Visually Grounded Concept Learning from Streaming Data

Xisen Jin
xisenjin@usc.edu

Abstract

In this project, we study visually grounded concept learning from streaming data. We setup an environment to simulate the scenario when the visual data comes in non-stationary distribution over time, similar to the environment where children see and learn. We focus on online continual learning algorithms, where the goal is to alleviate catastrophic forgetting, defined as the degraded performance on previously learned tasks. We experiment with existing replay buffer based continual learning algorithms, and also experiment with a regularization based approach which is free of replay buffers. The latter approach learns a Hypernetwork to generate parameters for downstream classifiers according to learned task embeddings and regularize the change of generated parameters at training. Experiments show that the replay buffer based approaches yield most competitive resistance to catastrophic forgetting, while the Hypernetwork based approach is also effective without storing any prior input examples. We also propose an approach to build connects between task embeddings and word embeddings to capture conceptual similarity of tasks.

1 Introduction

Language is visually grounded in experience. Children language learning cannot be separated from their environments: for example, children learn nouns to refer to visual objects, and learn adjectives in order to describe different objects. The natural compositionality of visual objects (e.g. an apple that appears small and red) help children to acquire compositional semantics in language (e.g. a small red apple) with light effort. By looking at some examples, children learn what "small" and "red" means, which enable them to imagine images from text, or describe new objects. It inspires researchers to incorporate visual information into neural networks for more robust language learning of models. Learning multimodal semantics of language is shown to be helpful to various downstream tasks such as image captioning, visually grounded question answering, and visually grounded commonsense reasoning.

In this project, we study visually grounded language learning from *streaming* data. The setting is more practical, where the model can learn from massive amount of streaming data without storing them. However, it introduces additional challenges, such as catastrophic forgetting, more difficult generalization, and that the model looks at data only once. Given these settings, we study the problem as an *online continual learning* problem, where the distribution of the data in the stream may change over time.

The organization of the report is as follows. First, we formally introduce the task of continual learning and summarize existing continual learning algorithms. Next, we summarize existing works on multi-modal language learning. Then we introduce preliminary experimental results and future research plan.

2 Formulation and Methods

2.1 Task formulation

Continual learning assumes the input-label pairs (X, y) arrive in a sequential manner. The data is drawn from a task distribution $\mathcal{T}$, which itself also changes over time. The key problem that continual learning algorithms try to handle is catastrophic forgetting. There are two types of forgetting, namely:

- *Class forgetting*. When the model was not presented with examples with certain labels for a long time, the model may never output the label again.

- *Degraded Generalization*. When the distribution of X regarding a certain label y change a lot, the model may fail to predict correct labels for input X from the old distribution.

In prior literature, these two challenges corresponds to two continual learning scenarios, namely class-incremental learning and domain-incremental learning (Hsu et al., 2018). In this project, we focus on domain incremental learning, where we handle the problem of degraded classification performance for old tasks in the stream.

We simplify our downstream task so that we could focus on online continual learning algorithms: given a continual stream of images, each task is to classify the color attribute of one given object. The object distribution in the stream changes over time and some old objects may never be visited again. The model is required to maintain classification accuracy on old objects while learning on new objects. Our base classifier employs ResNet to extract visual features, and perform classification over colors.

2.2 Replay Buffer based Continual Learning Algorithms

We first study replay buffer based continual learning algorithms. These algorithms maintains a fix-sized replay buffer to store portion of prior input examples. We employ the reservoir sampling techniques (Riemer et al., 2019) to select samples to store in the replay buffer. The sampling ensures that at any point the probability of drawing an example from the buffer is the same as drawing a sample from all previously seen data.

We employ Experience Replay (ER) and Average Gradient Episodic Memory (AGEM) algorithms to utilize stored examples at training.

Experience Replay. Experience Replay (ER) draws a sample mini-batch from the replay buffer and attaches the mini-batch to the current training batch from the data stream.

