Supplementray Materials: CTP: Towards Vision-Language Continual Pretraining via Compatible Momentum Contrast and Topology Preservation

Hongguang Zhu[1,2*] Yunchao Wei[1,2,3] Xiaodan Liang[3,4,5] Chunjie Zhang[1,2] Yao Zhao[1,2,3†]

[1]Institute of Information Science, Beijing Jiaotong University [2]Beijing Key Laboratory of Advanced Information Science and Network

[3]Peng Cheng Laboratory [4]Sun Yat-sen University [5]MBZUAI

Appendix Overview

This supplementary document mainly provides more information about our P9D dataset and implementation details of the baseline methods. Besides, we provide the pseudocode of CTP and more experimental studies.

A. P9D Dataset.

A.1. Dataset Split.

Figure 1 shows the quantity distribution of each subset of our P9D. The different subsets have a consistent quantity distribution across tasks, and this consistent distribution ensures the comprehensive and unified evaluation for pretraining. Different from the training set, the test set (cross-modal retrieval evaluation) and query set (multi-modal retrieval evaluation) need to be further filtered by humans. The filter criterion is that the text describes the image content as accurately as possible while ensuring that the test/query set is proportional to the training set for the same category.

A.2. Image-Text Examples.

Figure 2 shows some image-text examples. We show some images of same class and keep one described caption for simplicity. It shows that real-world web data is noisy and multi-domain mixing. There are prevalent and complicated situations in the web image domain, such as complex backgrounds, amorphous watermarks, irrelevant objects, and occlusion.

B. Details of Baseline Methods.

Because these baseline methods are originally proposed for continual learning on the image classification task. Thus, we re-implement them to adapt the setting of vision-language pretraining. In addition to the replacement of the main optimization loss, we present the implementation details of each comparison method as follows:

* This work was done when Hongguang Zhu worked as an research intern in Peng Cheng Laboratory. Email: kevinlight831@gmail.com

† Corresponding author: yzhao@bjtu.edu.cn

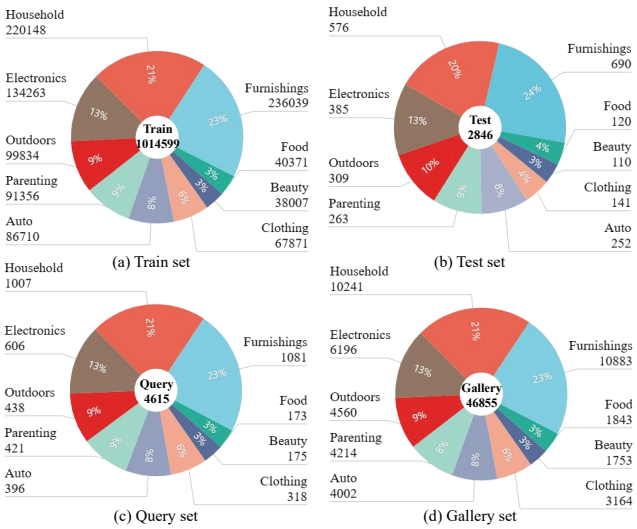

Figure 1: The quantity distribution of different task data is consistent for the four subsets of our P9D dataset.

B.1. Memory-Free methods

EWC [16] is the classical regularization methods. It maintains the old model parameters θ_{t-1} and an important matrix Ω with the same scale as the model. EWC builds an additional regularization loss to remember the old parameters according to the important matrix. Because the model θ_{t-1} at the last task stores the old knowledge, consolidating important parameters can fix the knowledge from being forgotten. The training loss can be formulated as:

$$\mathcal{L}_{EWC} = \mathcal{L}_{VLP} + \frac{1}{2}\lambda\Sigma_k\Omega_k(\theta_{t,k} - \theta_{t-1,k})^2, \quad (1)$$

where the $\theta_{t-1,k}$ denotes the k-th parameter after training last task data $\mathcal{D}_{t-1}$. Ω_k means the important weight of the k-th parameter and is calculated by the Fisher Information Matrix (FIM) in the EWC method.

SI [43] considers that the EWC is conducted at the end of each task and will ignore the optimization dynamics over the entire training trajectory. Thus, SI online estimates the

importance weight Ω_k by its contribution (backward gradient) to the total loss variation. However, this online strategy need to backpass the gradient twice for each iteration. In the re-implement, we store the gradient of each parameter by retaining the forward graph.

