1	600
Taxonomy+PT	1024	5e-5	0.5	500
Questions	1024	1e-4	0.9	50K
Questions+PT	512	5e-5	0.4	20K

Table 7: Best hyperparameters for all settings. PT: Pretraining

task. For example, top words in Scene such as 'sunny', 'room', 'outside' refer to the entire image, while Action words such as 'sport', 'playing', 'moving' refer to activities shown in the image.

B Implementation Details

Our implementation is based on the publicly available PyTorch codebase of UNITER (https://github.com/ChenRocks/UNITER). For the continual learning experiments, we train a UNITER-base model (86M parameters) on a cluster of NVIDIA V100 GPUs using a single node with 4 GPUs. Training on a sequence of 5 tasks requires on average $\sim$ 5 GPU hours. The main experiments (Table 4) require approximately a total of 200 GPU hours.

We first tune the batch size and learning rate with naive finetuning. Keeping these hyperparameters fixed, we then tune the continual learning hyperparameters (EWC, LwF λ). All hyperparameters are selected through grid search based on the maximum final accuracy as shown in Table 7. Initial results with a pretrained model on Taxonomy Domains showed that best performance is achieved with a mixing ratio of 3:1 of new and old data per batch. We keep this ratio constant for all experiments.

Each experiment is repeated five times with a different random seed and task order. The task orders used in our experiments are the following:

- **Diverse Domains**
- group 5, group 3, group 2, group 4, group 1
- group 1, group 2, group 5, group 3, group 4
- group 4, group 3, group 5, group 1, group 2
- group 3, group 1, group 4, group 2, group 5
- group 2, group 5, group 1, group 4, group 3
- **Taxonomy Domains**
- food, animals, sports, interior, transport
- transport, sports, food, animals, interior
- interior, animals, food, transport, sports
- animals, food, interior, sports, transport
- sports, interior, transport, animals, food
- **Question types**
- action, count, subcategory, scene, color
- color, subcategory, action, count, scene
- scene, count, action, color, subcategory
- subcategory, color, scene, action, count
- count, scene, color, subcategory, action

C Further CKA Results

Figure 10 provides detailed plots of the CKA similarity of the representations from all layers using a randomly initialized and a pretrained model. We plot the average CKA values from five task orders. Our results support the observations of Section 5.4. The change of CKA similarity corroborates that Question Types is the most challenging of the three settings. We also observe that representations of pretrained models remain more similar, especially representations from layers closer to the input (early layers) in Diverse and Taxonomy Domains which retain high similarity across training tasks. This indicates that early layers of the pretrained model have learned generic representations that transfer across tasks. Comparing the CKA results without pretraining for all settings, we see that in Diverse and Taxonomy Domains, the representations that change most continue to be those from the images. In Question Types, [CLS] token representations change most. Question word representations remain more similar than image representations of early layers (layers 0-7).

D Qualitative Results

Table 8 shows examples of predicted answers with different approaches. The two top examples are from two different task orders in Question Types, and the two bottom examples are from Taxonomy

