A Meta-transfer Learning framework for Visually Grounded Compositional Concept Learning

Anonymous ACL submission

Abstract

Humans acquire language in a compositional and grounded manner. They can describe their perceptual world using novel compositions from already learnt elementary concepts. However, recent research shows that modern neural networks lack such compositional generalization ability. To address this challenge, in this paper, we propose *MetaVL*, a meta-transfer learning framework to train transformer-based vision-and-language (V&L) models using optimization-based meta-learning method and episodic training. We carefully created two datasets based on MSCOCO and Flicker30K to specifically target novel compositional concept learning. Our empirical results have shown that *MetaVL* outperforms baseline models in both datasets. Moreover, *MetaVL* has demonstrated higher sample efficiency compared to supervised learning, especially under the few-shot setting.

1 Introduction

Acquiring language is the process of learning words from the surrounding environment. Humans acquire language in a compositional and grounded manner. They can combine words in novel ways to describe their perceptual world, although these novel compositions may have never been seen before. It would be desirable for intelligent systems to have such compositional generalization ability (Lake et al., 2017).

To address this issue, recent years have seen an increasing amount of work on grounded compositional concept learning (GCCL) which learns to describe perceptual world by composing novel concepts from previously learnt words. There are mainly two lines of work to formulate the GCCL problem. The first line of work studies compositional *attribute-object* pair learning and frames GCCL as a classification problem within the zero-shot learning (ZSL) framework(Misra et al., 2017; Nagarajan and Grauman, 2018). The second line frames GCCL as masked token prediction problem as proposed in (Jin et al., 2020; Surís et al., 2020). Our work follows the second line of formulation. Given a paired image-caption item with the target compositional concepts masked from the caption, models are expected to predict the masked concepts based on both linguistic and visual context. For example, as shown in Figure 1, suppose the models have learned primitive concepts such as *red* and *chair* from the training data, the models are expected to predict novel compositional concepts e.g., *red chair* in the testing data even though they have never appeared in the training data.

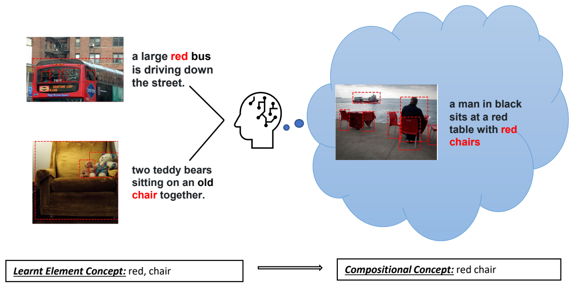


Figure 1: An illustration of Grounded Compositional Concept Learning(GCCL). For example, given concepts (red, bus) and (old chair) in the training data, the goal is to learn to predict novel compositional concept(red, chair) as masked token prediction at testing time.

By framing GCCL as a masked token prediction problem, current literature mainly employs transformer-based V&L models to solve the problem. Although self-supervised pre-training V&L models, such as VLBERT (Su et al., 2020) and LXMERT (Tan and Bansal, 2019), have achieved huge success and become the off-the-shelf encoding tools for downstream cross-modal applications, it has been recently noted that: 1) they are not data-efficient and typically require large amounts of fine-tuning data for satisfactory performance on the downstream tasks; and 2) pre-trained V&L models lack task-specific knowledge and ignore the discrepancy between pre-training tasks and downstream tasks which make it challenging to deploy such models in a low-resource setting. It is particularly challenging for our GCCL problem as the goal is to learn new compositional concepts which do not appear in the training data.

To address these issues, we propose a meta-transfer trained V&L model (*MetaVL*) for grounded compositional concept learning. Based on Model-Agnostic Meta-Learning(MAML) (Finn et al., 2017), MetaVL accumulates compositional knowledge by training through episodes. Each episode consists of a support set and a query set. Examples in the support set are used to learn element concepts, while examples in the query set are used to learn how element concepts are composed together to form a compositional concept. In addition, we combine MAML with transfer learning to exploit large-scale data through pre-training, similar to Sun et al. and Soh et al.. We further created two datasets based on MSCOCO and Flicker30K to specifically target novel compositional concept learning. Our empirical results have shown that *MetaVL* outperforms baseline models in both datasets. Moreover, *MetaVL* has demonstrated higher sample efficiency compared to supervised learning, especially under the few-shot setting.

The contributions of this work are the two folds. First, to the best of our knowledge, we are among the first to use the meta-learning framework on GCCL that achieves better performance compared to other transformer-based V&L models. It has demonstrated higher sample efficiency, especially under the few shot setting. Second, we have created two datasets, carefully curated for evaluating GCCL.These datasets will be made available to the community to support future research in this emerging area.