MAS [1] calculate Ω_k by a unsupervised way. Specifically, It accumulates important measures based on the sensitivity of predictive results (output features) to parameter changes. In our re-implement, we sum the norm of the visual, textual, and multi-modal features as the predictive result to calculate the importance.

RWalk [4] combines the regularization terms of SI [43] and EWC [16] to integrate their advantages. In each iteration, Rwalk simultaneously consolidates the parameter by considering the online importance weight from SI method and the offline important weight from EWC method.

AFEC [40] proposes to actively forget the old knowledge that interferes with the learning of new tasks for continual learning. Specifically, It introduces the extra forward-step trained model $\theta_t^\star$ as the expansion and collaboratively guides the update of the current model with the EWC method. Similar to EWC, the training loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{\mathcal{AFEC}} = \mathcal{L}_{VLP} + \frac{1}{2}\lambda\Sigma_k\Omega_k(\theta_{t,k} - \theta_{t-1,k})^2 \\ + \frac{1}{2}\lambda_e\Sigma_k\Omega_k^\star(\theta_{t,k} - \theta_{t,k}^\star)^2,\end{aligned} \tag{2}$$

where $\theta_t^\star$ is the parameter of forward-step trained model and λ_e is the FIM of $\theta_t^\star$.

LWF [23] aligns the representations of previous-step and current models for all new arriving data. We maintain one reference model whose parameters are copied from the previous-step trained model and align the image and text representation of the reference and current model by the cross-entropy loss for each iteration.

B.2. Memory-Buffer methods

For the memory updating processing, the replay buffer will delete some old samples and add some new samples according to the size of the new task data.

ER [5] is a popular sample selection strategy. It uses the reservoir sampling [38] randomly stores a fixed number of training samples for each input batch and each sample has the same probability of being replaced.

Kmeans [5] use the Kmeans clustering to process all samples of the current task and set the number of clusters to the number of corresponding replaced samples. Then the cluster-center samples are chosen to update the buffer.

MoF [32] is first proposed by ICARL [32] and selects samples that are closest to the feature mean of each class. Because vision-language pretraining has no class concept, we choose the samples that are closest to the multi-modal feature mean of the current task.

ICARL [32] perform knowledge distillation on both buffer samples and new samples. The sample selection strategy is Mean-of-Feature (MoF). In our implementation, we combine the LWF term to optimize the current model.

LUCIR [14] proposes to utilize a cosine classifier to avoid the influence of the biased classifier and encourage similar feature orientation of the new and previous-step models. In our implementation, we replace the regular projected linear layer with the cosine normalizing linear layer. In addition, we constrain that the similarity of same-modal embedding from new and old models is big as possible. However, the inter-class distance constraint of the original paper [14] cannot be re-implemented because there is no class label in vision-language pretraining.

C. Dataset Comparison.

In Table 1, we present the comparison of our P9D with popular datasets from the continual learning domain [8] and multi-modal domain [41]. We observe that traditional continual learning datasets have a small number of data samples with limited classes (mostly at the thousand level), and only for single-target class labeling without detailed text description. In addition, although existing multi-modal datasets contain a large number of web image-text pairs, their data are too noisy and mixed to conform to the data split for continual tasks. In contrast, our P9D contains abundant image-text pairs to support vision-language pretraining. Besides, Each task contains rich semantic concepts, and different generalized semantic domains. It can support the simulation of continual learning environments.

C.1. Class-Incremental Learning Datasets

Oxford Flowers [28], MIT Scenes [30], CUB200-2011 [39], Stanford Cars [17], FGVC-Aircraft [26], VOC Actions [10], Letters [7], SVHN [27]. Aljundi *et al*. [2, 1] propose to use a sequence of 8 highly diverse recognition tasks as continual tasks. This sequence is composed of 8 different topics, going from flowers, scenes, birds, and cars, to aircrafts, actions, letters, and digits.

CIFAR10/100 [19] consists of 60,000 32×32 color images in 10 classes, with 6,000 images per class. There are 50,000 training images and 10,000 test images. The CIFAR100 dataset has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class and the 100 classes can be grouped into 20 superclasses.