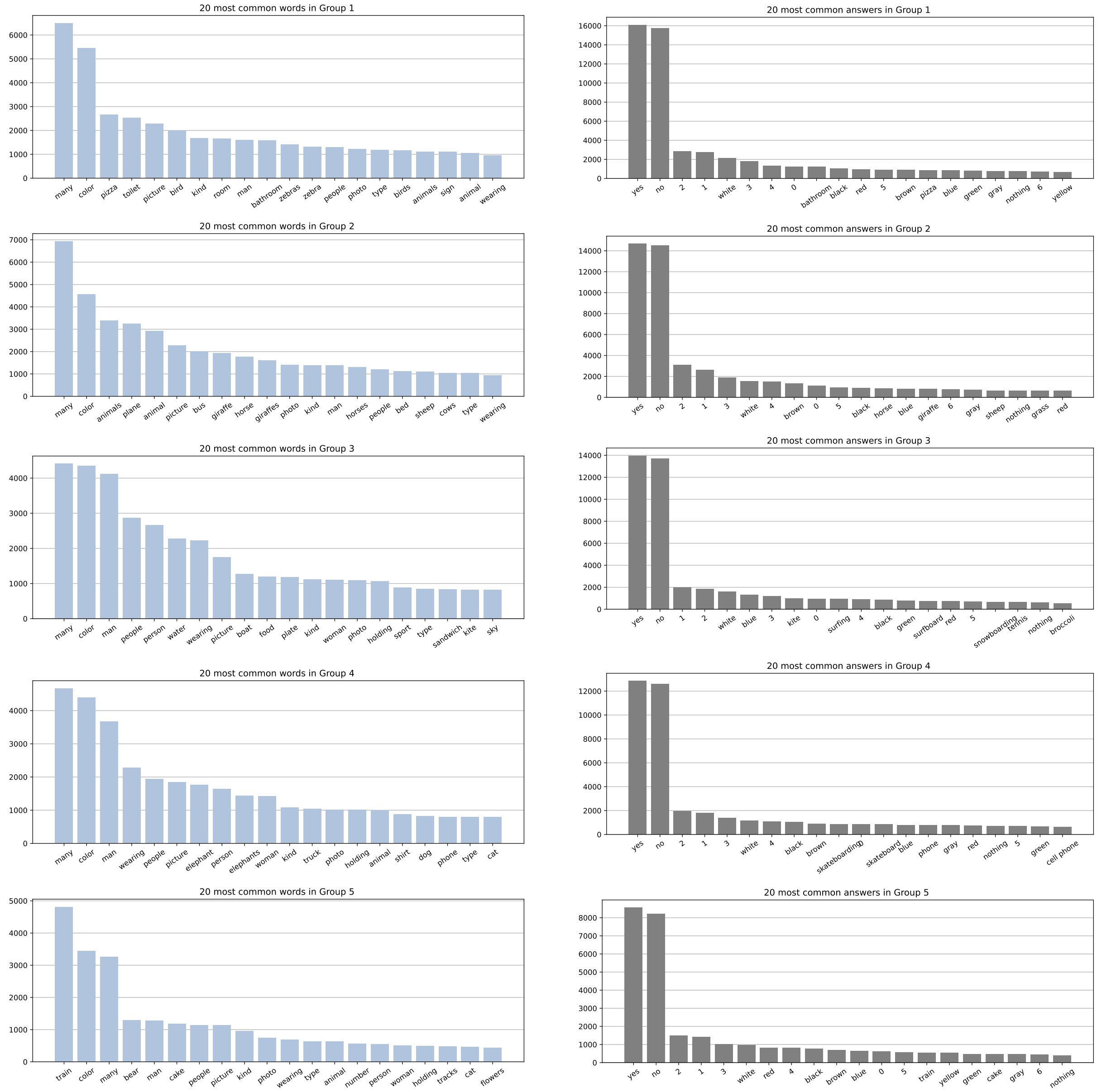


Figure 6: Most common words (left) and answers (right) per task Diverse Domains.

Domains. The model trained from scratch (column w/o PT) fails to retain knowledge from the corresponding training task. The pretrained model (column PT) is more resistant to forgetting and we observe that for the first and third images, it even manages to recover the correct answer during the training sequence. However, relying only on pretraining is insufficient, as the model still tends to change the predicted answer based on the most recent training task. Both EWC and ER combined with pretraining successfully retain previous knowledge.

Table 9 presents examples of the SBWT metric. Specifically, it compares SBWT for two pairs of predicted answers with the same initial reference answer. When the initial prediction (reference answer) is correct, and both compared answers are wrong, we observe that SBWT penalizes similar answers less than unrelated ones (see the first four rows of Table 9). Similarly, when one of the compared answers is partially correct (rows 5-8) according to the VQA accuracy metric, SBWT is less punishing compared to BWT, which in our examples would be -0.7. Finally, the last row shows an example of corrected compared answers, where the accuracy improvement is weighted with the semantic distance of reference and compared answers.