2 Related Work

2.1 Meta Learning

Meta learning, also known as *learning to learn*, aims to solve a low-resource problem by leveraging the learnt experience from a set of related tasks. Meta-learning algorithms deal with the problem of efficient learning so that they can learn new concepts or skills fast with just a few seen examples (few-shot setting) or even no seen examples (zero-shot setting). There are mainly three categories of meta-learning methods: 1) Metric-based methods learn a metric or distance function over tasks (Sung et al., 2018; Snell et al., 2017). 2) Model-based methods aim to design an architecture or a training process for rapid generalization across tasks (Ravi and Larochelle, 2016; Munkhdalai et al., 2018). 3) Optimization-based methods directly adjust the optimization algorithm to enable quick adaptation with just a few examples (Nichol et al., 2018; Finn et al., 2017). Meta learning has also been widely deployed in NLP field (Gu et al., 2018; Dou et al., 2019; Holla et al., 2020) recently to address the low-resource language processing problems.

2.2 Compositional Learning

Compositional learning is the key component of human intelligence and has been widely studied in the contexts of human-object interactions(HOI) (Kato et al., 2018; Hou et al., 2020), attribute-object learning (Nagarajan and Grauman, 2018; Misra et al., 2017), natural language processing (Lake, 2019; Nye et al., 2020) and language acquisition (Jin et al., 2020; Surís et al., 2020). Our work falls into the language acquisition category.

MetaVL has the similar problem formalization as (Jin et al., 2020) and (Surís et al., 2020), but different from their work. First, *MetaVL* focuses on compositional concept learning, not compositional phrase learning. Compositional concepts can be distributed in different parts of a sentence, not always in continuous phrase, which is a more rational and challenging compositional learning setting. Second, *MetaVL* adopts optimization-based meta-learning method to enhance the base V&L model's compositional ability instead of checking such compositional ability in continual setting. Surís et al. propose an episodic framework for grounded concept learning. Different from this work, *MetaVL* has a different learning setting and do not need to give a reference set in test time.

3 Problem and Dataset

3.1 Problem Formulation

Following Jin et al.'s work, we formulate GCCL as the grounded masked token prediction. In this setting, the training example is a four element tuple, $X = \left(\mathbf{x}_{\text{img}}, \mathbf{x}_{\text{bbox}}, \mathbf{x}_{\text{text}}, \mathbf{x}_{\text{label}}\right)$, where $\mathbf{x}_{\text{img}}$ and $\mathbf{x}_{\text{bbox}}$ are the image and the annotated bounding boxes, $\mathbf{x}_{\text{text}}$ is the related caption with the compositional concept $\mathbf{x}_{\text{label}}$ masked out. The models are expected to predict the masked compostional concepts $\mathbf{x}_{\text{label}}$ during test time. Different from (Jin et al., 2020) setting, the compositional concepts

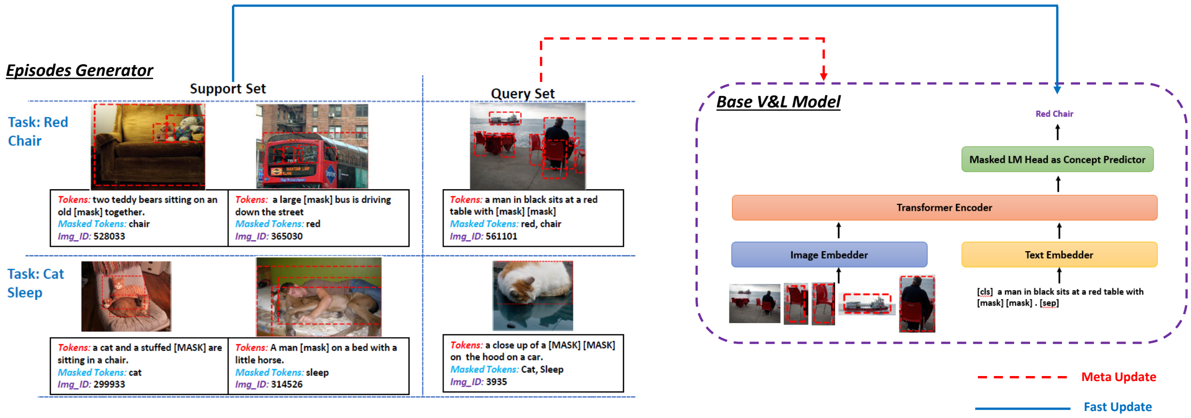


Figure 2: An illustration of *MetaVL*'s meta-learning process. Each episode is designed to teach the base V&L model to learn and compose the primitive concepts(i.e., "red", "chair") in the support set to recognize the compositional concept(i.e., "red chair") in the query set. The parameter updating within one episode happens in two levels: fast-update using element concepts from the support set and meta-update using the query set detailed in Section 5.3. Img_ID is from MSCOCO.

in GCCL do not need to be continuous phrases, which is a more realistic setting for compositonal learning (see Section 3.2). Moreover, we clarify the concept-related terms as follows:

- *Primitive or element concept* is the constituent of compositional concepts. It can be a verb, an adjective or a noun in GCCL. For example, *red* and *car* are element concepts regarding composiitonal concpet *red car*.
- *Compositional or pair concepts* refers to *adjective-noun* and *verb-noun* pairs in GCCL, including seen compositions and novel compositions based on whether we see them during the training time.