Average Gradient Episodic Memory. Average Gradient Episodic Memory is a fast approximation of Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato, 2017). It constrains that the gradient update of parameters do not interfere with any example in the memory. Whenever the model evaluate gradient g for the parameters on a mini-batch from the stream, the model projects the gradient to a direction in which the dot product with any gradient g_k evaluated on the replay buffer is greater or equal to zero. The projection is solved by finding a projected gradient $\tilde{g}$ by solving the following optimization problem.

$$\begin{aligned} \text{minimize} \quad & ||g - \tilde{g}|| \\ \text{s.t.} \quad & \langle g_k, \tilde{g} \rangle \geq 0, k \in 1..m \end{aligned} \tag{1}$$

where m is the size of the replay buffer. The formulation of preventing interference is plausible but the constraints are overly strong and solving this quadratic program at very integration of model training is not effective. Average Gradient Episodic Memory (AGEM) (Chaudhry et al., 2019a) tackles this problem by instead restricting the gradient update $\tilde{g}$ do not interfere with the average gradient g_{ref} evaluated on the replay buffer. It removes the need for solving QP problem, and is highly effective while maintaining almost the same performance as GEM. The optimization problem is formally written as,

$$\begin{aligned} \text{minimize} \quad & ||g - \tilde{g}|| \\ \text{s.t.} \quad & \tilde{g}^T g_{ref} \geq 0 \end{aligned} \tag{2}$$

The problem has an closed form optimal solution for the projected gradient $\tilde{g}$, via:

$$\tilde{g} = g - \frac{g^T g_{ref}}{g_{ref}^T g_{ref}} g_{ref} \tag{3}$$

2.3 Replay-free Approaches for Continual Learning

While replay buffer based approaches has shown promising performance at alleviating catastrophic forgetting, the performance is limited by the size of the replay buffer, which is potentially not scalable. Some researchers also argue that storing prior examples bypass the problem of improving inherent ability of long term memory of models. We explore online continual learning algorithms which do not requires a replay buffer. We study a state-of-the-art Hypernet based continual learning algorithm (von Oswald et al., 2019) which falls in this category.

Figure 1 shows the model architecture. Unlike ordinary neural network, the Hypernet model is trained to generate *weights* for another model. The classifier weight generator takes as input the object embedding, and generate classifier weights for color classification regarding that object. The

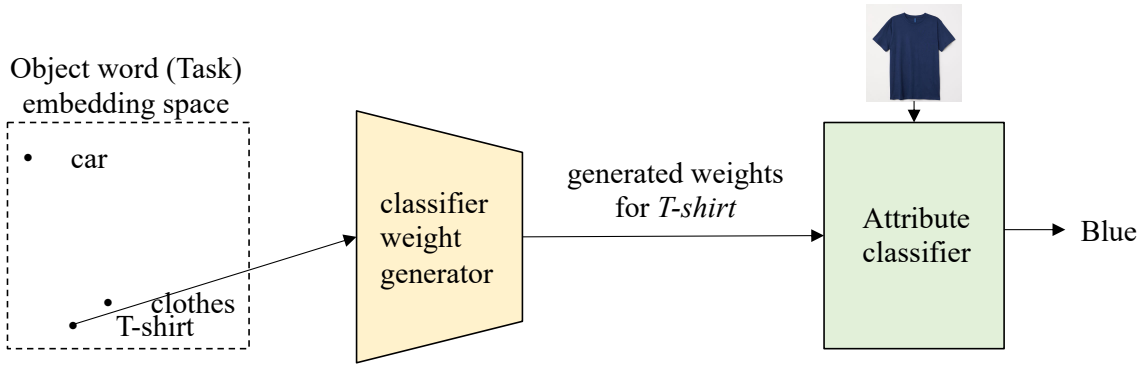


Figure 1: Hypernets for attribute classification. To perform classification, the classifier weight generator takes as input the embedding of the word ("T-shirt") and generate corresponding classification weights. The attribute classifier classify the image by loading the weights generated by classifiers.

weight generator is implemented as a 3-layer MLP.