MNIST [20] is a large handwritten digits dataset. It has 60,000 samples as the training set and 10,000 samples as

Dataset	Train Samples	Categories	Modal	Objects	Continual Task Split	URL
Popular Class-Incremental Learning Dataset						
Oxford Flowers [28]	2,040	102	image	single	yes	Link
VOC Actions [10]	3,102	11	image	single	yes	Link
MIT Scenes [30]	5,360	67	image	single	yes	Link
CUB200-2011 [39]	5,994	200	image	single	yes	Link
FGVC-Aircraft [26]	6,666	100	image	single	yes	Link
Letters [7]	6,850	52	image	single	yes	Link
Stanford Cars [17]	8,144	196	image	single	yes	Link
SVHN [27]	73,257	10	image	single	yes	Link
CIFAR10 [19]	50,000	10	image	single	yes	Link
CIFAR100 [19]	50,000	100	image	single	yes	Link
MNIST [20]	60,000	10	image	single	yes	Link
Tiny-ImageNet [8]	80,000	200	image	single	yes	Link
ImageNet-100 [34]	130,000	100	image	single	yes	Link
CORe50 [25]	120,000	50	image	single	yes	Link
Popular Multi-Modal Dataset						
Flickr30K [42]	29,000	–	image-text	multi	no	Link
COCO [24]	113,287	80	image-text	multi	no	Link
Visual Genome [18]	108K	–	image-text	multi	no	Link
FashionGen [33]	325,536	–	image-text	multi	no	Link
SBU [29]	875K	–	image-text	multi	no	Link
GQA [15]	1M	–	image-text	multi	no	Link
VQA v2.0 [11]	1.1M	–	image-text	multi	no	Link
CC3M [36]	3.1M	–	image-text	multi	no	Link
CC12M [3]	12M	–	image-text	multi	no	Link
YFCC-100M [37]	100M	–	image-text	multi	no	Link
LAION-400M [35]	400M	–	image-text	multi	no	Link
Our: P9D	1,014,599	3,814	image-text	multi	yes	–

Table 1: The overview of datasets about continual learning and vision-language pretraining domains. 'Categories' means the number of classes in the corresponding dataset and '–' means not mentioned. 'Objects' means the number of labeled/described objects in images. 'Continual Task Split' means the dataset contains different data chunks with discrepant semantic concepts and supports to simulate the continual environment. 'URL' means the hyperlink of corresponding dataset websites.

the test set.

Tiny-ImageNet [8] first used in the study of continual learning by Matthi *et al*. [8]. This is a subset of 200 classes from ImageNet [9] and the image size is rescaled to 64 $\times$ 64. Each class contains 500 samples subdivided into training (80%) and validation (10%), and 50 samples for evaluation.

ImageNet-100 (SubImageNet) [34] is a 100-class random sample subset of ImageNet. It contains 130,000 images for training and 5,000 images for testing.

CORe50 [25] is a collection of 50 objects collected in 11 distinct domains, where 8 of them (120,000 samples) are used for training, and the rest are used as a single test set (45,000).

C.2. Multi-modal Datasets

Flickr30K [42] is obtained by extending the corpus of Hodosh *et al*. [13] and the image topic contain everyday scenes and activities. There are 31,783 images associated with five manually annotated captions each, and 29,000 images are used for training.

COCO [24] is built based on MSCOCO dataset [24]. It consists of 123,287 images and each image is annotated with 5 captions. There are 113,287 training images, 5000 test images, and 5000 validation images. COCO and Flickr30K datasets are often used as the retrieval evaluation dataset for large-scale vision-language pretraining.

Visual Genome [18] is proposed to help to develop of visual understanding tasks (*i.e*. image caption and visual

question answering, *etc*.) by mining the relationships between objects. The dataset contains more than 108K images and each image has about 35 objects, 26 attributes, and 21 pairwise relationships.

FashionGen [33] contains 325,536 1360×1360 fashion images and each image has a paragraph-length caption as the description. Six different angles are photographed for all fashion items.

SBU [29] is collected and filtered from Flickr.com. It is usually used as the subset of vision-language pretraining [22, 21, 6].

GQA [15] is a balanced dataset with 1.7M samples which is mainly proposed for visual reasoning and compositional question answering.

VQA v2.0 [11] is proposed to reduce the language biases that existed in previous VQA datasets. It consists of around 1.1M image-question pairs and 13M corresponding answers based on 200K MSCOCO images.

CC3M [36] is a dataset annotated with conceptual captions and the image-text samples are mainly collected from the web. It contains about 3.3M image-description pairs.

CC12M [3] is a product of the urgent need for large-scale data with rapidly developing vision-language pre-training. The authors of CC3M relax the image-text filters and obtain the larger dataset CC12M.