3.2 Dataset Construction

Nikolaus et al. introduce novel compositional data split designed to evaluate the image-captioning models' compositional ability based on MSCOCO dataset. They select 24 pairs as novel compositions and remove images related to these 24 pairs from the training dataset. Then, they check whether current SoTA captioning models can generate captions containing the 24 pairs which are never seen during training time. Following Nikolaus et al.'s work, Jin et al. utilize the same data split to check current V&LModel's compositional ability on phrase learning under the continual learning setting. However, based on their extracting rules, most of the phrases are in the form of *article* + *noun*, like <u>*the car*</u> and <u>*a man*</u>, instead of the original *adj/verb-noun* pairs which may not be sufficient to evaluate compositional learning ability.

In order to evaluate and improve V&L model's compositional ability, we build our GCCL benchmarks *ComptCOCO* using Nikolaus et al.'s extracting rules and data split, but mask out the exact 24 held-out *adj/verb-noun* pairs from captions. Moreover, to verify MetaVL's compositional generalizing ability, we further use the same pairs and extracting rules to construct *ComptFlirck* from Flikr30k Entities (Plummer et al., 2015) with statistics in Table 3.

Concretely, we construct data items by scanning each image-caption pair in the captioning dataset. For the caption input, we parse the caption using Stanza (Qi et al., 2020), extract and mask verb-noun pairs and adj-noun pairs using the part-of-speech (POS) and dependency information following the extracting rules in Appendix B. For the image part, we use Detectron-2 (Wu et al., 2019)[1] to extract the image and regional features from the ground truth bounding boxes without any object label or attribute information. Here, each image-caption pair is transformed into a series of text tokens and visual tokens in addition with the extracted compositional concept's information, including the token indexes and the token labels.

[1]https://github.com/facebookresearch/detectron2

4 MetaVL Models

4.1 Base Model

We use V&L models as our base model to predict compositional concepts. We choose VLBERT(Su et al., 2020) and LXMERT(Tan and Bansal, 2019) in this work. Both models take the above visual and textual tokens as input and adopt a simple yet powerful stack of self-attention blocks (Vaswani et al., 2017) to extract fused multi-modal representation for each token. The difference is that VLBERT treats image and text jointly by a single self-attention encoder known as *single-stream* V&L model, while LXMERT is a *dual-stream* V&L model which processes each modality data separately before joint cross-modal information fusion (Bugliarello et al., 2020). We will compare the performance of such *single-stream* and *two-stream* V&L performance for GCCL in this work.

Given the above visual and textual tokens, after adding special tokens and masking out compositional concept, we obtain the input as $x = ([cls]t_1, ...[mask], ...t_l, [sep], v_{l+1}, ...v_N, [sep])$ where the compositional concepts are replaced with $[mask]$ tokens. The V&L model takes x and predict the masked tokens to conduct the compostional concept learning process. A V&L model f_θ in GCCL consists of two modules: a self-attention multimodal encoder e_ψ and a concept predicting head h_ϕ where $\theta = \psi \cup \phi$ and $f_\theta = h_\phi(e_\psi)$. f_θ accepts input x and calculates d-dimensional contextual representations v_i for each token using encoder e_ψ and use h_ϕ to do prediction using the masked token's representation $v_{[mask]}$.

In GCCL, V&L models are expected to learn compositional concepts $\mathbf{x}_{\text{label}}$ by learning both element concept meaning and composing rules from the training items. Moreover, V&L models in GCCL are trained from the scratch to 1) avoid having the novel concept knowledge by loading the pre-trained weights, 2) fair comparison with (Jin et al., 2020; Surís et al., 2020) and 3) simulate the language acquisition process.

4.2 Optimization-based Meta-Learning

In this section, we discuss two optimization-based meta-learning methods used in GCCL: MAML and FOMAML.

MAML. We employ MAML (Finn et al., 2017), an optimization-based meta-learning framework , to address the compostional learning problem. Generally, MAML attempts to learn how to learn model parameters across episodes[2]. In GCCL, MAML is trained on episodes $\mathcal{D}_i = \{\mathcal{D}_i^{sup}, \mathcal{D}_i^{qry}\}$ composed by support set $\mathcal{D}_i^{sup}$ which focus on element learning and query set $\mathcal{D}_i^{qry}$ which focus on composing learning. Intuitively, MAML encourages optimization on the element support examples to have a positive effect on the compositional query examples and balance the concept recognition ability between element concepts and compositional concepts. When given an episode, MAML conducts the following steps:

- *Initialization*. Create fast model by copying the meta model. The fast model can be treated as the task-specific model and learns the compositional concept in the current task.
- *Inner update(meta-train)*. Training fast model on the support set $\mathcal{D}_i^{sup}$ by a few gradient descent steps using Equation 1. In this step, MetaVL learns element concepts from task i and L is the cross-entropy loss function.

$$\hat{\theta} = \theta - \alpha \nabla_\theta \mathcal{L}_i \left(\theta, \mathcal{D}_i^{sup}\right) \tag{1}$$

- *Outer update(meta-test)*. Applying the fast-updated model on the query set $\mathcal{D}_i^{qry}$ and use the compositional loss on a batch of query sets to update parameters using Equation 2. In this step, MetaVL learns the composing rule by optimizing through gradient updating procedure.

$$\theta = \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i \left(\hat{\theta}, \mathcal{D}_i^{qry}\right) \tag{2}$$

FOMAML. The standard MAML needs to explicitly calculate gradients from θ' with respect to θ by differentiating through the optimizer and needs to calculate the Hessian matrix. FOMAML simplifies the MAML implementation as Equation 3 which doesn't treat θ' as a function of θ and assumes $\nabla_{\hat{\theta}} \sum_i \mathcal{L}_i \left(\hat{\theta}, \mathcal{D}_i^{qry}\right) \approx \nabla_\theta \sum_i \mathcal{L}_i \left(\hat{\theta}, \mathcal{D}_i^{qry}\right)$ (Finn et al., 2017). FOMAML ignores the Hessian matrix and is a first-order approximation of MAML. We compare its performance with FOMAML later.

$$\theta = \theta - \beta \nabla_{\hat{\theta}} \sum_i \mathcal{L}_i \left(\hat{\theta}, \mathcal{D}_i^{qry}\right) \tag{3}$$

[2]Task and episode have the same meaning in our *MetaVL* setting. We use them interchangeably in this paper.

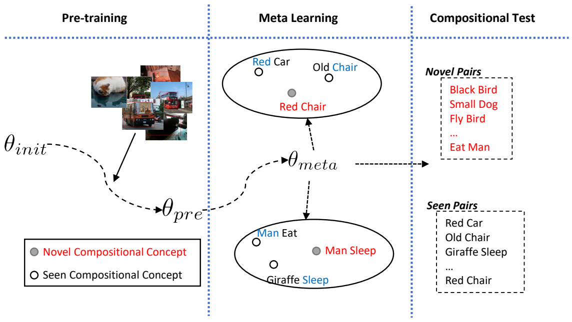


Figure 3: The meta-transfer learning framework for *MetaVL*. It includes three phases: pre-training phase, meta-learning phase and compositional test phase.

5 Meta-Transfer Training Pipeline

In conventional supervised learning, we usually assume the training items and the test items are from the same distribution. However, in GCCL, especially in the novel compositonal learning setting, this assumption does not hold. To address the compositional learning problems, we use meta-transfer pipeline to train MetaVL as (Sun et al., 2020; Soh et al., 2020). As shown in Figure 3, the overall meta-transfer training pipeline consists of three phases: 1) in transfer learning phase, we train *MetaVL* using all concepts, including element concepts and compositional concepts, to obtain the pre-trained parameters and transfer to the meta-training phase. 2) in meta-learning phase, we construct episodes to mimic the GCCL scenario and train MetaVL using MAML. 3) in the compositional test phase, we test MetaVL using both seen compositions and novel compositions. The meta-transfer training pipeline for MetaVL is detailed in Algirhtm 1.

5.1 Pre-training

At this phase, all training items are merged into a conventional training dataset D_T. The goal of the pre-training phase is to obtain relatively good parameters and equip the V&L models with basic ability to conduct concept recognition. Specifically, given an item $x^i = (\mathbf{x}^i_{\text{img}}, \mathbf{x}^i_{\text{bbox}}, \mathbf{x}^i_{\text{text}}, \mathbf{x}^i_{\text{label}})$, we randomly choose to mask out one single element concept or compositional concept corresponding to element concept learning or compositional concept learning. We use the cross-entropy loss as Equation 4 to update parameters in this phase where $x_i = (\mathbf{x}^i_{\text{img}}, \mathbf{x}^i_{\text{bbox}}, \mathbf{x}^i_{\text{text}})$ and $y_i = \mathbf{x}^i_{\text{label}}$.

$$L(\theta; D_T) = -\sum_{x_i, y_i \in D_T} \log P_\theta(y_i \mid x_i) \quad (4)$$

The pre-trained V&L model can be biased to frequent element and compositional concepts and lack compostional ability. Therefore, after pre-training, the parameters of θ are transferred to the next meta-learning phase to enhance the compositional ability.

Algorithm 1: Training MetaVL for GCCL.