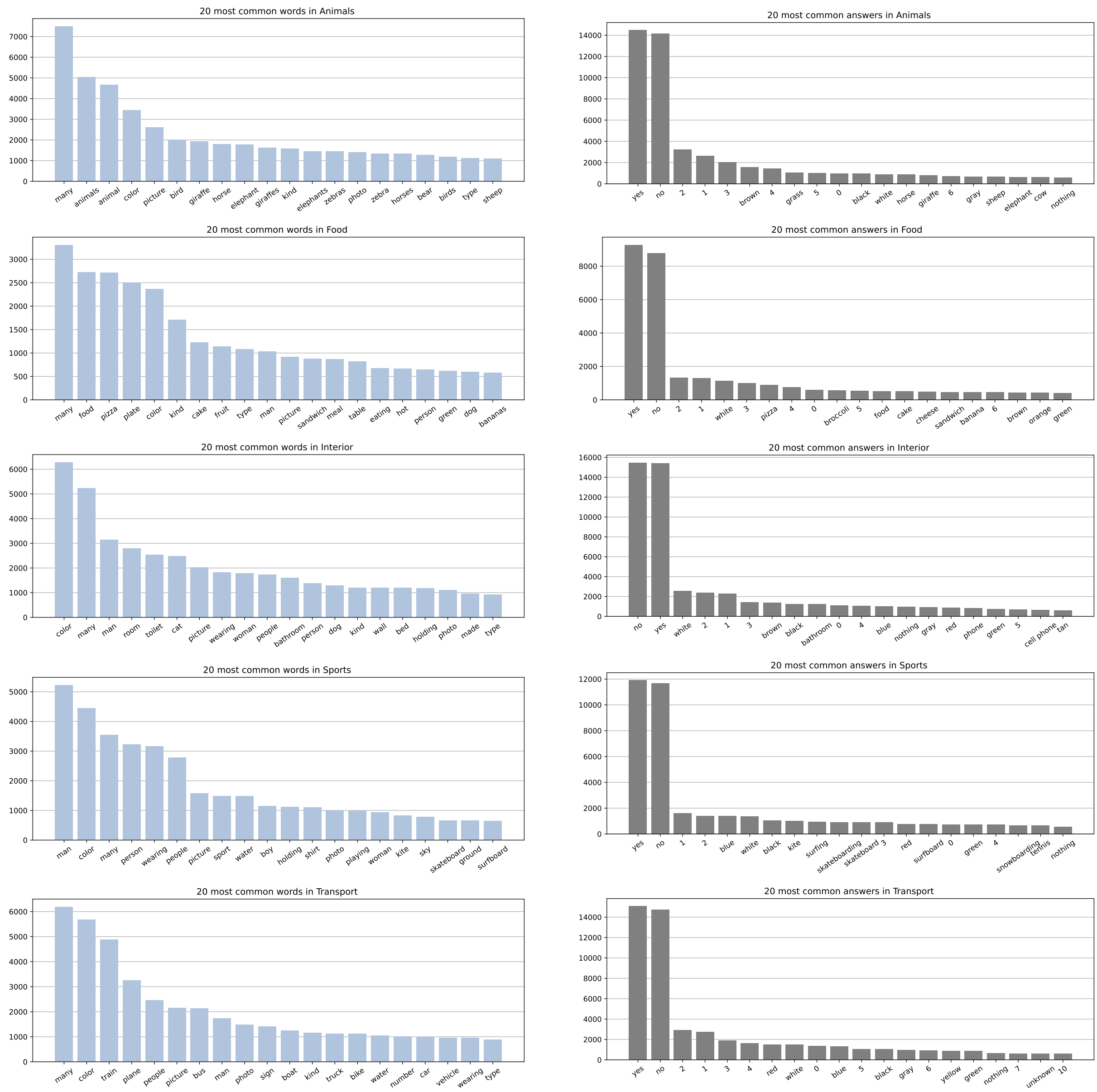


Figure 7: Most common words (left) and answers (right) per task Taxonomy Domains.