To prevent catastrophic forgetting, the model directly regularizes the change of generated classifier weights from the most recent snapshot of the generated weights for each task at the weight space. Denote the parameter of the weight generator $f(\cdot)$ as θ_h, and the task embedding for the t-th task as $\mathbf{e}^{(t)}$, and the current task as T, the model penalize the L2 distance between generated classifier weights and the snapshot of the generated weight (noted as Θ^*), after the estimated parameter update $\Theta_h + \Delta\Theta_h$. Note that at any point of training, the model keeps only one most recent snapshot of the generated weights, which is the main difference from replay buffers that store much earlier weights or examples. The additional regularization term added to the classifier loss is formally written as follows,

$$\mathcal{L}_{reg} = \sum_{t=1}^{T-1} ||f_h(\mathbf{e}^{(h)}, \Theta_h^*) - f_h(\mathbf{e}^{(h)}, \Theta_h + \Delta\Theta_h)||^2 \quad (4)$$

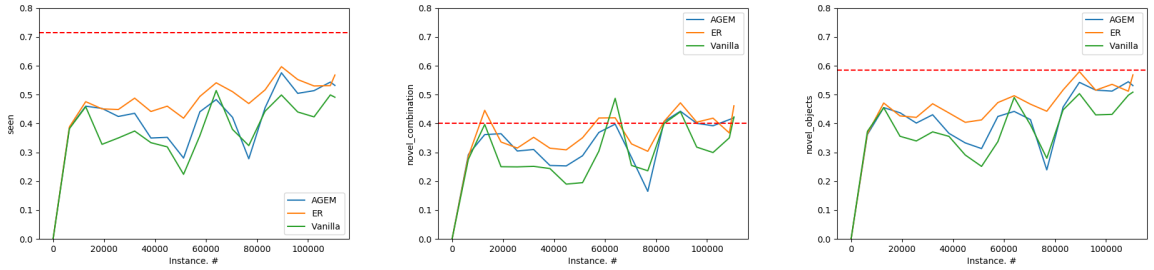
2.4 Learning Task Representations for Continual Learning

Since Hypernets generates classifier parameters according to task embeddings, similar parameters will be generated for similar tasks. The task embeddings will be gradually learned to capture task similarity. However, prior research has not fully exploit the task embedding space, simply keeping it fixed after training each task. Our key intuition is that contextually similar concepts (e.g. nouns) implies task similarity for downstream visual tasks (e.g. attribute classification for one given object). Given that hypernets generate model parameters according to task representations, we expect the generated parameters are similar for similar conepts. To test our hypothesis, we experiment with initializing the task embeddings in the Hypernets as Glove word embeddings. Glove embeddings can be regarded as an upperbound of the quality of word representations that can be learned from textual contexts associated with images (e.g. captions) in the dataset. This approach potentially reveal whether learning task representation from text associated with images will be helpful.

3 Experiments

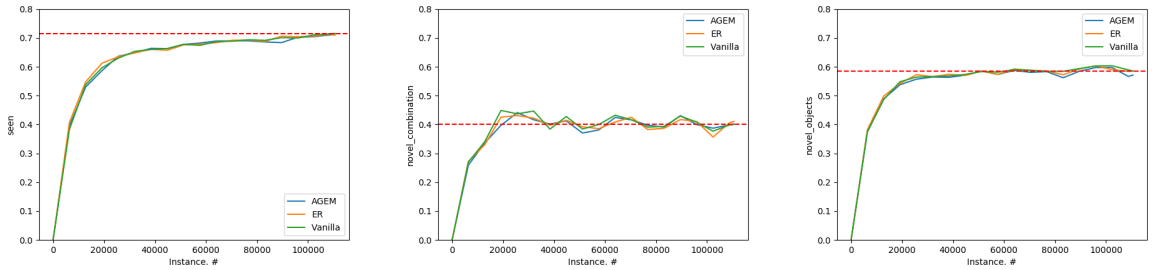
3.1 Experimental Setup

We use the GQA (Hudson and Manning, 2019) dataset for experiments. GQA is a large visual question answering dataset with scene graph and bounding box annoations. Each bounding box is associated with object labels and possibly some attribute labels. Figure 4 show an example in the GQA dataset. We train and test the model with most frequent 100 objects in the GQA dataset. We take the next 100 objects as "novel" objects to test on. For each object, we selected 80% of it associated colors as "seen" colors of that object for training and testing, the remaining 20% as "novel" colors of the object for testing only. We crop the image regions according to bounding box annotations, and feed into a ResNet-34 (He et al., 2016) classifier to make predictions. We plot the attribute prediction curve in Figures 2 and 3 in two task distribution settings, namely the *stationary* stream and *non-stationary* stream. In the non-stationary setting, the data stream is sorted by the objects.