Input: item $\left(\mathbf{x}^i_{\text{img}}\mathbf{x}^i_{\text{bbox}}, \mathbf{x}^i_{\text{text}}, \mathbf{x}^i_{\text{label}}\right)$, random initialized V&L model $f_\theta = h_\phi(e_\psi)$, meta-transfer learning parameters
Output: Optimized parameters $\theta = \psi \cup \phi$

```
/* Pre-train */
Pre-train (e_ψ, h_φ) using Eq.4 and obtain
   pre-trained parameters ψ_pre, φ_pre
/* Construct Episodes */
Construct Task Base T_i by sampling target
   compositional concepts, element concepts
   and related image-caption pairs described
   in Section 5.2
/* Model-Agnostic Meta-Learning */
while not done do
   for Each T_i do
      for Local Update Steps do
         // Meta Train on Sup-Set
         Compute ∇_ψ L_i(ψ), ∇_φ L_i(φ)
            on D_i^sup.
         Compute adapted parameters
            with gradient descent:
         ψ' = ψ − α∇_ψ L_{T_i}(θ)
         φ' = φ − α∇_φ L_{T_i}(φ)
      end
   end
   // Meta Test on Qry-Set
   Compute ∇_ψ' L_i(ψ), ∇_φ' L_i(φ) using
      batch of D_i^qry
   Update ψ and φ using either FOMAML
      or MAML.
end
/* Compositional Test */
Perform compositional concept recognition
   using meta-transfer updated parameters ψ
   and φ.
```

5.2 Episode Construction

Episode construction is one of the main challenges for meta-learning (Holla et al., 2020; Wang et al., 2021). Each episode in GCCL should be similar to the test environment and mimic the compositional learning process which requires both concept learning ability and concept composing ability. To build an compositional episode(*CompEpisode*), we first sample a *target compositional concept* from training dataset as virtual novel compositional concept, then we sample K items for the selected concept and mask the pair concepts and these K items make up the query set. For the support set, for each element concept in the selected compositional concept, we sample K items for each element concept and mask out the element concepts from the captions. Notably, we control the selected compositional concepts in support set not appearing in the query set to mimic the novel compositional learning setting. Then each episode has $3K$ items within which $2K$ items in the support set with element concepts masked and K items in the query set with compositional concepts masked as shown as Episode Generator in Figure 2 where K is set to 1 in this example. We define K, the item number in the query set, as the *shot number* in GCCL and we will study its effect in experiment section.

5.3 Meta-Learning

In this phase, we further fine-tune MetaVL using MAML and *CompEpisodes*. *MetaVL*'s meta-learning occurs at two levels including local update on the support set and meta update on a batch of query sets.

Intuitively, meta-learning's above bi-level optimization (Rajeswaran et al., 2019) encourages the optimization in the support set to have a positive effect on the query set as well. In GCCL setting, that means MetaVL learns pamameter θ not only beneficial to element concept recognition but also beneficial to compositional concept recognition.

5.4 Inference

At test time, we only focus on compositional concept prediction. Given an test item $(\mathbf{x}^i_{\text{img}}, \mathbf{x}^i_{\text{bbox}}, \mathbf{x}^i_{\text{text}})$, *MetaVL* predicts the masked compositional concepts using the meta-transfer trained θ without fine-tuning nor reference set using $\hat{y} = \arg\max p(y|x_{img}, x_{bbox}, x_{text})$ which is different from Surís et al.'s setting. Because the compositional concepts can be either novel pairs or seen pairs during test time, we report the performance under both settings.

6 Experiments

We created two datasets to evaluate the performance of *MetaVL*. This section gives detailed evaluation and analysis.

6.1 Dataset

Two datasets are created for GCCL as follows:
CompCOCO is constructed from COCO-captions's 2014 split version. COCO-captions has 103175 training images and 15112 validation images in the 2014 split (Lin et al., 2014; Chen et al., 2015). Because MSCOCO does not provide test data, we use the validation data as the testing data in *CompCOCO*. Furthermore, we randomly sampled 500 instances from the training set as the validation set. Moreover, we did some minor synonym modifications described in the Appendix A to extract more clean concepts.
CompFlickr is constructed from Flickr30k Entities (Plummer et al., 2015). Flickr30k contains $276k$ manually annotated bounding boxes for $31,783$ images and a total of $158,915$ English captions (five per image). We use the given train/val/test split in our experiment.

6.2 Implementation Details

We use *pytorch* on NVIDIA 2080Ti to implement all models and use Higher[3] to implement MAML and FOMAML. The learning rate in pre-training phase is $1e-4$ and in meta-learning is set to $5e-5$ for both inner updates and outer updates. Due to V&LModel's scale and computing resource limitation, we set inner update to 1 in our MAML's implementation.

6.3 Evaluation Metrics.

To measure the GCCL performance, we use accuracy as our primary metric. We also report Perplexity (PPL) (Mikolov et al., 2011) as in Jin et al.'s work. PPL measures the uncertainty about *MetaVL*'s compositional prediction and is calculated as $PPL(W) = -\frac{1}{N}\log P(W)$. Lower PPL is preferred.