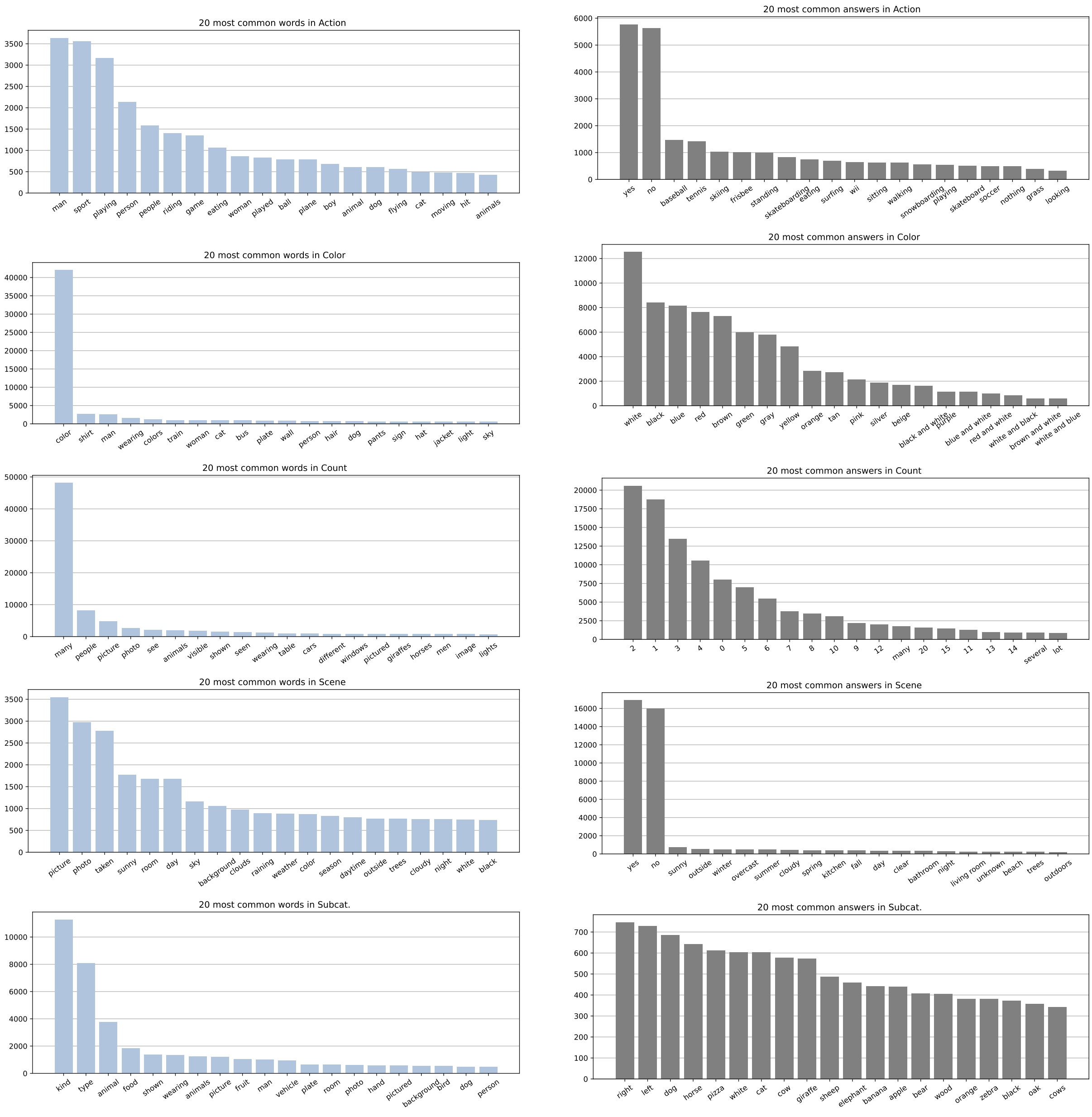
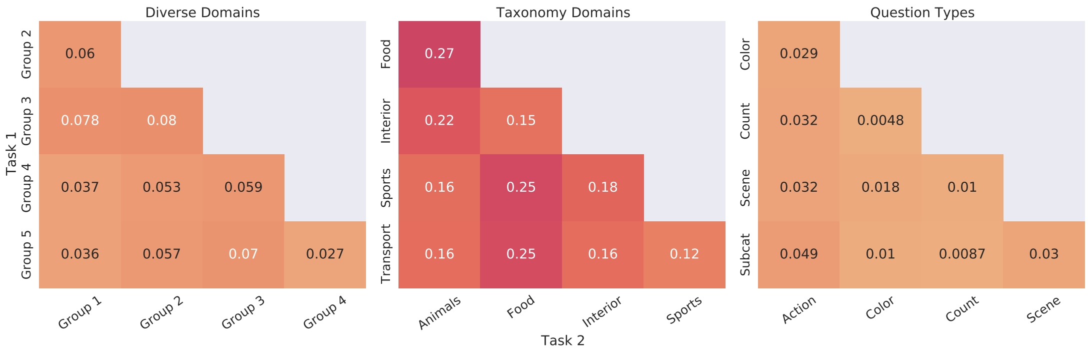
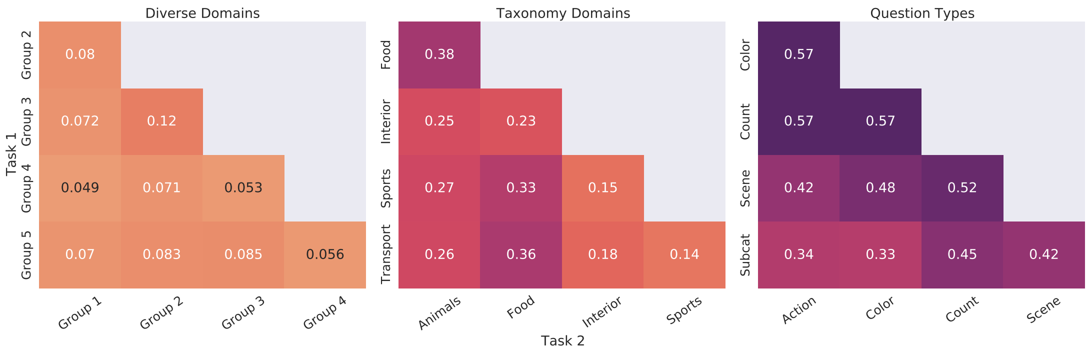