(a) Seen combination (b) Novel combination (c) Novel objects

Figure 2: Online continual color classification accuracy curves at **non-stationary** setting. x-axis notes for training instances seen so far in the stream. We report the performance of Vanilla, ER, and AGEM methods in green, orange, and blue curves respectively. The red dotted line notes for the performance obtained in batch learning setting (upper-bound).



(a) Seen combination (b) Novel combination (c) Novel objects

Figure 3: Online continual color classification accuracy curves at **stationary** setting. x-axis notes for training instances seen so far in the stream. We report the performance of Vanilla, ER, and AGEM methods in green, orange, and blue curves respectively. The red dotted line notes for the performance obtained in batch learning setting (upper-bound).

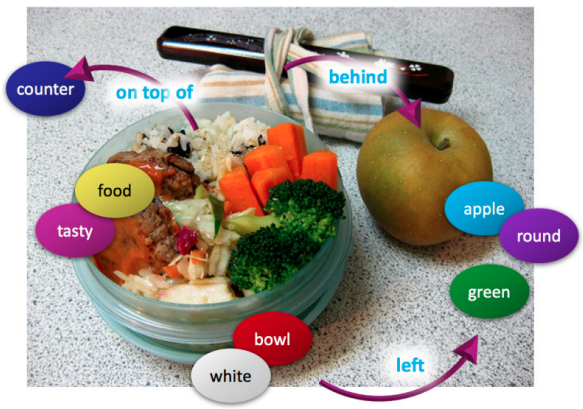


Figure 4: An example in the GQA dataset. At current stage, we only utilize bound box, object and color attribute annotations.

3.2 Results of Continual Learning Algorithms with Replay Buffers

In both of settings the data are visited only once. We plot the accuracy achieved by standard multi-epoch training setting in dotted lines in all the figures. We report the accuracy on seen combination of objects and attributes, novel combination of seen objects and attributes, and novel objects. The plot compares three continual learning algorithms, namely Vanilla, AGEM, and ER. A replay buffer with a capacity of 100 examples if used by default.

The results show that the model achieve only at most 72% accuracy for color attribute prediction, which is surprising, as color is one of the most easy visual concepts. We regard the performance as a consequence of noisy annotation. Nevertheless, the degraded performance on novel combinations and novel objects imply the concept of colors do not generalize perfectly to the inputs out of the training data distribution.

The continual learning algorithms, namely ER and AGEM, improve the accuracy on seen combinations by a large margin in non-stationary set-

ting. It also improves the accuracy on novel objects by a moderate amount in non-stationary setting. However, it does not show improvement for novel combinations. Actually, the performance of the model in non-stationary setting is equally bad compared to the stationary setting. It implies the compositional generalization is itself a challenging task even in stationary setting. We also see in stationary setting, no online continual learning algorithms make difference. It implies that continual learning algorithms do not naturally improve generalization of concepts in the stationary setting.

We also test the performance with different replay buffer size. The experiments show that the performance improves as the capacity of the replay buffer increases. We notice that even for a small replay buffer, e.g., 20 examples in total, already make a lot of difference.

3.3 Results of Hypernets

Table 1 shows experimental results on Hypernets. At our prelimiary experiments on Hypernets, we fix the feature extractor weights as (1) that pretrained at offline setting (noted as Offline classifier features), or (2) that pretrained on Imagenet classification, noted as Imagenet features. We use the classifier weight generator to generate weights for final 3 layers. We also assume ground truth objects are provided and feed the object labels into all comparators. Table 4 shows comparison between Hypernets, Vanilla, and ER in this setting. Experiments show while Hypernets performs not as well as ER, it outperforms the method where no regularization is imposed. It implies that the regularization of Hypernets indeed help. However, we find that Glove features are by no means helpful to the performance. It is probably because the disimilarity between text embedding space and the task embedding space. Also, since the generated parameters are very high-dimensinal, similarity encoded in the embedding space may be diluted in this case, i.e., similar task embeddings to not ensures similar classifiers at functiomn space.