6.4 Baselines

We use two baselines in this evaluation. The first baseline is the **pre-trained baseline**. It is exactly the off-line baseline as in Jin et al.. It is also the pre-trained model for MetaVL. The second baseline is a meta-learning baseline **Reptile** (Nichol et al.,

[3]https://github.com/facebookresearch/higher

2018) to demonstrate the importance of episode construction in GCCL. Reptile is another first-order optimization-based meta-learning method. It updates parameters using $\theta \leftarrow \theta + \epsilon(\theta^{(k)} - \theta)$ where $\theta^{(k)}$ is the inner updated parameters after k steps. Different from the MAML setting, it does not require tasks to have a query set. This makes it easier in task construction.

6.5 Main Results

We report the performance under both seen compositions and novel compositions in this section.

Seen Compositions. Table 1 shows the performance of different models under the seen setting (i.e., predicting compositional concepts that have appeared in the training set). From the table, we can see that *MetaVL*, including FOMAML and MAML, outperforms conventional pre-trained V&L models. This suggests that *MetaVL*, through optimizing the V&L model towards compositional generalization, captures a representation which is beneficial for compositional learning.

In contrast, while Reptile works well on few-shot learning, it does not improve the performance in GCCL. One reason is that Reptile does not have a query set in their episode construction. Therefore, it cannot capture how concepts are composed through the query set as in MetaVL. In fact, query sets are particularly important as they accumulate knowledge on how element concepts are composed together for learning compositional concepts.

	V&L-Model	VLBERT		LXMERT	
	Metric	Accu.↑	PPL↓	Accu.↑	PPL↓
COCO	Pre-Train	0.5975	1.7421	0.6158	1.5632
	Reptile	0.5962	1.7831	0.5998	1.7625
	FOMAML	0.6137	**1.6995**	0.6290	**1.5183**
	MAML	**0.6201**	1.7046	**0.6429**	1.5738
Flickr	Pre-Train	0.5573	2.3632	0.5889	1.7631
	Reptile	0.5488	2.3575	0.5800	1.7701
	FOMAML	0.5717	1.9956	0.6081	1.7258
	MAML	**0.5863**	**1.8741**	**0.6107**	**1.7022**

Table 1: Results on Seen Compositional Concept.

Novel Compositions. As shown in Table 2, *MetaVL* improves the performance on the novel setting compared to pre-trained model and Reptile. However, compared with seen compositions (i.e., Table 1), the performance on novel pairs drops significantly across the board. Taking VLBERT on *CompCOCO* as an example, the accuracy drops by about 18%. This indicates the compositional generalization is still a very difficult task for current V&L models.

	V&L-Model	VLBERT		LXMERT	
	Metric	Accu.↑	PPL↓	Accu.↑	PPL↓
COCO	Pre-Train	0.4180	2.2990	0.4222	2.1157
	Reptile	0.4017	2.3001	0.4239	2.1163
	FOMAML	0.4312	2.1936	0.4483	2.7818
	MAML	**0.4593**	**1.9897**	**0.4728**	**2.015**
Flickr	Pre-Train	0.4758	2.3918	0.5213	2.0497
	Reptile	0.4689	2.4102	0.5173	2.1546
	FOMAML	**0.5145**	2.0013	0.5376	1.9983
	MAML	0.5014	**1.8452**	**0.5719**	**1.6778**

Table 2: Results on Novel Compositional Concept.

Moreover, from Table 1 and Table 2, we can see the following interesting results: 1) LXMERT (two-stream V&L Model) has better performance compared with VLBERT (single-stream V&L Model) on both benchmarks which is worth further probing. 2) MAML outperforms its first-order approximation FOMAML. Hessian matrix may bring additional information for compositonal learning in MetaVL.

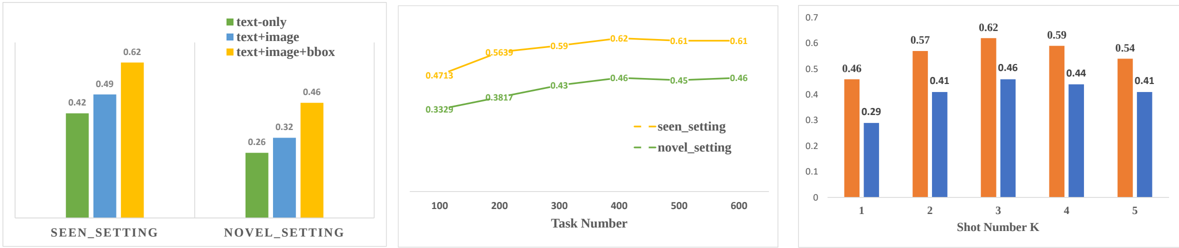
6.6 Ablation Study

Effect of Visual Input. In GCCL, one interesting question is how much visual input helps concept learning. To answer this question, we compare three configurations: 1) *Text-only Prediction:* zeroing-out all visual tokens and only keep the text tokens as input; 2) *Text + Image Prediction:* zeroing-out all bounding box tokens and keep text tokens and the whole image token as input; and 3) *Text + Image + BBox Prediction:* keep all text and visual information as described earlier.