YFCC-100M [37] totally contains 100 million media objects (99.2 million photos, 0.8 million videos) collected from Flickr.com.

LAION-400M [35] is filtered using pre-trained CLIP [31] and contains 400 million image-text pairs.

D. Algorithm

The Alg. 1 shows the training pipeline of our CTP in the task data $\mathcal{D}_t \in \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_T\}$.

E. More Experiments.

E.1. Momentum Setting.

For the first task ($t = 0$), There is no previous-step model and the current model has not adapted to the product domain. Thus, the momentum m of the first task is set to 0.995, and we keep it the same for all ablation studies. Because we find that the training loss oscillates and fails to converge if m is 0.9 in the first task.

For the following tasks, we set m as 0.9 and we also do the parameter-sensitive study about m. The results of Table 2 show the training on the $\{1, 2, .., T\}$ task is not sensitive to the setting of compatible momentum m. We suspect this is due to the fact that the model accepts parameters from both the previous-step and current models and is less prone to biased updates. In addition to the main vision-language pretraining loss, the compatible momentum contrastive loss is an auxiliary loss for continual learning. Thus, the model is more robust to momentum parameter selection and does not easily collapse [12].

Method	TR@1	IR@1	Rm	mAP
only θ^{t-1}	41.95	42.23	63.57	62.06
only θ^{t}	40.41	40.34	62.32	62.32
m=0.7	43.64	43.04	64.83	62.95
m=0.8	43.68	43.11	64.74	62.36
m=0.9	43.43	43.39	64.87	62.64
m=0.99	43.50	43.01	64.75	62.23
m=0.995	44.27	42.23	65.04	61.63

Table 2: The results of momentum selection experiment.

E.2. Reverse Task Order.

In the main text, all experiments are conducted in the default task order. To study the impact of task order on the performance ranking, we supplement a check experiment with the reversed task order [1]. The Table 3 shows the results of all baselines and our method in the reversed task order.

The result shows that although there are some changes in the ranking of some methods with similar performance, the overall performance ranking is still consistent with the performance ranking of default task order. Additionally, our method CTP exhibits superior performance in both continual learning scenarios (Memory-Free and Memory-Buffer), even when the order of tasks is changed. It indicates that the task order can affect the performance value of final result but not the performance ranking of our method. Our method consistently outperforms in different task order settings.

F. License

Our P9D dataset is released under CC BY-NC-SA 4.0 license and can freely be used for non-commercial purposes. The collection of data has obtained permission from the relevant websites. Once a conflict of interest, our group reserves all the rights for the final explanation.

References

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

[1] Electronics, Outdoor, Parenting, Auto, Clothing, Beauty, Food, Furnishings, and Household

Algorithm 1 Pseudocode of CTP in a PyTorch-like style.

```
# F, M, R: training, momentum, and reference (previous-task) model
# m, t, q_v, q_t: momentum, temperature, visual and textual queues

M.params, R.params = F.params, F.params # initialize momentum and reference model
for (image, text) in loader: # load a minibatch with N image-text pairs
   feat_v, feat_t = F.get_featuere(image, text)
   multimodal_out = F.multimodal_fusion(image, text, text.mask)
   S_i2t, S_t2i = feat_v @ feat_t.T, feat_t @ feat_v.T # cross-modal similarity
   S_i2i, S_t2t = feat_v @ feat_v.T, feat_t @ feat_t.T # same-modal similarity

   ita_loss = CE(S_i2t/t,eye_like(S_i2t))+CE(S_t2i/t,eye_like(S_t2i)) # image-text contrastive loss
   mlm_loss = CE(multimodal_out, labels=text.labels, mask=text.mask) # mask language modeling loss
   loss = ita_loss/2 + mlm_loss # conventional loss of the current task

   # compatible momentum update
   M.params = m*M.params+(1-m)/2*F.params+(1-m)/2*R.params
   feat_vm, feat_tm = M.get_featuere(image, text)
   multimodal_out_m = M.multimodal_fusion(image, text, text.mask)
   enqueue(q_v, feat_vm.detach(), q_t, feat_tm.detach()) # enqueue current features
   S_i2t_m, S_t2i_m = feat_v @ q_t.T, feat_t @ q_v.T

   # compatible momentum contrast
   ita_loss_m = CE(S_i2t_m/t, eye_like(S_i2t_m)) + CE(S_t2i_m/t, eye_like(S_t2i_m))
   mlm_loss_m = CE(multimodal_out,labels=multimodal_out_m.logits,mask=text.mask)
   loss += ita_loss_m/2 + mlm_loss_m
   dequeue(q_v, q_t) # dequeue earliest features

   # topology preservation
   feat_vr, feat_tr = R.get_featuere(image, text)
   S_i2t_r, S_t2i_r = feat_vr @ feat_tr.T, feat_tr @ feat_vr.T
   S_i2i_r, S_t2t_r = feat_vr @ feat_vr.T, feat_tr @ feat_tr.T
   loss_sm = CE(S_i2i/t,S_i2i_r/t,mask=eye_like(S_i2i))+ CE(S_t2t/t,S_t2t_r/t,mask=eye_like(S_t2t))
   loss_cm = CE(S_i2t/t, S_i2t_r/t) + CE(S_t2i/t, S_t2i_r/t)
   loss += loss_sm + loss_cm

   # parameter update
   loss.backward()
   update(F.params)
```