We also find that the learning rate affects the performance in online continual learning significantly. For example, when the learning rate is increased from the default value $5e-5$ to $2e-4$, the performance of Hypernets decrease to 0.522, while the performance of Hypernets + Glove decrease to 0.455. However, the classification accuracy of the offline classifer remained the same. We could conjecture that the difference orginates from the online setting - the model is trained for next tasks when they are not converge on previous tasks with a small learning rate. As a result, the model could slowly learn classifiers that are generalizable to all the objects. In contrast, with a high learning rate, the model learn rapidly on new tasks and meanwhile forget rapidly on old tasks. This should remind us we should be careful about the "online" setting, as prior continual learning literatures usually assume the model is trained on tasks until convergence. In these literatures, the tasks come in a stream, but the model is trained in batch seteting within each task.

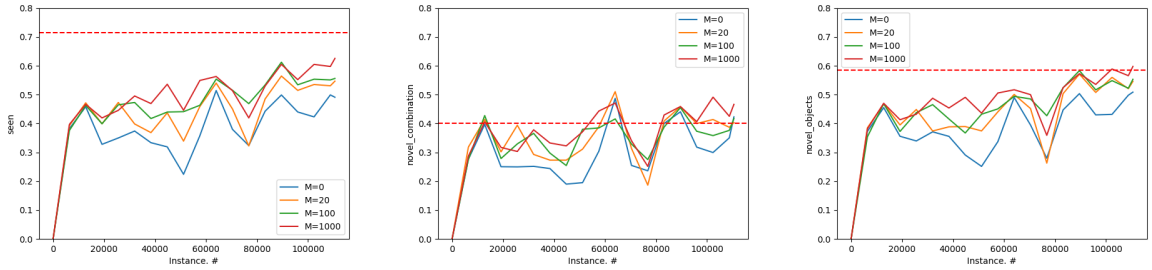
3.4 Discussion

We list some conclusion of the experiments, and discuss the next steps of the research in this section.

Generalization of concepts is close between offline and online settings. From Figure 2, we see the model performance in offline and online settings do not differ a lot for novel combinations and novel objects. It questions us whether we should rely on online continual learning algorithms to improve generalization. Perhaps to improve generalization, we should find a corresponding algorithm in the offline setting and modify it into the online setting.

Simple baselines perform rather well. We see the simple baselines like experience replay achieve the best performance in reducing catastrophic forgetting. Actually, this finding is confirmed by recent papers. Probably we should instead focus on which samples to store in the memory.

Generative Models are One of the Kernel Techniques of Continual Learning. Generative models are naturally resistant to catastrophic forgetting as they can generate previously seen examples , which can be easily taken as pseudo training data to train on. The state-of-the-art replay buffer free continual learning algorithm Deep Generative Replay (DGR) (Shin et al., 2017) follows this paradigm, which train a GAN simultaneously with the classifier on the data stream to model the data distribution. Hypernets can also be interpreted as a generative model in that it generates parameters of classifier instead of underlying data. The task embedding in the Hypernets is very similar to the latent variable in generative models such as Vari-



(a) Seen combination (b) Novel combination (c) Novel objects

Figure 5: Attribute prediction curve of Experience Replay algorithm in different replay buffer size settings

	Imagenet features	Offline classifier features
ER	-	0.657
Hypernets	0.320	0.556
Hypernets + Glove	0.301	0.532
Hypernets w/o regularization	-	0.516
Offline	-	0.730

Table 1: Results on Hypernets based continual learning. Offline classifier features notes for classification over ResNet features extracted from a trained classifier in offline setting for analytical purpose.

ational Auto Encoder (VAE). However, existing Hypernet do not generate parameters in a way like generative models, which typically requires a sampling step such as VAE. It is possible that bayesian generative models like VAE can GAN can be integrated into hypernets, which provides stronger probabilistic interpretation of task-conditioned hypernets, and further improve the quality of the learned continuous task embedding.