Figure 4a shows the importance of visual input for *MetaVL* in GCCL. We can see that without visual input, the accuracy drops from 0.62 to 0.42 on seen compositions and drops from 0.46 to 0.42 on seen compositions. Moreover, better contextual information as given by the bounding boxes helps MetaVL better learn compositional concepts.

Effect of Number of Episodes used for Learning. We examine how the number of episodes (i.e., tasks) used for learning in MetaVL may affect the outcome. From Figure 4b, we can see the trend that at the beginning the accuracy increases as *MetaVL* trained on more tasks, reaches the peak at about 400 episode and keeps stable afterward even trained on more episodes for both the seen and novel compositions.

Effect of Shot Number K in Each Episode. The number of examples (i.e., in the support set and the query set) in each episode may affect the learn-



(a) Visual information importance for VL-BERT on *CompCOCO* dataset.

(b) Effect of task number for VLBERT on *CompCOCO* dataset.

(c) Effect of shot number for VLBET on *CompCOCO* dataset.

Figure 4: Ablation study for *MetaVL*'s performance.

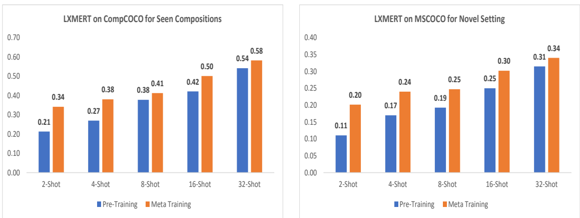


Figure 5: Data efficiency comparison between Supervised-Learning and Meta-Learning for compositional Concept Learning.

ing outcome. More training examples within one episode may introduce ambiguity, as *red* in *red wine* and *red car* have different meanings. We varied different numbers of training examples in each episode, i.e., K described in Section 5.2. Our results have shown that 32 examples in our setting has best performance (i.e, meaning the support set has 32 object concepts and 32 verb/adjective concepts and the query set has 32 compositional concepts).

7 Meta-Learning Efficiency

One key advantage of meta learning is its ability to learn how to learn a task through a small number of examples. In this section, we study the data efficiency of meta-learning compared with the conventional V&L model through supervised training in the compositional leraning setting. We select 400 tasks as our training data and change the shot number for each task. In this setting, meta-trained and supervised-trained models access the same set of data items. The difference is that MetaVL organized the data items into *CompEpisodes* and supervised-trained model learn from all the items. Fig. 5 shows that in both seen and novel settings, MetaVL achieves better compositional ability compared to supervised-learning. Empirically, meta-learning has demonstrated a higher sample efficiency as shown by the learning curves. Meta learning is consistently better than conventional supervised learning as it can leverage its past experience to solve new tasks. The difference is more significantly under the few shot setting (e.g., 2-shot setting).

8 Conclusion

In this paper, we propose *MetaVL*, a meta-transfer trained V&L model, for grounded compositional concept learning. It builds upon current V&L models and MAML to learn how to compose element concepts together to form compositional concepts. Our empirical results on two datasets have shown that *MetaVL* consistently outperforms conventional V&L models for GCCL. However, GCCL is still a challenging open problem. Many problems remain. Our future work will explore more cognitively plausible models and explicitly address the grounding ability in concept learning.

References

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. 2018. Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. *arXiv preprint arXiv:2004.14355*.

Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer.

Xisen Jin, Junyi Du, Arka Sadhu, R. Nevatia, and X. Ren. 2020. Visually grounded continual learning of compositional phrases. In *EMNLP*.

Keizo Kato, Yin Li, and Abhinav Gupta. 2018. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251.

Brenden M Lake. 2019. Compositional generalization through meta sequence-to-sequence learning. *arXiv preprint arXiv:1906.05381*.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

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 *European conference on computer vision*, pages 740–755. Springer.

Tomáš Mikolov, Anoop Deoras, Stefan Kombrink, Lukáš Burget, and Jan Černockỳ. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Twelfth annual conference of the international speech communication association*.

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.

Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri, and Adam Trischler. 2018. Rapid adaptation with conditionally shifted neurons. In *International Conference on Machine Learning*, pages 3664–3673. PMLR.

Tushar Nagarajan and Kristen Grauman. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185.