Figure 8: Most common words (left) and answers (right) per task in Question Types.

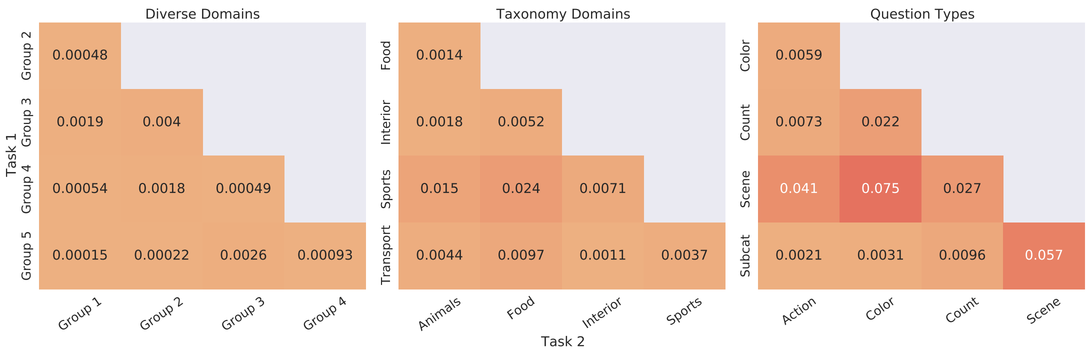
(a) Divergence of answer distributions.



(b) Cosine distance of image embeddings.



(c) Cosine distance of question embeddings.



(d) Cosine distance of joint embeddings.

Figure 9: Dissimilarity measures between task pairs.

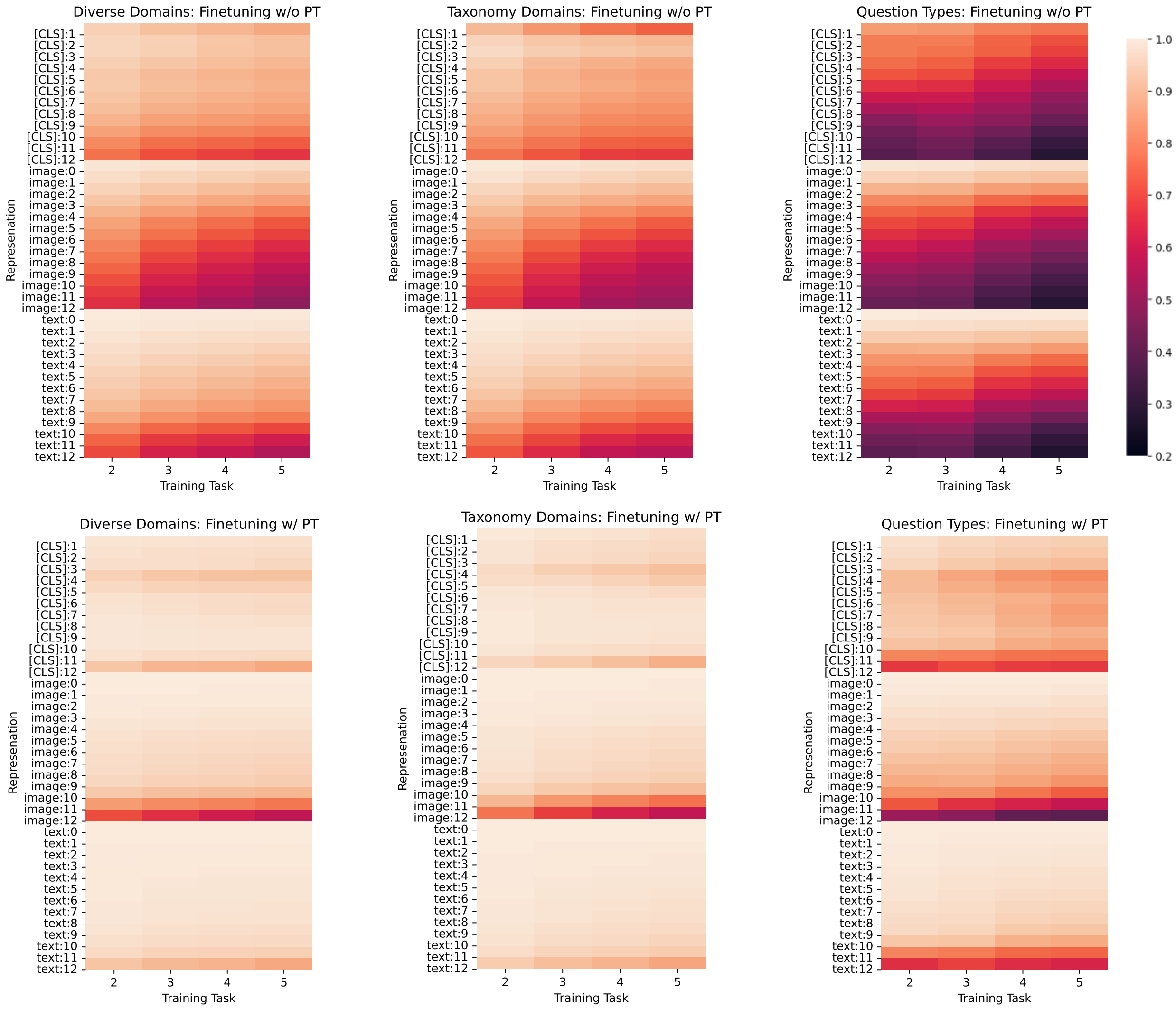
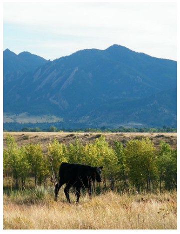