4 Related Works

4.1 Continual Learning

Existing continual learning algorithm can be categorized into regularization based approaches and replay buffer based approaches. Regularization based approaches include Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), Learning without Forgetting (Li and Hoiem, 2017) (LwF), etc. Deep Generative Replay (DGR) based approches also belong to this category as they do not store prior input examples, However, according to empirical evaluation, the performance gain of regularization based approaches are limited except DGR. Replay buffer based continual learning algorithms has shown strong performance. ER and GEM are popular algorithms that fall in this category. Multiple findings of (Chaudhry et al., 2019b; Hsu et al., 2018) show that ER approach is simple yet very effective. Meta Experience Replay (Riemer et al., 2019) considers the implicit connection between the first order meta-learning algorithm Reptile (Nichol et al., 2018) and dot product maximization between gradients, and runs Reptile algorithm on the examples in the replay buffer at each iteration.

Most of the prior works assume knowing the current task identities, or even knowing the number of the tasks beforehand. Some approaches use a *ring buffer* (Lopez-Paz and Ranzato, 2017), which allocate a fix memory per class per task, and store examples in a FIFO manner. However, in more practical scenarios, the identity of the tasks may be not known. Some approaches therefore run online k-mean algorithm over the feature outputs right before the classification layer (Aljundi et al., 2019), and store k examples that has the closest feature representation to k centroids respectively. The algorithm can be understood as storing prototype examples into the memory. However, the performance is degraded where the examples for different tasks are different by a large magnitude. Other approaches include *gradient based sample selection* (Aljundi et al., 2019), which maximize the diversity of examples stored in the replay buffer according to their gradients, either by solving an optimization problem or by greedy approximation. (Aljundi et al., 2019) show

that the gradient based sample selection is most-effective. (Chaudhry et al., 2019b) show cluster-based sample selection performs promisingly.

4.2 Concept Learning from Visual Data

The general goal of learning concepts from visual data has been explored with different tasks. Phrase grounding (Chen et al., 2018) is task where the model predict a bounding box from the input image that is referred to by the phrase. There exists prior work on weakly-supervised phrase grounding, where the model is only trained with image and phrase (or caption) pairs, without any bounding box annotations. Visual Question Answering is also interpreted as learning visual concepts (Mao et al., 2019), where the model learns nouns (shapes), adjectives (colors), and sentence parsing without parsing annotations.

4.3 Compositional Generalization

While in current experiments we only focused on evaluation of compositional generalization ability of model without explicitly tacking the problem, we regard this ability being relevant as it is one of importance abilities that even state-of-the-art models do not perform satisfactorily but is easy to humans. Existing research on compositional generalization has been studied without the context of continual learning. There are two major down stream tasks to measure compositional generalization. The first line of work is attribute-object classification. The models are usually designed to compose separate classifier of objects and attributes (Misra et al., 2017; Purushwalkam et al., 2019). The second line of works is compositional image captioning (Mao et al., 2015; Nikolaus et al., 2019).

5 Conclusion

In this project, we studied the problem of visually grounded concept learning from streaming data. We showed replay buffer based approaches performs the best in handling catastrophic forgetting, and the promising performance of hypernets for continual learning. We propose to initialize task embeddings as pretrained word embeddings to better capture task similarity, but at current timestep we did not get positive results. The experimental results on color prediction on novel combination also implies the inherent challenge of compositional generation, even in batch learning setting. Further study will be done for hypernets and compositional generalization. We will also test our approaches in other downstream tasks, such us image captioning.

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. 2019. Gradient based sample selection for online continual learning. In *NeurIPS 2019*.

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 K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. 2019b. Continual learning with tiny episodic memories. *arXiv preprint arXiv:1902.10486*.

Kan Chen, Jiyang Gao, and Ram Nevatia. 2018. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4050.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yen-Chang Hsu, Yen-Cheng Liu, and Zsolt Kira. 2018. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*.

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.

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

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes,

words, and sentences from natural supervision. In *International Conference on Learning Representations*.

Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE international conference on computer vision*, pages 2533–2541.

Ishan Misra, Abhinav Gupta, and Martial Hebert. 2017. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*.

Johannes von Oswald, Christian Henning, João Sacramento, and Benjamin F Grewe. 2019. Continual learning with hypernetworks. *arXiv preprint arXiv:1906.00695*.

Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. 2019. Task-driven modular networks for zero-shot compositional learning. *arXiv preprint arXiv:1905.05908*.

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*.

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