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. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

Maxwell I Nye, Armando Solar-Lezama, Joshua B Tenenbaum, and Brenden M Lake. 2020. Learning compositional rules via neural program synthesis. *arXiv preprint arXiv:2003.05562*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. 2020. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Qianru Sun, Yaoyao Liu, Zhaozheng Chen, Tat-Seng Chua, and Bernt Schiele. 2020. Meta-transfer learning through hard tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

D. Surís, D. Epstein, H. Ji, S. Chang, and C. Vondrick. 2020. Learning to learn words from visual scenes. *European Conference on Computer Vision (ECCV)*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Ran Wang, Siyu Long, Xinyu Dai, Shujian Huang, Jiajun Chen, et al. 2021. Meta-lmtc: Meta-learning for large-scale multi-label text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8633–8646.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

A Modified MSCOCO Synonym

In order to extract more compositional concepts, we modify drier's synonym list as : hair drier, hairdryer, hair dryer, blow dryer, blow drier

B Extracting Rules

We use exact extracting rules of (Nikolaus et al., 2019) to extract verbs and adjectives for *CompCOCO* and extract adjetives for *CompFlickr*

B.1 Adj-Noun Pair Extracting Rule

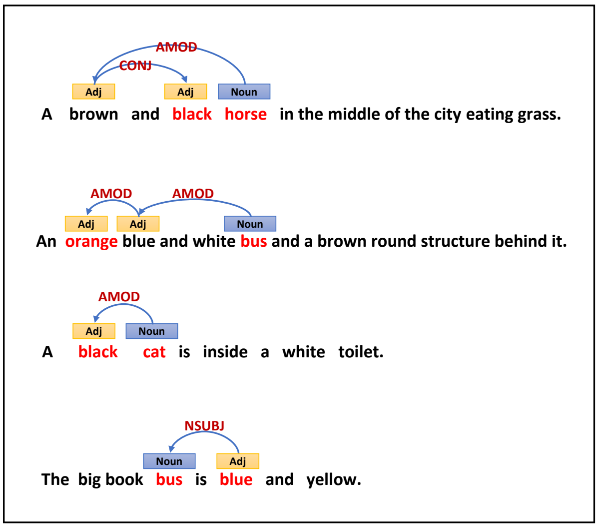


Figure 6: Rules to extract adj-noun pairs.

B.2 Verb-Noun Pair Extracting Rule

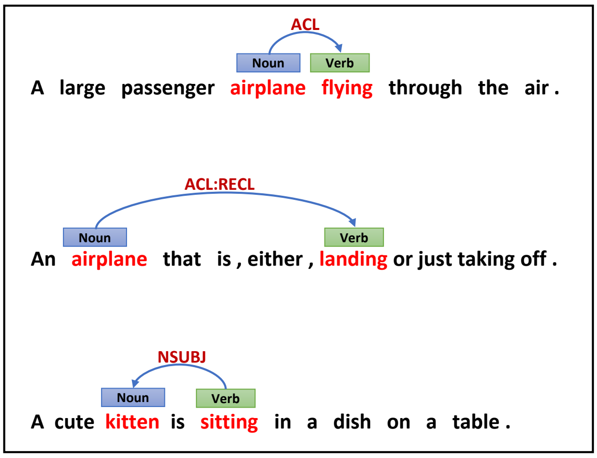


Figure 7: Rules to extract verb-noun pairs.

C Statistics of Novel Pairs

	MSCOCO				Flickr30K					
	Train Img.	Train Caps.	Test Img.	Test Caps.	Train Img.	Train Caps.	Val Img.	Val Caps.	Test Img.	Test Caps.
black bird	205	323	122	190	17	24	0	0	2	3
small dog	681	1067	316	481	360	612	11	12	17	33
white boat	373	261	196	134	69	85	0	0	3	8
big truck	417	601	191	288	28	38	0	0	1	1
eat horse	212	378	106	187	2	2	0	0	0	0
stand child	1288	1556	577	741	1048	1475	38	57	26	36
white horse	264	500	151	300	51	100	3	4	4	8
big cat	184	216	103	108	0	0	0	0	1	1
blue bus	276	506	143	243	11	16	0	0	0	0
small table	261	296	134	154	48	54	1	1	1	1
hold child	1328	1860	664	992	835	1289	27	37	35	60
stand bird	532	831	260	406	13	24	0	0	0	0
brown dog	613	878	291	430	934	1838	31	61	29	58
small cat	252	325	149	183	2	3	0	0	0	0
white truck	262	420	121	175	35	42	2	2	2	2
big plane	967	1345	357	494	5	5	0	0	0	0
ride woman	595	674	300	330	266	537	8	17	9	23
fly bird	245	526	132	283	29	53	0	0	0	0
black cat	840	1760	448	940	15	27	0	0	1	1
big bird	215	291	123	169	24	34	0	0	0	0
red bus	566	1212	232	474	11	20	0	0	1	1
small plane	481	833	158	279	13	20	0	0	0	0
eat man	555	698	250	314	153	272	4	5	5	10
lie woman	301	388	144	194	145	278	1	2	4	8

Table 3: Novel Pair Statistics for both *CompCOCO* and *CompFlickr*. For fair comparation, we use the same 24 pairs to verify the compositional generalization.