Figure 10: CKA similarity of the representations of all layers. Representations are indexed with 0-12 where 0 corresponds to representations from the input embedding layer and 12 from the transformer layer closest to the output. Deeper colors indicate lower similarity. We observe that representations of models trained from scratch (top row) remain less similar than pretrained models (bottom row). For pretrained models, mostly representations from the top two layers change evidently.

What is the horse doing?

Task	w/o PT	PT	PT+EWC	PT+ER
Action	jumping	jumping	jumping	jumping
Count	two	one	jumping	jumping
Subcat.	riding	jump	jumping	jumping
Scene	cold	jumping	jumping	jumping
Color	black	black	jumping	jumping



What color is the cow?

Task	w/o PT	PT	PT+EWC	PT+ER
Color	black	black	black	black
Subcat	black	black	black	black
Action	zero	yes	cow	black
Count	one	one	black	black
Scene	green	green	black	black

What is orange?

Task	w/o PT	PT	PT+EWC	PT+ER
Food	carrots	carrots	carrots	carrots
Animals	birds	carrots	carrots	carrots
Sports	nothing	kites	carrots	carrots
Interior	chair	carrots	carrots	carrots
Transport	nothing	tomato	carrots	carrots

What type of bird is this?

Task	w/o PT	PT	PT+EWC	PT+ER
Interior	dog	owl	owl	owl
Animals	pigeon	pigeon	pigeon	pigeon
Food	turkey	pigeon	pigeon	pigeon
Transport	not sure	duck	pigeon	seagull
Sports	zero	seagull	pigeon	seagull

Table 8: Examples of the evolution of predicted answers with different approaches. Column Task shows the order of the training tasks. The bold task corresponds to the task of the sample.

Reference		**Compared Answer 1**			**Compared Answer 2**		
Answer	**Acc**	**Answer**	**Acc**	**SBWT**	**Answer**	**Acc**	**SBWT**
skateboarding	1	skateboard	0	-0.164	black	0	-0.836
snowboarding	1	skiing	0	-0.134	winter	0	-0.529
breakfast	1	sandwich	0	-0.340	one	0	-0.855
food	1	meat	0	-0.320	toothbrush	0	-0.832
skateboarding	1	skateboard	0.3	-0.115	skateboard	0	-0.164
carrots	1	carrot	0.3	-0.093	three	0	-0.818
sheep	1	goat	0.3	-0.197	white	0	-0.676
cloudy	1	overcast	0.3	-0.151	gray	0	-0.577
black	0	black and white	1	0.136	brown	1	0.269

Table 9: Comparison of the SBWT metric of two answers with respect to the same reference answer. We verify that semantically more similar answers have higher SBWT.