CE: cross entropy loss; **eye_like**: create an identity matrix with the same size as input.

Methods	Cross-modal Retrieval							Multi-modal Retrieval		
	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	Rm	mAP@1	mAP@5	mAP@10
JointT	60.72	86.05	91.74	61.98	86.82	91.85	79.86	64.07	70.09	67.33
Memory-Free										
SeqF	37.81	64.69	74.00	38.05	64.23	74.46	58.87	61.86	68.08	65.18
SI [43]	38.51	64.41	74.84	38.90	65.14	74.14	59.32	61.56	67.63	64.69
MAS [1]	39.81	66.97	75.86	40.86	66.87	75.93	61.05	61.65	67.78	65.23
EWC [16]	39.11	67.96	77.09	41.46	68.73	77.30	61.94	62.17	68.11	65.22
AFEC [40]	40.13	68.17	77.62	41.57	68.69	76.99	62.19	61.60	67.66	64.96
LWF [23]	41.18	67.29	76.39	39.81	67.81	76.32	61.31	61.76	68.12	65.20
RWalk [4]	39.04	67.85	78.00	40.20	68.94	77.65	61.95	62.40	68.43	65.60
Our:CTP	**45.96**	**73.47**	**80.85**	**44.98**	**72.34**	**80.25**	**66.31**	61.08	67.20	64.19
Memory-Buffer										
MoF [32]	43.92	72.28	80.99	45.01	73.19	81.03	66.07	61.37	67.65	64.79
LUCIR [14]	45.36	72.91	80.92	45.61	73.68	80.74	66.54	61.89	67.92	65.31
ER [5]	44.59	72.87	81.20	45.92	73.05	80.89	66.42	62.32	68.35	65.42
Kmeans [5]	45.19	74.42	81.83	46.03	73.05	80.67	66.53	62.73	68.35	65.39
ICARL [32]	47.33	75.65	83.63	47.61	75.90	83.24	68.89	62.54	68.54	65.87
Our:CTP+ER	**51.05**	**79.20**	**86.23**	**51.30**	**78.60**	**85.45**	**71.97**	61.58	67.65	64.78

Table 3: The final cross-modal and multi-modal retrieval performance comparison when conducting the reversed task order.

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

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.

[4] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajan-

than, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.

[5] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and M Ranzato. Continual learning with tiny episodic memories. 2019.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision.*, 2020.

[7] Teófilo Emídio De Campos, Bodla Rakesh Babu, Manik Varma, et al. Character recognition in natural images. *VISAPP (2)*, 7(2), 2009.

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

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Computer Vision and Pattern Recognition*, 2009.

[10] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.

[11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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*, pages 6904–6913, 2017.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[13] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 2013.

[14] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 831–839, 2019.

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[16] 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.

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

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

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. 2021.

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

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[25] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR, 2017.

[26] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[27] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[28] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[29] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[30] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] 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.

[33] Negar Rostamzadeh, Seyedarian Hosseini, Thomas Boquet, Wojciech Stokowiec, Ying Zhang, Christian Jauvin, and Chris Pal. Fashion-gen: The generative fashion dataset and challenge. *arXiv preprint arXiv:1806.08317*, 2018.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[35] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[37] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[38] Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 1985.

[39] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[40] Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391, 2021.

[41] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *arXiv preprint arXiv:2302.10035*, 2023.

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

[43] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

Figure 2: Some examples of our dataset. The first and second rows are the corresponding image-text pair. For simplicity, the rest rows show the images from the same class.