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.

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, \( X = (x_{\text{img}}, x_{\text{bbox}}, x_{\text{text}}, x_{\text{label}}) \), where \( x_{\text{img}} \) and \( x_{\text{bbox}} \) are the image and the annotated bounding boxes, \( x_{\text{text}} \) is the related caption with the compositional concept \( x_{\text{label}} \) masked out. The models are expected to predict the masked compositional concepts \( x_{\text{label}} \) during test time. Different from (Jin et al., 2020) setting, the compositional concepts
in GCCL do not need to be continuous phrases, which is a more realistic setting for compositional 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 compositional concept *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&L Model’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 *the car* and *a man*, 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 ComptFlickr 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 (Bougliarello 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 compositional concept learning process. A V&L model $f_\theta$ in GCCL consists of two modules: a self-attention multimodal encoder $e_\psi$ and a concept predicting head $h_\phi$ where $\theta = \psi \cup \phi$ and $f_\theta = h_\phi(e_\psi)$. $f_\theta$ accepts input $x$ and calculates $d$-dimensional contextual representations $v_i$ for each token using encoder $e_\psi$ and use $h_\phi$ to do prediction using the masked token’s representation $v_{[mask]}$.

In GCCL, V&L models are expected to learn compositional concepts $x_{label}$ 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 compositional learning problem. Generally, MAML attempts to learn how to learn model parameters across episodes\(^2\). In GCCL, MAML is trained on episodes $D_i \equiv \{D_{i}^{sup},D_{i}^{qry}\}$ composed by support set $D_{i}^{sup}$ which focus on element learning and query set $D_{i}^{qry}$ 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 $D_{i}^{sup}$ 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.

  \[
  \hat{\theta} = \theta - \alpha \nabla_\theta \mathcal{L}_i (\theta, D_{i}^{sup})
  \]

- **Outer update(meta-test).** Applying the fast-updated model on the query set $D_{i}^{qry}$ 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.

  \[
  \theta = \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i (\hat{\theta}, D_{i}^{qry})
  \]

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

\[
\theta = \theta - \beta \nabla_\theta \sum_i \mathcal{L}_i (\hat{\theta}, D_{i}^{qry})
\]

\(^2\)Task and episode have the same meaning in our MetaVL setting. We use them interchangeably in this paper.
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 compositional 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-learning 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 Algorithm 1.

5.1 Pre-training

At this phase, all training items are merged into a conventional training dataset. 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 \( x^i = (x^i_{img}, x^i_{bbox}, x^i_{text}, x^i_{label}) \), 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 \( x_i = (x^i_{img}, x^i_{bbox}, x^i_{text}) \) and \( y_i = x^i_{label} \).

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

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

Algorithm 1: Training MetaVL for GCCL.

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

Output: Optimized parameters \( \theta = \psi \cup \phi \)

/ * Pre-train */

1 Pre-train \((e_\psi, h_\phi)\) using Eq.4 and obtain pre-trained parameters \( \psi_{pre}, \phi_{pre} \)

/ * Construct Episodes */

2 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 */

3 while not done do

4 for Each \( T_i \) do

5 for Local Update Steps do

6 // Meta Train on Sup-Set

7 Compute \( \nabla_\psi L_i(\psi), \nabla_\phi L_i(\phi) \) on \( D^sup_i \).

8 Compute adapted parameters with gradient descent:

9 \( \psi' = \psi - \alpha \nabla_\psi L_T(\theta) \)

10 \( \phi' = \phi - \alpha \nabla_\phi L_T(\phi) \)

end

11 // Meta Test on Qry-Set

12 Compute \( \nabla_\psi L_i(\psi), \nabla_\phi L_i(\phi) \) using batch of \( D^qry_i \)

13 Update \( \psi \) and \( \phi \) using either FOMAML or MAML.

14 end

/ * Compositional Test */

15 Perform compositional concept recognition using meta-transfer updated parameters \( \psi \) and \( \phi \).

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 parameter $\theta$ 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 $(x_{img}^i, x_{bbox}^i, x_{text}^i)$, MetaVL predicts the masked compositional concepts using the meta-transfer trained $\theta$ without fine-tuning nor reference set using $\hat{y} = \arg\max_{y} p(y|x_{img}, x_{bbox}, x_{text})$ 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
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.

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 compositional learning in MetaVL.

**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.

<table>
<thead>
<tr>
<th>V&amp;L-Model</th>
<th>VLBERT</th>
<th>LXMERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>Accu.↑</td>
<td>PPL↓</td>
</tr>
<tr>
<td>COCO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Train</td>
<td>0.3975</td>
<td>1.7421</td>
</tr>
<tr>
<td>Reptile</td>
<td>0.5962</td>
<td>1.7831</td>
</tr>
<tr>
<td>FOMAML</td>
<td>0.6137</td>
<td>1.6995</td>
</tr>
<tr>
<td>MAML</td>
<td>0.6201</td>
<td>1.7046</td>
</tr>
<tr>
<td>Flickr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Train</td>
<td>0.3573</td>
<td>2.3632</td>
</tr>
<tr>
<td>Reptile</td>
<td>0.5488</td>
<td>2.3575</td>
</tr>
<tr>
<td>FOMAML</td>
<td>0.5717</td>
<td>1.9956</td>
</tr>
<tr>
<td>MAML</td>
<td>0.5863</td>
<td>1.8741</td>
</tr>
</tbody>
</table>

Table 1: Results on Seen Compositional Concept.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accu.↑</th>
<th>PPL↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>COCO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Train</td>
<td>0.4180</td>
<td>2.2990</td>
</tr>
<tr>
<td>Reptile</td>
<td>0.4017</td>
<td>2.3001</td>
</tr>
<tr>
<td>FOMAML</td>
<td>0.4312</td>
<td>2.1936</td>
</tr>
<tr>
<td>MAML</td>
<td>0.4593</td>
<td>1.9897</td>
</tr>
<tr>
<td>Flickr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-Train</td>
<td>0.4758</td>
<td>2.3918</td>
</tr>
<tr>
<td>Reptile</td>
<td>0.4689</td>
<td>2.4102</td>
</tr>
<tr>
<td>FOMAML</td>
<td>0.5145</td>
<td>2.0013</td>
</tr>
<tr>
<td>MAML</td>
<td>0.5014</td>
<td>1.8452</td>
</tr>
</tbody>
</table>

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 compositional 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-
Table 4: Ablation study for MetaVL’s performance.

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

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 learning 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


Tushar Nagarajan and Kristen Grauman. 2018. At-
Tsendsuren Munkhdalai, Xingdi Yuan, Soroush Mehri,
Tomáš Mikolov, Anoop Deoras, Stefan Kombrink,
Tsung-Yi Lin, Michael Maire, Serge Belongie, James

Bryan A Plummer, Liwei Wang, Chris M Cervantes,
Juan C Caicedo, Julia Hockenmaier, and Svetlana


Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu,

Qianru Sun, Yaoyao Liu, Zhaozheng Chen, Tat-Seng


Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho,
and Victor OK Li. 2018. Meta-learning for low-

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis,


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 adjectives 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
| 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 comparison, we use the same 24 pairs to verify the compositional generalization.